INVITED PAPER

Global Navigation Satellite System Precise Positioning Technology

Nobuaki KUBO[†a)], ***Member***

SUMMARY In this study, the most recent topics related to the precise global navigation satellite system (GNSS) positioning technology are discussed. Precise positioning here means that the position can be estimated with centimeter-level accuracy. Technologies supporting precise GNSS positioning include an increase in the number of positioning satellites and the availability of correction data. Smartphones are now capable of centimeter-level positioning. For correction data, real-time kinematic positioning (RTK)-GNSS, which has primarily been used in surveying, and the new precise point positioning-real-time kinematic (PPP-RTK) and PPP, are garnering attention. The Japanese Quasi-Zenith Satellite System was among the first to broadcast PPP-RTK and PPP correction data free of charge. RTKLIB has long been popular for both real-time and post-processing precise positioning. Here, I briefly present a method for improving this software. Precise positioning technology remains crucial as the use of GNSS in highly reliable applications, such as advanced driver-assistance systems, autonomous drones, and robots, is increasing. To ensure precise positioning, improving multipath mitigation techniques is essential; therefore, key factors related to these techniques are discussed. I also introduce my efforts to develop software GNSS receivers for young researchers and engineers as a basis for this purpose. This study is aimed at introducing these technologies in light of the most recent trends.

key words: *GNSS, precise positioning, multipath mitigation, QZSS, PPP-RTK, PPP, RTK-GNSS, software GNSS receiver*

1. Introduction

Location-based applications of precise global navigation satellite system (GNSS) positioning are extremely diverse [1]. Examples include surveying, precision agriculture, precision construction machinery, precise monitoring of civil engineering construction, automated navigation using drones, advanced driver assistance systems, monitoring of crustal deformation, marine civil engineering, personal mobility, and smartphone navigation. While GNSS is generally effective outdoors, it cannot provide precise positioning indoors owing to insufficient radio signal reception. Positioning in transitional areas between outdoor and indoor environments is also challenging. Given the importance of location-based services in these difficult areas, I herein discuss methods for reducing GNSS errors. Although indoor positioning is not covered in this study, interested readers can refer to the work by Nishio [2] for more information.

This paper consists of seven sections. Section 1 introduces an outline of the paper. Section 2 discusses the status of global positioning satellites, including the global positioning system (GPS), Globalnaya Navigazionnaya Sputnikovaya Sistem (GLONASS), GALILEO, BeiDou Navigation Satellite (BDS), Quasi-Zenith Satellite System (QZSS), and Navigation with Indian Constellation (NavIC). Section 3 briefly introduces the basics of real-time kinematic positioning (RTK-GNSS), precise point positioning-real-time kinematic (PPP-RTK), and PPP. Section 4 outlines the Japanese QZSS, particularly its correction services to achieve centimeter-level positioning. Section 5 covers multipath mitigation techniques. Section 6 introduces the basics of GNSS receiver software. Section 7 presents RTKLIB, an open-source program package for GNSS positioning, and discusses a few approaches to improve its performance. The final section provides a summary of the paper and suggestions for further research.

2. Status of GNSS

In this section, the status of the satellite positioning system is discussed [3]. In the 1990s, the US GPS [4] and Russia's GLONASS [5] technologies were operational. As of 2023, six satellite positioning systems were in operation globally, spanning Europe and five non-European countries [6]–[9]. These multi-satellite positioning systems are collectively referred to as GNSS. Table 1 summarizes the operational status of each system in 2023. At this time, signals from 133 satellites were available. The term "global" in the operational status indicates that the satellite is designed to provide positioning services worldwide. For example, Japan's QZSS is termed a regional satellite because its service area covers Asia and Oceania. The Indian positioning satellites were not used for positioning calculations in this study; however, the other five systems were consistently in use.

Originally, positioning satellites were developed by the United States and the former Soviet Union for military purposes; China's BDS also has military applications. However, Europe, Japan, and India allegedly do not prioritize

Table 1 GNSS operational status (December 2023).

Country and name	Operated number and Bands
US GPS	32 (Global), L1/L2/L5
Russia GLONASS	24 (Global), G1/G2
EU GALILEO [6]	23 (Global), E1/E5a/E5AltBOC/E5b/E6
China BDS [7]	43 (Global), B1I/B2I/B3I//B1C/B2a/B2b
Japan QZSS [8]	4 (Regional), L1/L2/L5/L6
India NavIC [9]	7 (Regional), L5

Manuscript received April 2, 2024.
Manuscript revised June 11, 2024.
Manuscript publicized August 1, 2024.

[†]Tokyo University of Marine Science and Technology, Tokyo, 135-8533 Japan.
a) E-mail: nkubo@kaiyodai.ac.jp

DOI: 10.23919/transcom.2024EBI0001

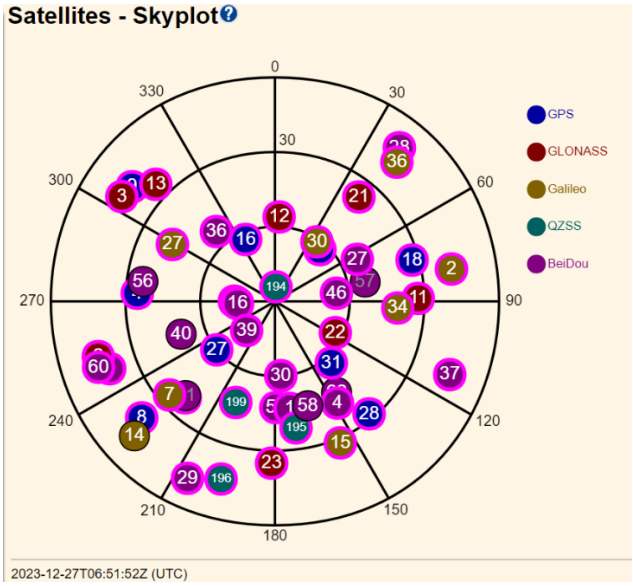


Fig. 1 Sky plots of GNSS (12/27/2023).

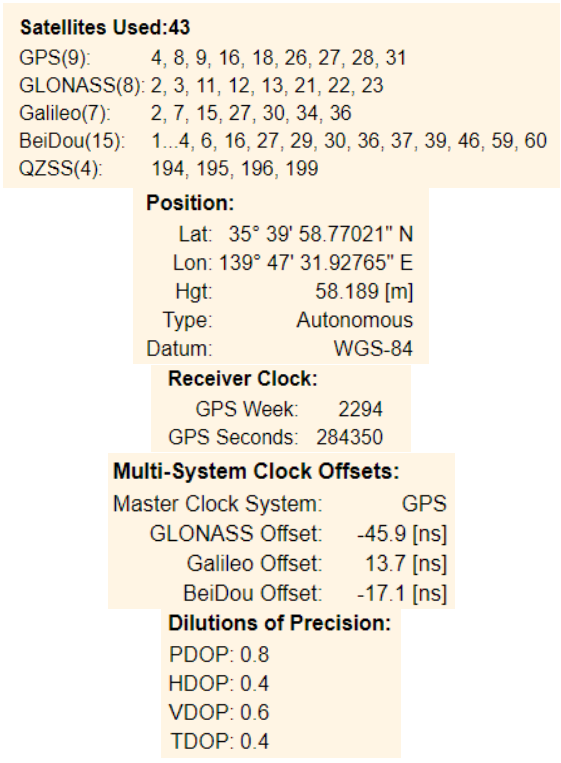


Fig. 2 Position Information (12/27/2023).

military use as their primary purpose. Interoperability, a crucial condition for user benefit, refers to the ability to use positioning satellites from different countries simultaneously. Open specifications for non-military signal formats enable general companies to develop GNSS receivers and utilize multiple systems from different countries to calculate the final position. The more satellites visible in the sky, the better the positioning accuracy, especially in urban environments with numerous obstacles. Figure 1 shows a GNSS sky plot obtained from the rooftop of a building on our campus in Tokyo. The GNSS receiver used is a Trimble Alloy.

Figure 2 outlines the positioning status. A total of 43 satellites were used for positioning. The type of position is "autonomous," meaning single-point positioning. The week represents the full GPS week number since the first epoch (January 6, 1980). The GPS seconds indicate epochs from 00:00:00 on Sundays (UTC) for one week. The clock offset represents the difference between GPS and other satellite systems. Because each system has its own master clock, these differences must be estimated when mixing positioning satellites from different constellations. When GPS and GLONASS are mixed, at least five satellites (one from each constellation) are required because the unknown parameters include the three positions (x, y, z), receiver clock offset, and clock offset between GPS and GLONASS. The dilution of precision (DOP) specifies error propagation as a mathematical effect of navigation satellite geometry. If the value is below 1, the positioning accuracy is higher than the accuracy of each satellite measurement.

In the BDS, more than 40 satellites operate simultaneously, combining second- and third-generation satellites. This development has been rapid in China over the last 5 to 10 years. Notably, the Russian GLONASS satellite transmits both FDMA-modulated and CDMA-modulated signals. With an increase in the number of such signals, full-scale service of CDMA signals is expected in the future. QZSS is a unique positioning satellite with a signal compatible with GPS, characterized by long stays at high elevation angles in the Asia-Oceania region. It is the first satellite in the world to transmit correction data, enabling centimeter-level accuracy, and its utilization is highly anticipated. The Japanese government's Strategic Headquarters for Space Policy plans to increase the number of QZSS satellites from 4 to 7 by 2024 and 2025, with a future goal of 11 satellites [10]. These 11 satellites will facilitate 24 h positioning in Japanese territory using only QZSS without GPS. Galileo also provides free-of-charge high-accuracy precise point positioning (PPP) corrections using its signal [11]. This operation is called a high-accuracy service (HAS, which offers improved real-time user positioning performance down to the decimeter level. GPS includes a mixture of satellites from different generations, with the oldest Block IIR satellite (PRN16) launched in 1997. Remarkably, 26 years have passed since then. The names of each generation of GPS satellites were NavStar I and II before Block IIA, followed by IIA, IIR, IIR-M, IIF, and III. Six GPS-III satellites have been launched.

In addition to the space segment described above, the satellite positioning infrastructure is divided into three major segments: the control segment, which manages the positioning satellites, and the user segment, which is used by the users. In terms of operating a positioning satellite, the role of the control segment is more important than that of the actual satellite. The potential of each country's positioning

satellites can be partially expressed by a signal in the space segment.

3. Basics of Precise Positioning

In this section, the error sources of GNSS and the three types of precise positioning: RTK-GNSS, PPP-RTK, and PPP, are discussed.

3.1 Measurements

GNSS pseudo-range and carrier phase measurements are expressed as follows:

$$P_{rov}^{sv1} = \rho_{rov}^{sv1} + c\,(dt_{sv1} - dT_{rov}) + ion_{rov}^{sv1} + tropo_{rov}^{sv1} + MP_{rov}^{sv1} + NOIS\,E_{rov}^{sv1} \quad (1)$$

$$\varphi_{rov}^{sv1} = \rho_{rov}^{sv1} + c\,(dt_{sv1} - dT_{rov}) - ion_{rov}^{sv1} + tropo_{rov}^{sv1} + N_{rov}^{sv1} + mp_{rov}^{sv1} + noise_{rov}^{sv1} \quad (2)$$

The superscript sv1 identifies the terms associated with the GNSS satellite, while the subscript rov denotes a term associated with the rover. Equation (1) indicates the pseudo-range measurements, and Eq. (2) indicates the carrier-phase measurements. In these equations, c is the speed of light, Dt is the satellite clock error, dt is the receiver clock error, ion is the ionospheric error, $tropo$ is the tropospheric error, MP is the pseudo-range multipath error, and mp is the carrier-phase multipath error. $NOISE$ is the pseudo-range noise and $noise$ is the carrier-phase noise. N is the integer ambiguity of the carrier phase measurement [12].

3.2 Error Sources

As shown in the equations, six major sources of error exist in GNSS observations. The first two errors originate from the satellite: the satellite position and clock errors. The satellite broadcasts parameters for estimating its position and clock errors, and the residual error after these calculations is the satellite error. The next two errors are related to the atmosphere: the ionospheric and tropospheric delays. Parameters for estimating the ionospheric delay are broadcasted and can be applied to reduce the error. If dual-frequency observations are available, an artificial observation that generates ionosphere-free coupling can be used to eliminate ionospheric delay. The ionospheric delay is strongly related to solar activity; it is minimal during periods of low solar activity but can cause significant errors when solar activity increases, with large differences between the delays corresponding to periods of strong and weak solar activity.

Ionospheric disturbances caused by solar flares make predictions difficult. The tropospheric delay is divided into dry and wet components. While the dry term can be estimated with high accuracy using model equations, accurately estimating the wet term is challenging, particularly during rainy and humid seasons. The error in the vertical direction is typically between 2.3 to 2.5 meters and does not vary significantly. However, during lightning events, satellite signals arriving from the direction of the lightning can be affected, complicating the achievement of RTK FIX solutions. The other two errors are the multipath error and receiver noise. Multipath error is particularly troublesome, and various methods have been proposed to mitigate it, with the best expected to become standard. The error due to receiver noise depends on the received signal strength, and the error can be estimated accurately using that strength [13]. Both multipath error and receiver noise are strongly related to the internal design of the GNSS receiver. Table 2 presents the six sources of error.

Table 2 Typical errors in GNSS measurements.

Source	Potential error size
Satellite clock model	1 m (rms)
Satellite ephemeris prediction	1 m (rms)
Ionospheric delay	2–10 m in zenith direction
Tropospheric delay	2.3–2.5 m in zenith direction
Multipath (Open sky)	Code: 0.5–1 m Carrier: 0.5–1 cm
Receiver noise	Code: 0.25–0.5 m (rms) Carrier : 1–2 mm (rms)

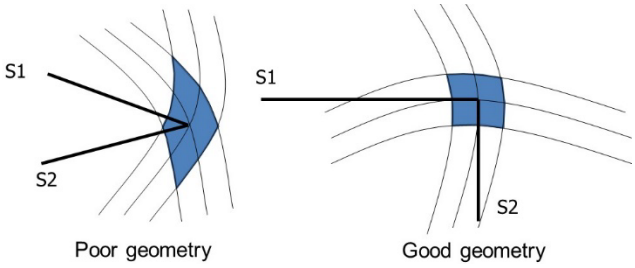


Fig. 3 Poor and good geometry.

Additionally, an error is induced by satellite geometry. Ideally, if the distance measurement from the satellite to the receiver is accurate, satellite geometry should not cause errors. In reality, however, measurement errors always occur, and their effect on positioning accuracy depends on the satellite geometry. To address this, a satellite geometry-induced degradation factor called the dilution of precision (DOP) has been defined in GNSS. Figure 3 illustrates that if satellites s1 and s2 are placed in the same direction, the positioning error exceeds the original ranging error. Conversely, if the satellites are placed far apart, the positioning error is closer to the original ranging error. For example, if an antenna is located near a wall, it can only receive signals from satellites on one side of the wall, resulting in increased vertical positioning errors. The DOP can be estimated during the positioning operation using the least-squares method and is uniquely determined by the satellite geometry at any given time. The actual GNSS error is the measurement error multiplied by the DOP. The sources of measurement error are listed in Table 2. For example, the horizontal error is the measurement error multiplied by the horizontal dilution of precision (HDOP). The vertical error is the measurement error multiplied by the vertical dilution of precision (VDOP).

The 3D position error is the measurement error multiplied by the position dilution of precision (PDOP).

In urban areas, the HDOP can exceed 10 owing to narrow streets and high-rise buildings. In such cases, a 2–3 meter position error from measurement inaccuracies without multipath interference can potentially become a 20–30 meter error due to the DOP. Additionally, the VDOP can easily exceed 10 in these environments, leading to significant height errors. An error ellipse can estimate the direction in which the error spreads, which is crucial for assessing actual errors. For further details, refer to the study by Fan and Ma [14].

3.3 Stand-Alone Positioning

Standalone positioning primarily uses pseudo-range measurements without correction data from any external source. This method relies on the satellite's broadcast information for positioning. In standalone positioning, carrier-phase measurements can help reduce noise in pseudo-range measurements. However, without correcting for satellite clock and atmospheric errors, the positioning error can be substantial. The satellite clock error, even with an atomic clock, must be corrected using broadcast information; otherwise, the position error can exceed 1 km. After correcting for satellite clock errors with broadcast parameters, the remaining errors depend on the accuracy of the ionospheric and tropospheric delay estimations. As shown in Table 2, these errors range from 2 m to several meters vertically and are estimated using an obliquity factor based on the elevation angle. When errors are estimated from these models, standalone positioning accuracy can be within a few meters when the multipath error and receiver noise are minimal. However, if atmospheric errors are not considered, the horizontal positional error may still be a few meters, but the vertical error can exceed 10 m. This discrepancy occurs because positioning satellites are always above the user's position; while horizontal errors may cancel out owing to signals from various azimuths, vertical errors do not cancel out. Both ionospheric and tropospheric delays contribute to the original true ranges.

GPS has been popular since its development, but estimating the receiver clock error has been problematic. The initial GNSS inventors proposed estimating the receiver clock error during positioning, meaning that 3D positions and receiver clock errors would be estimated simultaneously. The least-squares method is used to estimate the unknown position in three dimensions and the receiver clock errors. A typical standalone result after 24 h is presented in Fig. 4. Data were obtained from the rooftop of our laboratory using a Trimble Alloy receiver in February 2024. The analysis software used was RTKLIB with default parameters. The GPS, QZSS, GLONASS, and GALILEO systems were used.

In multi-GNSS, positioning is performed simultaneously with satellites that have time standards different from those of GPS. For example, when using GPS and GLONASS in standalone positioning, the difference between the GPS and GLONASS time standards must be estimated simultaneously. Therefore, in the case of GPS + GLONASS positioning, an additional unknown (the time difference between GPS and GLONASS) is added to the four unknowns (three position coordinates and the receiver clock error), necessitating observation values from at least five satellites. Five GPS satellites alone are insufficient; at least one satellite from each system is required. Thus, for GPS + GLONASS + QZSS + GALILEO + BDS positioning, at least seven satellites are needed, with at least one from each system. Although the discussion above focuses on GPS, positioning can also be performed using QZSS or GLONASS alone. However, because GPS is widely used worldwide, many receivers use the GPS time as the basis for positioning.

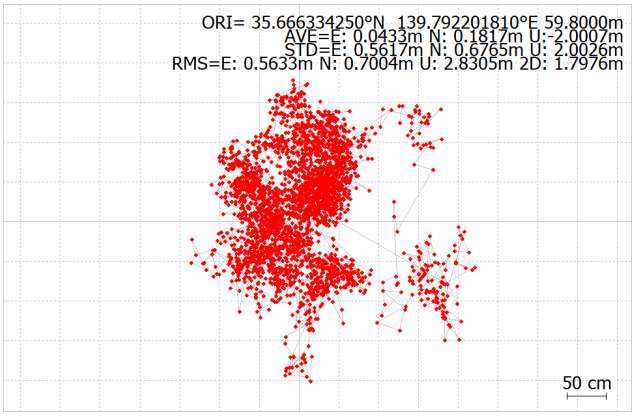


Fig. 4 Typical result of stand-alone positioning.

3.4 DGNSS

The differential global navigation satellite system (DGNSS) uses pseudo-range as the main observation and correction data from a nearby base station to determine the position of the user station. Of the six errors mentioned earlier, the satellite position error can be almost eliminated because both the base and user stations use the same satellite, and the satellite clock error can be completely eliminated using correction data from the same time. In practice, a slight time delay occurs when the user station uses the correction data from the base station, so the satellite clock error must be corrected for that time. The satellite clock error's time variation can be approximated by a linear function over a short period of approximately 10 s, which helps to nearly eliminate the time difference. Ionospheric and tropospheric delays can be almost completely eliminated if the correction data from the base station can be obtained within approximately 1 min, provided the baseline is not too long. This is because the errors due to these factors change slowly both spatially and temporally.

However, the tropospheric delay varies significantly with elevation angle, necessitating compensation for the error using the obliquity factor based on the elevation angle.

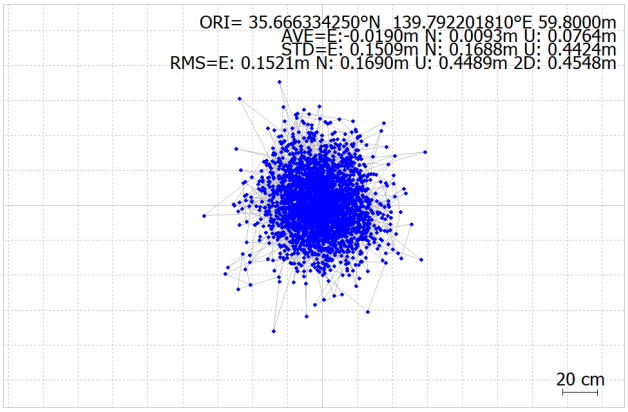


Fig. 5 Typical result of DGNSS (10 km).

For example, if the base and user stations are approximately 100 km apart, the elevation angle differs by approximately 1°, even if the same satellite is observed. The error in the tropospheric delay due to this 1° difference in elevation angle cannot be ignored. The same applies to the ionospheric delay, but its impact is smaller. Determining the baseline length limitations is complicated by the nature of atmospheric errors. When solar activity is low, a 100 km baseline length for DGNSS is feasible as long as an error of approximately 1 meter is acceptable. A typical DGNSS result obtained after 24 h is shown in Fig. 5. Data were collected from the rooftop of our laboratory using a Trimble Alloy receiver in February 2024. RTKLIB software, with default parameters, was used for analysis. The GPS, QZSS, GLONASS, and GALILEO systems were utilized, with the Ichikawa Station of GEONET serving as the base station. The baseline length was approximately 10 km.

There was a significant difference in accuracy between DGNSS with a baseline length of 10 km and that with a baseline length of 100 km. This difference is particularly noticeable in areas with a large spatial gradient of the ionosphere. At distances of 100 km or more, the offset effect of the spatial correlation of atmospheric errors may diminish, and the horizontal bias may exceed 0.1 m. These variations are due to natural atmospheric phenomena, making it difficult to estimate the magnitude of the error over a 100 km distance. The magnitude of the ionospheric delay is especially dependent on solar activity, which follows an 11-year cycle. When solar activity is high, the gradient concerning the horizontal distance can also be large, necessitating caution when using DGNSS. During localized ionospheric disturbances, using satellite signals that pass through the disturbance location can reduce the effectiveness of DGNSS. Table 3 summarizes the extent to which the DGNSS method can reduce various error factors.

3.5 RTK-GNSS

RTK-GNSS is a precise positioning method that uses carrier-phase measurements as the main observation value and obtains correction data from a nearby base station to estimate the user's position with centimeter-level accuracy. RTK-GNSS employs a double-difference observation technique. Figure 6 illustrates double-difference observations. Equation (3) provides the formula for generating the double-difference observation, which eliminates the clock error from both the satellite and receiver clocks. The notation follows the same conventions as in Eqs. (1) and (2).

Table 3 Typical error mitigations in DGNSS (10 km).

Source	Residual error after DGNSS
Satellite clock model	0.0 m (rms)
Satellite ephemeris prediction	Below 0.1 m (rms)
Ionospheric delay	Below 0.1 m in zenith direction Normal solar activity
Tropospheric delay	Below 0.1 m in zenith direction
Multipath (Open sky)	Uncorrelated between antennas Normally decimeter-level
Receiver noise	Uncorrelated between receivers Normally decimeter-level

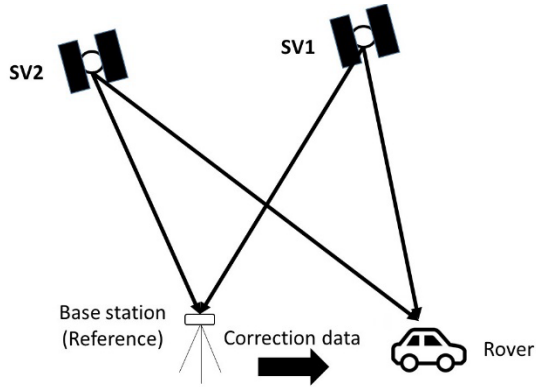


Fig. 6 Double-difference operation.

$$
\begin{aligned}
&\varphi_{rov\_ref}^{sv1\_sv2} = \left(\varphi_{rov}^{sv1} - \varphi_{ref}^{sv1}\right) - \left(\varphi_{rov}^{sv2} - \varphi_{ref}^{sv2}\right) \\
&= \rho_{rov}^{sv1} + c\left(dt_{sv1} - dT_{rov}\right) + ion_{rov}^{sv1} + tropo_{rov}^{sv1} + N_{rov}^{sv1} + mp_{rov}^{sv1} + noise_{rov}^{sv1} \\
&- \left[\rho_{ref}^{sv1} + c\left(dt_{sv1} - dT_{ref}\right) + ion_{ref}^{sv1} + tropo_{ref}^{sv1} + N_{ref}^{sv1} + mp_{ref}^{sv1} + noise_{ref}^{sv1}\right] \\
&- \left[\rho_{ref}^{sv2} + c\left(dt_{sv2} - dT_{rov}\right) + ion_{rov}^{sv1} + tropo_{rov}^{sv2} + N_{rov}^{sv2} + mp_{rov}^{sv2} + noise_{rov}^{sv2}\right] \\
&+ \left[\rho_{ref}^{sv2} + c\left(dt_{sv2} - dT_{ref}\right) + ion_{ref}^{sv2} + tropo_{ref}^{sv2} + N_{ref}^{sv2} + mp_{ref}^{sv2} + noise_{ref}^{sv2}\right] \\
&= \rho_{rov}^{sv1} - \rho_{ref}^{sv1} + \rho_{rov}^{sv2} - \rho_{ref}^{sv2} + N_{rov}^{sv1} - N_{ref}^{sv1} + N_{rov}^{sv2} - N_{ref}^{sv2} \\
&+ \left(ion_{rov}^{sv1} + tropo_{rov}^{sv1}\right) - \left(ion_{ref}^{sv1} + tropo_{ref}^{sv1}\right) \\
&- \left(ion_{rov}^{sv2} + tropo_{rov}^{sv2}\right) + \left(ion_{ref}^{sv2} + tropo_{ref}^{sv2}\right) \\
&+ \left(mp_{rov}^{sv1} + noise_{rov}^{sv1}\right) - \left(mp_{ref}^{sv1} + noise_{ref}^{sv1}\right) \\
&- \left(mp_{rov}^{sv2} + noise_{rov}^{sv2}\right) + \left(mp_{ref}^{sv2} + noise_{ref}^{sv2}\right) \qquad (3)
\end{aligned}
$$

As seen from the equation, atmospheric error is negligible if the baseline is short. Estimating the integer ambiguity using only carrier-phase measurements is challenging; however, with this technique, the integer ambiguity can be easily estimated using the integer least squares method, resulting in centimeter-level accuracy. Consider a case where the base and rover stations are in the same position (antenna). The baseline vector is zero, and in Eq. (3), only the integer ambi-

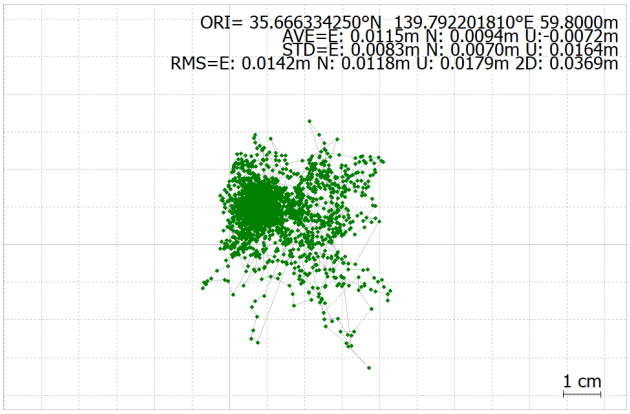


Fig. 7 Typical result of RTK-GNSS (10 km).

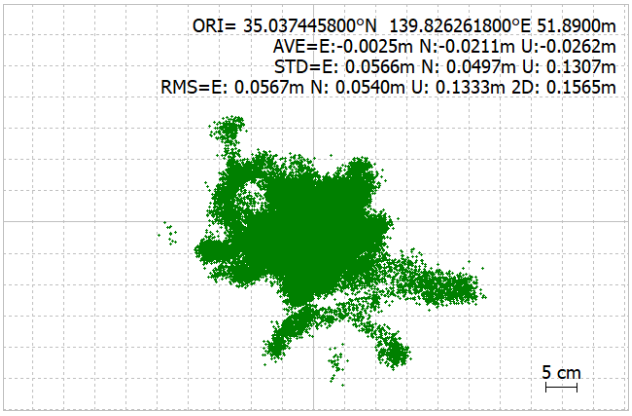


Fig. 8 Typical result of wide-lane RTK (70 km).

Table 4 Typical error mitigations in RTK-GNSS (10 km).

Source	Residual error after RTK-GNSS
Satellite clock model	0.0 m (rms)
Satellite ephemeris prediction	0.0 m (rms)
Ionospheric delay	Below 0.01 m (rms)
Tropospheric delay	Below 0.01 m (rms)
Multipath (Open sky)	Uncorrelated between antennas Normally below 0.01 m
Receiver noise	Uncorrelated between receivers Normally mm-level

guity of the double difference of the carrier phase remains. Consequently, the integer ambiguity is known. Additionally, the multipath error is completely eliminated, leaving only the noise of the carrier-phase measurement. Thus, several important points can be learned by observing the zero-baseline double difference. Typical RTK-GNSS results after 24 h are shown in Fig. 7. Data were obtained from the rooftop of our laboratory using a Trimble Alloy receiver in February 2024. RTKLIB software was used for analysis, with parameters set to default. GPS, QZSS, GLONASS, and GALILEO were used. The base station is the same as that used in the DGNSS test. The fixed rate was 91.6%. Table 4 summarizes the degree to which the RTK-GNSS method can reduce various error factors. Under normal conditions, a bias or error of approximately 1 cm can be expected for every 10 km baseline extension, presumably allowing an RTK-FIX solution to be obtained.

Consequently, the baseline vector from the base station to the user station can be estimated with centimeter-level accuracy. This is the essence of RTK-GNSS. As the baseline length increases, atmospheric errors become significant, with errors exceeding approximately 1 cm at 10 km. This is an estimate. When the baseline length exceeds 30 km, the amount of error due to atmospheric delay becomes a significant obstacle, complicating the determination of the integer ambiguity. If the atmospheric error can always be corrected to 1 cm even with a 100 km baseline, long baseline RTK becomes feasible. However, the increasing atmospheric error with longer baselines complicates long baseline RTK. VRS [15] and PPP-RTK have been developed to improve this situation.

PPP-RTK will be described later in this paper. Generally, short-baseline RTK is used for baseline lengths up to approximately 10 km, medium-baseline RTK for lengths up to 100 km, and long-baseline RTK for lengths exceeding 100 km. To smoothly estimate the integer ambiguity in RTK for medium and long baselines, the concept of wide lanes is employed. A wide lane indicates that the virtual wavelength exceeds that of a single L1 or L2 band. For example, if a virtual signal of L1–L2 is generated, the frequency is approximately 348 MHz and the virtual wavelength is approximately 86 cm, which facilitates the estimation of integer ambiguity. Additionally, for baselines exceeding 100 km, the Melbourne–Wübbena linear combination [16] can be used, enabling the accurate estimation of wide-lane integers even over long distances. However, even if wide-lane integers can be estimated accurately, the atmospheric error in the double-difference observation of the carrier phase measurement cannot be eliminated. Therefore, the position estimated using the wide-lane integers contains certain errors, complicating medium-length baseline RTK and making long-baseline RTK relatively difficult to achieve. While atmospheric errors and baseline length have been discussed, multipath errors also remain. Multipath errors are particularly troublesome because they are difficult to predict and mitigate. Multipath errors affect RTK performance regardless of the baseline length. The biggest challenge for short-baseline RTK is multipath error. Typical RTK-GNSS results for 24 h using a wide lane are shown in Fig. 8. Data were obtained from the rooftop of our laboratory using a u-blox F9P receiver in February 2024. Modified RTKLIB software was used for analysis, with the mask angle set to 30°. GPS, QZSS, GLONASS, BDS, and GALILEO were used. The base station was the Tateyama station in Chiba Prefecture, with a baseline length of approximately 70 km. The fixation rate was 93.9%.

3.6 PPP-RTK

PPP-RTK provides centimeter-level positioning accuracy with fewer base stations than conventional RTK. The ac-

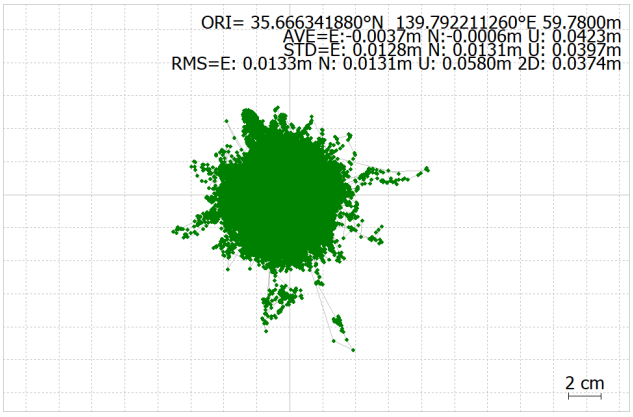


Fig. 9 Typical result of PPP-RTK using QZSS correction.

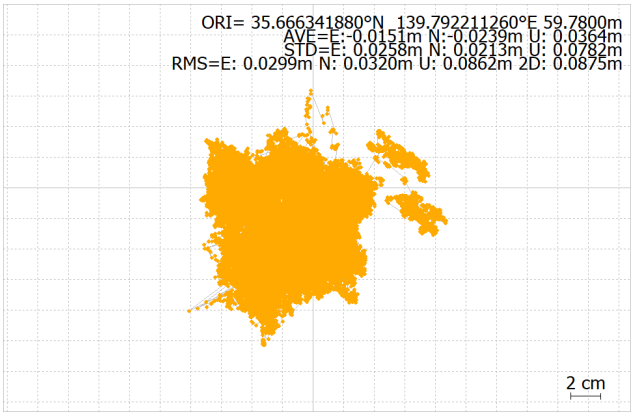


Fig. 10 Typical result of PPP using QZSS correction.

tual accuracy is approximately 1 cm RMS in the horizontal direction, which is slightly less optimized than RTK. PPP-RTK is a correction method for the Japanese centimeter-level augmentation service (CLAS) and the centimeter-level correction method of QZSS. Specifically, GNSS observation data from multiple base stations with a mid-baseline of approximately 50 km are collected, and correction data are generated and broadcast. The expertise of each PPP-RTK service provider is incorporated into generating correction data to minimize loss of accuracy and data volume. The correction data include the satellite's precise orbit, precise clock, and ionospheric and tropospheric delays. The ionospheric delay is corrected by generating a grid with latitude and longitude and providing a vertical ionospheric delay on the grid. Similar correction data are provided for the tropospheric delay, reducing the volume of data to be transmitted. Although centimeter-level positioning cannot be achieved instantaneously as with RTK, it can be determined within approximately 1 min. It is called PPP-RTK because the correction data are broadcast separately for the satellite's precise orbit, clock, and atmospheric delays, similar to PPP, while the user-side positioning is determined by integer ambiguity, as in RTK. If user-side positioning uses the generation of a double-difference observation, it is also similar to RTK. Several companies have announced PPP-RTK services. For details of the PPP-RTK method, refer to the article by engineers at Mitsubishi Electric Corporation, the service provider [17], [18]. Typical PPP-RTK results after 24 h are shown in Fig. 9. Data were obtained from the rooftop of our laboratory using a Core AsteRx4 receiver in February 2024. The fixation rate was 99.3%.

3.7 PPP

In contrast to conventional RTK and PPP-RTK, PPP generally uses only precise information on the satellite's orbit and clock as correction data, with the receiver's positioning software estimating the rest to achieve centimeter-level accuracy. The ionospheric-free combination is the most common method, although a method that does not use an ionospheric-free combination and instead uses ionospheric estimation to converge has also been devised. Starting with the model, the tropospheric delay can be estimated more accurately because the error model for tropospheric delay is precise. The dry delay can be estimated at the centimeter level using this error model. Integer ambiguity determination also exists in PPP positioning, known as precise point positioning ambiguity resolution (PPP-AR). If PPP-AR is used, it can achieve the same accuracy as PPP-RTK. In contrast, PPP without PPP-AR generally achieves a horizontal accuracy of less than 10 cm. The convergence time for normal PPP is approximately 15–20 min to achieve a horizontal accuracy of 10 cm. However, Trimble RTX, a commercial PPP service, achieves convergence in 2–3 min, as confirmed using the rooftop antenna at our laboratory and a Trimble NetR9 receiver. This performance, achieved with a correction amount of below 1 kbps, suggests there may be a key factor in shortening convergence time. In commercial services, because the base station and user receiver can be specified, their hardware biases are known, reducing the need to consider them as unknown quantities. Often, the same model is used, allowing full use of multi-GNSS, securing the number of satellites. General open-source PPP faces difficulties in convergence time, requiring approximately 15–20 min to obtain a horizontal positioning solution of less than 10 cm owing to a shortage of precise atmospheric delay information. Therefore, using PPP for mobile applications is challenging. However, as long as satellite signals are not interrupted by obstacles, centimeter-level positioning can be maintained.

Typical PPP results for 24 h are shown in Fig. 10. Data were obtained from the rooftop of our laboratory using an MSJ 3008-GM4-QZS receiver in February 2024, with correction data from QZSS [17], [19]. For example, Fig. 11 shows the PPP results for crustal movement during the Noto earthquake. The open-source software MADOCALIB was used in this study [20]. PPP positioning using Geospatial Information Authority of Japan (GSI) GEONET data during a large earthquake can accurately estimate the crustal deformation at the antenna locations to within centimeters, regardless of the baseline length. The results were com-

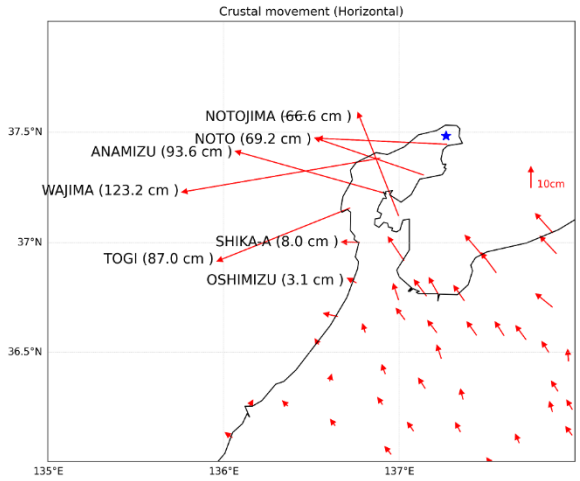


Fig. 11 PPP results at Noto earthquake (2024/1/1).

Table 5 Characteristics of RTK/PPP-RTK/PPP.

	RTK-GNSS	PPP-RTK	PPP
Horizontal Accuracy	Below 1 cm	1–2 cm	2–5 cm
Correction	Mainly 4G/5G	Satellite	Satellite
Density of Base Stations	10–20 km	50–100 km	Approx. 1,000 km
Convergence	Instantaneous	Below 1 min.	15–20 min.
Convergence (Short Gap)	Below 3–5 sec.	Below 10 s	Below 5 min.
Availability (Open Sky)	Over 99%	Over 99% (depends on solar activity)	Over 99% (convergence time is required)
Availability (Semi-urban)	Over 90%	60–70%	Below 50%
Availability (Dense-urban)	60–70%	10–30%	Impossible

Correction data must be received.

pared with the official GSI results, and the differences were within a few centimeters. In RTK, this accuracy is not always possible because the base station moves. If a base station is located over 100 km from the earthquake location, a sophisticated software is required. Table 5 summarizes the characteristics of RTK-GNSS, PPP-RTK, and PPP, all of which provide centimeter-level augmented positioning. For more detailed information about correction messages for RTK/PPP-RTK/PPP, refer to this article [21].

4. Quasi Zenith Satellite System (QZSS)

QZSS, also known as Michibiki, is a Japanese positioning satellite system with a unique orbit. The first satellite, launched in 2010, was recently retired and replaced by a new satellite launched in 2021. The second, third and fourth satellites were launched in 2017. Currently, a four-satellite constellation is operated by the Cabinet Office. QZSS adopted a quasi-zenith orbit ahead of other countries, and at the time of its launch, RTK performance in urban areas was greatly improved using GPS and QZSS in the zenith direction. Although the effect is less visible now owing to the increase in the number of other positioning satellites, the advantage of being at the zenith remains significant. QZSS

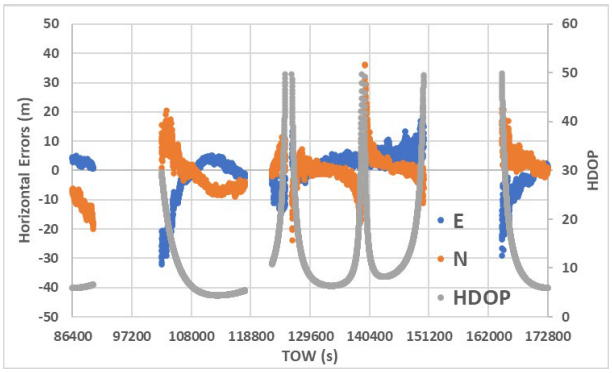


Fig. 12 QZSS only stand-alone positioning.

not only is a complement to GPS but also has a positioning augmentation aspect, broadcasting SLAS and CLAS correction data for free since 2018. SLAS is equivalent to DGNSS with a horizontal accuracy of less than 1 m. CLAS is equivalent to PPP-RTK and was the first to freely broadcast PPP-RTK correction data worldwide. PPP has also been broadcasting free of charge on a trial basis and will be officially operational soon, under the name MADOCA-PPP. Although not widely known, QZSS provides a public regulated service that transmits concealed and encrypted signals, usable only by government-approved users to avoid signal jamming and spoofing. A similar signal exists for Galileo satellites, which do not have a military signal.

4.1 QZSS as a Positioning Complement

As a positioning complement, QZSS maintains compatibility with GPS by broadcasting almost the same signals as GPS, making the composition of GPS + QZSS easy to understand. Even with the current four-satellite system, positioning with QZSS alone is possible for approximately 15 h a day in the vicinity of Tokyo, with an actual horizontal accuracy of less than 30 m. Figure 12 shows the positioning results using only four QZSS satellites, with observation data acquired by a u-blox F9P receiver using an antenna on the laboratory roof. Evidently, the horizontal accuracy depends on the DOP, a degradation factor of satellite geometry. Looking at the improvement of RTK-GNSS in dense urban areas, the fix rate can increase approximately 5–10% by simply adding QZSS to other GNSSs because at least one or two QZSS satellites stay at a high elevation angle. Units 5, 6, and 7 of QZSS will be launched within the next 1–2 yrs. The expected satellite configuration at the time of the constellation for the 11 satellites is shown in Fig. 13 [22].

4.2 QZSS as Positioning Augmentation

As mentioned previously, SLAS, CLAS, and PPP are the most common types of positioning augmentation. As the CLAS and PPP test results were already shown in the previous section, Fig. 14 shows the 24 h horizontal results of the SLAS acquired at the rooftop of the laboratory in Febru-

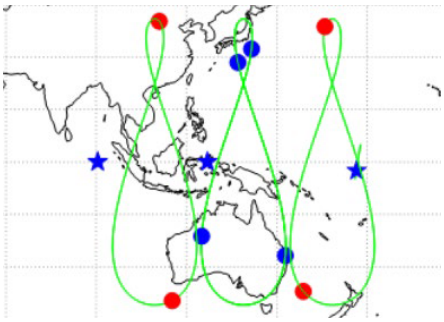


Fig. 13 Eleven satellites constellation for future QZSS.

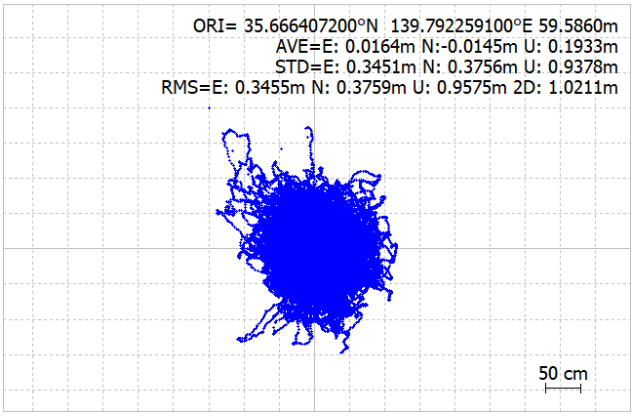


Fig. 14 Typical result of SLAS using QZSS correction.

ary 2024. The data were acquired using a u-blox F9P receiver. The base station for SLAS was located in Hitachiota, Ibaraki, Japan. Except for the ionospheric effects of solar activity, the results were stable. The horizontal RMS of SLAS was approximately 50–60 cm when the results were stable. Please refer to detailed long-term results of these three methods [23]. For PPP, I collaborated with overseas universities, mainly in Southeast Asia, to conduct long-term evaluations using the same receivers as those in my laboratory. These results are almost as accurate as those obtained in my laboratory, and they are available at the Laboratory Home Page website [24]. All the previous results were verified. Because solar activity is expected to reach a maximum in the near future, GNSS positioning is expected to be more difficult, and the performance of SLAS and CLAS will be especially noteworthy. The solar activity forecast provided by NASA is shown in Fig. 15 [25].

5. Multipath Error Mitigation

In this section, several methods for mitigating multipath errors in pseudo-range measurements, which represent the weakest aspect of the GNSS, are introduced. Although the multipath error in carrier-phase measurements is also important, its magnitude is only a few centimeters; therefore, it is not presented here [26]. The impact of multipath errors in Doppler frequency measurements is illustrated below [27].

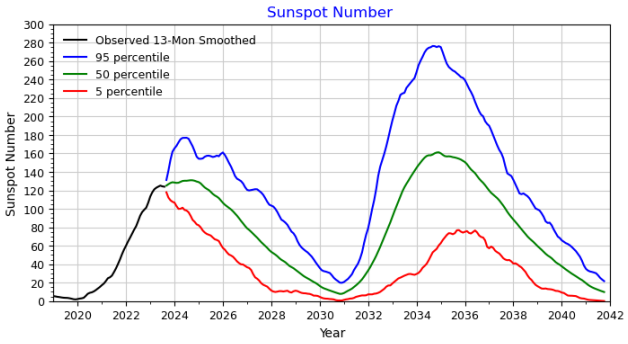


Fig. 15 Solar cycle forecast by NASA.

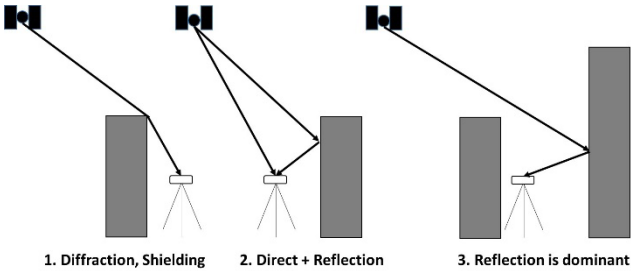


Fig. 16 Three typical patterns of multipath.

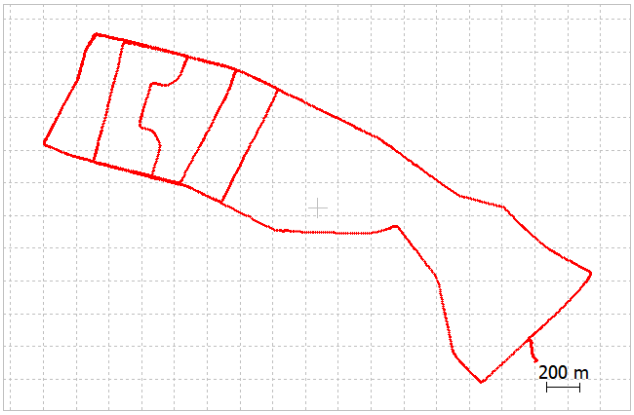


Fig. 17 Ground track produced using smartphone.

Figure 16 depicts the three patterns of multipath cases. The third case in these patterns, where the level of the reflected signal exceeds that of the direct signal, is a major source of error and is termed a non-line-of-sight (NLOS) signal.

Subsequently, I assessed the test results obtained by placing a smartphone (Google Pixel 6) in front of the car navigation system on the dashboard of the vehicle. The smartphone results consist of the NMEA data acquired by the GNSS Logger (an Android application provided by Google), and I did not investigate whether the results were output in real time. Figure 17 illustrates the ground track of the test route, while Fig. 18 presents the actual horizontal errors. These errors were estimated using the Applanix POSLVX system, a compact, fully integrated, turnkey position and orientation system that utilizes integrated inertial

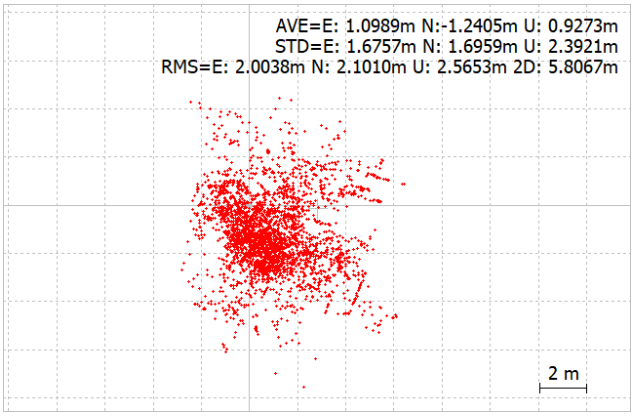


Fig. 18 Horizontal errors by smartphone.

technology to generate stable, reliable, and repeatable positioning solutions for land-based vehicle applications. Notably, the maximum error is only approximately 8 m, with a root mean square (RMS) of 2.9 m, even when using only the smartphone.

The Nihonbashi and Marunouchi areas boast numerous high-rise buildings, presenting a challenge for achieving accuracy with the commercial receivers I evaluated. Remarkably, the accuracy achieved in these areas surpassed expectations. One reason for this outcome is that in such environments, the signal often includes reflections, complicating location determination with GNSS receivers alone, despite these receivers having antennas installed on the vehicle roof, which typically provides a more optimized environment than smartphones. There are several potential explanations for these results. First, smartphones rely on positioning by mobile base stations, and urban areas are abundant in mobile base station antennas, integrated with their positioning services. Second, smartphones utilize the IMU sensor built into them, integrating it with positioning data. Thirdly, they leverage the map-matching function of Google Maps. Lastly, they incorporate information from 3D maps. I believe there are various methods for mitigating large multipath errors by utilizing additional information, even though doing so with GNSS alone is challenging. Details on the Google Smartphone Decimeter Challenge, a competition focusing on accuracy using smartphone observation data, can be found in the report by Chow et al. [28]. Additionally, Odolinski et al. demonstrated the feasibility of RTK-GNSS using smartphone observation data (Google Pixel 6) [29].

Below, I introduce several multipath error mitigation methods for GNSS receivers. I cannot cover all mitigation methods here; therefore, for more information on reducing multipath errors, please refer to the report by Smolyakov et al. [30]. Recently, factor graph optimization (FGO) has also been employed to enhance accuracy in challenging areas [31].

5.1 Mitigation Using GNSS Antenna

The antenna serves as the initial receiver of signals from the satellite, and mitigating multipath effects here is crucial. One approach is to minimize the impact of multipath signals originating from the lower hemisphere of the antenna. Because positioning satellite signals invariably come from above the user, receiving signals from the upper hemisphere is essential. Conversely, signals arriving from the bottom of the antenna are considered multipath. Hence, my goal was to suppress signals originating from the bottom of the antenna as much as possible. Various methods have been developed to achieve this. One widely used method employs an antenna design known as a choke ring, particularly favored for GNSS base stations requiring high accuracy. Another approach utilizes a ground plane to suppress multipath signals. Figure 19 illustrates two antenna examples: one featuring a ground plane on the left and a choke ring on the right. To assess multipath errors originating from the bottom of the antenna, we evaluated the accuracy of pseudo-range output from the GNSS receiver under two conditions: with the antenna placed on the ground in an open sky and when elevated to a height of approximately 1 m. Accuracy is consistently increased when the antenna is grounded, as receiving signals from the bottom of the antenna becomes nearly impossible. The code-minus-carrier (CMC) metric offers a method for determining multipath errors [32]. By subtracting the carrier phase from the code (pseudo-range) and compensating for ionospheric effects, the multipath error can be observed in the pseudo-range. Another strategy leverages the transmission of satellite signals with right-handed circular polarization, which may shift to left-handed circular polarization depending on reflection from buildings. To counter this issue, some antennas are designed to receive right-handed circularly polarized signals [33]. Recently, compact and lightweight helical antennas have garnered attention for their application in drones and other fields. These antennas were evaluated for their multipath resistance performance and were found to outperform standard patch antennas [34].

Fig. 19 Two types of GNSS antennas.

5.2 Mitigation Using Signal Processing

Several methods have been developed to mitigate multipath effects in the signal processing section of GNSS receivers. Among these are techniques known as strobe and

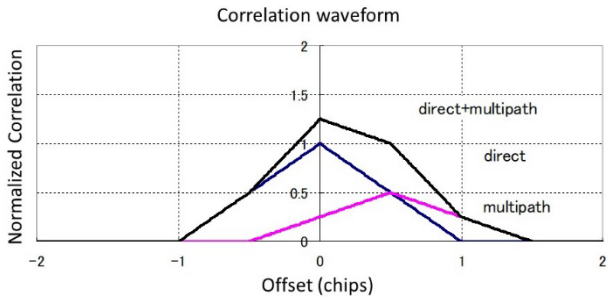


Fig. 20 Correlation waveform by direct + multipath.

pulse-aperture correlators, as detailed in specific reports [35], [36]. The strobe correlator operates by capturing the peak from the correlation waveform through multiple correlation points within a delay-locked loop (DLL). If the bandwidth of the GNSS receiver allows for signals from the satellite to be received adequately, the top of the correlation waveform can be sharply observed. For instance, with GPS L1-C/A signals, which exhibit a triangular correlation waveform, the apex of the triangle can be observed more precisely. Figure 20 illustrates the correlation waveform obtained from both the direct and one multipath signals, with a delay of approximately 150 m (0.5 chip) and a received signal level approximately half that of the direct signal. The final correlation waveform observed at the receiver closely resembles this waveform. The timing of the direct signal, unaffected by multipath, can be determined by extracting two correlation values at each end of the apex of the triangle. However, this method exhibits imperfections when the multipath delay is too close to the apex of the triangle, typically within 15 m. Moreover, if the level of the reflected signal exceeds that of the direct signal, this method becomes ineffective. A novel approach for capturing the distortion of the code rather than the correlation waveform has been proposed [37].

Another signal processing method for accurately estimating the timing of direct signal tracking involves separating the direct signal from the multipath signal using the correlation waveform, known as MEDLL [38]. Additionally, a method utilizing maximum likelihood estimation for separating direct signals from multipath signals has been presented [39]. In contrast to the strobe correlator, this method entails an estimation process and is not widely employed in commercial receivers, although it represents an important concept. If the receiver possesses sufficient computing power, continuously monitoring the correlation waveforms of the satellite signals once they are tracked is preferable. This approach enables seamless tracking even if the received signal level of the multipath surpasses that of the direct signal, as the buried location of the direct signal can be estimated on the basis of continuously observed correlation waveforms over time. Therefore, when the signal level of the direct signal returns to its original level, tracking can proceed smoothly.

5.3 Mitigation Using GNSS Measurements

Finally, let us delve into mitigation methods that utilize measurement data. Many of the techniques discussed thus far have been employed in survey-class receivers and GNSS units with additional computing power. The methods outlined in this section, however, are tailored for inexpensive survey-class receivers, constrained by hardware limitations. Below are several methods for mitigating multipath errors, especially concerning pseudo ranges.

5.3.1 Satellite Selection Based on Signal Level and Signal Tracking Information of Receiver

One established approach involves using signal levels to determine the usability of satellites and assigning weights to each satellite for positioning. A mask angle to establish the minimum elevation angle at which a satellite can be used is also often used. While adjusting the minimum elevation angle can be beneficial in RTK ambiguity resolution, it may be constrained by the mask angle or signal level due to partial ambiguity, which removes ambiguity from satellites with large covariance values [40]. Furthermore, information from the receiver's internal tracking loop, which can indicate the accuracy of the phase-locked loop (PLL), frequency-locked loop (FLL), or delay-locked loop (DLL) tracking, can be leveraged. For instance, in urban environments, differences in positioning performance are observed while using satellites locked by the PLL or DLL. In most cases, position and velocity accuracy are higher when using satellites with a locked PLL. However, owing to fewer available satellites, the positioning rate may slightly decrease. Therefore, detailed information about the tracking loop should be incorporated into positioning operations. Nevertheless, such information is sometimes unavailable; for example, RINEX, a common format for GNSS observation data, does not provide detailed tracking status information. To effectively utilize tracking status information for positioning, a thorough understanding of the positioning program and the receiver's tracking loop is essential. Note that PLL locks imply that, for instance, in the GPS L1-C/A signal, the carrier phase is locked to approximately 1/100th of a 19-centimeter wavelength, which generally indicates reasonably high-quality observations. However, in scenarios involving driving through high-rise buildings, proper signals are often obstructed, leading to significant errors in pseudo ranges even if the carrier phase is locked. In cases where only multipath signals are received, the carrier phase is likely to correspond to the reflected signals, even if it is locked.

5.3.2 Relationship between Speed and Multipath

According to our test results, strong multipath signals are more likely to be received when the speed is very low or zero. This aligns with the tracking loop process: when in

motion, the relationship between surrounding buildings and satellite signals changes rapidly, meaning if a direct signal is received, the multipath effect will be less pronounced. In a tracking loop, a continuous signal of approximately 20 ms or more is often integrated. However, if the signal pauses for more than a few seconds, a stable multipath signal may emerge, even in the presence of a direct signal. In such scenarios, the impact of multipath signals seems significant. If other sensors can detect a stop, accurately estimating the position becomes easier, with inertial measurement unit sensors serving the same purpose.

5.3.3 Carrier Smoothing

A smoothing method replaces pseudo-range noise with the time-difference value of the carrier phase measurement, which can mitigate multipath errors and pseudo-range noise. Typically, when the carrier phase is locked, the error due to multipath is only a few centimeters [26], considerably more optimized than pseudo-range accuracy. Setting the time constant for carrier smoothing is crucial; a time constant of 100 s, for instance, can significantly reduce the effect of multipath signals, which fluctuate over approximately 10 s. However, this method may not be universally applicable, as a 100 s time constant is suitable for relatively open environments, not urban driving conditions. Nonetheless, a similar effect to carrier smoothing can be implemented within the receiver. The accuracy of the receiving frequency derived from the PLL or FLL is notably superior to that derived from the DLL; therefore, changing the input frequency of the DLL's loop filter to a PLL- or FLL-derived receiving frequency can substantially reduce noise caused by the DLL, leading to a significant improvement in pseudo-range accuracy. While pseudo-range remains crucial for absolute measurement, this method is effective in reducing noise in distance information. Adjusting the DLL noise bandwidth to 2 Hz, for instance, provides a reciprocal filtering effect of 0.5 s. Some receivers allow manual adjustment of this noise bandwidth, with a default setting often around 0.05 Hz. The design of the PLL or DLL loop filter is critical, requiring significant ingenuity and expertise from receiver developers to produce accurate pseudo-ranges. Additionally, the Kalman filter can be employed in the receiver's internal tracking loop to precisely and continuously track the signal.

5.3.4 Integration Velocity and Position Information

With the global introduction of high-sensitivity receivers around 2000, their position trajectories have become smoother compared to conventional receivers. They often utilize velocity information derived from the time difference of carrier phase (TDCP) or Doppler frequency. Indeed, we observed a considerable improvement in position solutions by incorporating velocity obtained from TDCP or Doppler frequency and integrating it with position information using a Kalman filter with varying weights [41]. This enhancement is primarily attributed to the fact that carrier phase and Doppler frequency are less susceptible to multipath effects, resulting in smaller errors compared to pseudo-range measurements. While survey receivers typically achieve similar accuracy in velocity estimation from Doppler frequency and TDCP, in low-cost receivers, TDCP-derived velocity appears to be slightly more optimized. Doppler frequency is essentially an integral part of carrier-phase measurement, hence their principles are similar. However, the time interval plays a crucial role. Utilizing TDCP with a 1 s interval yields velocity accuracy of less than 1 cm/s, whereas velocity estimated from Doppler frequency over a few tens of milliseconds interval achieves accuracy in the range of a few cm/s. Although velocity and position are measured in different units, the accuracy of position derived from pseudo-range is often at the meter level or higher, especially in urban areas. The Kalman filter effectively integrates both velocity and position information. However, using TDCP becomes impractical in the presence of cycle slips.

6. Software GNSS Receivers

Typically, GNSS researchers focus on improving positioning accuracy using observation data output from commercial receivers. However, conducting in-depth research necessitates the development of GNSS receivers. Despite the challenge faced by students and researchers in developing GNSS receivers with performance comparable to commercial ones, the concept of software GNSS receivers has emerged, with textbooks being published and software GNSS receivers widely adopted among university researchers [42]. The source code is often available on GitHub, and manufacturers test them on FPGA and other devices before eventually selling them as chips in ASICs. Implementing current multi-GNSS and multi-frequency receivers on a PC is challenging due to computational speed requirements; however, Ifen, Germany, introduced a PC-compatible front-end software GNSS receiver. Additionally, Fraunhofer launched the Goose project, an FPGA-based GNSS receiver, providing researchers with an easy-to-use platform [43]. Figure 21 provides an overview of the process following the reception of signals from positioning satellites by the antenna.

Once the signal passes through a high-frequency section known as the front end and is converted into digital data, development of a GNSS receiver on a PC becomes feasible. Initially, in signal acquisition, the receiver captures the signal coarsely, estimating the approximate code correlation and Doppler frequency. Subsequently, in the signal-tracking stage, a finer resolution is employed to track the code correlation and Doppler frequency, while also per-

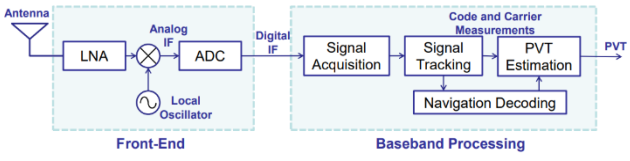


Fig. 21 Overview of the process of GNSS-SDR.

forming carrier-phase tracking and decoding of navigation messages. This stage also involves estimating pseudo-range and carrier-phase measurements, which constitute the foundation of GNSS observation data. Finally, the receiver estimates its position using the observed measurements.

Various methods for position estimation have been delineated. External correction data may be necessary in some cases. Since the GNSS receiver generates pseudo-range and carrier-phase measurements crucial for positioning estimation, comprehending the portion responsible for generating these observations is pivotal for improving position accuracy. For instance, understanding the operation of the correlation in the signal-tracking segment is essential if multipath mitigation techniques are to be implemented using the correlation part. It is imperative to elucidate the mechanism for determining the accuracy of pseudo-range and carrier-phase measurements, as well as the generating part within the receiver. Notably, the accuracy, reliability, and availability levels of pseudo-range, Doppler frequency, and carrier-phase measurements may vary among different commercial receivers. Understanding the software GNSS receiver would help elucidate the distinctions among commercial receivers. The potential advantages of a software GNSS receiver include:

- Exploration of particular signal characteristics, new GNSS signals
- Easy to customize source code compared to FPGA
- New signal processing development
- Specific receiver development (scintillation monitor etc.)
- Deep integration with other sensors
- Suitable as an education tool of communication
- Excellent for prototyping
- LEO positioning
- Spoofing and jamming

From 2021 to 2023, Tokyo University of Marine Science and Technology, Chubu University, and Chiba Institute of Technology, commissioned by the Ministry of Education, Culture, Sports, Science, and Technology, spearheaded research and engineering education, predominantly focusing on software GNSS receivers to advance Aerospace Science and Technology. The GNSS domain demands a cadre of young individuals well-versed in positioning algorithms and GNSS receiver signal processing, particularly as GNSS undergoes imminent changes. While GNSS has conventionally served convenience purposes like car and pedestrian navigation, it has also found application in ship and airplane navigation, spanning both civil and military domains. In the foreseeable future, GNSS is poised to be instrumental in highly reliable applications such as autonomous vehicles and civilian drones.

To expedite satellite positioning technology, emphasis on the final phase of the positioning section, as well as the signal processing segment of the receiver, as depicted in Fig. 21, is imperative, as highlighted by Kubo et al. [44]. All teaching materials, actual programs, and content scrutinized in the developmental tasks are accessible. In the competition conducted during the program's culminating year, the student demonstrating the highest proficiency in capturing, tracking, and decoding positioning satellite signals from the intermediate frequency (IF) data acquired by the front end was honored. The source code utilized by the victors is readily available.

7. Open Source Program and Some Modifications

RTKLIB, an open-source program for precise positioning developed by Takasu [45], has been an integral tool in our lectures and seminars for over 15 years, serving as a valuable resource for learning GNSS positioning and conducting research and development (R&D). The RTK performance of commercial receivers has notably advanced with the availability of low-cost dual-frequency receivers. Derived programs like rtklibexplorer [46] have also found application. Here, we provide a brief overview of an enhanced RTK algorithm tailored for vehicles in urban settings, along with experimental findings [47].

7.1 Three Modifications

Our modification to RTKLIB involved replacing the entire relative position function, a core feature responsible for resolving both float and fixed solutions in RTK-GNSS. This entailed utilizing the original RTKLIB source code, excluding the relative positioning component, which is denoted as "relpos."

First, we implemented a satellite selection approach employing subsets of GNSSs. While partial ambiguity resolution can offer enhanced efficacy, precautions must be taken against erroneous fixes. Second, we introduced a smoothed floating solution to mitigate large errors commonly encountered in pseudo-range measurements within urban environments. This involved integrating the pseudo-range-based position with velocity data derived from carrier phase or Doppler frequency using a loosely coupled Kalman filter. Additionally, we incorporated the concept of an Adaptive Kalman filter [48]. Thirdly, we adopted ambiguity resolution techniques leveraging velocity information derived from Doppler frequency or time difference of carrier-phase measurements [49]. These strategies represent combinations of conventional methods, with our future focus aimed at reducing the incidence of incorrect fixes.

7.2 Precise Positioning Challenge

The Institute of Positioning, Navigation, and Timing of Japan sponsored the Precise Positioning Challenge in FY2023, where GNSS observation data from low-cost receivers in downtown Tokyo and Nagoya were made publicly available for analysis, with the accuracy of the positioning algorithm results duly verified. Kubo et al. [47] have reported some of the outcomes of this data analysis. We eagerly anticipate the involvement of young enthusiasts in

such endeavors. Please refer to The Institute of Positioning, Navigation, and Timing of Japan website [50] for further details.

8. Conclusion

In this paper, I have delved into the technology behind achieving precise positioning with GNSS. Initially, I highlighted the current era of multi-GNSS and provided an overview of the fundamentals of precise positioning technology. Subsequently, I elucidated the characteristics and correction data of QZSS, a Japanese positioning satellite. Following this, I explored multipath mitigation technology, recognized as pivotal for achieving precise positioning, underscoring the importance of comprehending GNSS receivers for this purpose. I also outlined the fundamentals of software GNSS receivers. Lastly, I introduced RTKLIB, a globally recognized open-source program for precise positioning, along with some modifications to its functionality. Regrettably, due to space constraints, I was unable to discuss the integration of GNSS with other sensors in this paper. However, I would be delighted if students and young engineers could acquaint themselves with the integration technology of GNSS with IMU and other sensors. Additionally, it is noteworthy that issues concerning interference and spoofing of GNSS signals have surfaced in recent years. I hope that some of you will engage in addressing these challenges both theoretically and practically.

Acknowledgments

I wish to express my sincere appreciation for the students in the lab who prepared a few materials for the present study.

References

[1] N. Kubo, “Satellite Positioning and Location Information,” Nikkan Kogyo Shimbun, 2018.
[2] N. Nishio, “Indoor Positioning and Location Information,” Nikkan Kogyo Shimbun, 2018.
[3] N. Kubo, “Satellite positioning, current and future,” Robot. Soc. Jpn., vol.37, no.7, pp.567–572, 2019.
[4] United States Coast Guard, “GPS Constellation,” https://www.navcen.uscg.gov/gps-constellation, accessed Dec. 27. 2023.
[5] “GLONASS Constellation Status,” https://www.glonass-iac.ru/en/GLONASS/, accessed Dec. 27. 2023.
[6] European GNSS Service Centre, “Galileo Constellation Status,” https://www.gsc-europa.eu/system-service-status/constellation-information, accessed March 21. 2024.
[7] China Satellite Navigation Office, “BDS Constellation Status,” https://www.csno-tarc.cn/en/system/constellation, accessed March 21. 2024.
[8] Cabinet Office, National Space Policy Secretariat, “QZSS Operation Information,” http://qzss.go.jp/en/, accessed Dec. 27. 2023.
[9] Indian Space Research Organization, Department of Space, “NavIC,” https://www.isro.gov.in/SatelliteNavigationServices.html, accessed Dec. 27. 2023.
[10] Cabinet Office, National Space Policy Secretariat, “National Space Policy,” https://qzss.go.jp/info/archive/keikaku_230713.html, accessed Dec. 27. 2023.
[11] European Union Agency for the Space Programme, “Galileo High Accuracy Service,” https://www.euspa.europa.eu/european-space/galileo/services/galileo-high-accuracy-service-has, accessed Dec. 27. 2023.
[12] P. Misra and P. Enge, Global Positioning System Signals, Measurements, and Performance, 2nd ed., Ganga-Jamuna Press, Lincoln, 2006.
[13] E. Kaplan and C. Hegarty, Understanding GPS/GNSS: Principles and Applications, 3rd ed., Artech House Publishers, London, 2017.
[14] J. Fan and G. Ma, “Characteristics of GPS positioning error with non-uniform pseudorange error,” GPS Solut., vol.18, no.4, pp.615–623, 2014.
[15] G. Wübbena, A. Bagge, and M. Schmitz, “RTK Networks Based on Geo++® GNSMART - Concepts, Implementation, Results,” Proc. 14th International Technical Meeting of the Satellite Division of The Institute of Navigation (ION GPS 2001), Salt Lake City, UT, pp.368–378, Sept. 2001.
[16] S. Sugimoto and R. Shibasaki, GPS Handbook, Asakura-Shoten, Tokyo, 2010.
[17] Cabinet Office, National Space Policy Secretariat, “National Space Policy,” https://qzss.go.jp/technical/download/ps-is-qzss.html, accessed March 21. 2024.
[18] “Transistor Technology,” CQ Publishing Co., Ltd., Jan. 2022.
[19] K. Kawate, Y. Igarashi, H. Yamada, K. Akiyama, M. Okeya, H. Takiguchi, M. Murata, T. Sasaki, S. Matsushita, S. Miyoshi, M. Miyoshi, and S. Kogure, “MADOCA: Japanese precise orbit and clock determination tool for GNSS,” Adv. Space Res., vol.71, no.10, pp.3927–3950, 2023.
[20] Cabinet Office, and National Space Policy Secretariat, “National Space Policy,” https://qzss.go.jp/technical/dod/madoca/madoca_testlibrary.html, accessed March 21. 2024.
[21] R. Hirokawa, I. Fernández-Hernández, and S. Reynolds, “PPP/PPP-RTK open formats: Overview, comparison, and proposal for an interoperable message,” NAVIGATION: Journal of the Institute of Navigation, vol.68, no.4, pp.759–778, Dec. 2021.
[22] S. Kogure, “‘QZSS from conception to realization of a 7-satellites, toward an 11-satellites’, QBIC 28th Social Implementation Promotion WG,” Sept. 2023.
[23] Laboratory Home Page, “GNSS TUTOR, CLAS/PPP/SLAS,” https://www.denshi.e.kaiyodai.ac.jp/gnss_tutor/clas_ppp_slas.html, accessed March 22. 2024.
[24] Laboratory Home Page, “GNSS TUTOR, MADOCA PPP,” https://www.denshi.e.kaiyodai.ac.jp/gnss_tutor/madoca.html, accessed March 22. 2024.
[25] NASA Solar Cycle Progression and Forecast, https://www.nasa.gov/solar-cycle-progression-and-forecast/, accessed May 31. 2024.
[26] J. Dickman, Z. Zhu, and C. Bartone, “Carrier phase multipath error characterization and reduction in single aircraft relative positioning,” GPS Solut., vol.14, no.2, pp.141–152, 2010.
[27] J. Fang, L. Zhang, G. Zhang, B. Xu, and L.T. Hsu, “Modeling and validation of GNSS multipath-induced Doppler tracking error,” Proc. 2024 International Technical Meeting of the Institute of Navigation, Long Beach, California, pp.931–940, Jan. 2024.
[28] A. Chow, D. Orendorff, M. Fu, M. Khider, S. Dane, and V. Gulati, 2023, Google Smartphone Decimeter Challenge 2023, Kaggle, https://kaggle.com/competitions/smartphone-decimeter-2023
[29] R. Odolinski, H. Yang, L.T. Hsu, M. Khider, G. Fu, and D. Dusha, “Evaluation of the multi-GNSS, dual-frequency RTK positioning performance for recent android smartphone models in a phone-to-phone setup,” Proc. 2024 International Technical Meeting of the Institute of Navigation, Long Beach, California, pp.42–53, Jan. 2024. https://doi.org/10.33012/2024.19575.
[30] P.D. Groves, Z. Jiang, M. Rudi, and P. Strode, “A portfolio approach to NLOS and multipath mitigation in dense urban areas,” Proc. 26th International Technical Meeting of the Satellite Division of the Institute of Navigation (ION GNSS+ 2013), Nashville, TN, 3231–3247, 2013.
[31] T. Suzuki, “First place award winner of the smartphone decimeter

challenge: Global optimization of position and velocity by factor graph optimization," Proc. 34th International Technical Meeting of the Satellite Division of the Institute of Navigation (ION GNSS+ 2021), St. Louis, Missouri, pp.2974–2985, Sept. 2021.
[32] B.W. Parkinson and J.J. Spilker, Jr., "Global positioning system: Theory and applications, vol.I, progress in astronautics and aeronautics," American Institute of Aeronautics and Astronautics, Washington DC, 1996.
[33] HEXAGON NovAtel, "High Performance GPS Pinwheel Antenna," https://novatel.com/support/support-materials/technical-papers, accessed March 22. 2024.
[34] T. Ozeki and N. Kubo, "Performance evaluation of low-cost GNSS antennas," J. Appl. Surv. Technol., 34, 2023.
[35] L. Garin, F. van Diggelen, and J.-M. Rousseau, "Strobe & edge correlator multipath mitigation for code," Proc. 9th International Technical Meeting of the Satellite Division of the Institute of Navigation (ION GPS 1996), Kansas city, MO, pp.657–664, Sept. 1996.
[36] HEXAGON NovAtel, "Theory and Performance of the Pulse Aperture Correlator," https://novatel.com/support/support-materials/technical-papers, accessed March 23. 2024.
[37] Y. Tachita, "Study of multipath effects on GPS positioning," Tokyo University of Marine Science & Technology Doctor thesis, 2001.
[38] HEXAGON NovAtel, "Performance Evaluation of the Multipath Estimating Delay Lock Loop," https://novatel.com/support/support-materials/technical-papers, accessed March 23. 2024.
[39] N. Kubo, "A study on GPS multipath mitigation and its practicability for high precise positioning," The University of Tokyo Doctor thesis, 2015.
[40] S. Verhagen, P.J.G. Teunissen, H. van der Marel, and B. Li, "GNSS ambiguity resolution: Which subset to fix?," Proc. International Global Navigation Satellite Systems Society IGNSS Symposium, 2011.
[41] M. Higuchi and N. Kubo, "Achievement of continuous decimeter-level accuracy using low-cost single-frequency receivers in urban environments," Proc. 29th International Technical Meeting of the Satellite Division of the Institute of Navigation (ION GNSS+ 2016), Portland, oregon, pp.1891–1913, Sept. 2016.
[42] K. Borre, I. Fernández-Hernández, J.A. L'opez-Salcedo, and M.Z.H. Bhuiyan, GNSS Software Receivers, Cambridge University Press, 2022.
[43] M. Overbeck, F. Garzia, A. Popugaev, O. Kurz, F. Förster, W. Felber, A. Ayaz, S. Ko, and B. Eissfeller, "GOOSE — GNSS receiver with an open software interface," Proc. 28th International Technical Meeting of the Satellite Division of the Institute of Navigation (ION GNSS+ 2015), Tampa, Florida, pp.3662–3670, Sept. 2015.
[44] N. Kubo, T. Ebinuma, and T. Suzuki, "Tutorials of software GNSS receiver," https://gnss-learning.org/, accessed March 23. 2024.
[45] T. Takasu, "RTKLIB," https://www.rtklib.com/, accessed March 23. 2024.
[46] T. Everett, "Rtklibexplorer," https://rtklibexplorer.wordpress.com/, accessed March 23. 2024.
[47] N. Kubo, T. Ozeki, and T. Komori, "Modified RTK program for kinematic urban condition," Proc. 2024 International Technical Meeting of the Institute of Navigation, Long Beach, California, pp.846–857, Jan. 2024.
[48] T. Tominaga and N. Kubo, "Adaptive estimation of measurement noise to improve the performance of GNSS single point positioning in dense urban environment," J. Inst. Position. Navig. Timing Jpn., vol.8, no.1, pp.1–8, 2017.
[49] N. Kubo, "Advantage of velocity measurements on instantaneous RTK positioning," GPS Solut., vol.13, no.4, pp.271–280, 2009.
[50] The Institute of Positioning, Navigation and Timing of Japan, "Precise Positioning Challenge," https://www.denshi.e.kaiyodai.ac.jp/challenge/, accessed March 23. 2024.

Nobuaki Kubo earned his B.S. and M.S. degrees in Electrical Engineering from Hokkaido University in 1996 and 1998, respectively. He commenced his professional journey with NEC Corporation in 1998 before transitioning to Tokyo University of Marine Science and Technology (TUMSAT) in 2001. Subsequently, he obtained his Ph.D. from the University of Tokyo in 2005. In 2008-2009, he pursued further studies on GNSS at Stanford University. Presently, he serves as a Professor at TUMSAT.

IEICE TRANS. COMMUN., VOL.E107–B, NO.11 NOVEMBER 2024

PAPER

High Stability Code Tracking for Band-Limited DSSS Systems

Zhiwei LU[†a)], **Yiwen JIAO**[†b)], ***and*** **Yudi CHEN**[†c)], ***Nonmembers***

SUMMARY In this paper, we study the problem of high stability code tracking for band-limited direct sequence spread spectrum (DSSS) systems. In band-limited DSSS systems carrying critical applications, high stability is required in addition to low error variance for code tracking. Therefore, we propose a high stability code tracking method for band-limited DSSS systems, which constructs frequency domain vectors from the received signal, reduces the dimension of the vectors by frequency domain integration and dump, and estimates the time-delay error by the subspace method. We also give a closed-form expression for the steady-state time-delay error variance of the proposed method, which can be used to analyze the error variance performance theoretically and design proper band-limited DSSS systems. The theoretical analysis and simulation results show that the proposed method is able to enhance both the maximum and linear code tracking ranges, thus realizing high stability code tracking, and has constant error variance performance and appropriate computational complexity.

key words: *band-limited DSSS systems, code tracking, high stability, code tracking range, frequency domain*

1. Introduction

Direct sequence spread spectrum (DSSS) systems are extensively used in the field of satellite telemetry, tracking and command (TT&C), navigation and communication [1], [2], due to its robust anti-interference ability, convenience of multiple access, and simultaneous communication and ranging capabilities. Band-limited DSSS systems use band-limited pulse waveforms such as root raised cosine (RRC), Gaussian, and other optimized designed pulse waveforms, and have advantages in spectrum efficiency, spectrum leakage, and ranging accuracy, compared with square pulse waveform systems.

Code tracking is the fine time synchronization process. In band-limited DSSS systems carrying critical applications, the stability is equally important in addition to the error variance performance in code tracking. In order to achieve higher stability, i.e., lower probability of loss of lock, when the error variance performance is the same, the code tracking method needs to have a larger maximum and linear code tracking ranges. However, the existing code tracking methods based on correlation delay-locked loop (DLL) have relatively small maximum and linear code tracking ranges, which are difficult to meet the high stability requirement for high demand band-limited DSSS systems. Therefore, the aim of this paper is to propose a high stability code tracking method with a larger maximum and linear code tracking ranges under the condition of constant error variance performance and appropriate computational complexity.

The code tracking problem has been extensively studied for square pulse waveform DSSS systems over the past few decades [3]–[5]. The first all-digital code tracking method applied to band-limited DSSS systems was proposed in [6], which used digital matched filter, interpolator and a 2 times decimator to generate the required data of forward and backward branches. To reduce the complexity of code tracking, [7], [8] use parallel matched filter to improve the matched filter part, [9], [10] use fast Fourier transform (FFT) to improve both the matched filter and interpolator parts and [11] realizes the matched filter, interpolator, and code despreading simultaneously in the frequency domain, achieving further complexity reduction. In addition, many methods based on Kalman filter [12] and adaptive loop bandwidth methods [13] were proposed to further improve the error variance performance of code tracking by improving the loop filter part. To cope with complex channel conditions, researchers have also improved the performance of code tracking in multipath environment [14] and narrowband interference [15], etc. Although these studies have improved the code tracking under different conditions, they are still based on the correlation DLL, which limits the maximum and linear code tracking ranges and the probability of loss of lock, as well as the stability. Simultaneously using the outputs of multiple correlators to obtain the error estimation can extend the maximum code tracking range, but additional noise accumulation inevitably degrades the error variance performance. Therefore, [16], [17] selects the optimal two out of multiple correlation outputs for time-delay variance estimation, i.e., the selection DLL method, but the presence of selection leads to having a severe loss of error variance performance under low signal-to-noise ratios (SNRs).

This paper presents a code tracking method for band-limited DSSS systems in order to achieve higher stability with constant error variance performance. The time-delay error adjustment and correlation are implemented in frequency domain to obtain the frequency domain vectors. In order to reduce the computational complexity of the subsequent processing, we realize the dimension reduction of the vectors by frequency domain integration and dump. Afterwards, in order to make full use of the frequency domain

Manuscript received December 5, 2023.
Manuscript revised May 28, 2024.
Manuscript publicized August 1, 2024.

[†]Key Laboratory of Intelligent Space TTC&O, Ministry of Education, Space Engineering University, Beijing 101416, China.
a) E-mail: hgdluzw@163.com
b) E-mail: jiaoyiwen1985@163.com (Corresponding author)
c) E-mail: chenyudi9438@163.com

DOI: 10.23919/transcom.2023EBP3201

Table 1 Variable list.

Notations	Definition
a_i, c_m	Data symbol and spread spectrum code
$p_T(t), p_R(t)$	Transmitted and local pulse waveforms
T_c, T_s, T	Chip period, sampling period and date symbol period
τ, d	Time delay and normalized time delay with $d = \tau/T_s$
N, N_s, N_b	Length of the spread spectrum code, number of points for a single data period and the reduced number after integration and dump
$H_k^{FD}(d)$	Frequency domain fractional delay filter
G_k^{TX}, G_k^{RX}	Transmitted and local frequency domain waveforms
$P_N(f)$	Frequency response of the Nyquist raised cosine (RC) filter
$\mathbf{X_n}$	Frequency domain vector
$\hat{\mathbf{R}}_n$	Estimated covariance matrix
$\hat{\mathbf{E}}_N$	Estimated noise subspace
$\varepsilon, \hat{\varepsilon}$	True and estimated values of normalized time-delay error

vectors to achieve a larger maximization and linear code tracking range, as well as to improve the stability, we use the modified root-MUSIC subspace algorithm to estimate the time-delay error. Finally the estimated time-delay error is passed through the loop to realize the code tracking. The frequency domain integration and dump is to ensure that the proposed method has appropriate computational complexity, which exploits the property that the autocorrelation of the band-limited DSSS signals aggregates at zero point during the code tracking process, and realizes the dimension reduction of the vectors in frequency domain in a way similar to low-pass filtering, thus the subsequent time-delay error estimation can be implemented with small length frequency domain vectors. We also analyze the code tracking range, steady-state error variance performance and computational complexity of the proposed method, and simulate the proposed method focusing on the first two.

The rest of the paper is organized as follows. We first model the received signal in Sect. 2. Then, in Sect. 3, we propose the high stability code tracking method. In Sect. 4, Sect. 5 and Sect. 6, we analyze the code tracking range, steady-state error variance performance and computational complexity of the proposed method, respectively. Simulation results are shown in Sect. 7. Finally, Sect. 8 gives the conclusion. The major variables adopted in this paper is listed in Table 1 for ease of reference.

2. Signal Model

We adopt DSSS binary phase-shift keying (BPSK) modulation with RRC pulse shaping in this paper, and the transmitted signal can be modeled as [6]

$$s(t) = \sum_{m=-\infty}^{\infty} a_{\lfloor mT_c/T \rfloor} c_m p_T(t - mT_c), \tag{1}$$

where $\lfloor \bullet \rfloor$ is the rounding toward negative infinity operation, a_i is the data symbol with a period of T, $c_m = \pm 1$ is the spread spectrum code with chip period T_c and cycle length N, and $p_T(t)$ is the pulse waveform of the transmit signal with its Fourier transform $P_T(f) = T_c\sqrt{P_N(f)}$, where $P_N(f)$ is the frequency response of Nyquist RC filter.

We assume that the spread spectrum code period is equal to the data symbol period, i.e., $T = NT_c$ in this paper, and the method proposed here can be extended to the cases easily where T is an integer multiple of the code period.

In code tracking, we assume that the receiver has completed acquisition and the received signal is still affected by carrier phase, time delay, and noise. As a result, the complex baseband representation of the received signal can be expressed as

$$\begin{aligned} r(t) &= e^{j\theta} s(t - \tau) + w(t) \\ &= e^{j\theta} \sum_{m=-\infty}^{\infty} a_{\lfloor mT_c/T \rfloor} c_m p_T(t - mT_c - \tau) + w(t), \end{aligned} \tag{2}$$

where θ is the carrier phase, uniformly distributed in $[-\pi, \pi)$, τ is the initial time delay, $w(t)$ is the complex Gaussian white noise with power spectral density (PSD) N_0/P, where P is the power of the intermediate frequency signal.

Code tracking is the fine time synchronization process. Before this, the receiver has already completed the rough estimation of τ through acquisition, and the initial time delay τ is generally much smaller than the spread spectrum code period, i.e., $\tau \ll NT_c$.

The received signal after sampling can be expressed as

$$r_\ell = e^{j\theta} \sum_{m=-\infty}^{\infty} a_{\lfloor mT_c/T \rfloor} c_m p_T\left[(\ell - d)T_s - mT_c\right] + w_\ell, \tag{3}$$

where $r_\ell = r(\ell T_s)$, $w_\ell = w(\ell T_s)$, T_s is the sampling period, and $d = \tau/T_s$ is the normalized initial time delay. We make N_s be the number of sampling points of a single data period, so that $T = NT_c = N_s T_s$.

3. Proposed Method

The proposed code tracking method accomplishes time-delay adjustment and correlation in the frequency domain, dimension reduction of vectors by frequency domain integration and dump, and time-delay error estimation by a modified subspace algorithm, after which code tracking is achieved through the loop. It is able to fully utilize the generated frequency domain vectors to achieve time-delay error estimation instead of the two forward and backward correlation values in the correlation DLL. The structure of the proposed method is illustrated in Fig. 1.

The proposed method contains three main components. The first is the frequency domain vector generation, which completes time-delay adjustment of the received signal based on the previously feedback time delay, realizes correlation with the local waveform, and vector dimension reduction through frequency domain integration and dump. The second is the time-delay error estimation, which estimates the

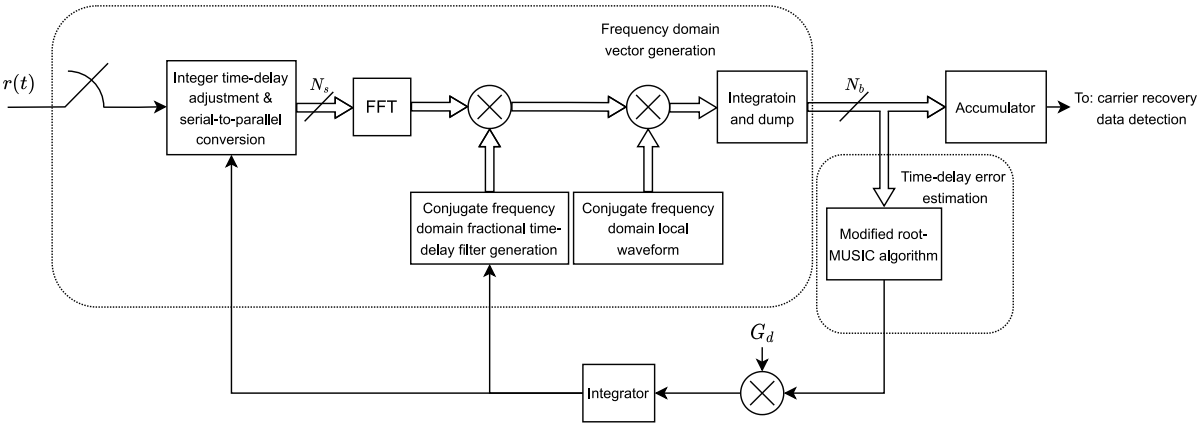


Fig. 1 Structure of the proposed code tracking method.

time-delay error using the dimension-reduced vector through the modified root-MUSIC subspace algorithm. The third is the time delay generation, which involves a multiplier and an integrator, used to obtain the estimated value of time delay for the next code tracking period.

In addition, following the frequency domain vector generation, an accumulator can be used to generate the signal for carrier recovery and data detection.

3.1 Frequency Domain Vector Generation

The frequency domain vector generation component primarily completes time-delay adjustment, correlation with the local waveform, and vector dimension reduction.

In order to make the data symbols almost constant within a single processing cycle, i.e., data period, the time-delay adjustment is divided into two parts: integer delay adjustment and fractional delay adjustment. Integer delay adjustment is completed by shifting the time domain signal after sampling directly, while fractional delay adjustment is realized by conjugate multiplication of signal and the fractional delay filter in the frequency domain.

To achieve correlation between the received signal and the local waveform, we perform a conjugate multiplication in the frequency domain. The local waveform includes information about the spread spectrum code and the pulse waveform and can be generated in advance and remain unchanged. By employing correlation, we complete the despreading and matched filtering of the received signal simultaneously, thereby enabling the accumulation of energy of the useful signal in the time domain.

The correlation output of the band-limited DSSS signal is concentrated near zero in the time domain. Similar to low-pass filtering, we can use integration and dump in the frequency domain to filter out the invalid data in time domain. The integration and dump greatly reduces the length of the frequency domain vectors, allowing the subsequent estimation of the time-delay error realized efficiently, making the computational complexity of the proposed code tracking method appropriate.

We assume that in the n-th data period, the feedback normalized time delay is $\hat{d}_n$, with its initial value $\hat{d}_0 = 0$. We adjust the integer delay $\lfloor \hat{d}_n \rfloor$ by shifting the received signal in the time domain. The fractional delay $\hat{d}_n - \lfloor \hat{d}_n \rfloor$ is adjusted by frequency domain fractional delay filter. Since time domain shifting does not affect the statistical properties of the noise, we can assume the noise term unchanged. Since both the time delay and the pulse waveform period are much smaller than the data symbol period, $\tau \ll T$ and $T_c = T/N$, for simplicity we ignore the effect between different data symbols. The received signal of the n-th data period, after integer delay adjustment and serial-to-parallel conversion, is given by

$$r_{\ell,n} = a_n e^{j\theta} \sum_{m=0}^{N-1} c_m p_T \left[\left(\ell - d + \lfloor \hat{d}_n \rfloor \right) T_s - m T_c \right] + w_{\ell,n}$$
$$\text{for} \quad \ell = 0, 1, \cdots, N_s - 1. \tag{4}$$

The received signal can be transformed into frequency domain after N_s-pt FFT. Subsequently, the frequency domain representation of (4) can be expressed as

$$R_{k,n} = a_n e^{j\theta} H_k^{FD} \left(d - \lfloor \hat{d}_n \rfloor \right) G_k^{TX} + W_{k,n}$$
$$\text{for} \quad k = 0, 1, \cdots, N_s - 1, \tag{5}$$

where $H_k^{FD}(d)$ is the frequency domain fractional delay filter with length N_s [11], [18], [19], which can be used to model the time delay in frequency domain, and its expression is

$$H_k^{FD}(d) = \begin{cases} 1, & k = 0 \\ e^{-j\frac{2\pi}{N_s}kd}, & k = 1, 2, \cdots, N_s/2 - 1 \\ cos(\pi d), & k = N_s/2 \\ e^{j\frac{2\pi}{N_s}(N_s - kd)}, & k = L_s/2 + 1, \cdots, N_s - 1, \end{cases} \tag{6}$$

and where G_k^{TX} is the frequency domain waveform of the transmitted signal, which contains the information of the spread spectrum code and the pulse waveform, and can be expressed as

$$G_k^{TX} = \mathrm{FFT}_{N_s}\left[\sum_{m=0}^{N-1} c_m p_T(\ell T_s - mT_c)\right]. \quad (7)$$

After the integer delay adjustment has been completed, the fractional delay is given by $d - \lfloor \hat{d}_n \rfloor$. This is also reflected in (5). To adjust the fractional time delay, we perform a conjugate multiplication of $R_{k,n}$ with a frequency domain fractional delay filter, denoted by $H_k^{FD}(\hat{d}_n - \lfloor \hat{d}_n \rfloor)$. The signal after fractional delay adjustment can be expressed as

$$\begin{aligned} R_{k,n}^{\alpha} &= \left[H_k^{FD}(\hat{d}_n - \lfloor \hat{d}_n \rfloor)\right]^* R_{k,n} \\ &= a_n e^{j\theta} H_k^{FD}(\varepsilon_n) G_k^{TX} + W_{k,n}, \end{aligned} \quad (8)$$

where $\varepsilon_n = d - \hat{d}_n$ represents the normalized time-delay error. The frequency domain fractional delay filter almost does not affect the power spectrum of the noise, but only causes a phase change. Therefore, we can assume that in (8), the noise term remains unchanged.

After the time delay adjustment, $R_{k,n}^{\alpha}$ still contains information about the spread spectrum code and pulse waveform in G_k^{TX}. To accumulate the energy of the useful signal, we perform the correlation by conjugate multiplication of $R_{k,n}^{\alpha}$ with a local frequency domain waveform, denoted by G_k^{RX}. This operation effectively performs the code despreading and matched filtering. The local frequency domain waveform is given by

$$G_k^{RX} = \mathrm{FFT}_{N_s}\left[\frac{T_s}{N}\sum_{m=0}^{N-1} c_m p_R(\ell T_s - mT_c)\right], \quad (9)$$

where $p_R(t)$ is the local pulse waveform, and its Fourier transformation is $P_R(f) = \sqrt{P_N(f)}$. After completing the correlation, the result can be expressed as

$$\begin{aligned} R_{k,n}^{\beta} &= \left(G_k^{RX}\right)^* R_{k,n}^{\alpha} \\ &= a_n e^{j\theta} H_k^{FD}(\varepsilon_n) G_k + \tilde{W}_{k,n}, \end{aligned} \quad (10)$$

where $G_k = \left(G_k^{RX}\right)^* G_k^{TX}$, and $\tilde{W}_{k,n} = \left(G_k^{RX}\right)^* W_{k,n}$. Through the derivation presented in Appendix A, G_k can be expressed as

$$G_k = \frac{T_c}{T_s} P_N\left(\frac{k}{N_s T_s}\right). \quad (11)$$

We define $\tilde{\mathbf{W}}_n = \left[\begin{array}{cccc} \tilde{W}_{0,n} & \tilde{W}_{1,n} & \cdots & \tilde{W}_{N_s-1,n} \end{array}\right]^T$, and the noise covariance matrix $\overline{\tilde{\mathbf{W}}_n \tilde{\mathbf{W}}_n^H}$ is given in Appendix B.

When correlation is completed, the time domain signal associated with $R_{k,n}^{\beta}$ becomes concentrated around the zero point, and its offset is related to the time-delay error ε_n. If the frequency domain form of a time domain signal has non-zero values only in the low frequency part of the signal, we can realize a reduction in the data rate of the time domain signal by using time domain integration and dump, i.e., low-pass filtering and extraction operations. Because the time domain and frequency domain are two symmetrical representations, we can also use frequency domain integration and dump to reduce the length of the frequency domain signal, i.e., to remove the invalid time domain data. The result can be expressed as

$$\begin{aligned} X_{k,n} &= \frac{1}{N_I}\sum_{m=kN_I-N_I/2}^{kN_I+N_I/2-1} R_{m,n}^{\beta} \\ &\approx a_n e^{j\theta} H_k^{FD}(\varepsilon_n)\tilde{G}_k + V_k \\ &\quad \text{for} \quad k = 0, 1, \cdots, N_b - 1, \end{aligned} \quad (12)$$

where N_I is the integration and dump number, and N_b is the length of output signal, which satisfies $N_b = N_s/N_I$. $H_k^{FD}(\varepsilon_n)$ is the frequency domain fractional delay filter with length N_b here. The RC pulse function's attenuation characteristic and (11)–(12) allow us to express the signal waveform approximately as

$$\tilde{G}_k \approx \frac{T_c}{T_s} P_N\left(\frac{k}{N_b T_s}\right). \quad (13)$$

We define $\mathbf{V}_n = \left[\begin{array}{cccc} V_0 & V_1 & \cdots & V_{N_b-1} \end{array}\right]^T$, and the noise covariance marix $\overline{\mathbf{V}_n \mathbf{V}_n^H}$ is also given in Appendix B.

We rewrite (12) to simplify the expression as

$$\begin{aligned} X_{k,n} &= A_k(\varepsilon_n) F_n + V_k \\ &\quad \text{for} \quad k = 0, 1, \cdots, N_b - 1, \end{aligned} \quad (14)$$

where

$$\begin{aligned} F_n &= a_n e^{j\theta}, \\ A_k(\varepsilon_n) &= H_k^{FD}(\varepsilon_n)\tilde{G}_k, \end{aligned} \quad (15)$$

and the vectorial form is

$$\mathbf{X}_n = \mathbf{A}(\varepsilon_n) F_n + \mathbf{V}_n, \quad (16)$$

where

$$\begin{aligned} &\mathbf{X}_n = \left[\begin{array}{cccc} X_{0,n} & X_{1,n} & \cdots & X_{N_b-1,n} \end{array}\right]^T, \\ &\mathbf{A}(\varepsilon_n) = \mathrm{diag}\{\tilde{\mathbf{G}}\}\mathbf{H}^{FD}(\varepsilon), \\ &\tilde{\mathbf{G}} = \left[\begin{array}{cccc} \tilde{G}_0 & \tilde{G}_1 & \cdots & \tilde{G}_{N_b-1} \end{array}\right]^{\mathbf{T}}, \\ &\mathbf{H}^{FD}(\varepsilon) = \left[\begin{array}{cccc} H_0^{FD}(\varepsilon) & H_1^{FD}(\varepsilon) & \cdots & H_{N_b-1}^{FD}(\varepsilon) \end{array}\right]^T. \end{aligned} \quad (17)$$

The frequency domain vector $\mathbf{X}_n$ can be used for two purposes. First, it can generate the signal for carrier recovery and data detection through an N_b point accumulator. Second, it can be used to estimate the time-delay error ε_n.

3.2 Time-Delay Error Estimation

In order to fully utilize the frequency domain vectors, we use a modified root-MUSIC subspace algorithm to estimate the time-delay error. The parameter estimation of the root-MUSIC algorithm can be divided into two parts, one is the

covariance matrix estimation, and the other is the eigendecomposition and polynomial root finding.

In covariance matrix estimation, the root-MUSIC algorithm uses the method of averaging the sample covariance matrices, which is not suitable for code tracking with feedback. To address this limitation, we propose a new method for covariance matrix estimation by using RC (resistor-capacitor) integral filter, since it can be used in the presence of feedback and the sum of coefficients of the filter is 1. The system function of the RC (resistor-capacitor) integral filter can be expressed as

$$H(s) = \frac{1}{1 + s\tau_1}, \tag{18}$$

where τ_1 is the time constant. By replacing the differential operator s with $(1-z^{-1})/T$, we can obtain the system function in the z domain as

$$H(z) = \frac{T}{T + \tau_1 - \tau_1 z^{-1}}. \tag{19}$$

Then at the n-th data period, the estimated covariance matrix can be expressed as

$$\hat{\mathbf{R}}_n = \frac{\tau_1}{T + \tau_1}\hat{\mathbf{R}}_{n-1} + \frac{T}{T + \tau_1}\mathbf{X}_n\mathbf{X}_n^H, \tag{20}$$

where the initial value $\hat{\mathbf{R}}_{-1} = 0$. After eigendecomposition, (20) can be expressed as

$$\hat{\mathbf{R}}_n = \hat{\mathbf{E}}_S\hat{\Lambda}_S\hat{\mathbf{E}}_S^H + \hat{\mathbf{E}}_N\hat{\mathbf{\Lambda}}_N\hat{\mathbf{E}}_N^H, \tag{21}$$

where $\hat{\mathbf{E}}_S$ is the estimated signal subspace, which is a vector in this paper, and $\hat{\Lambda}_S$ is the corresponding eigenvalue. $\hat{\mathbf{E}}_N$ is the estimated noise subspace, $\hat{\mathbf{\Lambda}}_N$ is the corresponding eigenvalue matrix. The signal subspace $\hat{\mathbf{E}}_S$ is eigenvector that corresponds to the largest eigenvalue of the covariance matrix. Conversely, the noise subspace $\hat{\mathbf{E}}_N$ is composed of the eigenvectors that correspond to the remaining eigenvalues and $\hat{\mathbf{E}}_N$ is a matrix of $N_b \times (N_b - 1)$.

We can use the estimated noise subspace matrix $\hat{\mathbf{E}}_N$ to construct the parameter spectrum of the time-delay error estimation, that is

$$S(\varepsilon) = \frac{1}{\mathbf{A}^H(\varepsilon)\hat{\mathbf{E}}_N\hat{\mathbf{E}}_N^H\mathbf{A}(\varepsilon)}. \tag{22}$$

Since the signal subspace and the noise subspace are nearly orthogonal, the parameter spectrum, denoted by $S(\varepsilon)$, attains its maximum value when the time-delay error ε is optimally estimated.

In order to skip the search process and improve performance at the same time, we estimate the time-delay error by finding the root of $S^{-1}(\varepsilon)$. By substituting (13), (15), $S^{-1}(\varepsilon)$ can be rewritten as

$$\begin{aligned} S^{-1}(\varepsilon) &= \mathbf{A}^H(\varepsilon)\hat{\mathbf{E}}_N\hat{\mathbf{E}}_N^H\mathbf{A}(\varepsilon) \\ &= \left[\mathbf{H}^{FD}(\varepsilon)\right]^H \mathrm{diag}^H\{\tilde{\mathbf{G}}\}\hat{\mathbf{E}}_N\hat{\mathbf{E}}_N^H\mathrm{diag}\{\tilde{\mathbf{G}}\}\mathbf{H}^{FD}(\varepsilon), \end{aligned} \tag{23}$$

According to (17) and (6), $\mathbf{H}^{FD}(\varepsilon)$ can be decomposed as

$$\mathbf{H}^{FD}(\varepsilon) = \mathbf{C}\mathbf{E}(\varepsilon), \tag{24}$$

where the phase vector

$$\mathbf{E}(\varepsilon) = \begin{bmatrix} e^{j\frac{2\pi}{N_b}\frac{N_b}{2}\varepsilon} & e^{j\frac{2\pi}{N_b}\left(\frac{N_b}{2}-1\right)\varepsilon} & \cdots & e^{-j\frac{2\pi}{N_b}\frac{N_b}{2}\varepsilon} \end{bmatrix}^T, \tag{25}$$

and the transformation matrix

$$\mathbf{C} = \begin{bmatrix} & & & & 1 & & & & \\ & & & & & 1 & & & \\ & & & & & & \ddots & & \\ & & & & & & & 1 & \\ 1/2 & & & & & & & & 1/2 \\ & 1 & & & & & & & \\ & & \ddots & & & & & & \\ & & & 1 & & & & & \end{bmatrix}_{N_b \times (N_b+1).} \tag{26}$$

Thus (23) can be further expressed as

$$S^{-1}(\varepsilon) = [\mathbf{E}(\varepsilon)]^H\mathbf{C}^H\mathrm{diag}^H\{\tilde{\mathbf{G}}\}\hat{\mathbf{E}}_N\hat{\mathbf{E}}_N^H\mathrm{diag}\{\tilde{\mathbf{G}}\}\mathbf{C}\mathbf{E}(\varepsilon). \tag{27}$$

We define

$$\mathbf{B} = \mathbf{C}^H\mathrm{diag}^H\{\tilde{\mathbf{G}}\}\hat{\mathbf{E}}_N\hat{\mathbf{E}}_N^H\mathrm{diag}\{\tilde{\mathbf{G}}\}\mathbf{C}, \tag{28}$$

and let $B_{-N_b/2,-N_b/2}$ denote the first element in the upper-left corner of the matrix $\mathbf{B}$. We can rewrite $S^{-1}(\varepsilon)$ as follows

$$\begin{aligned} S^{-1}(\varepsilon) &= \sum_{m_1=-N_b/2}^{N_b/2}\sum_{m_2=-N_b/2}^{N_b/2} e^{j\frac{2\pi}{N_b}m_1\varepsilon}B_{m_1,m_2}e^{-j\frac{2\pi}{N_b}m_2\varepsilon} \\ &= \sum_{m=-N_b}^{N_b} b_m e^{j\frac{2\pi}{N_b}m\varepsilon}. \end{aligned} \tag{29}$$

Additionally, b_m represents the sum of the m-th diagonal elements in the matrix $\mathbf{B}$ and is given by the following equation

$$b_m = \sum_{m_1-m_2=m} B_{m_1,m_2}. \tag{30}$$

According to (29), we can define polynomial

$$D(z) = \sum_{m=-N_b}^{N_b} b_m z^m. \tag{31}$$

By finding the root of $D(z)$ closest to the unit circle, the time-delay error estimation can be obtained. Assume that the root closest to the unit circle is

$$z = z_1. \tag{32}$$

The time-delay error estimation, corresponding to the n-th estimated covariance matrix $\hat{\mathbf{R}}_n$, can be expressed as

$$\hat{\varepsilon}_n = \frac{N_b}{2\pi} \arg(z_1). \tag{33}$$

3.3 Time Delay Estimation and Parameter Settings

In this subsection, we obtain the time delay estimation using a loop. The time-delay error estimation $\hat{\varepsilon}_n$ has been obtained, and we use it as input, passing it through a multiplier and an integrator to calculate the time delay estimation, which is then used as the feedback.

The estimated time delay for the $n+1$ code tracking period, can be expressed as

$$\hat{d}_{n+1} = \hat{d}_n + G_d \hat{\varepsilon}_n, \tag{34}$$

where G_d is the coefficient of the multiplier for two purposes. First, it consists the gain factor of the integrator. Second, it can controls the loop bandwidth.

The damping coefficient affects the oscillation characteristics and response speed of the tracking system, and in the RC (resistor-capacitor) integral filter tracking loop, the damping coefficient $\zeta = \sqrt{T/G_d\tau_1}/2$ [20]. The loop bandwidth is the bandwidth of the loop's equivalent ideal narrow-band filter, which affects the loop's noise resistance. In the RC (resistor-capacitor) integral filter tracking loop, the loop bandwidth $B_L = G_d/4T$ [20]. In order to balance the oscillation characteristics and response speed of the tracking system, we set the damping coefficient to a typical value, i.e., $\zeta = 0.707$, and keep the tracking loop in a critical damping state. Based on the known damping coefficient ζ and the required loop bandwidth B_L, we can calculate the multiplication factor G_d and the time constant τ_1 of the RC (resistor-capacitor) integral filter required for loop operation as

$$\begin{aligned} G_d &= 4TB_L, \\ \tau_1 &= \frac{1}{8B_L}. \end{aligned} \tag{35}$$

In the proposed band-limited DSSS code tracking method, the setting of the sampling period T_s needs to satisfy the Nyquist sampling theorem, and it is best to meet the requirements of efficient FFT at the same time, that is, the number of sampling points in a single data period is a power of 2. When the spread spectrum code adopts the M code or the Gold code, its length satisfies $N = 2^r - 1$, where r is the number of the generation registers. The number of sampling points in a single data period and the sampling period can be set as

$$\begin{aligned} N_s &= 2(N+1), \\ T_s &= \frac{T}{N_s} = \frac{NT_c}{2(N+1)}. \end{aligned} \tag{36}$$

In contrast, the correlation DLL code tracking method proposed in [6] uses $N_s = 2N$ so that the forward and backward branches can be generated by a decimation of 2.

4. Code Tracking Range Analysis

The code tracking curve refers to the relationship between the time-delay error output of the code tracking method and the true time-delay error of the input signal with open loop and noiseless conditions, and its two mian observations are the maximum and linear code tracking ranges. The linear code tracking range is the error range in which the error output is linear or nearly linear with the actual error, and the maximum code tracking range is the error range in which the positive and negative signs of the error output are the same as the actual error. A larger maximum and linear code tracking ranges imply a lower probability of loss of lock i.e. higher stability when the steady-state error variance performance is the same.

In this section, we analyze the maximum and linear code tracking ranges of the proposed method and compare them with the correlation DLL and selection DLL methods [6], [17]. The comparison is done using a combination of theoretical and numerical analyses, where the selection DLL is only analyzed numerically due to the complexity of the theoretical analysis.

The proposed code tracking method uses subspace algorithm to estimate the time-delay error in frequency domain, and its maximum code tracking range is determined by the length of frequency domain vectors. At the same time, since the used subspace algorithm is asymptotically unbiased, the linear code tracking range can be considered equivalent to the maximum code tracking range. Referring to (16) and (17), the length of the frequency domain vector $\mathbf{X}_n$ is N_b. As there are two time-delay error directions, positive and negative, the normalized maximum and linear code tracking ranges of the proposed method is

$$\varepsilon_P = \frac{N_b}{2}. \tag{37}$$

Therefore, the code tracking curve of the proposed method can be expressed as

$$\eta_P(\varepsilon) = \begin{cases} \varepsilon, & -\varepsilon_P \le \varepsilon \le \varepsilon_P \\ 0, & \text{else.} \end{cases} \tag{38}$$

The code tracking curve of the correlation DLL method is given in [6], which can be expressed as

$$\eta_D(\varepsilon) = \frac{T_c}{A}\left[p^2\left(\varepsilon T_s - \frac{T_c}{2}\right) - p^2\left(\varepsilon T_s + \frac{T_c}{2}\right)\right], \tag{39}$$

where $p(t)$ is the RC pulse waveform with frequency response $P(f) = T_c P_N(f)$, and A is a known scale factor.

The code tracking curves for the proposed code tracing method, the correlation DLL method, and the selection DLL method are shown in Fig. 2, where the solid lines are the theoretical results, the dashed lines are the numerical results, and N_{corr} is the number of correlation outputs used in the selection DLL method. The signal pulse is a root raised cosine waveform with the roll-off factor $\alpha = 0.25$. The selection algorithm in the selection DLL method determines

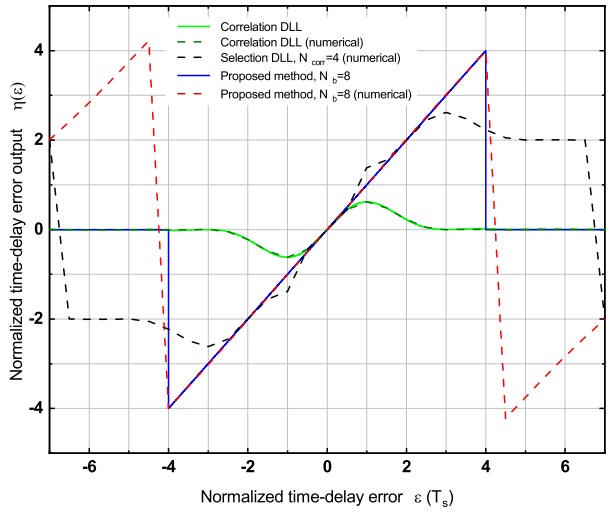


Fig. 2 Code tracking curves of the proposed method, the correlation DLL method and the selection DLL method.

first the branch with maximal energy and then its neighbor with highest energy. We can discover that the maximum and linear code tracking ranges of the proposed method and the selection DLL method are larger than that of the correlation DLL method, which means that the proposed method and the selective DLL method have lower probability of loss of lock and higher stability under the same steady-state error variance performance. The figure only shows the code tracking curve of the proposed method when $N_b = 8$. When N_b takes a larger value, such as 16 or 32, the code tracking range will be increase. As N_{corr} becomes larger, the code tracking range of the selection DLL method can likewise be further increased.

5. Steady-State Error Variance Analysis

In this section, we first analyze the output of time-delay error estimation and then perform loop analysis to obtain the expression of the steady-state time-delay error variance $\overline{|\varepsilon_n|^2}$.

The frequency domain vector is (16), which can be rewritten as

$$\mathbf{X}_n = \mathbf{A}(\varepsilon_n)F_n + \mathbf{V}_n. \tag{40}$$

Then the covariance matrix can be expressed as

$$\begin{aligned}\mathbf{R}_n &= \overline{\mathbf{X}_n\mathbf{X}_n^H} \\ &= \overline{(\mathbf{A}(\varepsilon_n)F_n + \mathbf{V}_n)(\mathbf{A}(\varepsilon_n)F_n + \mathbf{V}_n)^H} \\ &= \mathbf{A}(\varepsilon_n)\overline{F_nF_n^H}\mathbf{A}^H(\varepsilon_n) + \overline{\mathbf{V}_n\mathbf{V}_n^H} \\ &= \mathbf{A}(\varepsilon_n)\mathbf{A}^H(\varepsilon_n) + \overline{\mathbf{V}_n\mathbf{V}_n^H} \\ &= \mathbf{A}(\varepsilon_n)\mathbf{A}^H(\varepsilon_n) + \mathbf{E}\mathbf{\Lambda}\mathbf{E}^H,\end{aligned} \tag{41}$$

where the orthonormal matrix $\mathbf{E}$ and noise eigenvalue matrix $\mathbf{\Lambda}$ are

$$\begin{aligned}\mathbf{E} &= [\mathbf{S}_1, \mathbf{S}_2, \cdots, \mathbf{S}_{N_b}], \\ \mathbf{\Lambda} &= \mathrm{diag}(\lambda_1, \lambda_2, \cdots, \lambda_{N_b}).\end{aligned} \tag{42}$$

The first eigenvector in the orthonormal matrix $\mathbf{E}$ can be defined as

$$\mathbf{S}_1 = \frac{\mathbf{A}(\varepsilon_n)}{\|\mathbf{A}(\varepsilon_n)\|_2}. \tag{43}$$

Thus, the first noise eigenvalue can be expressed as

$$\begin{aligned}\lambda_1 &= \overline{\mathbf{S}_1^H\mathbf{V}_n\left(\mathbf{S}_1^H\mathbf{V}_n\right)^H} \\ &= \mathbf{S}_1^H\overline{\mathbf{V}_n\mathbf{V}_n^H}\mathbf{S}_1.\end{aligned} \tag{44}$$

The sum of residual noise eigenvalues can be expressed based on the matrix analysis theory as

$$\sum_{i=2}^{N_b}\lambda_i = \mathrm{tr}(\overline{\mathbf{V}_n\mathbf{V}_n^H}) - \lambda_1, \tag{45}$$

where the specific expression of the noise covariance matrix $\overline{\mathbf{V}_n\mathbf{V}_n^H}$ is given in Appendix B. The estimation of the covariance matrix is given by (20), which can be rewritten as

$$\begin{aligned}\hat{\mathbf{R}}_n &= \frac{\tau_1}{T+\tau_1}\hat{\mathbf{R}}_{n-1} + \frac{T}{T+\tau_1}\mathbf{X}_n\mathbf{X}_n^H \\ &= \sum_{i=0}^{n}\frac{T\tau_1^i}{(T+\tau_1)^{i+1}}\mathbf{X}_{n-i}\mathbf{X}_{n-i}^H.\end{aligned} \tag{46}$$

When the system is in a steady state, it can be assumed that $\mathbf{X}_n$ has the same statistical characteristics at different times n. Applying the central limit theorem, we can assume that $\hat{\mathbf{R}}_n$ obeys a complex Wishart distribution with its degree of freedom M expressed as

$$\begin{aligned}\frac{1}{M} &= \lim_{n\to+\infty}\sum_{i=0}^{n}\left[\frac{T\tau_1^i}{(T+\tau_1)^{i+1}}\right]^2 \\ M &= 1 + \frac{2\tau_1}{T}.\end{aligned} \tag{47}$$

Therefore, the error variance in the signal zeros of the modified root-MUSIC algorithm can be expressed as [21]

$$\overline{|\Delta z_n|^2} = \frac{(\lambda_1^s + \lambda_1)\sum_{i=1}^{N_b}\lambda_i}{M(\lambda_1^s)^2}\frac{\left|\mathbf{A}^H(\varepsilon_n)\mathbf{S}_1\right|^2}{\left(\mathbf{S}_1'\right)^H\mathbf{P}_N\mathbf{S}_1'}, \tag{48}$$

where $\mathbf{S}_1'$ is the derivative of $\mathbf{S}_1$ with respect to $2\pi\varepsilon_n/N_b$. Additionally, since the time-delay error ε_n is approximately equal to 0 in steady state, we can substitute $\varepsilon_n = 0$ into (48) when calculating the theoretical value. In this paper, the phase of root can be expressed with normalized time-delay error ε_n as $2\pi\varepsilon_n/N_b$. Therefore, we can express the variance of the estimated normalized time-delay error using the error variance in signal zeros with $\hat{\varepsilon}_n = \varepsilon_n + \tilde{N}_n$ as [21]

$$\overline{\left|\tilde{N}_n\right|^2} = \frac{N_b}{8\pi^2}\overline{|\Delta z_n|^2}. \tag{49}$$

We can observe that $\tilde{N}_n$ is not independent at different times n due to the use of RC (resistor-capacitor) integral filter, which is inconvenient for further processing, but we can model the noise term approximately as white noise passes through an RC (resistor-capacitor) integral filter. The time-delay error estimation can be expressed as

$$\begin{aligned}\hat{\varepsilon}_n &= \varepsilon_n + \tilde{N}_n \\ &= \varepsilon_n + N_n * h_n,\end{aligned} \tag{50}$$

where h_n represents the impulse response of the RC (resistor-capacitor) integral filter. Then, the digital PSD of N_n can be expressed as

$$S_N(e^{j\omega}) = \frac{2\pi\overline{\left|\tilde{N}_n\right|^2}}{\int_{-\pi}^{\pi}\left|H(e^{j\omega})\right|^2 \mathrm{d}\omega}, \tag{51}$$

where $H(e^{j\omega})$ is the discrete-time Fourier transform of h_n.

Subsequently, we get the final expression of the steady-state time-delay error variance $\overline{|\varepsilon_n|^2}$. Since the time-delay error is given by $\varepsilon_n = d - \hat{d}_n$. Referring to (34) and (50), we can get the time-delay error as

$$\begin{aligned}\varepsilon_{n+1} &= d - \hat{d}_{n+1} \\ &= \varepsilon_n - G_d\hat{\varepsilon}_n \\ &= \varepsilon_n - G_d\left(\varepsilon_n + N_n * h_n\right).\end{aligned} \tag{52}$$

In order to get the variance of ε_n, we can represent ε_n and N_n in their respective frequency domain forms as [22]

$$\begin{aligned}\varepsilon_n &= \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{jn\omega}\mathrm{d}\xi_\varepsilon\left(e^{j\omega}\right), \\ N_n &= \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{jn\omega}\mathrm{d}\xi_N\left(e^{j\omega}\right),\end{aligned} \tag{53}$$

where $\xi_\varepsilon\left(e^{j\omega}\right)$ and $\xi_N\left(e^{j\omega}\right)$ are their respective orthogonal incremental processes. By substituting (53) into (52), when the code tracking is in a steady state, we can get

$$\begin{aligned}\varepsilon_n &= \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{jn\omega}\mathrm{d}\xi_\varepsilon\left(e^{j\omega}\right) \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{-G_dH(e^{j\omega})}{e^{j\omega}-1+G_d}e^{jn\omega}\mathrm{d}\xi_N\left(e^{j\omega}\right).\end{aligned} \tag{54}$$

Since $\overline{\mathrm{d}\xi_N\left(e^{j\omega_1}\right)\mathrm{d}\xi_N^*\left(e^{j\omega_2}\right)} = 2\pi\delta_{\omega_1,\omega_2}S_N\left(e^{j\omega_1}\right)\mathrm{d}\omega_1$, and the time-delay error variance $\overline{|\varepsilon_n|^2} = \overline{\varepsilon_n\varepsilon_n^*}$, we can get the expression of the final steady-state time-delay error variance as

$$\begin{aligned}\overline{|\varepsilon_n|^2} &= \frac{S_N(e^{j\omega})}{2\pi}\int_{-\pi}^{\pi}\left|\frac{G_dH(e^{j\omega})}{e^{j\omega}-1+G_d}\right|^2\mathrm{d}\omega \\ &= \frac{N_b\overline{|\Delta z_n|^2}}{8\pi^2\int_{-\pi}^{\pi}\left|H(e^{j\omega})\right|^2\mathrm{d}\omega}\int_{-\pi}^{\pi}\left|\frac{G_dH(e^{j\omega})}{e^{j\omega}-1+G_d}\right|^2\mathrm{d}\omega.\end{aligned} \tag{55}$$

The above equation provides a closed-form expression for the steady-state error variance of the proposed code tracking method. By substituting (48) into the above equation, we can calculate the steady-state time-delay error variance theoretically.

6. Computational Complexity Analysis

In this section, we analyze the computational complexity of the proposed code tracking method and compare it with the correlation DLL method and the selection DLL method [6], [17]. Specifically, we use the number of complex multiplications needed to process the received signal over a single data period as an indicator of their computational complexities. Due to the very small computational complexity, the time delay generation part is not accounted for all three methods.

The computational complexity of each part of the proposed method is shown in Table 2. N_s is the number of sampling points in a single data period, and N_b is the length of frequency domain vector after integration and dump. The overall computational complexity of the proposed code tracking method is the sum of each part, which can be expressed as

$$C_P = \frac{N_s}{2}\log_2(N_s) + 2N_s + 10N_b^3 + 4N_b^2 + 2N_b. \tag{56}$$

For the correlation DLL method, we use FFT to accelerate the matched filter part by default, and the interpolator adopts the cubic interpolation. Then the computational complexity required by the correlation DLL method can be expressed as

$$C_D = 2N_s\log_2(2M) + 6N_s + 2N, \tag{57}$$

where M is the length of the matched filter and N is the length of the spread spectrum code. Similarly, for the selection DLL method, the computational complexity can be expressed as

$$C_S = 2N_s\log_2(2M) + 6N_s + N_{corr}N, \tag{58}$$

where N_{corr} is the number of correlation outputs used in the selection DLL method.

We can perform a specific comparative analysis of the computational complexity of the three methods by substituting typical values. We set the length of the spread spectrum code $N = 1023$. In the proposed method, we set the number of sampling points in a single data period $N_s = 2(N+1)$. In the correlation DLL method and the selection method we set $N_s = 2N$, the length of matched filter $M = 16$. At the same time, we set N_b and N_{corr} so that the proposed method and the selected DLL method have approximate maximum and linear code tracking ranges. When the number of frequency domain vector after integration and dump $N_b = 8$ in the proposed method and the correlation number $N_{corr} = 4$ in the selection DLL method, the computational complexity ratio of the proposed method to the correlation DLL method is $C_P/C_D = 0.60$, the ratio of the proposed method to the selection DLL method is $C_P/C_S = 0.56$. When $N_b = 16$

Table 2 Computational complexity of each part of the proposed code tracking method.

Frequency domain vector generation	Time-delay error estimation			
	Covariance estimation	Eigen decomposition	Matrix **B** generation	Polynomial rooting
$\frac{N_s}{2}\log_2 N_s + 2N_s$	$3N_b^2$	N_b^3	$(N_b-1)N_b^2 + 2N_b(N_b+1)$	$(2N_b)^3$

and $N_{corr} = 8$, the computational complexity ratios are $C_P/C_D = 1.65$ and $C_P/C_S = 1.40$, respectively. When $N_b = 32$ and $N_{corr} = 16$, the ratios are $C_P/C_D = 9.98$ and $C_P/C_S = 7.07$, respectively.

We can find out that the proposed method's computational complexity is comparable to that of the correlation DLL method and the selection DLL method when $N_b = 8$ or $N_b = 16$. Additionally, as the value of N_b increases, the computational complexity of the proposed method increases significantly. In summary, we recommend using relatively small N_b when using the proposed method for code tracking in band-limited DSSS systems. If a larger maximum and linear code tracking ranges is required, a larger N_b can be chosen.

7. Simulations

In this section, we perform three simulations to analyze and verify the proposed code tracking method for band-limited DSSS systems. First, we perform simulation to analyze the advantages of the proposed method in terms of the maximum and linear code tracking ranges. Second, we verify the theoretical steady-state error variance closed-form expression. Finally, we compare the proposed method with the correlation DLL method and the selection DLL method [6], [17], in terms of steady-state error variance performance, using Monte Carlo method.

First, we perform simulation to analyze the advantages of the proposed method in terms of the maximum and linear code tracking ranges. We examine the differences between the proposed method, the correlation DLL method and the selection DLL method in their code tracking results under various initial time delays τ. We set the length of the spread spectrum code $N = 1023$, chip rate $1/T_c = 10.23$ MHz, roll-off factor of the RRC pulse waveform $\alpha = 0.25$, code period equal to data period, i.e., $T = NT_c = N_sT_s$, loop bandwidth $B_L = 10$ Hz, and SNR $= -10$ dB. We consider five different initial time delays as $\tau = T_s$, $\tau = 2T_s$, $\tau = 3T_s$, $\tau = 3.9T_s$, and $\tau = 7.9T_s$ where T_s is the sampling period of the proposed method. For the proposed method, we set sampling rate $1/T_s = 20.48$ MHz, $N_b = 16$ when $\tau = 7.9T_s$, and $N_b = 8$ for other values of τ. For the correlation DLL method, we set sampling rate $1/T_s = 20.46$ MHz, length of matched filter $M = 16$, and use cubic interpolater to adjust fractional time delay. For the selection DLL method, we set the correlation number $N_{corr} = 8$ when $\tau = 7.9T_s$, $N_{corr} = 4$ for other values of τ, other parameters are consistent with the correlation DLL method, and the selection algorithm determines first the branch with maximal energy and then its neighbor with highest energy. The code tracking results are presented in Fig. 3.

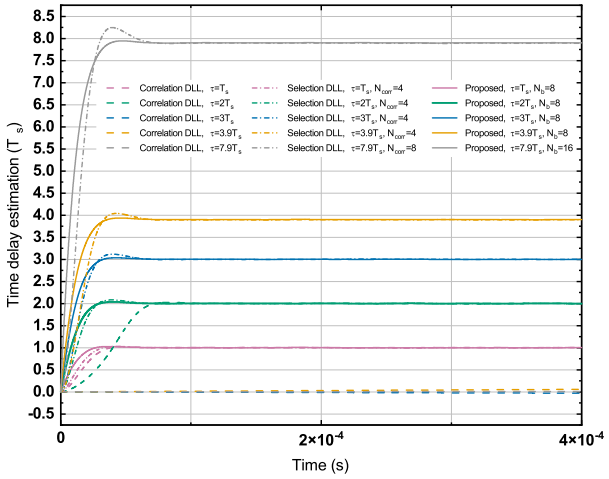


Fig. 3 Code tracking results of the three methods under different initial time delays.

In Fig. 3, the solid lines represent the code tracking results of the proposed method, the dashed lines represent the code tracking results using the correlation DLL method and the dashdotted lines represent the code tracking result of the selection DLL method. The x-axis is the time, united by second, and the y-axis is the estimation of the time delay, united by the sampling period of the proposed method.

For an initial time delay $\tau = T_s$, all three methods establish a steady state quickly, and at this time, τ falls within the approximate linear code tracking range of all three methods. For an initial time delay $\tau = 2T_s$, the proposed method and the selection DLL method can establish a steady state faster. At this time, τ falls within the nonlinear code tracking range of the correlation DLL method, but falls within the linear code tracking range of the other two methods still. When we set the initial time delay to be $\tau = 3T_s$, $\tau = 3.9T_s$ and $\tau = 7.9T_s$, the correlation DLL method fails to track as the three initial time delays exceed its maximum tracking range, but the proposed method and the selection DLL method can track effectively still. Therefore, the proposed method and the selection DLL method both have larger maximum and linear code tracking ranges compared to the correlation DLL method, which means lower probability of loss of lock and higher stability for the same steady-state error variance performance.

Second, to verify the accuracy of the theoretical steady-state error variance closed-form expression of the proposed method in Sect. 5, we compare the theoretical and simulation results under different SNRs. We set $N_b = 8$, the loop bandwidths $B_L = 10$ Hz and $B_L = 100$ Hz, and the other parameters remain unchanged. The steady-state error variance results are presented in Fig. 4.

The x-axis of the Fig. 4 is SNR of the received signal,

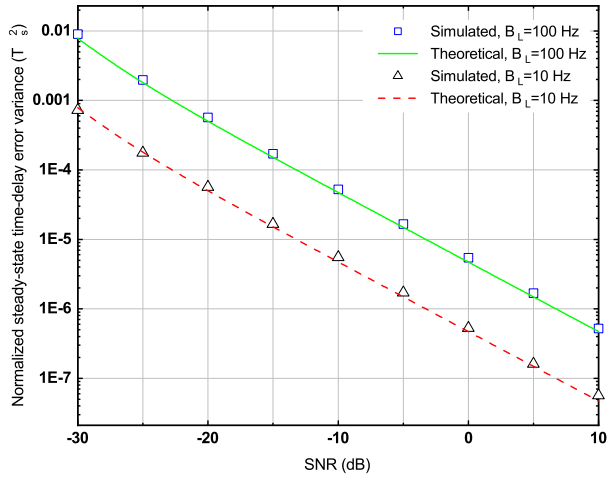


Fig. 4 Theoretical and simulated steady-state time-delay error variances.

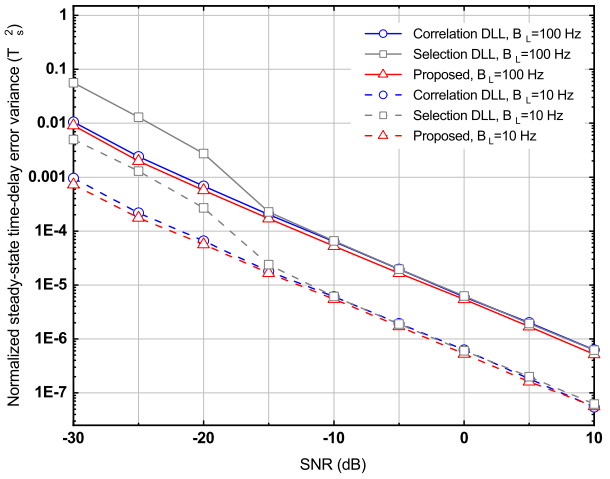


Fig. 5 Steady-state time-delay error variances of the proposed method, the correlation DLL method and the selection DLL method.

which is SNR = $10\lg(PT_s/N_0)$. The y-axis is the normalized steady-state time-delay error variance. The theoretical results are calculated by (55), and the simulation result corresponding to each SNR under each loop bandwidth B_L is obtained by performing 50 Monte Carlo simulations, and the duration of the steady-state tracking used in each simulation is set to 1 second to obtain a stable error variance result.

As shown in Fig. 4, the theoretical and simulated steady-state time-delay error variances bear again a good agreement. The results demonstrate that the variance of the steady-state delay error decreases as the loop bandwidth decreases and the SNR of the received signal increases. Additionally, the simulation results are slightly greater than the theoretical values, which may be attributed to the approximations used in the intermediate steps during the theoretical derivation process.

Finally, we compare the steady-state time-delay error variance performance of the proposed code tracking method, the correlation DLL method and the selection DLL method [6], [17]. The parameters settings remain unchanged and the results are shown in Fig. 5. The error variance corresponding to each SNR under each method is obtained by performing 50 Monte Carlo simulations with the duration of the steady-state tracking used in each simulation is set to 1 second. The y-axis has been normalized using the sampling period of the proposed method.

As shown in Fig. 5, the correlation DLL method performs slightly worse than the proposed code tracking method under different SNRs and loop bandwidths. The reason may be that the correlation DLL method uses cubic interpolation for delay adjustment, which has a larger error than the frequency domain delay adjustment method. Besides that, under low SNRs, the steady-state time-delay error variance performance of the selection DLL method is worse than the other two methods. This is due to the fact that the probability of selecting the wrong correlation outputs is higher for the selective DLL method under low SNRs, leading to deteriorated error variance performance. Therefore, the steady-state time-delay error variance performance of the proposed method is similar to that of the correlation DLL method, and the selected DLL method has a severe performance deterioration under low SNRs.

In summary, compared with the correlation DLL method, the proposed code tracking method in this paper has advantages in maximum and linear code tracking ranges, which means lower probability of loss of lock and higher stability, with constant error variance performance and appropriate computational complexity. In contrast, the selection DLL method has a severe error variance performance deterioration under low SNRs.

8. Conclusion

This paper proposes a high stability code tracking method for band-limited DSSS systems, which has a larger maximum and linear code tracking ranges, constant error variance performance and appropriate computational complexity. Then we analyze the code tracking range, steady-state error variance and computational complexity of the proposed method, and give a closed-form expression for the steady-state error variance, which can be used to analyze the error variance performance theoretically of the proposed method and design proper band-limited DSSS systems. Beside, simulations are also made to analyze and verify the proposed method. The results of both the theoretical analysis and simulations show that the proposed code tracking method has a larger maximum and linear code tracking ranges which means lower probability of loss of lock, constant error variance performance and appropriate computational complexity, and it can be used in band-limited DSSS systems and improve its stability.

References

[1] A. Modenini and B. Ripani, "A tutorial on the tracking, telemetry, and command (TT&C) for space missions," IEEE Commun. Surveys Tutss, vol.25, no.3, pp.1510–1542, 2023.
[2] K. Tomita, M. Okumura, and E. Okamoto, "Demonstration of chaos-

based radio encryption modulation scheme through wired transmission experiments," IEICE Trans. Commun., vol.E106-B, no.8, pp.686–695, Aug. 2023.
[3] K.J. Quirk and M. Srinivasan, "PN code tracking using noncommensurate sampling," IEEE Trans. Commun., vol.54, no.10, pp.1845–1856, 2006.
[4] W. Liu, Y. Hu, T.H. Hsieh, J. Zhao, and S. Wang, "Quinary offset carrier modulations for global navigation satellite system," IEICE Trans. Commun., vol.E104-B, no.5, pp.563–569, May 2021.
[5] J.W. Betz and K.R. Kolodziejski, "Generalized theory of code tracking with an early-late discriminator part I: Lower bound and coherent processing," IEEE Trans. Aerosp. Electron. Syst., vol.45, no.4, pp.1538–1556, 2009.
[6] R. De Gaudenzi, M. Luise, and R. Viola, "A digital chip timing recovery loop for band-limited direct-sequence spread-spectrum signals," IEEE Trans. Commun., vol.41, no.11, pp.1760–1769, 1993.
[7] F.J. Harris and M. Rice, "Multirate digital filters for symbol timing synchronization in software defined radios," IEEE J. Sel Areas Commun., vol.19, no.12, pp.2346–2357, 2001.
[8] Z. Gao, M. Zhou, P. Reviriego, and J.A. Maestro, "Efficient fault-tolerant design for parallel matched filters," IEEE Trans. Circuits Syst. II, Exp. Briefs, vol.65, no.3, pp.366–370, 2017.
[9] M. Srinivasan, C.C. Chen, G. Grebowsky, and A. Gray, "An all-digital, high data-rate parallel receiver," JPL TDA Progress Report, vol.42, p.131, 1997.
[10] C. Wang, C. Lin, Q. Chen, B. Lu, X. Deng, and J. Zhang, "A 10-gbit/s wireless communication link using 16-QAM modulation in 140-GHz band," IEEE Trans. Microw. Theory Techn., vol.61, no.7, pp.2737–2746, 2013.
[11] Z. Lu and Y. Jiao, "Efficiently all-digital code tracking for band-limited DSSS systems," IEEE Commun. Lett., vol.27, no.2, pp.686–690, 2023.
[12] J.H. Won, "A novel adaptive digital phase-lock-loop for modern digital gnss receivers," IEEE Commun. Lett., vol.18, no.1, pp.46–49, 2013.
[13] I. Cortés, J.R. Van Der Merwe, A. Rügamer, and W. Felber, "Adaptive loop-bandwidth control algorithm for scalar tracking loops," 2020 IEEE/ION Position, Location and Navigation Symposium (PLANS), pp.1178–1188, IEEE, 2020.
[14] X. Chen, F. Dovis, S. Peng, and Y. Morton, "Comparative studies of GPS multipath mitigation methods performance," IEEE Trans. Aerosp. Electron. Syst., vol.49, no.3, pp.1555–1568, 2013.
[15] H. Wang, Q. Chang, Y. Xu, and X. Li, "Adaptive narrow-band interference suppression and performance evaluation based on code-aided in GNSS inter-satellite links," IEEE Syst. J., vol.14, no.1, pp.538–547, 2019.
[16] A. Wilde, "Extended tracking range delay-locked loop," Proc. IEEE International Conference on Communications ICC'95, pp.1051–1054, IEEE, 1995.
[17] A. Wilde, "Correlation branch selection for extended tracking range delay-locked loops," European Transactions on Telecommunications, vol.9, no.1, pp.57–64, 1998.
[18] S.C. Pei and Y.C. Lai, "Closed form variable fractional time delay using FFT," IEEE Signal Process. Lett., vol.19, no.5, pp.299–302, 2012.
[19] M. Blok, "Comments on "closed form variable fractional time delay using FFT"," IEEE Signal Process. Lett., vol.20, no.8, pp.747–750, 2013.
[20] F.M. Gardner, Phaselock Techniques, John Wiley & Sons, 2005.
[21] B.D. Rao and K.S. Hari, "Performance analysis of root-music," IEEE Trans. Acoust., Speech, Signal Process., vol.37, no.12, pp.1939–1949, 1989.
[22] P.J. Brockwell and R.A. Davis, Time Series: Theory and Methods, Springer Science & Business Media, 2009.

Appendix A: Derivation of the Frequency Domain Waveform G_k

We set $g_\ell = \mathrm{IFFT}_{N_s}[G_k]$, and get the expression of G_k by solving g_ℓ.

Finding the solution for g_ℓ is generally a simplifying process of the correlation operation. We use the corresponding relationship between frequency domain conjugate multiplication and time domain circular convolution and $[\cdot]_{N_s}$ to represent the N_s points cycle extension of a signal. Then g_ℓ can be represented as (A· 1).

Because the pulse waveforms $p_R(t)$ and $p_T(t)$ are approximately equal to 0 when $|t| > T/2$, where $T = NT_c = N_sT_s$. The finite accumulations about i and infinite accumulations about x_1, x_2 in (A· 1) can be rewritten as an infinite accumulation about i. Meanwhile, the discrete convolution $p_R(iT_s) * p_T(iT_s) = p(iT_s)/T_s$ corresponds to the continuous convolution $p_R(t) * p_T(t) = p(t)$. $p_T(t)$ is the RRC pulse waveform of the transmitted signal, $p_R(t)$ is the local RRC pulse waveform, and $p(t)$ is the RC pulse waveform, with the Fourier transform $P(f) = T_cP_N(f)$. Thus, (A· 1) can be further simplified to (A· 2).

Based on the orthogonality property of spread spectrum code, the latter term in (A· 2) is approximately equal to 0. Therefore, (A· 2) can be further simplified as

$$g_\ell \approx p(\ell T_s). \tag{A· 3}$$

At the same time, the Fourier transform of $p(t)$ is known to be $P(f) = T_cP_N(f)$, then G_k can be expressed finally as

$$\begin{aligned} G_k &= \mathrm{FFT}_{N_s}[g_\ell] \\ &= \mathrm{FFT}_{N_s}[p(\ell T_s)] \\ &= \frac{T_c}{T_s} P_N\left(\frac{k}{N_sT_s}\right). \end{aligned} \tag{A· 4}$$

Appendix B: Derivation of the Frequency Domain Noise Covariance Matrices $\overline{\tilde{\mathbf{W}}_n\tilde{\mathbf{W}}_n^H}$ and $\overline{\mathbf{V}_n\mathbf{V}_n^H}$

We define $\tilde{w}_{\ell,n} = \mathrm{IFFT}_{N_s}\left[\tilde{W}_{k,n}\right]$, then $\tilde{w}_{\ell,n}$ can be expressed as (A· 5), where $w'(t) = p_R(t) * w(t)$.

Furthermore, given that the PSD of $w(t)$ is N_0/P, and the frequency response of $p_R(t)$ is $P_R(f) = \sqrt{P_N(f)}$, thus the PSD of $w'(t)$ is $N_0P_N(f)/P$.

As the spread spectrum code is uncorrelated with the noise term in (A· 5), the PSD of $\tilde{w}(t)$ can be represented by $N_0P_N(f)/(NP)$. Assuming the noise bandwidth of the baseband signal is $1/T_s$, we have $\tilde{w}_{\ell,n} = \tilde{w}[(\ell + nN_s)T_s]$. Thus, the digital PSD of $\tilde{w}_{\ell,n}$ is

$$S_{\tilde{w}}(e^{j\omega}) = \frac{N_0}{PNT_s} P_N\left(\frac{\omega}{2\pi T_s}\right). \tag{A· 6}$$

Then we can get that

$$
\begin{aligned}
g_\ell &= \mathrm{IFFT}_{N_s}\left[G_k\right] \\
&= \mathrm{IFFT}_{N_s}\left[\left(G_k^{RX}\right)^* G_k^{TX}\right] \\
&= \frac{T_s}{N}\sum_{i=0}^{N_s-1}\left[\sum_{m_1=0}^{N-1} c_{m_1} p_R\left(iT_s - m_1T_c\right)\right]_{N_s}\left[\sum_{m_2=0}^{N-1} c_{m_2} p_T\left[(i+\ell)T_s - m_2T_c\right]\right]_{N_s} \\
&= \frac{T_s}{N}\sum_{i=0}^{N_s-1}\sum_{m_1=0}^{N-1} c_{m_1}\left[p_R\left(iT_s - m_1T_c\right)\right]_{N_s}\sum_{m_2=0}^{N-1} c_{m_2}\left[p_T\left[(i+\ell)T_s - m_2T_c\right]\right]_{N_s} \\
&= \frac{T_s}{N}\sum_{i=0}^{N_s-1}\sum_{m_1=0}^{N-1} c_{m_1}\sum_{x_1=-\infty}^{\infty} p_R\left[iT_s - (m_1+x_1N)T_c\right]\sum_{m_2=0}^{N-1} c_{m_2}\sum_{x_2=-\infty}^{\infty} p_T\left[(i+\ell)T_s - (m_2+x_2N)T_c\right] \\
&= \frac{T_s}{N}\sum_{m_1=0}^{N-1}\sum_{m_2=0}^{N-1} c_{m_1}c_{m_2}\sum_{i=0}^{N_s-1}\sum_{x_1=-\infty}^{\infty} p_R\left[iT_s - (m_1+x_1N)T_c\right]\sum_{x_2=-\infty}^{\infty} p_T\left[(i+\ell)T_s - (m_2+x_2N)T_c\right]. \qquad (A\cdot 1)
\end{aligned}
$$

$$
\begin{aligned}
g_\ell &= \frac{T_s}{N}\sum_{m_1=0}^{N-1}\sum_{m_2=0}^{N-1} c_{m_1}c_{m_2}\sum_{i=-\infty}^{\infty} p_R\left(iT_s - m_1T_c\right) p_T\left[(i+\ell)T_s - m_2T_c\right] \\
&= \frac{1}{N}\sum_{m_1=0}^{N-1}\sum_{m_2=0}^{N-1} c_{m_1}c_{m_2}\, p\left[\ell T_s + (m_1-m_2)T_c\right] \\
&= p\left(\ell T_s\right) + \frac{1}{N}\sum_{m_1=0}^{N-1}\sum_{\substack{m_2=0 \\ m_2\neq m_1}}^{N-1} c_{m_1}c_{m_2}\, p\left[\ell T_s + (m_1-m_2)T_c\right]. \qquad (A\cdot 2)
\end{aligned}
$$

$$
\begin{aligned}
\tilde{w}_{\ell,n} &= \mathrm{IFFT}_{N_s}\left[\tilde{W}_{k,n}\right] \\
&= \mathrm{IFFT}_{N_s}\left[\left(G_k^{RX}\right)^* W_{k,n}\right] \\
&= \frac{T_s}{N}\sum_{i=0}^{N_s-1}\left[\sum_{m_1=0}^{N-1} c_{m_1} p_R\left(iT_s - m_1T_c\right)\right]_{N_s} w\left[(i+\ell+nN_s)T_s\right] \\
&= \frac{T_s}{N}\sum_{i=0}^{N_s-1}\sum_{m_1=0}^{N-1} c_{m_1}\sum_{x_1=-\infty}^{\infty} p_R\left[iT_s - (m_1+x_1N)T_c\right] w\left[(i+\ell+nN_s)T_s\right] \\
&= \frac{T_s}{N}\sum_{m_1=0}^{N-1} c_{m_1}\sum_{i=-\infty}^{\infty} p_R\left[iT_s - (m_1+x_1N)T_c\right] w\left[(i+\ell+nN_s)T_s\right] \\
&= \frac{1}{N}\sum_{m_1=0}^{N-1} c_{m_1} w'\left[(\ell+nN_s)T_s\right], \qquad (A\cdot 5)
\end{aligned}
$$

$$
\overline{\tilde{W}_{k,n}\tilde{W}_{k,n}^*} = \frac{N_sN_0}{PNT_s}P_N\left(\frac{k}{N_sT_s}\right). \qquad (A\cdot 7)
$$

Since the noise at different frequency indices is uncorrelated, the frequency domain noise covariance can be expressed as

$$
\overline{\tilde{\mathbf{W}}_n\tilde{\mathbf{W}}_n^H} = \frac{N_sN_0}{PNT_s}\begin{bmatrix} P_N(0) & & & \\ & P_N\left(\frac{1}{N_sT_s}\right) & & \\ & & \ddots & \\ & & & P_N\left(\frac{N_s-1}{N_sT_s}\right) \end{bmatrix}. \qquad (A\cdot 8)
$$

The digital PSD of $v_{\ell,n}$, where $v_{\ell,n} = \mathrm{IFFT}_{N_b}\left[V_{k,n}\right]$, is approximately equal to that of $\tilde{w}_{\ell,n}$, thus we can get that

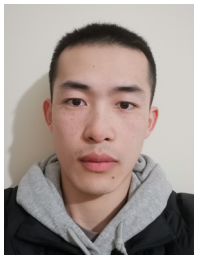
$$
\overline{V_{k,n}V_{k,n}^*} = \frac{N_bN_0}{PNT_s}P_N\left(\frac{k}{N_bT_s}\right). \qquad (A\cdot 9)
$$

Then the frequency domain noise covariance after integration and dump can be expressed as

$$
\overline{\mathbf{V}_n\mathbf{V}_n^H} = \frac{N_bN_0}{PNT_s}\begin{bmatrix} P_N(0) & & & \\ & P_N\left(\frac{1}{N_bT_s}\right) & & \\ & & \ddots & \\ & & & P_N\left(\frac{N_b-1}{N_bT_s}\right) \end{bmatrix}. \qquad (A\cdot 10)
$$

Zhiwei Lu received the B.S. degree from Space Engineering University, Beijing, China, in 2019, where he is currently pursuing the Ph.D. degree in information and communication engineering. His research interests include space TT&C technology, spread spectrum technology, and signal processing.

Yiwen Jiao received the B.S. degree in electronic and information technology from the Beijing University of Aeronautics and Astronautics, in 2008, and the M.S. and Ph.D. degrees from the Institution of Command and Technology of Equipment, Beijing, in 2010 and 2015, respectively. He is currently working as an Associate Professor at Space Engineering University, Beijing, China. His research interests include space TT&C technology, high-speed signal processing, and advanced signal processing.



Yudi Chen received the B.S. degree from Space Engineering University, Beijing, China, in 2018, where he is currently pursuing the Ph.D. degree. His research interests include TT&C of spacecraft, target tracking and signal processing.

PAPER

Adaptability of Breadcrumbs Routing to Geographical Locality of Content Popularity in Information Centric Networks

Yusaku HAYAMIZU[†a)], ***Member and*** **Miki YAMAMOTO**[††], ***Fellow***

SUMMARY In Information-Centric Networking (ICN), different routing and caching schemes have been proposed to efficiently utilize in-network caches and reduce network traffic. Most of them assume that the popularity distribution of user-requested content is homogeneous. However, the actual popularity distribution measured on the Internet is reported to possess spatial and temporal localities, which can heavily affect caching performance in ICN. Breadcrumbs (BC) routing is a key solution to mitigate performance degradation due to spatial locality because of its ability to flexibly discover cached contents in the off-path. In this paper, we deeply investigate the spatial effects of BC by revealing where utilized cached contents are located, how BC discovers these contents, what kind of contents are found, and how BC fill in the locality gap of content popularity. We also focus on another time-dimension perspective, i.e., the temporal locality of content popularity, and conduct a comprehensive study of how BC routing can be adapted to the spatiotemporal locality of content popularity in ICN.

key words: *ICN, CCN, NDN, breadcrumbs, routing, caching, content popularity, locality*

1. Introduction

Various research communities have studied Information-Centric Networking (ICN), which includes Content-Centric Networking (CCN) [1] and Named-Data Networking (NDN) [2], to support efficient content distribution [1]–[4] and address the problem of content traffic explosion [5]. In ICN, content can be provided from anywhere in the network, that is, not only from servers but also from *in-network cache(s)*, thus enabling users to quickly retrieve their requested contents with low overhead. Therefore, a crucial technical challenge is *cache control, i.e., caching*, and many effective approaches have been designed and proposed thus far [3].

Locality has been reported to exist in the popularity of content, and it can be classified into *spatial locality* and *temporal locality* [6]–[8]. Spatial locality refers to the existence of a spatial or geographical bias in the popularity distribution of traffic demand. For example, a popular content in an area is not demanded at all in other areas, or a content that is not demanded in an area is frequently demanded in another. Simply put, global content popularity distribution differs from local content popularity distribution in a specific area. This paper focuses on this spatial locality of content popularity and calls it *geographical locality*. Brodersen et al.'s [6] trend analysis of more than 20 million views of YouTube content over one year revealed that for about 50% of all videos, more than 70% of views were from one geographic area. This case shows a large gap between global content popularity distribution and the local distribution in that area. Different traffic control schemes [9]–[12] assume that the popularity distribution of content is *homogeneous*, meaning both global and local content popularity distributions follow identical Zipf-like distributions [13]. Therefore, the effectiveness of these proposed approaches would be suspicious in an environment where the spatial locality of content popularity exists.

Another key solution to effectively utilize in-network cached content is the routing method of request packets, as request routing triggers cache hits in ICN. In the original ICN [1], [2], request packets are transferred through the shortest path from a user to a server based on the installed routing table on each router; thus, cache utilization is limited only to the shortest path, i.e., the *on-path*. However, intermediate routers have another option to forward request packets to the outside of the shortest path, the *off-path*, for better cache utilization. This routing method, also known as off-path routing [9], is reported to produce high cache hit performance by cooperatively combining routing and caching policies in ICN [9], [14]. Considering a situation in which a content popularity gap exists between the local and global areas, cached content in each area might substantially differ. When we apply blinkered on-path routing, there would be little possibility of cache hits. In this sense, off-path routing is expected to expand the scope of cache discovery and makes it possible to improve cache utilization in the case of locality in cache popularity.

This paper investigates the impact of content popularity locality on caching performance in ICN. According to the locality of content popularity, we focus on spatial locality because of the strong relation between content routing and the spatial distribution of cached contents. Regarding routing policies, we specifically adopt shortest-path routing (SPR) and breadcrumbs (BC) for performance evaluation. SPR is a widely used on-path routing policy, and BC is a promising off-path routing method that enables best-effort content routing toward cached content outside the on-path. For caching decisions, transparent en-route cache* (TERC) and Fix(p) are applied because each of these approaches is

Manuscript received May 22, 2024.
Manuscript publicized August 1, 2024.

[†]Network Research Institute at the National Institute of Information and Communications Technology (NICT), Koganei-shi, 184-8795 Japan.
[††]Faculty of Engineering Science, Kansai University, Suita-shi, 564-8680 Japan.
a) E-mail: hayamizu@nict.go.jp

DOI: 10.23919/transcom.2024EBP3085

*Some studies call this caching decision strategy "cache everything everywhere."

reported to show high cache hit performance with BC and SPR, respectively [14].

Our detailed evaluation reveals two findings. First, spatial locality leads to poor performance for SPR. In SPR, content requests from a regional area with spatial locality are just routed on the shortest path to the content server. This content search regulation degrades the cache hit performance of the content requests from this area. Second, BC provides good cache hit performance even in the spatial locality case. The content search capability of BC's off-path routing makes content requests find locally popular content cached at a router outside the area with geographical locality. This produces good cache hit performance for locally popular content. Therefore, BC can fill the locality gap of content popularity and generate high cache hit performance in a heterogeneous environment compared to SPR as will be described in Sect. 3. In this paper, we define *adaptability* as ability to adapt to the variation of content popularity and to keep high cache hit performance. Also, Sect. 3.2 not only discusses the one-dimensional aspect, i.e., content being obtained from the on-path or the off-path, but also conducts a simulation analysis to precisely investigate where and how BC is used in the network, what kind of cached contents are finally discovered, and why cache hit performance is improved.

In this paper, we also discuss another dimension of content locality: temporal locality. In Sect. 4, we focus on the time-dimensional aspect and confirm the adaptability of BC to the temporal locality of content popularity in Sect. 4. Content popularity has temporal locality when the popularity of content significantly changes over time. For example, content that is not previously demanded becomes frequently requested at a certain period and vice versa. It is reported that the characteristics of this phenomenon vary depending on content attributes and that the variation patterns of such popularity can be diverse [15]. We evaluate performance in a situation where content popularity changes temporally and reveal that TERC performs well because popular contents are quickly spread all over the network caches. BC shows good performance when combined with TERC, which means BC also works well even in the temporal locality case.

The remainder of this paper is organized as follows. Section 2 introduces our research background, CCN/NDN and BC routing. Section 3 investigates the adaptability of BC to the spatial locality of content popularity and validates the effectiveness of BC routing through detailed simulation analysis. Section 4 sheds light on another factor of content locality: temporal locality. Section 6 discusses some remaining issues and concludes this paper.

2. Background Technology

2.1 CCN/NDN

CCN/NDN adopts a receiver-driven communication model, where communication begins from a user's content request. The user sends a content request packet called *Interest*, where the name of the desired content is described, and communication is performed by the response packet, i.e., *Data* packet transmission, which corresponds to the Interest packet. The Forwarding Information Base (FIB), which is a routing table in a CCN/NDN router, is used to transmit an Interest packet, which is then transferred to content server(s) based on FIB direction. During Interest transmission, the interface at which the Interest packet arrives is recorded in the Pending Interest Table (PIT), which will be used for later Data packet transfer. When the Interest packet finally reaches the server, the corresponding Data packet is returned along the PIT trail. During data transmission, each intermediate router on the transmission path temporarily stores the Data packet in its own Content Store as in-network cache.

In CCN/NDN, Interest-forwarding methods are classified into on-path routing and off-path routing. On-path routing, also known as SPR, transfers Interest packets on the shortest path from a user to a server. Meanwhile, off-path routing is performed to effectively utilize cached content outside the shortest path, and several effective routing methods have been proposed thus far [9]–[12]. From a network load viewpoint, this paper focuses on the BC routing method, which is a lightweight best-effort routing strategy for leveraging off-path-cached content as described in the next section.

2.2 Breadcrumbs: Best-Effort Off-Path Routing

BC has been proposed as a routing method to effectively use off-path caches in networks [16]. Figure 1 shows an example of a BC operation. During data download, each router records the download direction as breadcrumbs (bc)[†]. This bc refers to the temporary route information that indicates the data download direction where the data may exist as in-network caches in the off-path. The sequence of these pointers constructed by data download is called the bc trail. When a user's request occasionally arrives at a router holding the corresponding bc, the request-forwarding policy switches from SPR- to BC-forwarding mode. In the BC-forwarding mode, the request is forwarded toward the bc's direction at each router. This guidance repeats until the request finds the cached content. If no cached content is found, the bc trail

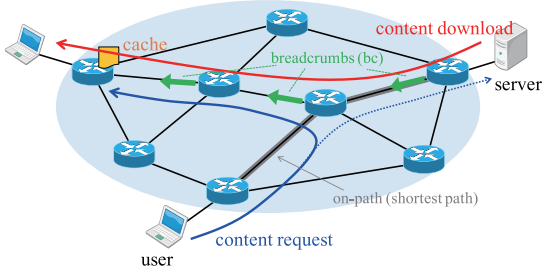


Fig. 1 In-network guidance by BC forwarding.

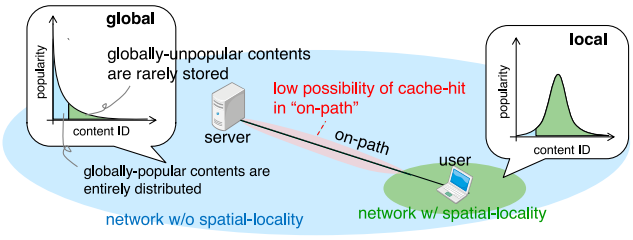
[†]This paper distinguishes "BC," the routing method name, from "bc," the routing pointer information held by routers.

is immediately invalidated, and the request is forwarded to the origin server. In this manner, BC utilizes other users' download information as a path to discover off-path caches. Thus, BC may effectively discover caches stored outside the on-path with low network traffic overhead.

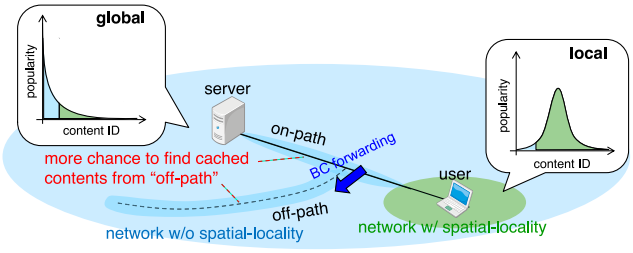
3. Adaptability to Spatial Locality

3.1 Hypothesis

First, we briefly explain the kind of adaptability BC possesses to the regional characteristics of content popularity. Figure 2 shows the difference in cache usage characteristics when SPR and BC are used as routing policies. Figure 2(a) shows that in SPR, the cache discovery range is highly limited to the shortest path (on-path), which tends to store popular contents in that area. When the request comes from another area where the popularity distribution largely differs from the global one, one would expect only a few cache hits because the requested content could be globally unpopular and rarely stored in the on-path. Meanwhile, when the routing policy uses BC, a wide range of cached content in the network can be discovered because the request is likely to be forwarded to the off-path (BC forwarding) as shown in Fig. 2(b). The cache search range expands to the off-path alongside the on-path, with a high possibility that relatively diverse cached contents stably exist. Therefore, even if the request is not satisfied at the on-path, the content would be found in the off-path, and the user may acquire the content from the cache.



(a) SPR



(b) BC

Fig. 2 Difference in cache retrieval locations between SPR and BC.

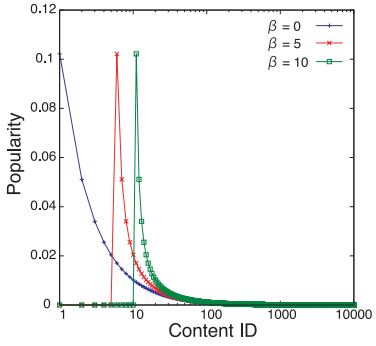


Fig. 3 Popularity distribution with/without geographical locality of contents requested by users.

3.2 Simulation Analysis

3.2.1 Model

To reveal the effects of BC routing, we conducted simulation analysis using a network simulator. We used ndnSIM [17], a popular ICN network simulator developed based on ns-3 [18]. The network model is the Barabási–Albert (BA) [19] model with 100 routers, with one user connected to each router. The content count is 10000, which are randomly arranged on five servers without duplication; that is, each server manages 2000 contents. Each router has a cache size of five contents, and each user's request generation time interval follows an exponential distribution with an average of 1.0 [sec].

Because of the difficulty in modeling the geographical locality of the content popularity distribution in a regional area, we simply shifted the popularity rank of the Zipf distribution with reference to [20] because of the lack of actual trace data. Figure 3 shows the popularity distribution of the contents requested by the users. The popularity distribution of request generation in non-regional areas follows the Zipf's law ($\alpha = 1.0$). In a regional area, the popularity ranks of content files were determined by shifting a rank order on the content IDs by the geographical locality parameter β. The regional area is called the network with geographical locality and is denoted as *area x* shown as the green network in Fig. 2. It is reported that for 50% of YouTube videos, at least 70% of views come from a single area [6]. Thus, in the initial evaluation, we assume a situation in which the network size of this area x is small, i.e., 1. For parameter setting of the network size for the area x, we tried several parameters for the network size to evaluate performance. However, there are no significant differences on cache hit performance. It is reported that the size of network with geographic localities is small in the existing study [6], so that we just selected the small parameter as 1. Users' geographic distance has also been reported to be highly associated with users' social interaction; therefore, a wide geographic distance would cause content localization [6]. Therefore, we set the location of area x as the router with the lowest betweenness centrality in this evaluation. Note that we focused only on the spatial locality of content popularity in this evaluation; the temporal aspect of this popularity will be discussed in Sect. 4.

With regard to cache policies, we adopted TERC, a typical cache decision policy widely used in ICN, and Least Recently Used (LRU), which is also a standard cache replacement method. Regarding the routing methods for con-

tent requests, we used SPR, the shortest path selection commonly used as on-path routing, and the BC method, which refers to best-effort off-path routing as mentioned previously. Throughout the evaluations, we focused only on cache hit performance for users that exist in area x; that is, cache hit count and other performance metrics are measured for the requests transmitted by users in area x. For the other performance metrics, fairness and network load are important factors in computer networking. It is well studied that traffic congestion is expected to be alleviated by utilizing BC routing [21]. However, the main focal point of this paper is evaluating/analyzing the ability to discovery cached contents, and thus we do not focus on the transport aspect such as throughput and content retrieval time. Also, BC routing causes extension of path length and its performance have evaluated in the work [22]. Such path stretch would also occur in the locality scenario of this paper.

3.2.2 Results

Overall Cache Hit Performance: Figure 4 shows the cache hit rates of SPR and BC in an environment where content popularity has spatial locality. With the increase in the geographical locality parameter β, which represents the strength of geographical locality, the on-path cache hit rate of SPR significantly decreases. When BC is used for the request routing policy, the on-path cache hit rate decreases as well. However, because the off-path cache hit rate does not decrease substantially, the high rate is maintained. These results show that BC can fill the popularity gap between the regional and nonregional areas while maintaining a high cache hit performance.

To analyze the adaptability of the BC method to spatial locality, we investigated the cache hit performance of each request routing policy with and without the locality. We used β as a parameter to quantify the locality and evaluate the performance for $\beta = 0, 10$. Figure 5 shows the on-path cache hit characteristics of SPR. The cache hit counts in the overall network (Fig. 5(a)), area x (Fig. 5(b)), and area $\bar{x}$ (Fig. 5(c)) are shown. Comparing cache hit characteristics with ($\beta = 10$) and without ($\beta = 0$) spatial locality, as shown in Fig. 5(b), we see little difference in the number of cache hits in area x. For both $\beta = 0$ and $\beta = 10$, cache hit performance was not affected by locality existence because the popular contents in each area are stably cached in each network. On the other hand, as shown in Fig. 5(c), focusing on the most popular contents, the on-path cache hits in area $\bar{x}$ significantly decreased when $\beta = 10$. Since the popularity distribution of area x (i.e., $\beta = 0$) completely differed from that of area $\bar{x}$ (i.e., $\beta = 10$), requests/contents generated in area x were less likely to be cached in the on-path in area $\bar{x}$. Therefore, the cache hits significantly decrease because of the popularity gap due to spatial locality.

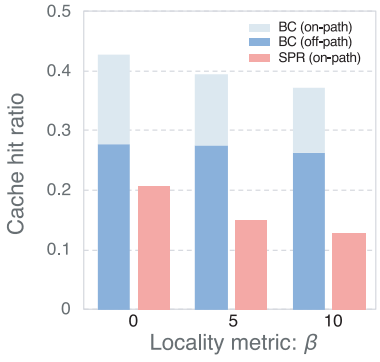


Fig. 4 Cache hit characteristics for spatial locality.

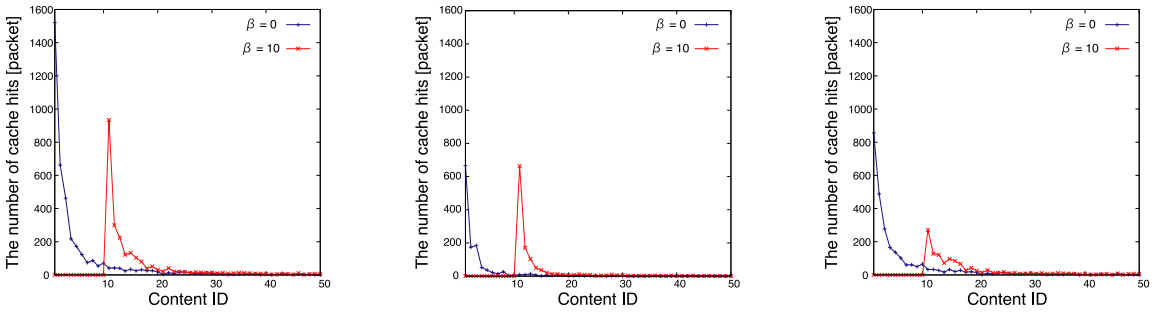
Figure 6 shows the on-path cache hit characteristics of the BC method. The cache hit counts in the overall network (Fig. 6(a)), area x (Fig. 6(b)), and area $\bar{x}$ (Fig. 6(c)) are shown. The on-path cache hit characteristics are similar to those of the SPR method, and the total on-path cache hit count slightly decreases. However, as shown in Fig. 7, the BC method increased more cache hits from the off-path and displayed comparable performance with the nonspatial locality case. The BC method flexibly explores cached content not only from the on-path but also from the off-path. Even if on-path cache hits are slightly reduced, many off-path-cached contents can be discovered by BC guidance. Hence, the BC method works well even in environments with regional differences in content popularity, and a high cache hit performance can be obtained by utilizing the cached contents in the off-path.

Performance Analysis: To investigate why the BC method can maintain its high performance, we measured the cached content distribution in the network. Figure 8 shows the characteristics of cache residence time ratio. The cache residence time ratio for content k is defined as follows:

$$\frac{t_i^k}{\sum_{i \in \mathcal{V}} \sum_{k \in \mathcal{K}} t_i^k}. \tag{1}$$

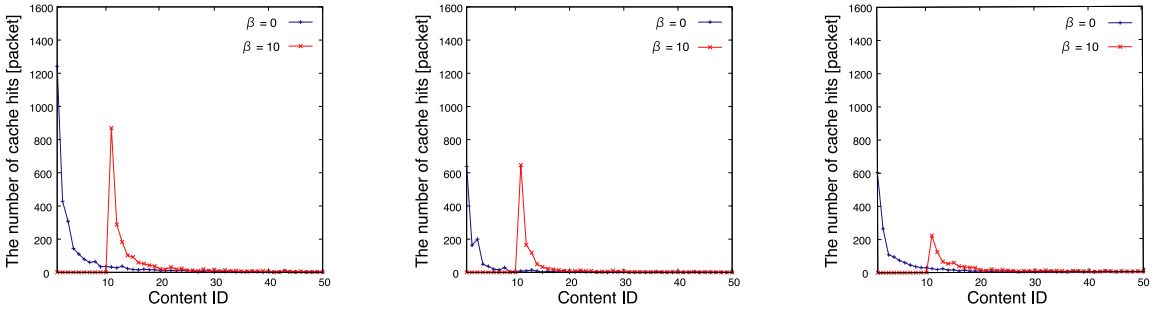
$\mathcal{V}$ refers to the set of routers in the network, and $\mathcal{K}$ denotes the set of contents. t_i^k is the cache residence time for content k in node i. The vertical axis shows content IDs sorted in descending order of content popularity. The performance for contents whose IDs are 101–10000 is omitted because of the extremely few samples due to their low popularity. The horizontal axis shows node IDs sorted in descending order of betweenness centrality. In both SPR and BC, popular contents with small content IDs are not stably stored in core routers with high betweenness centrality. In contrast, at the edge of the network, popular contents are stored in routers with low centrality. Only in the router (node ID 99) connected to the regional area, the contents (IDs 11–20) are likely to be stored in the network because these contents are popular in area x.

Figure 9 shows the distribution of cache hit locations for SPR and BC. The results for contents whose IDs are 1–10 and 101–10000 are omitted because of low sample count. With SPR, cache hits occur only in the on-path; hence, regional users' requests are mostly satisfied at (response data are replied from) the router (node ID 99) connected to the



(a) Overall network (b) area x (c) area $\bar{x}$

Fig. 5 SPR on-path cache hit performance.



(a) Overall network (b) area x (c) area $\bar{x}$

Fig. 6 BC on-path cache hit performance.

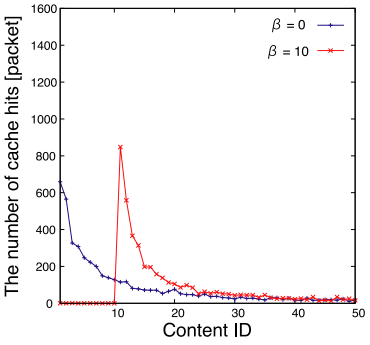
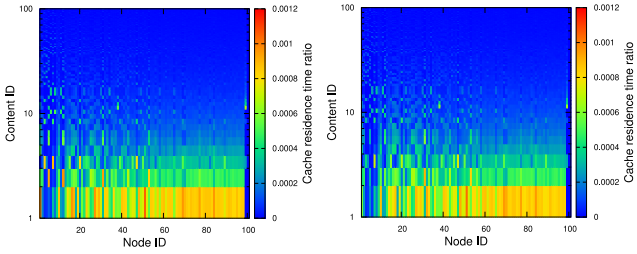


Fig. 7 BC off-path cache hit performance.



(a) SPR (b) BC

Fig. 8 Cached content distribution.

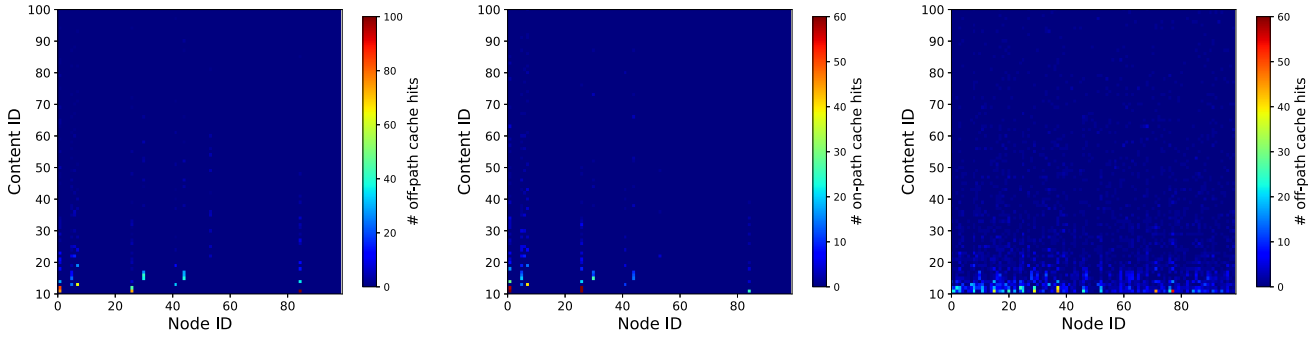
regional area[†]. If a request was not satisfied at the router, the request is forwarded to the origin servers. In the nonregional network, i.e., area $\bar{x}$, the request might be satisfied at the core router with a high outdegree number with high betweenness centrality. In contrast, when BC is used for the routing policy, the on-path cache hit locations are similar to those in SPR as shown in Fig. 9(b). However, off-path cache hits work well in the BC method, and high-/middle-popularity contents are replied by the in-network cache distributed in the whole network as shown in Fig. 9(c).

Figure 10 shows the BC forwarding rate, which refers to the ratio of the number of BC forwarding initiations for content k at node i to the total number of BC forwarding in the whole network for content k. This metric implies the possibility of BC forwarding initiation. These results show that BC forwarding is mostly initiated at the core network nodes (node IDs 1–10) with high betweenness centrality, and requests are distributed to the whole network through off-path routing. As described in Fig. 8, various contents could be found at the edge side, and they are replied. As a result, BC can effectively utilize in-network cached contents even in an environment where a popularity gap exists between regional and nonregional areas.

3.3 Topological Effects

Previously, we revealed the adaptability of BC to spatial locality in content popularity. Although it is reported that BC well performs in the hierarchical topology like [23], perfor-

[†]Since we intend to focus on cache hit performance in area x, we do not show this point in Fig. 9(a), that is, the number of cache hits at the node where ID 99 is excluded.



(a) SPR (on-path) (b) BC (on-path) (c) BC (off-path)

Fig. 9 Cache hit location distribution.

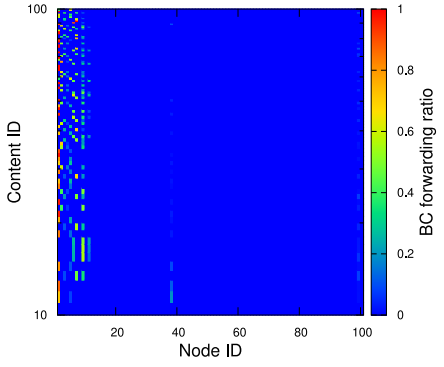
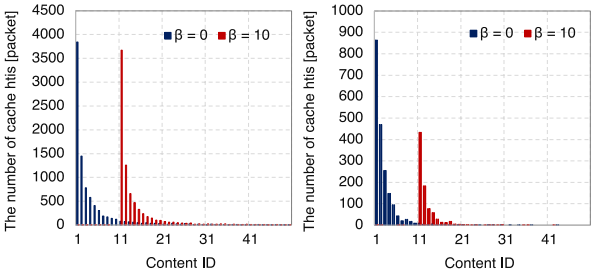


Fig. 10 BC forwarding rate.



(a) On-path (b) Off-path.

Fig. 11 Cache hit performance in the tree topology.

mance evaluation in a spatial locality scenario have not been conducted well. To investigate whether there exist topological effects on our hypothesis, we next evaluated cache hit performance in a hierarchical tree topology. We used a 4-branch tree topology with depth 3 for the evaluation. The number of routers is 85 and one user is connected to each leaf node, and thus totally 64 users are in the network. The server is placed to the root of the tree. Other simulation parameters are same with the previous model.

Figure 11 shows cache hit performance of BC in the tree topology. As shown in Fig. 11(a), on-path cache hits occur even in a spatial locality scenario, i.e. $\beta = 10$. Compared with the BA model (corresponding to Fig. 6(a)), on-path cache hits frequently occur in the tree model. Figure 11(b) shows cache hit performance in the off-path. Compared with Fig. 7, off-path cache hit performance is slightly lower than the BA model; however, similar cache hit characteristics are observed in the tree model. Therefore, we conclude that BC has the adaptability to spatial locality in content popularity not depending on topological differences.

4. Adaptability to Temporal Locality

Section 3 focused on the spatial locality aspect of content popularity. This section discusses the temporal locality dimension and investigates how BC can be adapted to an environment where content popularity distribution dynamically changes. Note that we focused only on temporal locality and did not introduce spatial characteristics in this environment, described in Sect. 3, for a separate investigation of effects.

4.1 Evaluation Model

Content requests follow a Zipf distribution at the beginning of the simulation, and their popularity ranks will be inverted at time t [s] to observe the effects of a drastic popularity change. The performance metric is the average cache hit ratio measured every τ [s]. In this evaluation, $t = 300$, $\tau = 2$ are used. For evaluation methods, SPR and BC are adopted as routing policies, and for caching policies, TERC and Fix(p) [24], [25] are used. Fix(p) is a caching policy in which a router decides whether to store the received content based on a fixed probability p. Fix(p) is reported to perform well when combined with the SPR routing policy [14]. For parameter setting, we set $p = 0.01$ as recommended by Rossini et al. [9]. It is well-known that BC performs poorly when combined with Fix(p) [14], so we will only focus on the results of three combinations: $\{SPR, TERC\}, \{SPR, Fix(\mathrm{p})\}, \{BC, TERC\}$.

4.2 Results

Figure 12 shows the characteristics of the average cache hit ratio. For the combinations that use TERC as their caching policy, that is, $\{SPR, TERC\} and \{BC, TERC\}$, TERC can

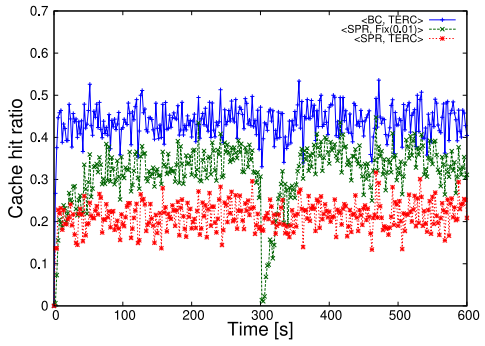
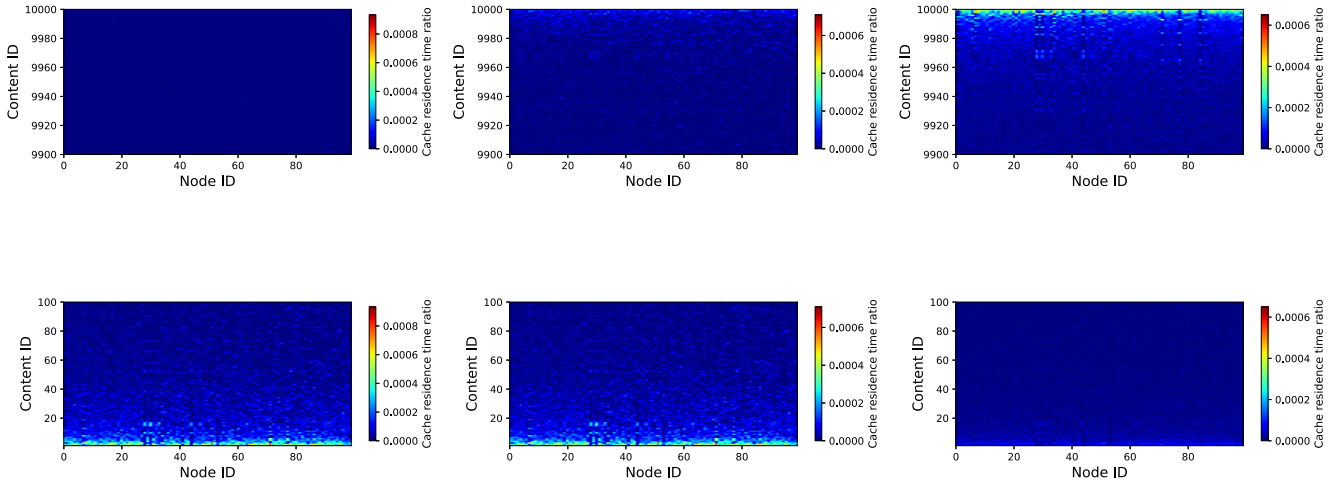


Fig. 12 Average cache hit ratio vs. time.

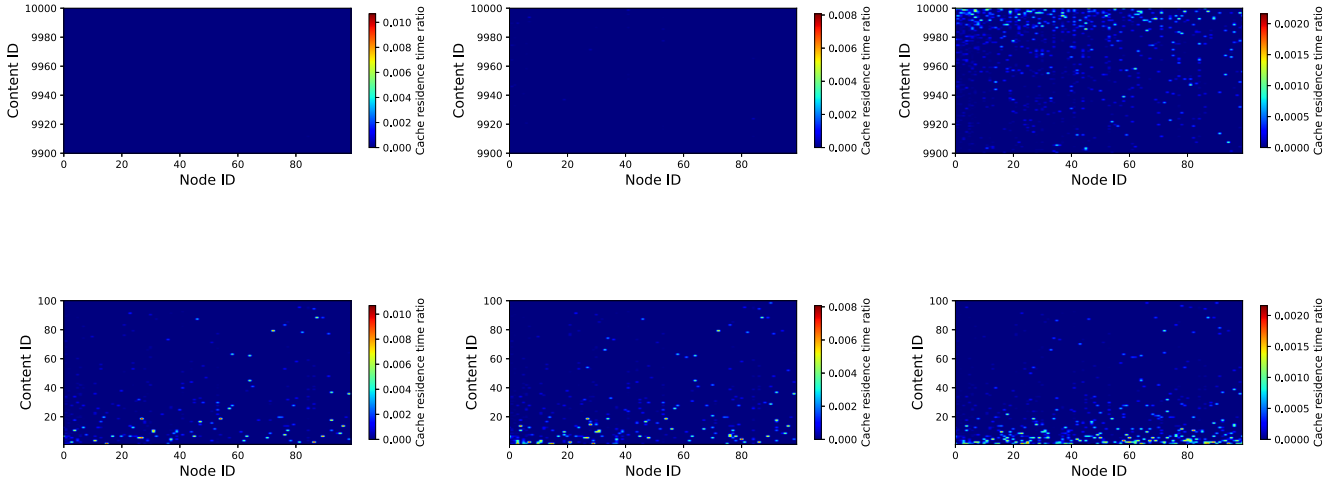
quickly replace the stored cached contents and adapt to popularity change at $t = 300$ [s]. In contrast, Fix(p) holds popular contents to the network caches for a long time and therefore cannot refresh outdated cached contents and experiences a long convergence time. This means Fix(p) has lower adaptability to temporal locality. Because of the cache hit degradation during this period, the average cache hit ratio of Fix(p) is lower than that of TERC. Therefore, TERC can adapt to the temporal locality of content popularity as opposed to Fix(p).

Figures 13 and 14 show the time transitions of cached contents in the whole network with {BC,TERC} and {SPR,Fix(p)}. Figures 13(a), 13(b), and 13(c) are the cache residence time ratios of {BC,TERC} at $t = 300$, 400, and



(a) t=300 (b) t=400 (c) t=2000

Fig. 13 Time transition of cached contents with {BC,TERC}.



(a) t=300 (b) t=400 (c) t=2000

Fig. 14 Time transition of cached contents with {SPR,Fix(p)}.

2000, respectively. Figure 13(a) shows that highly popular contents at that time, i.e., content IDs 1–10, are stored in the network before popularity change at $t = 300$. At $t = 400$, popularity is inverted, and the popular contents at that time, whose IDs are 9990–10000, are likely to be stored in the network. However, the elapsed time after the popularity change is short; therefore, the relative cache residence time ratio remains low as in Fig. 13(b). At $t = 2000$, cached contents are completely refreshed, and popular contents at that time, i.e., content IDs 9990–10000, are stably stored in the network as shown in Fig. 13(c).

With {SPR, Fix(p)} ($p = 0.01$), caching opportunities are roughly reduced to about one hundredth in the whole network, and cached contents are infrequently updated/replaced compared to TERC. Popular contents at that time are likely to be stored in the network for a long time as shown in Fig. 14(a). Even when popularity has changed, cached contents are rarely refreshed for a short period as in Fig. 14(b). At $t = 2000$, shown in Fig. 14(c), the elapsed time after popularity change becomes quite long, and cached contents are slightly replaced. Nevertheless, the outdated contents, whose IDs are 1–100, remain in the network.

As a traffic control scheme in ICN, a highly stable cache hit function is required for routing and caching policies that mitigate the temporal volatility of content popularity. The combination $\{BC, TERC\}$ can adapt to the temporal locality of content popularity and maintain a high cache hit performance in this situation. Hence, this combination produces a higher cache hit performance from a temporal locality viewpoint.

5. Related Work

M. Garetto et al., [26] focused only on the temporal locality and revealed that it is possible to accurately capture the behavior of caching networks under the environment where content popularity has "temporal locality" into a finite population analytical model.

S. Traverso et al., [27], [28] tried to study not only on the temporal locality but also geographical locality. For the temporal locality, they proposed Shot Noise Model (SNM), that enables users to natively capture the dynamics of content popularity. They additionally mentioned modeling of the geographical locality. However, due to the limited available data set, they have concluded that accounting for geographical locality in a traffic model can be a difficult task due to cultural/social effects. Thus, they shifted their main focus to the temporal locality of content popularity, while our main focus of this paper is the geographical locality.

A. Montazeri et al., [29] focused on geographic locality and developed an algorithm that can create synthetic traffic for geographical areas that possess Zipf's law properties. The proposed algorithm simply and randomly divided the global Zipf distribution into multiple local Zipf distributions. The concept and motivation for modeling the geographical locality of contents are same with us; however, the modeling is not based on the actual trace data and cannot be used for the evaluation with realistic scenarios.

6. Conclusion

In previous years, most ICN traffic control schemes have been proposed and evaluated based on the assumption that the content popularity of the entire network is homogeneous, that is, with no geographical/temporal locality characteristics. However, reports have shown that the actual popularity distribution measured on the Internet has spatial or temporal locality. This paper analyzed the performance of a BC routing method in an environment where the spatial locality of content popularity exists. BC routing distributes requests to the whole network to forward them outside the regional area and can fill the popularity gap of content locality for better cache utilization. We also focused on another aspect, i.e., the temporal locality of content popularity, and found that BC is highly adaptable to temporal locality when combined with TERC, which exhibits tolerance for temporal locality. Future work would benefit from dataset-driven experiments in environments where spatial and temporal localities coexist.

Acknowledgments

This work was supported in part by JSPS KAKENHI grant numbers 23K21662 and 21H03434.

References

[1] V. Jacobson, D.K. Smetters, J.D. Thornton, M.F. Plass, N.H. Briggs, and R.L. Braynard, "Networking named content," Proc. ACM CoNEXT'09, Rome, Italy, pp.1–12, Dec. 2009.
[2] L. Zhang, A. Afanasyev, J. Burke, V. Jacobson, K. Claffy, P. Crowley, C. Papadopoulos, L. Wang, and B. Zhang, "Named data networking," ACM SIGCOMM Comput. Commun. Rev., vol.44, no.2, pp.66–73, July 2014.
[3] M. Yamamoto, "A survey of caching networks in content oriented networks," IEICE Trans. Commun., vol.E99-B, no.5, pp.961–973, May 2016.
[4] H. Asaeda, A. Ooka, K. Matsuzono, and R. Li, "Cefore: Software platform enabling content-centric networking and beyond," IEICE Trans. Commun., vol.E102-B, no.9, Sept. 2019.
[5] Cisco, "Cisco Visual Networking Index: Forecast and Trends, 2017–2022" https://twiki.cern.ch/twiki/pub/HEPIX/TechwatchNetwork/HtwNetworkDocuments/white-paper-c11-741490.pdf, accessed on 6 March, 2022.
[6] A. Brodersen, S. Scellato, and M. Wattenhofer, "YouTube around the world: Geographic popularity of videos," Proc. WWW'12, Lyon, France, pp.241–250, April, 2012.
[7] N. Kamiyama and M. Murata, "Modelling popularity dynamics of YouTube videos," IEICE Technical Report, NS2015-81, Sept. 2015.
[8] N. Kamiyama, "Cache control based on diversity of content," IEICE Technical Report, IN2016-90, Jan. 2017.
[9] G. Rossini and D. Rossi, "Coupling caching and forwarding: Benefits analysis and implementation," Proc. ACM ICN'14, Paris, France, pp.127–136, Sept. 2014.
[10] R. Chiocchetti, D. Rossi, G. Rossini, G. Carofiglio, and D. Perino, "Exploit the known or explore the unknown?: Hamlet-like doubts in ICN," Proc. ACM ICN'12, Helsinki, Finland, pp.7–12, Aug. 2012.
[11] R. Chiocchetti, D. Perino, G. Carofiglio, D. Rossi, and G. Rossini, "INFORM: A dynamic INterest FORwarding mechanism for information centric networking," Proc. ACM ICN'13, Hong Kong, China,

pp.9–14, Aug. 2013.
[12] S. Eum, K. Nakauchi, M. Murata, Y. Shoji, and N. Nishinaga, "CATT: Potential based routing with content caching for ICN," Proc. ACM ICN'12, Helsinki, Finland, pp.49–54, Aug. 2012.
[13] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and Zipf-like distributions: Evidence and implications," Proc. IEEE INFOCOM'99, New York, NY, USA, pp.126–134, March 1999.
[14] Y. Hayamizu, A. Shibuya, and M. Yamamoto, "The combination effect of cache decision and off-path cache routing in content oriented networks," IEICE Trans. Commun., vol.E102-B, no.5, pp.1010–1018, May 2019.
[15] K. Nagata, N. Kamiyama, and M. Yamamoto, "Cache policy based on popularrity dynamics of YouTube video content," Proc. IEEE CCNC'20, Las Vegas, NV, USA, Jan. 2020.
[16] E. Rosensweig and J. Kurose, "Breadcrumbs: Effcient, best-effort content location in cache networks," Proc. IEEE INFOCOM'09, Rio de Janeiro, Brazil, pp.2631–2635, April 2009.
[17] A. Afanasyev, I. Moiseenko, and L. Zhang, "ndnSIM: NDN simulator for NS-3," NDN, Technical Report, NDN-0005, 2012.
[18] ns-3, network simulator 3. http://www.nsnam.org/
[19] R. Albert and A. Barabási, "Statistical mechanics of complex networks," Rev. Mod. Phys., vol.74, no.1, pp.47–97, 2002.
[20] H. Yamanaka, Y. Teranishi, Y. Hayamizu, A. Ooka, K. Matsuzono, R. Li, and H. Asaeda, "User-centric in-network caching mechanism for off-chain storage with blockchain," IEEE ICC'22, May 2022.
[21] Y. Hayamizu, M. Yamamoto, E. Rosensweig, and J.F. Kurose, "Multiple-Breadcrumbs: A new in-network guidance for off-path cache in cache networks," IEICE Trans. Commun., vol.E101-B, no.12, pp.2388–2396, Dec. 2018.
[22] A. Shibuya, Y. Hayamizu, and M. Yamamoto, "Cache decision policy for Breadcrumbs in CCN," Proc. 2016 IEEE Globecom Workshops, Washington, DC, USA, pp.1–6, 2016.
[23] H. Nakajima, T. Tsutsui, H. Urabayashi, M. Yamamoto, E. Rosensweig, and J.F. Kurose, "Performance evaluation of partial deployment of an in-network cache location guide scheme, Breadcrumbs," IEICE Trans. Commun., vol.E99-B, no.1, pp.157–166, Jan. 2016.
[24] N. Laoutaris, S. Syntila, and I. Stavrakakis, "Meta algorithms for hierarchical Web caches," Proc. IEEE ICPCC'04, pp.445–452, Aug. 2004.
[25] S. Arianfar, P. Nikander, and J. Ott, "On content-centric router design and implications," Proc. ACM ReArch'10, Nov. 2010.
[26] M. Garetto, E. Leonardi, and S. Traverso, "Efficient analysis of caching strategies under dynamic content popularity," 2015 IEEE Conference on Computer Communications (INFOCOM), Hong Kong, China, pp.2263–2271, 2015.
[27] S. Traverso, M. Ahmed, M. Garetto, P. Giaccone, E. Leonardi, and S. Niccolini, "Temporal locality in today's content caching: Why it matters and how to model it," ACM SIGCOMM Comput. Commun. Rev., vol.43, no.5, pp.5–12, Oct. 2013.
[28] S. Traverso, M. Ahmed, M. Garetto, P. Giaccone, E. Leonardi, and S. Niccolini, "Unravelling the impact of temporal and geographical locality in content caching systems," IEEE Trans. Multimedia, vol.17, no.10, pp.1839–1854, Oct. 2015.
[29] A. Montazeri and D. Makaroff, "Geographically-distinct request patterns for caching in information-centric networks," 2017 IEEE 42nd Conference on Local Computer Networks (LCN), Singapore, pp.579–582, 2017.

Yusaku Hayamizu received his B.E., M.E., and Ph.D. in engineering from Kansai University in 2014, 2016, and 2019, respectively. He currently works as a researcher at the National Institute of Information and Communications Technology (NICT). His research interests include computer networks, information-centric networks, traffic control, in-network computing, and network softwarelization. Dr. Hayamizu is the recipient of best paper awards from the 2017 IEEE CQR Workshop and the 2018 IEEE LANMAN Symposium. He is a member of the ACM and the IEEE.

Miki Yamamoto received his B.E., M.E., and Ph.D. in communications engineering from Osaka University in 1983, 1985, and 1988. He joined the Department of Communications Engineering at Osaka University in 1988. He moved to the Department of Electrical Engineering and Computer Science of Kansai University in 2005, where he is a professor. He visited the University of Massachusetts at Amherst in 1995 and 1996 as a visiting professor. His research interests include content oriented networks, high speed networks, wireless networks, and the evaluation of performance of these systems. Dr. Yamamoto has served many technical and geographical activities in IEEE, including Kansai Section chair, an Executive Committee member of Region 10, and Technical Program Co-Chair of ICC 2019 CQRM Symposium. He has received the best paper award of IEEE CQR Workshop 2017 and IEEE LANMAN workshop 2018. He is a senior member of IEEE, a member of ACM and IPSJ, and a fellow of IEICE.

IEICE TRANS. COMMUN., VOL.E107–B, NO.11 NOVEMBER 2024

PAPER

Spatial Anomaly Detection Using Fast xFlow Proxy for Nation-Wide IP Network

Shohei KAMAMURA[†a)], ***Senior Member***, **Yuhei HAYASHI**[††], ***Nonmember***, ***and*** **Takayuki FUJIWARA**[††], ***Member***

SUMMARY This paper proposes an anomaly-detection method using the Fast xFlow Proxy, which enables fine-grained measurement of communication traffic. When a fault occurs in services or networks, communication traffic changes from its normal behavior. Therefore, anomalies can be detected by analyzing their autocorrelations. However, in large-scale carrier networks, packets are generally encapsulated and observed as aggregate values, making it difficult to detect minute changes in individual communication flows. Therefore, we developed the Fast xFlow Proxy, which analyzes encapsulated packets in real time and enables flows to be measured at an arbitrary granularity. In this paper, we propose an algorithm that utilizes the Fast xFlow Proxy to detect not only the anomaly occurrence but also its cause, that is, the location of the fault at the end-to-end. The idea is not only to analyze the autocorrelation of a specific flow but also to apply spatial analysis to estimate the fault location by comparing the behavior of multiple flows. Through extensive simulations, we demonstrate that base station, network, and service faults can be detected without any false negative detections.

key words: *correlation analysis, IP networks, telecommunication traffic, xFlow protocol*

1. Introduction

For the sustainable operation of wide-area internet protocol (IP) networks as a social infrastructure, it is crucial to accurately detect and recover from anomalies. In this paper, an anomaly is defined as a decrease in the traffic observed in a network. A Decrease in the traffic is primarily caused by failures in the Internet services used or of the networks that connect to the services. Since anomalies in a network represent changes in the behavior of communication traffic, we can diagnose the occurrence of anomalies by measuring the state of the traffic. Communication traffic consists of individual flows, which are defined as a collection of packets identified by information in a transmission control protocol (TCP) and IP packet header called a 5-Tuple. Technically, the flows can be collected using flow measurement protocols such as NetFlow [1] or IPFIX [2].

We proposed the Fast xFlow Proxy, which enables fine-grained measurement of flows in large-scale wide-area IP networks [3], and devised an algorithm to detect anomalies such as service failures and rapid increases in traffic caused by cyber-attacks [4]. The results from flow measurement protocols are usually presented as macroflows superimposed of various communication flows. As these flows have different characteristics, it is difficult to accurately detect anomalies in individual flows. The Fast xFlow Proxy can decompose these macro-observed flows into individual flows in real time at ultra-high speeds (100 Gbps). Thus, even if packets are encapsulated by segment routing using multiprotocol label switching (SR-MPLS) [5] or the layer 2 tunneling protocol (L2TP) [6], the internal packet header information can be referred to and fine-grained flow statistics can be obtained. Once individual flows are observed in fine-grained by the Fast xFlow Proxy, an analysis of whether the current behavior of the flow is similar compared to its normal state can be performed. We demonstrated that this simple autoregressive analysis based on correlation values can detect anomalies with high accuracy [4].

The existing analysis algorithm [4] targets fault detection on the over the top (OTT) side, which we refer to as the *back side*. For example, video distribution servers and cloud services exist on the back side. Therefore, existing algorithms cannot detect faults in *front side* access networks, such as optical or mobile access, or in their own networks. Instead of fault detection by measuring flows, a direct method for detecting faults is to monitor the alarm from the equipment of each network [7]. However, monitoring multiple pieces of equipment from various networks complicates the network operations. Additionally, the access network may not be owned by the operator operating the network, and it may not be possible to immediately obtain fault alarms. Therefore, a method that can centrally detect faults on the fronts side, own network, and back side by simply monitoring its own network is needed.

Therefore, in this study, we extend the existing anomaly detection algorithm [4] and propose a method for detecting anomalies not only in the back side, where the OTT exists, but also in the front side, which constitutes the access network and in the own network. Our key idea is a spatial analysis of flows: our algorithm not only analyzes the autocorrelation of a specific flow, but also analyzes the correlation between flows. Thus, the fault location is estimated. Figure 1 illustrates an example of the spatial analysis. The flow X from user 1 to service A is observed in its core network. If the shape of flow X changes significantly from its previous shape, then the autocorrelation value decreases. Thus, an anomaly was detected in flow X. At this point, it is unclear which fault is responsible for the anomaly: a fault on

Manuscript received December 18, 2023.
Manuscript revised March 28, 2024.
Manuscript publicized May 14, 2024.

[†]Faculty of Science and Technology, Seikei University, Musashino-shi, 180-8633 Japan.
[††]NTT Network Innovation Center, NTT Corporation, Musashino-shi, 180-8585 Japan.
a) E-mail: shohei-kamamura@st.seikei.ac.jp

DOI: 10.23919/transcom.2023EBP3208

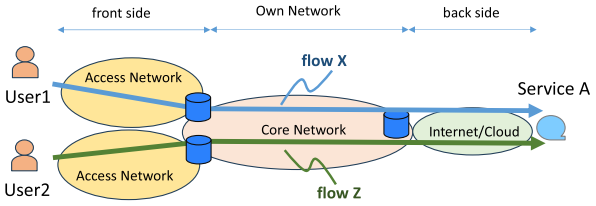


Fig. 1 Example of spatial analysis comparing behavior between flows.

the front side, own network, or the back side. To detect the fault location, our algorithm observes another flow Z that uses service A. If the correlation between flows X and Z is high, both flows encounter an anomaly, and the suspected point is the own network or the back side, which is a common passage point. Conversely, if the correlation between flows X and Z is low, the suspected point is on the front side where only X passes. Thus, the aim of our algorithm is to narrow down the suspected points by conducting a spatial analysis between flows in addition to an existing autoregressive analysis.

This study makes two major contributions to the existing literature. First, this study provides upward compatibility with existing algorithms [4]. Contrarily, the previous algorithm [4] only detects the presence of anomalies, the algorithm proposed in this study can centrally detect faults on the front side, its own network, and the back side fault location by simply monitoring traffic flows in its own network. Users may contact telecommunications carriers without knowing the cause of the end-to-end service failure. In such cases, using our proposed algorithm, the network operator can promptly inform users of the causes of the faults. It was demonstrated that base station, service, and network failures can be detected with no detection error (False Negative) if the noise in the flow reaches approximately 30%. The second contribution is the feasibility of the proposed algorithm in realistic cases. To reduce the false negatives to zero, it is necessary to analyze multiple flows iteratively. We demonstrated that even in the largest scenario, the computation time is assumed to be trivial at approximately 20 s, and the storage consumption is 170 MB. Additionally, we demonstrated that our proposals will scale up to 10 times the current scale in preparation for future networks, such as Beyond 5G.

The remainder of this paper is organized as follows. Section 2 introduces related works, an overview of the Fast xFlow Proxy, and the problem statement. Section 3 describes the proposed spatial anomaly detection algorithm. Section 4 presents numerical results. Finally, we conclude the paper in Sect. 5.

2. Related Works and Problem Statement

2.1 Related Works

Traffic monitoring techniques are categorized into those that observe the total amount of traffic per network interface, and those that observe end-to-end traffic. Representative technologies for total traffic monitoring include the classical pull-type simple network management protocol (SNMP) [8] and the recently introduced push-type network telemetry [9]. End-end traffic monitoring can be classified into a flow measurement approach [1], [2], [10], [11] and a detailed packet capturing approach called deep packet inspection (DPI) [12]. We focus on the flow measurement approach because the granularity of information obtained from the overall traffic monitoring is too coarse, whereas that of the DPI is detailed; however, too much information is expensive to manage. The flow measurement approach counts the statistics of flows, which are defined as a collection of packets identified by information in a TCP/IP packet header called a 5-Tuple. The flow measurement approach is not limited to simple statistical methods, but also uses probabilistic data structures based on the hash function to compress data volumes, such as the Bloom Filter [13] and Sketch structure [14].

The flow measurement approach is highly scalable; however, in large-scale carrier networks, packets may be encapsulated to provide user authentication or virtual private network (VPN) services [5], [6]. In such cases, the information obtained by the flow measurement protocols is outer encapsulated information, which leads to a rough granularity of information. To solve this problem, we developed the Fast xFlow Proxy that strips off the outer header information and enables the observation of inner flows with arbitrary granularity [3].

2.2 Fast xFlow Proxy and its Capabilities

The Fast xFlow Proxy [3] enables the observation of flows with arbitrary granularity in large wide-area IP networks. A Fast xFlow Proxy was deployed between the IP router and the commercial flow collectors [15]. If the user has given permission for detailed monitoring of their traffic, the Fast xFlow Proxy can be configured to remove outer IP headers in advance. Once the Fast xFlow Proxy receives the flows, it forwards the inner header information to the flow collectors. At this point, the flows were observed to have arbitrary granularity. The flow collector performs statistical processing based on the inner header information observed by the IP router. The hardware component of the Fast xFlow Proxy was implemented using field programmable gate array network interface cards (FPGA NICs) [16], and the software component was implemented using data plane development kit (DPDK) [17]. These implementations enable ultra-high-speed processing at 100 Gbps [3].

Fine-grained flow information from the Fast xFlow proxy can be applied to anomaly detection [4] and traffic engineering [18], [19]. In this study, we focus on anomaly detection. Since flow information is time-series data, data analysis algorithms for time-series data are applicable [20]–[27]. For example, autoregressive analysis, such as ARIMA, can be used to predict the next fluctuation and compare it with the actual measured value to detect anomalies [20], [21]. Although many advanced analysis algorithms use machine learning [22], [23], simple algorithms that do not use pre-

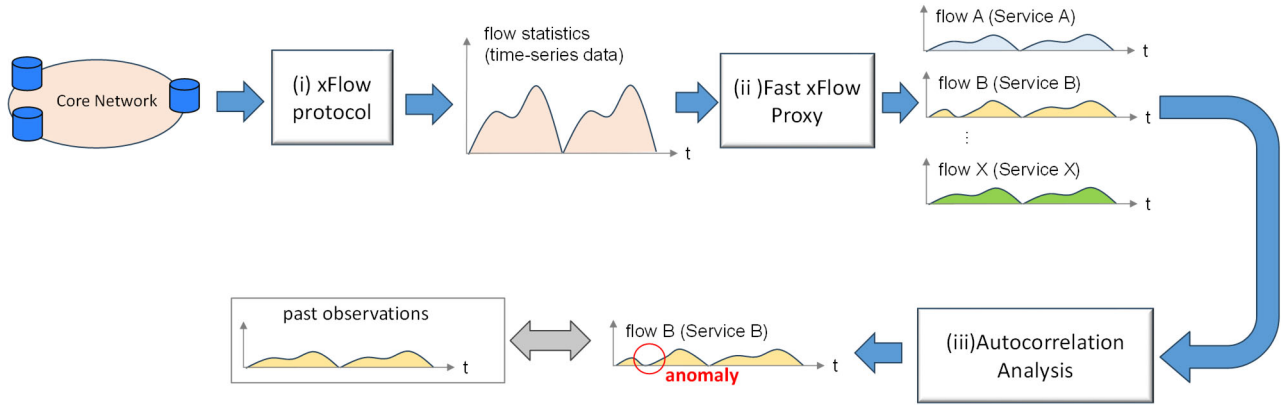


Fig. 2 Anomaly detection using the Fast xFlow Proxy and autocorrelation analysis. (i) The system obtains flow statistics from the network, (ii) decompose them to a fine granularity using the Fast xFlow Proxy, and then (iii) use autocorrelation analysis to detect anomalies.

diction are also effective for data with simple and periodic patterns. For example, FFT [24] and SR [25] are based on signal-based analyses. NormA-SJ [26] and SAND [27] are based on data mining. They defined their unique distances and compared the lengths of the distances to detect anomalies. These algorithms detect anomalies by creating a low-dimensional latent space of normal time-series data and comparing it with the observed values.

We propose an anomaly detection algorithm based on autoregressive correlation analysis. Since the Fast xFlow Proxy can observe flows with fine granularity [4], the fluctuation of the flows becomes obvious, enabling accurate anomaly detection. Figure 2 presents an overview of this process. (i) The proposed system obtains flow statistics from the network, (ii) decomposes them into a fine granularity using the Fast xFlow Proxy, and (iii) applies autocorrelation analysis to detect anomalies. We demonstrate that it can detect faults and sudden increases in traffic with an accuracy of over 80% in real time.

2.3 Problem Statement

Existing anomaly detection algorithms for time-series data determine whether an anomaly exists in the flow and do not specify the point of fault that causes the anomaly. Several suspicious points exist: 1) the front side, such as the access network, 2) the own network, and 3) the back side, such as the Internet or cloud services. Therefore, it is crucial to identify the fault location in addition to the occurrence of anomalies to quickly perform fault management operations.

In this study, we defined the problem of not only detecting anomalies but also finding where the anomalies occur: on the front side, own network, and back side. As previously mentioned, combining different network management systems is assumed to be operationally complex. Therefore, we focus on the existing technology [4] that can detect anomalies simply by monitoring the traffic in the core network. We then extend it to detect fault locations.

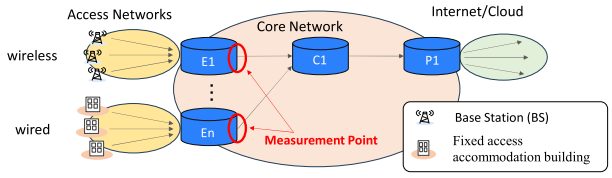
3. Spatial Anomaly Detection Algorithm

The input of our algorithm is the communication flows measured at the edge routers, and our spatial anomaly-detection algorithm then computes the autocorrelation and correlation between the flows to decide where the fault occurs: the front side, own core network, or back side. In the following sections, we define the network model. We then present an anomaly-detection method based on correlation values. Subsequently, an algorithm for identifying fault locations based on correlations is presented. We assumed that faults at a single location and multiple faults are outside the scope of this study.

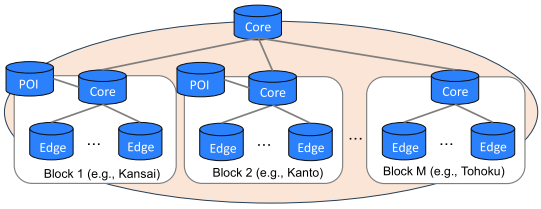
3.1 Network Model

We first explain the end-to-end communication from users to Internet services using the network model illustrated in Fig. 3(a). In Fig. 3(a), the traffic flows of users arrive at access networks, such as the radio base station (BS) or fixed access accommodation buildings. Subsequently, traffic flows through the access network into the core network operated by telecommunications carriers. The flows then reach the Internet services via an Internet connection point called the point of the interface (POI) (router P1 in Fig. 3(a)). Routers E1 and En located at the boundaries of the core network are called edge routers, while the center router C1, which specializes in high-capacity forwarding, is called the core router. Note that this study is applicable only to the core network that relays between the access network and the Internet. More specifically, that users under the edge router on the core network or any Internet services are assumed to always go through the edge router. It is also assumed that packets are encapsulated by outer headers. Therefore, the method is not applicable to arbitrary transit networks such as the Internet.

Figure 3(b) represents the core network topology model. The core network comprises multiple blocks. Each block



(a) Entire network model. User traffic goes from the fixed or wireless access network through the core network to the Internet and cloud network.



(b) Core network topology model. Edge routers consist of thousands of routers and core routers consist of hundreds of routers.

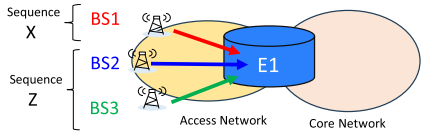
Fig. 3 Network model: (a) entire network model and (b) core network model.

corresponds to a physical region. The blocks are defined to accommodate a roughly uniform number of users. In principle, the same hierarchical tree structure is used within the blocks. Edge routers consist of thousands of routers and core routers consist of hundreds of routers. The POI router may not be present in every block, but the traffic toward the POI can be observed by the edge router in any block, regardless of the location of the POI.

Flow measurements were performed at the core-network side interface of the edge routers (E1 to En in Fig. 3(a)) located at the boundary between the access and core networks. In our previous study, it was shown that flow measurement at all edge routers is sufficiently practical with our scale-out system [3]. In this study, measurements were performed by selecting arbitrary edge routers based on the intended purpose. For example, in Fig. 3(a), if the correlation among multiple BSs is to be analyzed, the analysis should be performed on the results measured at E1. If it is desirable to analyze the correlation between wireless and wired trends on a more macroscopic level, the correlation between the results observed at E1 and En should be analyzed. In the actual network, the distribution of traffic generation may differ between measurement blocks. However, for the sake of simplicity here, the traffic distribution difference between blocks constitutes only the amount of flow based on the number of users.

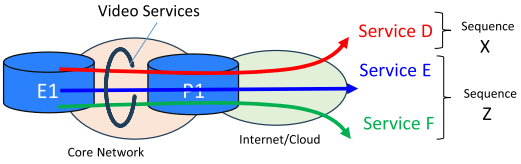
The measured flows are decomposed into individual flows using the Fast xFlow Proxy, as shown in Fig. 2. At this point, it is necessary to define what each flow represents. Therefore, after decomposing the flows into fine granularities, the Fast xFlow Proxy defines the meaning of the flows using (a) the front table and (b) the back table in Fig. 4. The front table links which BSs the measured flows originate from. This can be identified by comparing the source IP

Flow ID	Source IP Address:Port	Destination IP Address:Port	Base Station
Flow A	A.A.A.A:AA	ANY	BS1
Flow B	B.B.B.B:BB	ANY	BS2
Flow C	C.C.C.C:CC	ANY	BS3
..	..		..



(a) Example of front table and related sequences (X and Z)

Flow ID	Source IP Address:Port	Destination IP Address:Port	Service
Flow D	ANY	D.D.D.D:DD	Video Service D
Flow E	ANY	E.E.E.E:EE	Video Service E
Flow F	ANY	F.F.F.F:FF	Video Service F
..		..	..



(b) Example of back table and related sequences (X and Z)

Fig. 4 (a) Front table and (b) back table and their related sequences. The front table is used to map flows to base stations in wireless case, and the back table is used to map flows to the services.

address of the measured flow with that of the BS. The BS IP address is provided by the access network operator. The back table links which Internet services the flow is going to. This can be identified by comparing the destination IP address of the measured flow with the IP address of the Internet services. Generally, the IP address of Internet services changes dynamically; however, some products [28] that refer to the Internet routing registry (IRR) or domain name system (DNS) can be used to resolve the IP address of Internet services.

3.2 Anomaly Detection Based on Correlation Analysis

In our analysis, the measured flow is regarded as a sequence, and anomaly detection is performed by analyzing the correlation of three types of sequences: X, the targeted sequence being measured; Y, the past trend of the targeted sequence; and Z, a sequence similar to the targeted sequence. The flows that compose Z are defined for each case where the front table and back table is referred to. Figure 4 also illustrates an example of the relationship between sequence X and Z when each table is referenced. (a) When referring to the front table, if flows originating from BS1 is defined as a targeted sequence X, flows originating from BS2 and BS3 under the same edge router are defined as flows that compose Z. (b) When referring to the back table, if flows destined for

Service D is defined as a targeted sequence, flows destined for another service of the same genre as the targeted sequence is defined as flows that compose Z. For Internet services, for example, video, web conferencing, SNS, games, storage services, etc. can be defined as genres.

First, the analysis data for the targeted flow were measured as the sequence X^{raw} as follows:

$$X^{raw} = \left(x_1^{raw}, x_2^{raw}, \ldots, x_n^{raw}\right) \in \mathbb{R}^n. \tag{1}$$

Then, X^{raw} is normalized as follows.

$$X = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n. \tag{2}$$

where,

$$x_i = \frac{x_i^{raw}}{\| X \|} = \frac{x_i^{raw}}{\sqrt{\sum_{j=1}^{n} \left(x_j^{raw}\right)^2}}. \tag{3}$$

Each element x_i in sequence X represents the normalized amount of traffic per unit of time. For example, if one data point is observed every hour and the analysis is performed for one day, n is set to 24. Similarly, our algorithm produces the regulated sequence Y, which is T weeks before target date X, as follows:

$$Y^T = \left(y_1^T, y_2^T, \cdots, y_n^T\right) \in \mathbb{R}^n, \quad T = 1, 2, \cdots T_m. \tag{4}$$

Our algorithm computes the average $\overline{Y}$ of multiple observations to suppress the effect of singularities of previous observations.

$$Y = \overline{Y} = (\overline{y_1}, \overline{y_2}, \cdots, \overline{y_n}) \in \mathbb{R}^n, \tag{5}$$

where $\overline{y_l}$ is as follows:

$$\overline{y_l} = \frac{1}{T_m} \sum_{k=1}^{T_m} y_l^k. \tag{6}$$

The sequence Z to be spatially analyzed are then defined as follows:

$$Z^S = \left(z_1^S, z_2^S, \ldots, z_n^S\right) \in \mathbb{R}^n, \quad S = 1, 2, \ldots, S_m. \tag{7}$$

Our algorithm computes the average $\overline{Z}$ of multiple observations to suppress the effect of singularities of observations.

$$Z = \overline{Z} = (\overline{z_1}, \overline{z_2}, \ldots, \overline{z_n}) \in \mathbb{R}^n, \tag{8}$$

where $\overline{z_l}$ is as follows:

$$\overline{z_l} = \frac{1}{S_m} \sum_{k=1}^{S_m} z_l^k. \tag{9}$$

Based on the above mathematical models, the occurrence of the $X - Y$ error, which is an error on the time axis, and the occurrence of the $X - Z$ error, which is an error on the spatial axis, are evaluated by the algorithm shown in Fig. 5. In Fig. 5, sequences Y and Z are created based on the given sequence X (line 2-3). Subsequently, the similarity between X and Y is obtained, and if it is below a threshold, the $X - Y$ error is set (lines 4-5). Although various methods can be applied to this similarity comparison function, we adopted the effective correlation computation proposed in [4] as follows:

```
01:  For incoming stream X
02:      Create Ȳ = (ȳ1, ȳ2, ..., ȳn) ∈ R^n
03:      Create Z̄ = (z̄1, z̄2, ..., z̄n) ∈ R^n
04:      IF s1 := Similarity(X, Ȳ) < Th1
05:          X-Y_Error := true
06:      IF s2 := Similarity(X, Z̄) < Th2
07:          X-Z_Error := true
08:      SetAlert(X-Y_Error, X-Z_Error, s1, s2)
09:  Exit()
```

Fig. 5 Pseudo code for anomaly detection and its location detection algorithm.

Case	Input		Output Alert Type		
	X-Y Error	X-Z Error	Own Network Failure	Targeted Sequence X Failure	Other Similar Sequences Failure
(a)	-	-	-	-	-
(b)	✘	-	Suspected	-	-
(c)	✘	✘	-	Suspected	-
(d)	-	✘	-	-	Suspected

Fig. 6 Anomaly location detection table. $X - Y$ and $X - Z$ errors are computed by algorithm in Fig. 5. Then alert type is determined by this table.

$$Similarity\left(X, \overline{Y}\right) = \frac{\frac{1}{n}\sum_{t=1}^{n}(x_t - \overline{x})(y_t - \overline{y})}{\sigma_x \cdot \sigma_y}, \tag{10}$$

where $\overline{x}$ and $\overline{y}$ denote the means of X and $\overline{Y}$, respectively, and σ_x and σ_y denote their standard deviations, respectively. Correlation is a standardized value close to +1 when two sequences are correlated and close to zero when they are not correlated. Subsequently, the similarity between X and Z is computed, and if it is less than a threshold value, the $X - Z$ error is set (lines 6-7). Finally, based on the $X - Y$ error and $X - Z$ error values, the SetAlert function determines the occurrence and location of a fault (line 8). The implementation of SetAlert is described in the following section.

3.3 Detecting Anomalies and Their Location

By referring to Fig. 6, we explain the algorithm for detecting a fault and its location, which corresponds to the implementation of the SetAlert function described in the previous section. The results of the correlation computation and the values of $X - Y$ error and $X - Z$ error are provided as inputs. Note that as the degree of suspicion is highly dependent on

the threshold for determining the $X - Y$ and $X - Z$ errors, the algorithm records not only the binary values of the $X - Y$ and $X - Z$ errors but also the correlation values ($s1$ and $s2$ in Fig. 5).

If no error is present (case (a)), there is no variation in the flow in time or space; therefore, an alert is output as no fault has occurred. The case where only $X - Y$ error exists (case (b)) means that the correlation value between the targeted flow and other similar flows is high while the autocorrelation of the targeted flow is low. This means that the measured flow is failed together with other similar flows; an alert is output as a fault occurred in its own network.

Subsequently, we consider the case in which both $X - Y$ and $X - Z$ errors occur (case (c)). In this case, the correlation value between the targeted flow and its own past flow is low, and the correlation value between the targeted flow and other similar flows at the same time is also low. This case can be judged as one in which an anomaly occurs only in the targeted flow X, and an alert is output as a fault has occurred in the targeted sequence. A failure in sequence X implies a targeted BS or Internet service failure. Details are provided in the last paragraph of this section.

Finally, we consider the case in which only $X - Z$ error occur (case (d)). In this case, the autocorrelation of the targeted flow is high, but the correlation value between the targeted flow and other similar flows is low. Therefore, this case can be considered an anomaly that occurred in any of the other similar flows.

Our algorithm can be applied to different fault location decisions by switching the measured flows. Referring to the back table shown in Fig. 4, the targeted flow is a specific Internet service. Therefore, our algorithm determines whether a fault occurs in the Internet service, on its own network, or in other similar services. If the front table in Fig. 4 is referred to, the targeted flow is a specific BS or fixed-access accommodation building. Therefore, the proposed algorithm determines whether a fault occurs in a BS or building, its own network, or any other BS or building. As mentioned in Sect. 3.1, these flows can be observed collectively at the edge router. Therefore, by integrating the analysis results, it can be determined where the fault occurred: on the front side, core network, or back side.

4. Evaluation

4.1 Aims and Conditions

We evaluated whether our algorithm, which checks the appearance of $X - Y$ and $X - Z$ errors, can detect the fault location correctly. The occurrence of errors significantly depends on how the threshold value is set. To bring transparency to the evaluation, therefore, our evaluation measures the $X - Y$ and $X - Z$ correlation values themselves, and then sets the threshold values.

Table 1 presents the example of thresholds for determining $X - Y$ and $X - Z$ errors for noise of 0.01 and 0.3 in service-side flow monitoring. The method for calculating the threshold values is as follows. First, clustering is performed on the measured correlation values using the k-means method with $k = 2$ to create clusters with large and small values. If the difference between the mean values of the clusters is greater than the minute value δ, then the midpoint of the cluster mean value is used as the threshold value. On the other hand, if the difference between the cluster means is less than the minute value δ, the cluster cannot be classified into two clusters and becomes one cluster. Therefore, a value smaller than the cluster mean (e.g. the cluster mean minus δ) is set as the threshold value. Accordingly, δ should be determined to be smaller than the distance between the two clusters. In this evaluation, δ is set as 0.2. The distance between clusters varies based on the degree of anomaly. Thus, it is advisable to refer to the values that correspond to the time of failure or traffic increase, listed in [18].

Table 1 Example of threshold value under the service-side flow monitoring.

	noise=0.01		noise=0.3	
	With Fast xFlow Proxy	Without Fast xFlow Proxy	With Fast xFlow Proxy	Without Fast xFlow Proxy
X-Y Threshold	0.816	0.818	0.661	0.661
X-Z Threshold	0.812	0.792	0.693	0.688

The evaluation index is the error rate. The following metrics were used to compute the error rates: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). For example, if the Internet service is down but our algorithm could not detect it, it is judged to be an FN. Conversely, if the Internet service is not down but our algorithm regarded the service as down, it is judged as FP. The error rate was defined as follows:

$$Error\ Rate = 1 - \frac{TP + TN}{TP + TN + FP + FN}. \tag{11}$$

In our evaluation, each of the cases shown in Fig. 6 is generated using a traffic generator [29] and analyzed to check if the corresponding cases ((a), (b), (c), and (d)) can be accurately detected. In this evaluation, time-series data was generated for one month, and in the cases where anomalies occur, i.e., ((b), (c), and (d)), the amount of traffic at a targeted date and time was set to zero for a period of three hours. For example, in cases (c) and (d), the traffic volume of one observed sequence is zero at the targeted time, while in case (b), the traffic volume of all observed sequences is zero at the targeted time. The traffic fluctuates by noise in the range of 0.01 to 0.3 based on our operational experience [4], and the average of 100 measurements is plotted for each case.

When referring to the back table, which monitors Internet services, the cases corresponding to (a), (b), (c), and (d) in Fig. 6 are denoted by (a), (b), (c), and (d), respectively. When referring to the front table, which monitors BSs states, the cases corresponding to (a), (b), (c), and (d) shown in Fig. 6 are denoted by (a), (b), (c'), and (d'), respectively. In other words, by managing the status of (a), (b), (c), (d), (c') and (d'), it can be determined where the

fault occurred on the front side, core network or back side. Note that the anomalies in cases (c') and (d') were caused by a 3-hour failure of a specific BS as in cases (c) and (d). The evaluation compares the proposed method, which uses the Fast xFlow Proxy to monitor at high resolution, with the conventional method, which observes at 1/1000 sampling granularity under the assumption that the Fast xFlow Proxy is not available.

When calculating the $X - Y$ error, Y is the average of the last three weeks ($T = 3$). For sequence Z in calculating the $X - Z$ error, S_m was set to 5 for the back table and S_m was set to 50 for the front table. The former assumes that at most five similar services exist in video, web conferencing, and online storage. F or the latter, we assumed that there were approximately 250,000 BSs and 5,000 across-edge routers in Japan, and then 50 BSs were connected to each router. We evaluated our algorithm through computer simulations. The computing environment used was an Intel Core i7-10700 CPU 2.9 GHz (8-core) with 16 GB of memory.

4.2 Results for Service-Side Flow Monitoring

Figures 7 and 8 illustrate the $X - Y$ and $X - Z$ correlation values with the Fast xFlow Proxy when Internet service flows were monitored with reference to the back table. Figures 9 and 10 show the $X - Y$ and $X - Z$ correlation values without the Fast xFlow proxy. In each figure, an error is detected when low correlation values are plotted. For example, as defined in Fig. 6, cases (a) and (d) are where no $X - Y$ error actually occurs, whereas cases (b) and (c) are where $X - Y$ error occurs. In Fig. 7, it is evident that the correlation values for (a) and (d) are high, and those for (b) and (c) are low, which indicates that the results are as expected. For the $X - Z$ error in Fig. 8, we originally expected our algorithm to obtain a trend where the correlation values of (a) and (b), where no $X - Z$ error occurred, are high, and the correlation values of (c) and (d), where an error occurred, are low. However, the correlation value for (d) is high. This implies that it is difficult to detect failures of similar sequences. In the conventional method, as shown in Fig. 9, the correlation value is high in case (c), where an $X - Y$ error occurs. In addition, in Fig. 10, the correlation value is high for case (c), in which $X - Z$ error actually occurs. This indicates that conventional methods without the Fast xFlow Proxy cannot detect targeted service errors in addition to other similar service errors.

Based on the results shown in Figs. 7–10, the error rate is illustrated in Fig. 11 when the threshold value is set from 0.661 to 0.818 for both $X - Y$ and $X - Z$ errors. Errors in the proposed method include an FN where a fault occurs in another Internet service but is not detected. Errors in the conventional method also include an FN where a fault occurs in the targeted Internet service; however, it is not detected. This critical error in the conventional method is caused by low monitoring resolution. When the noise is 0.3, both the proposed and conventional methods generate FP, where the $X - Y$ error is judged to be present, although no failure

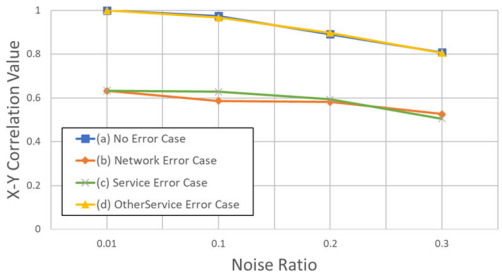


Fig. 7 $X-Y$ correlation values with the Fast xFlow Proxy during service-side flow monitoring.

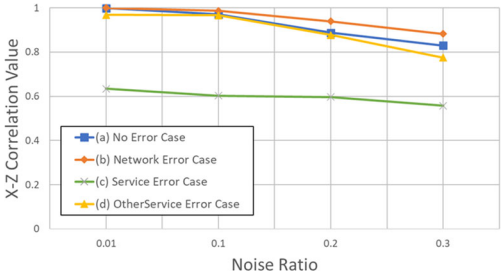


Fig. 8 $X-Z$ correlation values with the Fast xFlow Proxy during service-side flow monitoring.

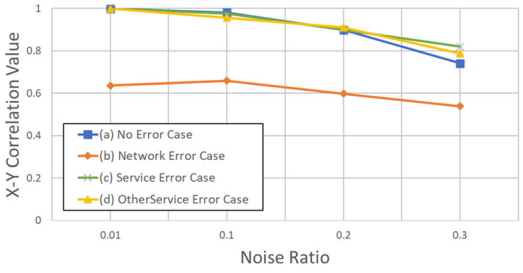


Fig. 9 $X - Y$ correlation values without the Fast xFlow Proxy during service-side flow monitoring.

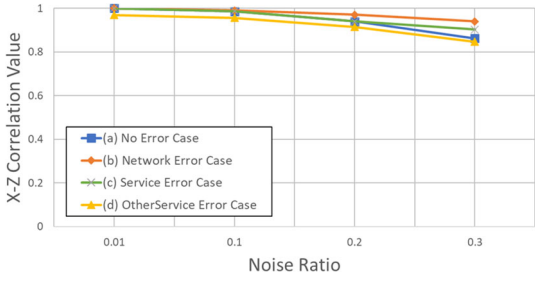


Fig. 10 $X - Z$ correlation values without the Fast xFlow Proxy during service-side flow monitoring.

occurs because there are some cases where the correlation value is slightly below the threshold value. Figure 12 shows the results of precision calculated by TP/(TP+FP) and recall calculated by TP/(TP+FN). For the precision, when the noise reaches 0.3, a few normal conditions are falsely detected as an anomaly. For the recall, the proposed algorithm is affected by the non-detection of the other service errors, while the

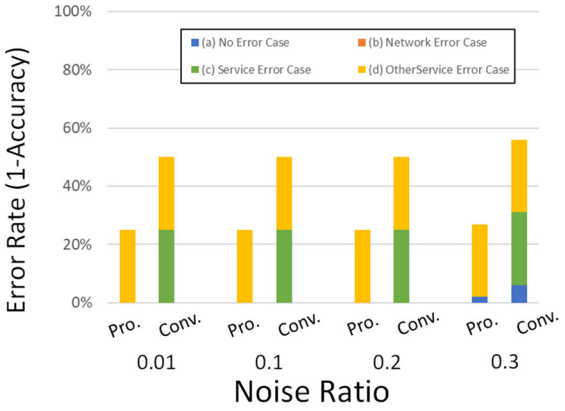


Fig. 11 Error Rate (1-Accuracy) occurring for each situation (a), (b), (c) and (d). Pro. represents the case with the Fast xFlow Proxy and Conv. represents the case without it.

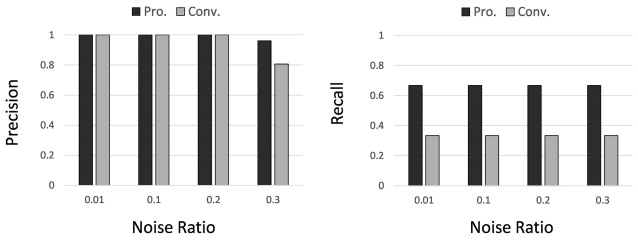


Fig. 12 Precision and recall during service-side flow monitoring. Pro. represents the case with the Fast xFlow Proxy and Conv. represents the case without it.

conventional algorithm is affected by the non-detection of the targeted service and other service errors.

4.3 Results for BS-Side Flow Monitoring

Figures 13 and 14 illustrate the $X-Y$ and $X-Z$ correlation values with the Fast xFlow Proxy when flows from the BSs are monitored with reference to the front table. Figures 15 and 16 show the $X-Y$ and $X-Z$ correlation values without the Fast xFlow Proxy. Based on the results shown in Figs. 13–16, the error rate is illustrated in Fig. 17 when the threshold value is set from 0.661 to 0.818 for both $X-Y$ and $X-Z$ errors. Note that the main difference between service-side monitoring and BSs monitoring is that the number of S_m was increased tenfold, from 5 to 50.

As shown in Fig. 17, the results show almost the same trend as for Internet service monitoring, with the proposed method being able to detect the targeted BS fault (c′) but not the other BS faults (d′). In addition, the conventional method cannot detect the targeted BS fault (c′) due to the reduced resolution. For the $X-Z$ error, it was expected that cases (c′) and (d′) would become more obvious because of the suppression of noise effects by increasing the number of S_m to 50; however, this did not have a sufficient impact to change the trend. The results for the precision and recall are shown in Fig. 18, and the trend is almost the same as in Fig. 12.

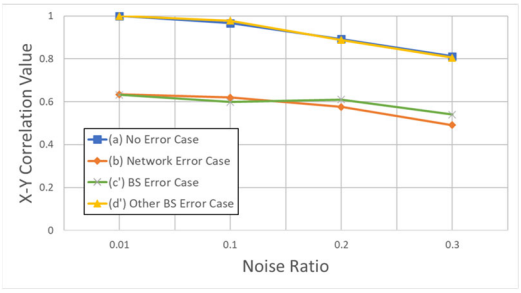


Fig. 13 $X-Y$ correlation values with the Fast xFlow Proxy during BS-side flow monitoring.

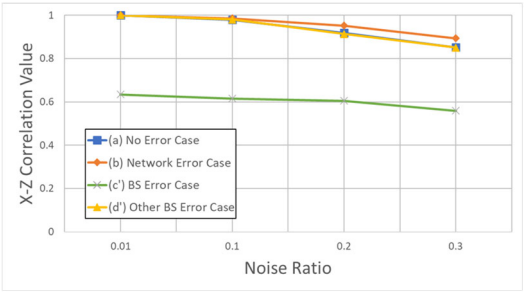


Fig. 14 $X-Z$ correlation values with the Fast xFlow Proxy during BS-side flow monitoring.

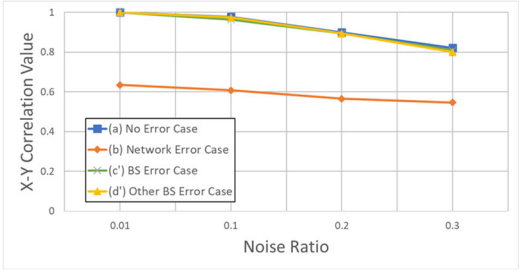


Fig. 15 $X-Y$ correlation values without the Fast xFlow Proxy during BS-side flow monitoring.

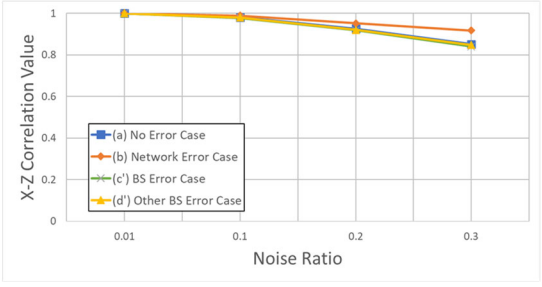


Fig. 16 $X-Z$ correlation values without the Fast xFlow Proxy during BS-side flow monitoring.

4.4 Results for Feasibility

The results of Sects. 4.2 and 4.3 show that the proposed algorithm can accurately detect faults in the targeted Internet

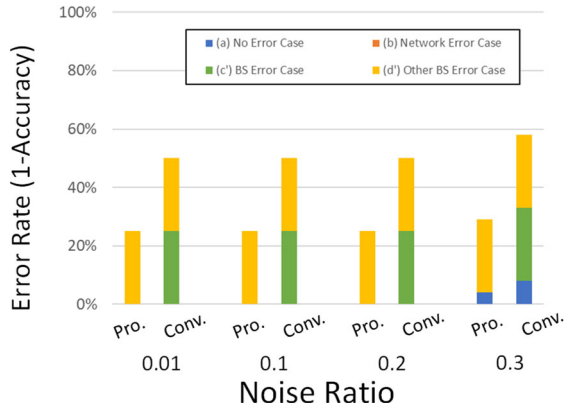


Fig. 17 Error Rate (1-Accuracy) occurring for each situation (a), (b), (c′) and (d′). Pro. represents the case with the Fast xFlow Proxy and Conv. represents the case without it.

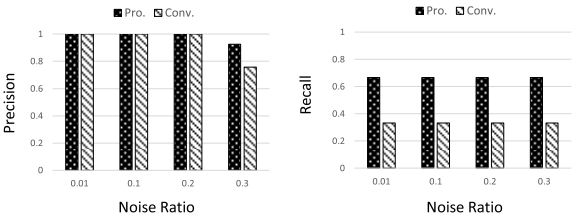


Fig. 18 Precision and recall during BS-side flow monitoring. Pro. represents the case with the Fast xFlow Proxy and Conv. represents the case without it.

service or BS, but it cannot detect faults in other similar services or BSs, resulting in FN errors. If all flows included in sequence Z can be treated as targeted flow, then anomalies in all monitored flows can be detected without error. However, it is necessary to analyze whether it is practical to extract the flow in sequence Z individually as the targeted flow X and analyzing them iteratively is practical. Therefore, in this subsection, we discuss computation time and storage capacity to analyze all the monitored flows in Z iteratively.

Figure 19 illustrates the results of the computation when the number of sequences S_m increases. In addition to the case in which the computation is performed on data from the past month, the computation times are shown for the past two and six months. Since the amount of data stored in the array only increases by a factor of 2–6, the computation time is almost independent of the period and is sufficiently fast.

Figure 19 shows that for the range of numbers of Internet services ($S_m = 5$) or BSs ($S_m = 50$) discussed in Sects. 4.2 and 4.3, the computation time is trivial, falling within 25 seconds. If the network size increases further, e.g. owing to 5G spreading or Beyond 5G, the number of BSs (S_m) will increase. Figure 19 also shows that S_m can be scaled up to 500 for actual conditions, where flows are observed once every five minutes.

Figure 20 shows the results of storage usage when the number of sequences S_m increases. We assumed that the flows were measured once every five minutes and stored in the database. The structure of a flow is composed of 5-

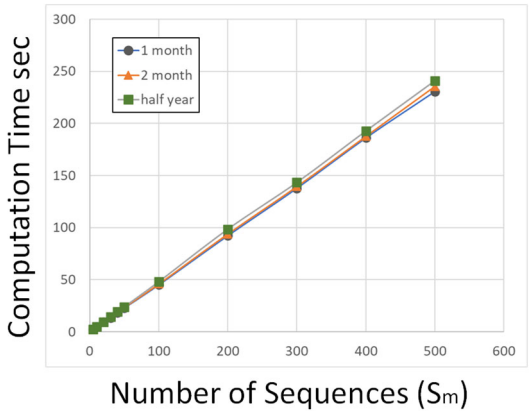


Fig. 19 Computation time when varying the number of sequences analyzed. Environment is 3-core Intel(R) Xeon(R) Gold 6212U CPU@2.40 GHz.

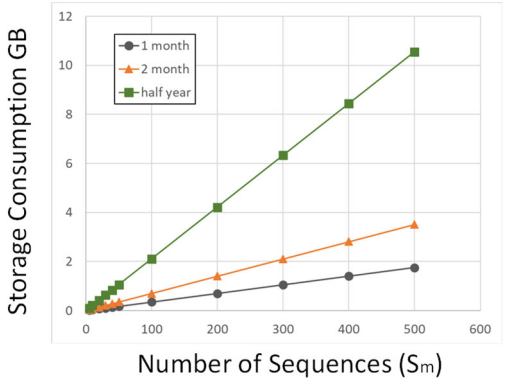


Fig. 20 Storage Consumption when varying the number of sequences analyzed.

Tuple, a measured interface, and live periods. From this structure, we define the size of each flow per unit time as approximately 423 bytes. Based on this data size, the size of storage required for one month, two months, and six months was computed, as illustrated in Fig. 20. For example, for one month of flow data, the data size for detecting the BS error with $S_m = 50$, which is the current simulation scenario, was small at approximately 170 MByte. When the number of analysis targets increases to 500 as the size of network increases, the data size for one month is 1.7 GByte, and 10.5 GByte for six months. We believe that given the current storage capacity, this amount is not considered a problem. Therefore, our algorithm can detect other similar service errors by iteratively analyzing the sequences in Z.

Note that even when S_m was increased up to 500, and the accuracy trend is similar to the results presented in Sects. 4.2 and 4.3, primarily because the increase of S_m only affected noise suppression.

5. Concluding Remarks

This study proposed a fault detection algorithm using the Fast xFlow Proxy, which enables fine-grained measurement

of communication traffic. Our algorithms not only analyze the autocorrelation of a specific flow but also the correlation between flows; thus, the fault location is estimated. Thus, we can centrally detect faults in the front-side access network, core network, and backside services by simply monitoring the traffic flows in its own network. Based on the evaluation, our algorithms can detect base station, service, and network failures with no detection errors if the noise in the flow reaches approximately 30%. Moreover, we demonstrated that the computation time and storage consumption are not problems at the scale we assume, and that it is possible to scale up to 10 times the current scale.

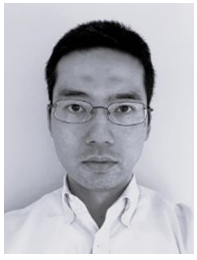
This study did not use actual traffic, but used pseudo-traffic generated by our traffic generator [29]. In the future, we plan to implement the proposed system and demonstrate its feasibility for anomaly detection in actual traffic. It was found that the storage consumption was greater than the computation time in the analysis system. Therefore, we plan to study more efficient flow measurement and storage algorithms. In particular, our assumption that traffic can be measured by selecting arbitrary edge routers from all edge routers may cause inefficiency in traffic management. Therefore, we will study the optimal placement of measurement points.

References

[1] B. Claise, "Cisco Systems NetFlow services export version 9," IETF RFC3954, Oct. 2004.
[2] B. Claise, B. Trammell, and P. Aitken, "Specification of the IP flow information export (IPFIX) protocol for the exchange of flow information," IETF RFC7011, Sept. 2013.
[3] S. Kamamura, Y. Hayashi, Y. Miyoshi, T. Nishioka, C. Morioka, and H. Ohnishi, "Fast xFlow proxy: Exploring and visualizing deep inside of carrier traffic," IEICE Trans. Commun., vol.E105-B, no.5, pp.512–521, May 2022.
[4] S. Kamamura, Y. Takei, M. Nishiguchi, Y. Hayashi, and T. Fujiwara, "Network anomaly detection through IP traffic analysis with variable granularity," IEEE Access, vol.11, pp.129818–129828, 2023.
[5] C. Filsfils, S. Previdi, L. Ginsberg, B. Decraene, S. Litkowski, and R. Shakir, "Segment routing architecture," IETF RFC8402, July 2018.
[6] W. Townsley A. Rubens, G. Pall, G. Zorn, and B. Palter, "Layer two tunneling protocol 'L2TP'," IETF RFC2661, Aug. 1999.
[7] A. Clemm, M.F. Zhani, and R. Boutaba, "Network management 2030: Operations and control of network 2030 services," J. Netw. Syst. Manage., vol.28, pp.721–750, 2020.
[8] J. Case, M. Fedor, M. Schoffstall, and J. Davin, "A simple network management protocol (SNMP)," IETF RFC 1157, May 1990.
[9] M. Yu, "Network telemetry: Towards a top-down approach," ACM SIGCOMM Computer Communication, vol.49, issue 1, pp.11–17, Feb. 2019.
[10] "sFlow," Available at: http://www.sflow.org/, accessed Dec. 13 2023.
[11] S. Kashima A. Kobayashi, and P. Aitken, "Information elements for data link layer traffic measurement," IETF RFC7133, May 2014.
[12] T. Bujlow, V. Carela-Español, and P. Barlet-Ros, "Independent comparison of popular DPI tools for traffic classification," Comput. Netw., vol.76, pp.75–89, 2015.
[13] Y. Li, R. Miao, C. Kim, and M. Yu, "FlowRadar: A better NetFlow for data centers," Proc. NSDI, pp.311–324, 2016.
[14] G. Xie, Q. Li, G. Duan, Y. Jiang, Z. Qi, S. Liu, and Q. Wang, "Efficient flow recording with InheritSketch on programmable switches," Proc. IEEE ICDCS 2023, July 2023.
[15] "Flowmon collector," Available at: https://www.flowmon.com/en/products/appliances/netflow-collector, accessed Dec. 13 2023.
[16] "Intel FPGA programmable acceleration card N3000 data sheet," Available at: https://www.intel.com/content/dam/www/programmable/us/en/pdfs/literature/ds/ds-pac-n3000.pdf, accessed Dec. 13 2023.
[17] "DPDK," Web Site, https://www.dpdk.org/, accessed Dec. 13 2023.
[18] S. Kamamura, "Dynamic traffic engineering considering service grade in integrated service network," IEEE Access, vol.10, pp.79021–79028, 2022.
[19] Z. Xu, F.Y. Yan, R. Singh, J.T. Chiu, A.M. Rush, and M. Yu, "Teal: Learning-accelerated optimization of WAN traffic engineering," Proc. ACM SIGCOMM, pp.378–393, Sept. 2023.
[20] F. Wang, M. Li, Y. Mei, and W. Li, "Time series data mining: A case study with big data analytics approach," IEEE Access, vol.8, pp.14322–14328, 2020.
[21] S. Saha, A. Haque, and G. Sidebottom, "An empirical study on Internet traffic prediction using statistical rolling model," Proc. Int. Wirel. Commun. Mob. Comput. (IWCMC), vol.2022, pp.1058–1063, 2022.
[22] P. Malhotra, L. Vig, G. Shroff, and P. Agarwal, "Long short term memory networks for anomaly detection in time series," Proc. European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN), April 2015.
[23] M. Munir, S.A. Siddiqui, A. Dengel, and S. Ahmed, "DeepAnT: A deep learning approach for unsupervised anomaly detection in time series," IEEE Access, vol.7, pp.1991–2005, 2019.
[24] F. Rasheed, P. Peng, R. Alhajj, and J. Rokne, "Fourier transform based spatial outlier mining," Proc. International Conference on Intelligent Data Engineering and Automated Learning (IDEAL), 2009.
[25] H. Ren, B. Xu, Y. Wang, C. Yi, C. Huang, X. Kou, T. Xing, M. Yang, J. Tong, and Q. Zhang, "Time-series anomaly detection service at Microsoft," Proc. International Conference on Knowledge Discovery and Data Mining (SIGKDD) 2019, pp.3009–3017, 2019.
[26] P. Boniol, M. Linardi, F. Roncallo, T. Palpanas, M. Meftah, and E. Remy, "Unsupervised and scalable subsequence anomaly detection in large data series," VLDB J., vol.30, no.6, 909–931, 2021.
[27] P. Boniol, J. Paparrizos, T. Palpanas, and M.J. Franklin, "SAND: Streaming subsequence anomaly detection," Proc. VLDB Endowment, vol.14, no.10, pp.1717–1729, 2021.
[28] "GenieAnalytics Deep Trace," https://www.genie-networks.com/genieanalytics-deep-trace/, accessed Dec. 13 2023.
[29] S. Kamamura, Y. Takei, M. Nishiguchi, Y. Hayashi, and T. Fujiwara, "Design and implementation of time-series traffic generator," 2023 IEICE General Conference, B-6-3, March 2023.

Shohei Kamamura received the B.E., M.E., and Ph.D. degrees from Waseda University, Tokyo, in 2004, 2006, and 2013, respectively. He joined the Nippon Telegraph and Telephone Corporation (NTT) in 2006. Since then, his research interests have included IP and optical network design, control, optimization, and network data mining. From 2017 to 2020, he was engaged in commercial development of SDN controller for Web conferencing at NTT Communications. Since 2021, he has been an associate professor in the Department of Computer and Information Science at Seikei University. He received the Young Engineer Award in 2010 and the ICM English Session Encouragement Award in 2010 from the IEICE, the Best Paper Awards of the ICC from the IEEE in 2014, the TELECOM System Technology Award from the Telecommunication Advancement Foundation in 2017, and the Distinguished Contributions Award in 2017 from the IEICE. He is a member of IEEE and IPSJ.

Yuhei Hayashi received the B.E. and M.E. degrees from Tokyo Institute of Technology, Tokyo in 2012, and 2014, respectively. He joined the Nippon Telegraph and Telephone Corporation (NTT) in 2014. Since then, his research interests have included network security, network visualization and optimization. Currently, he is a researcher at NTT.



Takayuki Fujiwara received the M.E. degree from Osaka University in 2006. He joined the Nippon Telegraph and Telephone Corporation (NTT) in 2006. His research interests include solving network design and flow optimization with mathematical methods.

PAPER

Energy-Efficient Power Management for O-RAN Base Stations Utilizing Pedestrian Flow Analytics and Non-Terrestrial Networks

Takaya MIYAZAWA[†a)], **Kentaro ISHIZU**[†], ***Members***, **Hitoshi ASAEDA**[†], **Hiroyuki TSUJI**[†], ***Fellows***, ***and*** **Hiroaki HARAI**[†], ***Member***

SUMMARY Recently, the open radio access network (O-RAN) architecture has been expected to enhance both the openness of network components and the intelligence of control functions as a promising RAN architecture for Beyond 5G (B5G)/6G networks. Meanwhile, the power consumption of base stations (BSs) in RAN is a serious problem that needs to be addressed owing to the recent increase in service types such as 4G-LTE, 5G, and local 5G, and it will be more remarkable in the future B5G era. However, a conventional RAN experiences energy wastage because it turns on the power of all BSs at all times, even in coverage areas that accommodate a small number of mobile terminals and low traffic. The O-RAN Alliance discusses the energy savings of BSs, but its standard specification lacks sufficient discussions on concrete models and protocols to realize highly energy-efficient power-on/off management of BSs. On the other hand, terrestrial network (TN) and non-terrestrial network (NTN) convergence has recently been considered in both academic research and standardization as an emerging technology for B5G networks. However, utilizing NTN capacities for BS power-on/off control of TN in the standard O-RAN architecture remains uninvestigated, although it has the potential to achieve higher energy efficiency. This study proposes a novel energy-efficient power management architecture for O-RAN BSs. The proposed power management architecture extends the traditional standard O-RAN architecture such that the pedestrian flow analytics results and NTN capacities can be effectively utilized to obtain a higher energy-saving effect for O-RAN BSs. Consequently, the proposed power-on/off control reduces the power consumption of O-RAN BSs while maintaining the continuity of communications, bitrate, and other metrics. We performed numerical calculations using real datasets of pedestrian flows in regional mesh areas. As a result, we proved that the proposed architecture reduces power consumption by up to 40% when the NTN can accommodate UEs' traffic of approximately 400 Mbps. In addition, we implemented pedestrian flow analytics and power control functions in the controllers. We verified the feasibility of the functions by demonstrating the power-on/-off of an O-RAN BS using a mobile network testbed.

key words: *O-RAN, base station, data analytics, power control, non-terrestrial network, testbed*

1. Introduction

Recently, mobile communication traffic has been increasing at an annual rate of approximately 20% to 30% [1], and this tendency is expected to continue in Beyond 5G (B5G)/6G era. This will require mobile network operators (MNOs) to efficiently accommodate constantly increasing traffic, and MNOs will be required to promptly and stably provide diverse communication services while maintaining and enhancing the quality of service (QoS) [2]. Consequently, it is crucial to enhance the flexibility of network components that transfers data between mobile terminals and mobile core networks and to fully automate network operations and management functions. To achieve this, open radio access networks (O-RANs) have attracted considerable attention recently [3]. The O-RAN Alliance discusses the specifications of O-RAN, which have two major characteristics: (1) "Openness" enabling MNOs to install multi-vendor network devices and promptly provide network services, and (2) "Intelligence" to achieve full automation of network operations and management.

Current mobile networks provide diverse service types that handle various frequency bands and coverage scales, such as the existing 4G-LTE, 5G, and local 5G services; consequently, the type and number of base stations (BSs) have continuously been increasing. Thus, the power consumption of BSs in a RAN is a serious problem that needs to be addressed [3]–[6]. In [6], KDDI Corp. reported that BSs were responsible for approximately 60% of a company's total energy consumption. When 6G services are commercialized in the future, this problem will become more notable because it is essential to deploy a large number of 6G BSs while maintaining existing BSs. Thus, the significance of energy-efficient technologies for RAN has increased in recent years. However, conventional RAN turns on the power of all BSs at all times, even in coverage areas where the number of mobile terminals is small and mobile traffic is low, which wastes power resources. Although the standard O-RAN specification [3] states that future RAN will require a framework to achieve higher energy efficiency, it lacks sufficient discussions on concrete models and protocols to reduce the power consumption of O-RAN BSs.

Meanwhile, a multilayered network realized by terrestrial network (TN) and non-terrestrial network (NTN) convergence has recently been discussed in both academic research and standardization as an emerging technology for B5G networks [7]–[10]. This convergence enables mobile networks to seamlessly cover various geographical areas, such as urban and rural areas, mountains, sea, and air, while maintaining QoS. However, utilizing NTN capacities for BS power-on/off control of TN even in the standard O-RAN specification remains uninvestigated, even though it has great potential for achieving higher energy efficiency in mobile systems.

In this study, we propose an energy-efficient power man-

Manuscript received December 18, 2023.
Manuscript revised March 22, 2024.
Manuscript publicized May 6, 2024.

[†]National Institute of Information and Communications Technology, Koganei-shi, 184-8795 Japan.
a) E-mail: takaya@nict.go.jp

DOI: 10.23919/transcom.2023EBT0009

agement architecture for O-RAN BSs to reduce the power consumption of O-RANs while maintaining the continuity of data communication and bitrates. The proposed management architecture extends the standard O-RAN architecture to realize BS power-on/off control utilizing the intelligent analytics results of pedestrian flow and NTN capacities. This is the world's first technology for O-RAN BSs to utilize both pedestrian flow analytics and NTN for energy-efficient power management of mobile systems.

The proposed architecture comprises three sequential processes.

(1) Pedestrian flow analytics for each BS coverage area:

A data analytics engine trains a machine learning (ML) model by using past data on pedestrian flow in each BS coverage area, and then predicts the pedestrian flow for a specific period. Here, the pedestrian flow means the time-varying number of mobile terminals in each BS coverage area. Furthermore, we assume that a BS coverage area corresponds to a regional mesh area of 500 meters square, which is called "mesh area" hereafter.

(2) Power-on/off scheduling for each specified time:

For each specific time, a power scheduling engine performs power-on/off scheduling for BSs in consideration of the pedestrian flow prediction results obtained by the process (1) in many mesh areas. The scheduling algorithm maximizes the number of BSs that can be turned off. Besides, in the algorithm, UEs existing in mesh areas in which the BSs are turned off can access NTN such as Low Earth Orbit (LEO) and High Altitude Platform Station (HAPS) to reduce the power consumption of TN while maintaining the continuity of data communications, the bitrates, and so on.

(3) Power-on/off control of O-RAN BSs:

Each O-RAN BS is dynamically turned on or turned off at each specified time in accordance with the power control scheduling result obtained by the process (2).

To the best of our knowledge, this is the first study to make this attempt to design, evaluate, and demonstrate the sequential processes of highly energy-efficient O-RANs. The contributions of this study are summarized as follows:

- For (1) Pedestrian flow analytics, the main challenge is how to accurately predict the pedestrian flow for every mesh area in a specific period to enable proactive power-on/off scheduling. Considering that it takes considerable time (e.g., several tens of minutes) to activate several network components constituting O-RAN, reactive power scheduling methods using past data are incapable of following the time variation of pedestrian flow. This makes it difficult to maintain the continuity of data communications and performances such as bitrates under the limited NTN capacities. To cope with the challenge, we apply and compare three existing ML algorithms for pedestrian flow analytics. We performed numerical calculations using real datasets of pedestrian flow in Koganei City and Shinjuku Ward in Tokyo, Japan, and evaluated the prediction accuracy of pedestrian flow. As a result, we proved that the application of long short-term memory (LSTM) [11]–[13] is the optimal solution and can reduce the prediction error by up to 48.9% compared with other algorithms and enhance the suitability of (2) power-on/off scheduling.
- For (2) power-on/off scheduling, the main challenge is how to create a scheduling model which can turn off as many BSs as possible considering the capacities and power consumption of NTN. Here, we target at LEO [14], HAPS [15], and balloon [16] as communication equipment constituting NTN, and estimate the available capacity and power consumption of each equipment. We formulated a power-on/off scheduling model and evaluated the relative power consumption of the proposed method in Koganei City and Shinjuku Ward in Tokyo, Japan. As an evaluation result, we proved that the proposed scheduling method can reduce the power consumption of O-RAN BSs by up to 40% when the NTN can accommodate UEs' traffic of around 400 Mbps in Koganei City and around 1.0 Gbps in Shinjuku Ward, respectively. Consequently, our architecture can highly contribute to energy savings for O-RAN BSs.
- For (3) power-on/off control of O-RAN BSs, at first, we implemented the functions of processes (1) and (2) into our controller, and connected the controller to an NICT B5G mobile testbed called "Mobile Network Development Environment" [17], [18]. Then, we made field experiments to verify the power-on/off control of the O-RAN BS. For B5G/6G networks, we target at the frequency bandwidth ranging from 4.8 GHz to 4.9 GHz which is within sub-6 GHz frequency bands and higher than the frequencies used in existing mobile systems such as 4G-LTE and 5G. We show captured images of control processes, the time required for BS power-on process, and measurement results of frequency spectrum displayed on an oscilloscope, all of which were obtained in the experiments. In this way, we verify the feasibility of the proposed sequential processes.

The remainder of this paper is organized as follows: Related work is introduced in Sect. 2. In Sect. 3, we discuss the proposed power-management architecture for O-RAN BSs. We present the models for pedestrian flow analytics and power control scheduling in Sect. 4 and the experimental system in Sect. 5. In Sect. 6, we evaluate the prediction accuracy of the pedestrian flow, the energy-saving effect obtained by our proposed power control scheduling, and the suitability of scheduling. We also present the results of field experiments using the B5G mobile testbed in Sect. 7. Finally, Sect. 8 concludes the paper and provides an overview of future work.

2. Related Work

2.1 Power Control of BSs in TN

The authors of [19]–[22] adopted game theory approaches to allocate power to each BS. For example, in [22], Xiao et al. applied reinforcement learning (RL) for power control in downlink data transmissions, in addition to formulating the power allocation problem as a zero-sum game. The application of game theory and RL may be effective for power control when the effective utilization of power resources is essential because of the limitation of the maximum amount of available power.

In [23], Ye et al. proposed a BS on/off switching control method utilizing a deep RL (DRL) technique and a deep deterministic policy gradient. They proved that it can be adopted in time-varying network environments while minimizing the power consumption of BSs, considering the network traffic. In [24], Wu et al. proposed traffic-aware dynamic BS sleep control utilizing DRL. They applied a convolutional neural network and LSTM to predict mobile traffic and an actor-critic RL with a deep neural network to minimize long-term power consumption while satisfying the QoS requirements of mobile users. They also demonstrated the control method by using a real dataset from Milan, Italy.

In [25], Tan et al. proposed a cooperative BS switching-off control method that mainly consists of a roaming-cost-based algorithm to search for feasible inactive BSs and a cooperative beamforming algorithm to optimize beamforming vectors among multiple MNOs. They proved that the proposed method could enhance the energy efficiency of the RAN while maintaining the economic profit of the MNOs.

However, all the above studies lack discussions on power-on/off control for components in O-RAN BSs. In addition, they did not consider utilizing both the pedestrian flow analytics results and NTN capacities to realize highly energy-efficient BSs, which is a novel approach in this study.

2.2 Operations and Management of O-RAN in TN

The authors of [26]–[28] proposed operations and management methods related to handover procedures for mobile users. For example, in [28], Cao et al. proposed a user access control method that utilizes a deep Q-network (DQN) and a mechanism to learn the parameters of multiple DQN models using federated DRL. Their proposed method can reduce the frequency of user handovers while maximizing throughput from a long-term perspective.

In [29], D'Oro et al. proposed an automated network intelligence orchestration framework called OrchestRAN which follows the standard O-RAN specification. This enables MNOs to instantiate data-driven services on demand with minimal control overhead and latency. In [30], Masur et al. introduced possible applications of ML techniques for O-RANs. For example, they stated that supervised learning techniques can be applied to user quality of experience optimization, traffic steering, and vehicle-to-X handover management, and DRL techniques can be applied to solve a spectrum sensing problem to realize random access from distributed IoT devices.

However, the above studies did not consider power management for O-RAN BSs to realize a highly energy-efficient RAN.

In [31], Giannopoulos et al. proposed the application of LSTM to predict traffic data in cells and applied a DRL to maximize the energy efficiency of an O-RAN. Although their study is partly similar to our study presented in this paper, they did not consider the utilization of NTN capacities for power control scheduling and did not implement a power-on/off control function for O-RAN BS's components to achieve higher energy efficiency. In addition, they did not use a network testbed, including a BS compliant with O-RAN, to experimentally verify the feasibility of the proposed method. Our study is more advanced in that, in addition to utilizing both pedestrian flow analytics results and NTN capacities for realizing a highly energy-efficient O-RAN, we demonstrate pedestrian flow prediction for each mesh area using real datasets measured in Tokyo, Japan, power-on/off scheduling for multiple mesh areas at each specific time, and power-on/off control using the NICT B5G mobile testbed.

2.3 TN and NTN Convergence

TN and NTN convergence technologies have rapidly gained attention, owing to their potential for seamless coverage in various geographical areas [7]–[10]. In [7], Liolis et al. introduced the project named "Satellite and terrestrial networks for 5G (SaT5G)," investigated the extension of network virtualization and softwarization capabilities to satellite network management, and proposed a technology to seamlessly integrate NTN into the management and orchestration architecture of 5G systems. In [8], Zhu et al. presented a comprehensive survey of various integrated satellite-terrestrial network architectures at multiple layers, ranging from the physical layer to the application layer. In [9], Kafle et al. presented scenarios and architectures for integrated and individual network control systems that can accommodate various control functions for the design and orchestration of end-to-end network services, performance monitoring, and resource control of both TNs and NTNs. In addition, TN and NTN integrated network control architectures have recently been discussed in the ITU-T [10].

However, all the above studies and standardization lack discussions on power management for O-RAN BSs to realize a highly energy-efficient RAN in TN.

2.4 Standardization

The O-RAN Alliance [3] states that ML techniques are applicable to the optimal control of O-RANs in consideration of the trade-off between energy efficiency and QoS. It also suggests that the ML model and scheduling algorithm should

be updated when the predetermined target value of energy saving is not achieved.

The O-RAN Plugfest, sponsored by the O-RAN Alliance, is an international event that shares the results of verification experiments for interoperability between different vendor BSs that are compliant with O-RAN specifications [32]. Although the event mainly focuses on interoperability, energy saving in O-RAN will be a significant topic because reduction of power consumption is crucial for O-RAN BSs, as mentioned in Sect. 1.

Energy saving for RAN has also been discussed in 3GPP [4], [5]. The technical specification in [4] states that MNOs should decide on a policy for power control of BSs. The technical report in [5] describes a sequence diagram for training ML models; however, it has no distinct characteristics and specifies no concrete algorithms and protocols.

There is no existing standard specification that utilizes both pedestrian flow analytics results and NTN capacities to realize highly energy-efficient O-RAN BSs.

3. Architecture

3.1 Traditional O-RAN Architecture

Figure 1 shows a schematic of the traditional O-RAN architecture standardized by the O-RAN Alliance [3], [33]–[40]. The O-RAN standard architecture separates a BS into several component—Central Unit (CU), Distributed Unit (DU), Radio Unit (RU), and RAN Intelligent Controller (RIC)—that are interconnected with each other [33]. CU, DU, and RU correspond to the data processing unit, wireless signal processing unit, and antenna unit, respectively. RIC is the control unit for the entire O-RAN and has many functions, such as resource control, QoS control, and traffic control. RIC is divided into two controllers: (1) Non-Realtime (Non-RT) RIC conducting control processes requiring more than one second [34], and (2) Near-Realtime (Near-RT) RIC conducting control processes requiring the range from 10 ms to 1 s [35].

The O-RAN specifications define several interfaces to interconnect the components: "A1" between the Non-RT RIC and Near-RT RIC [36], "E2" between the Near-RT RIC and two components (CU and DU) [37], "F1" between CU and DU [38], "O1" between the Non-RT RIC and three components (Near-RT RIC, CU, and DU) [39], and "Open Fronthaul" between DU and RU [40]. The User Element (UE) is a mobile terminal communicating directly with an RU and is connected to a mobile core via the O-RAN.

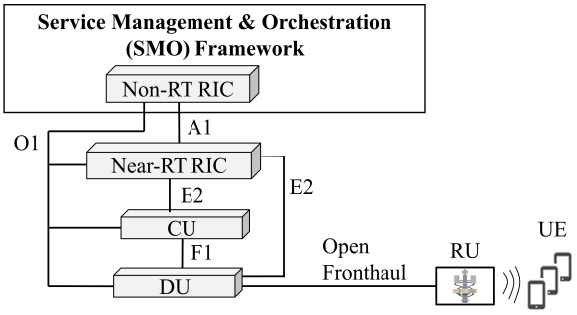


Fig. 1 O-RAN standard architecture.

3.2 Proposed Architecture

Figure 2 illustrates a schematic diagram of the proposed power management architecture, which extends the standard O-RAN architecture such that the pedestrian flow analytics results and NTN capacities can be effectively utilized to achieve higher energy efficiency. According to the specifications in [3], a Near-RT RIC is deployed in the BS where the corresponding CU and DU are deployed, whereas a Non-RT RIC can be deployed in a remote site, such as a mobile core. In the proposed architecture, because power-on/off scheduling is executed considering pedestrian flow in many mesh areas, the Non-RT RIC is deployed in the mobile core. In this study, a mesh area corresponds to the area covered by one O-RAN BS (i.e., a set of CU and DU), and one or more RUs are deployed in each mesh area.

The Non-RT RIC deploys a pedestrian flow analytics engine for each mesh area, which predicts the pedestrian flow in a specific period by analyzing the past data of the pedestrian flow received from the corresponding BS. In addition, the Non-RT RIC deploys a power scheduling engine that schedules the turning on or turning off BSs in advance for each specific future time based on the pedestrian flow prediction results in many mesh areas. Subsequently, in accordance with the scheduling results, it sends a power-on/off control message to the Near-RT RIC in every BS via the O1 or A1 interface. After receiving the control message from the Non-RT RIC, the Near-RT RIC powers the CU and DU on/off via the E2 interface. Although the interfaces of O1,

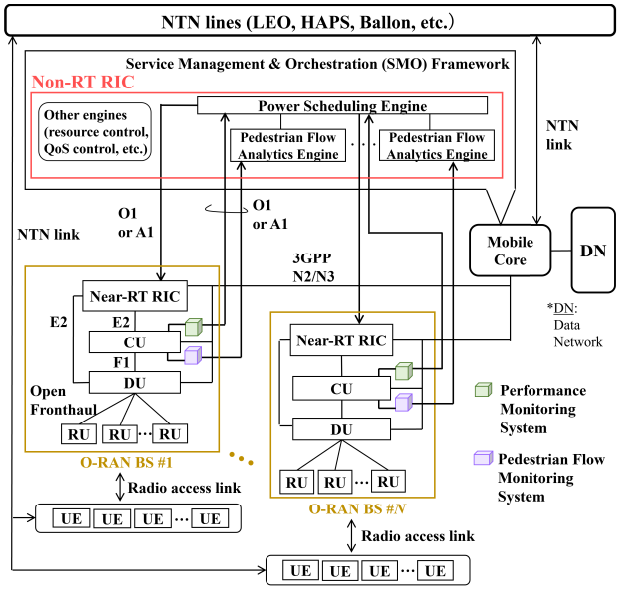


Fig. 2 A schematic diagram of proposed power management architecture for O-RAN BSs.

A1, and E2 are defined in the standard O-RAN specifications, no concrete models or protocols have been specified to realize their power control functions.

The Non-RT RIC in our proposed architecture can use any ML platforms that can deploy algorithms of time-series data prediction; for example, there are many existing platforms such as Google Cloud AI, Microsoft Azure AI, Amazon SageMaker, and IBM Watson. Besides, it is also possible for each MNO to construct its own platform, install ML functions, and analyze collected datasets. In this study, we made our own script by using Python's standard libraries such as tensorflow, tflearn, scikit-learn, and so on, and deployed our own ML platform.

Each O-RAN BS has two monitoring systems for pedestrian flow and communication. The pedestrian flow monitoring system monitors the number of UEs every 15 min in the corresponding mesh area and sends the data to the corresponding pedestrian flow analytics engine in the Non-RT RIC. The performance monitoring system monitors parameters such as the power consumption and bitrates every 15 min in the corresponding mesh area and sends the data to the power scheduling engine in the Non-RT RIC. We assume that the performance monitoring system and pedestrian flow monitoring system are deployed near or in the CU. The bitrate in each coverage area can be measured by analyzing the number of packets getting through the performance monitoring system by means of some tool such as Wireshark. Meanwhile, the number of mobile terminals for each 15 min can be estimated in each coverage area by means of a mechanism of location registration in mobile systems.

In the proposed architecture, each UE can access not only O-RANs in TN but also NTN lines. The link interface to NTN is the same as that between UEs and RUs in TN because we expect that the same type of O-RAN BSs will be deployed on NTN, such as LEO and HAPS, in the future. When UEs access NTN lines, it is also indispensable to deploy a wireless link between the mobile core and NTN so that the UEs can communicate with the data network (DN) (e.g., the Internet) via the mobile core.

Note that there may be UEs requiring a high QoS, such as a large communication capacity in the mesh areas where the BSs are turned off. In this case, if NTN can satisfy the QoS, the traffic from such UEs can be transferred to NTN; otherwise, it is necessary for such UEs to access another type of access point in TN, such as WiFi, 5G, or local 5G. This is also applicable to several cases where the UEs are incapable of accessing NTN; for example, severe weather and equipment failure.

Generally, there could be four types of structures of CU/DU/RU as illustrated in Fig. 3: (a) One CU is connected to one DU, and the DU can accommodate one RU, (b) one CU is connected to one DU, and the DU can accommodate multiple RUs, (c) one CU is connected to multiple DUs, and each DU can accommodate one RU, and (d) one CU is connected to multiple DUs, and each DU can accommodate multiple RUs. MNOs can select either one of the four cases for each geographical area flexibly depending on the situations such as pedestrian flow, geographical constraints, and performance requirements for RICs.

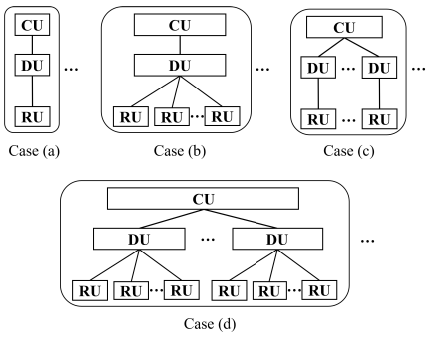


Fig. 3 Four types of structures of CU/DU/RU.

4. Pedestrian Flow Analytics and Power Control Scheduling

In this section, we target the data analytics of the pedestrian flow and BSs power-on/off scheduling, both of which are executed in a Non-RT RIC.

4.1 Pedestrian Flow Prediction in Each Mesh Area

It is well known that no ML algorithm can provide perfect generalizable solutions for any dataset or objective function. Therefore, the best method is to compare the prediction accuracies of multiple algorithms and determine the optimal solution for each case. In this study, we apply and compare three existing models for pedestrian flow prediction: (I) Average Values Selection (AVS), (II) least-squares Support Vector Regression (LS-SVR), and (III) LSTM. Many studies have applied SVR and LSTM as representative ML algorithms for time-series data analytics such as stock price prediction [41], [42], landslide displacement [43], and liver motion prediction [44]. Thus, in this study, we apply SVR and LSTM to analyze and predict the time-series pedestrian flow in each mesh area. Besides, we compare the prediction accuracies of SVR and LSTM with that of AVS to evaluate the advantageous effect of applying the ML algorithms.

Every model predicts the pedestrian flow for every T_{pred} h the next day by using the pedestrian flow dataset for the past D_{past} days.

(I) The AVS calculates the average value of the past (training) dataset for every T_{pred} hour within the period of 24 hours and regards a set of the average values as the predicted values of pedestrian flow in the next day.

(II) LS-SVR was proposed in 2005 [45] as an extension to the typical SVR [46] and has been applied in many research fields [47]–[49]. Although both typical SVR and LS-SVR have almost the same generalization ability, LS-SVR has several advantages over the typical SVR, such as lower computation costs, flexible selection of kernel functions, simplified maintenance, and

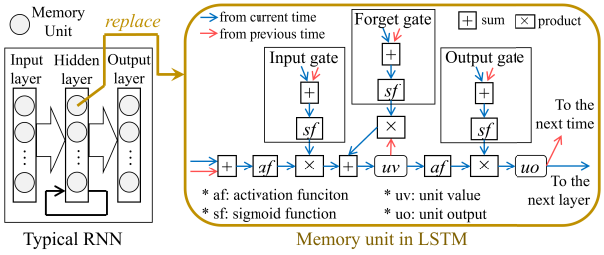


Fig. 4 A block diagram of LSTM.

a lower memory footprint. In particular, a lower computation cost is achieved by defining simple simultaneous linear equations for solving an optimization problem, whereas the typical SVR defines a complicated quadratic programming problem. The mathematical model of LS-SVR can be seen in [49], which uses a Radial Basis Function (RBF) kernel as a representative and the most commonly used kernel method to realize a nonparametric nonlinear regression [50]. In this study, the LS-SVR calculates the average value of the past dataset for every T_{pred} hour and then derives a nonlinear regression curve from the average values, which is regarded as the predicted values of pedestrian flow in the next day.

(III) LSTM [11] is a type of recurrent neural network (RNN) that contains a directed closed loop. RNNs generally suffer from a vanishing gradient problem, which restricts the amount of input data that can be reflected in the output. LSTM adaptation resolves this by selectively remembering or forgetting related or unrelated information, respectively. Owing to its ability to learn long-term data dependencies, LSTM is highly suitable for time-series data analytics. A block diagram of LSTM is illustrated in Fig. 4, where the RNN's memory units are replaced with the units for LSTM in hidden layers so that the new model can adjust the number of inputs carried forward from the previous layer. Hence, the number of unit values from the previous round is now applied in the new round, and the updated number is output to the next round. Hence, the LSTM propagates data on a long-term basis to support extended time-series data analytics [12], [13]. Unlike the AVS and LS-SVR, the LSTM predicts the next day's pedestrian flow by using all the past data without calculating the average value of the past dataset for every T_{pred} hour.

4.2 Power Control Scheduling for Mesh Areas

In this study, we assumed that an RU was deployed in each mesh area (i.e., each BS). We also assume that every RU runs at all times to maintain the stability of several devices, such as the control circuit and amplifier inside the RU. The CU and DU are dynamically turned on or off for every T_{sche} h. Table 1 lists the main parameters, for which the unit of power is W.

Table 1 Parameters.

Symbol	Definition
N	Number of mesh areas in a target city or ward.
$v_i(t)$	Predicted value of pedestrian flow in ith mesh area at time t. $i = 1, 2, 3, \cdots, N$.
$b_{i,j}(t)$	Bitrate of user j in ith mesh area at time t (Unit: bps)
$s_i(t)$	Power state of CU and DU in ith mesh area at time t, set to 0 if the state is ON; otherwise, set to1.
B_{NTN}	NTN capacity to accommodate UEs in OFF areas (Unit: bps)
$P_{\text{on}}(i)$	Total power consumption of CU/DU/RU in ith mesh area when the power state of CU and DU is ON.
$P_{\text{off}}(i)$	Power consumption of RU in ith mesh area when the power state of CU and DU is OFF.
P_{NTN}	Total power of NTN lines
$R(t)$	Relative power consumption at the time t in the proposed method

We define $b_{i,\text{total}}(t)$ as the sum of bitrates in the ith mesh area at time t, which is equal to $\sum_{j=1}^{v_i(t)} b_{i,j}(t)$ and assume that it does not exceed the maximum communication capacity of each BS. The proposed power control scheduling selects the value of $s_i(t)$ for each mesh area, such that the number of mesh areas with $s_i(t) = 1$ can be maximized under the condition that satisfies the following equation:

$$\sum_{i=1}^{N} s_i(t) \times b_{i,\text{total}}(t) \quad \leqq \quad B_{\text{NTN}}, \tag{1}$$

This is the same as the process for setting the value of $s_i(t)$ to 1 in ascending order of $b_{i,\text{total}}(t)$, satisfying Eq. (1).

Consequently, the relative power consumption in the proposed method, $R(t)$ ($t \in (T_{\text{sche}}, 2{\cdot}T_{\text{sche}}, 3{\cdot}T_{\text{sche}}, \cdots)$), can be expressed as

$$R(t) = \frac{P_{\text{prop}} + P_{\text{NTN}}}{P_{\text{conv}}} \tag{2}$$

P_{conv} is the total power consumption in the conventional method that turns on the BSs deployed on all mesh areas in TN at all times and is expressed as $\sum_{i=1}^{N} P_{\text{on}}(i)$.

P_{prop} is the total power consumption of BSs deployed on all mesh areas in TN when executing the proposed control scheduling, which is expressed by the following equation:

$$\sum_{i=1}^{N} \Big\{ s_i(t) \times P_{\text{off}}(i) + (1 - s_i(t)) \times P_{\text{on}}(i) \Big\} \tag{3}$$

4.3 Suitability of the Power Control Scheduling

We define suitability as a metric for evaluating how many results of the proposed power-control scheduling coincide with the correct scheduling results using datasets of actual measurement values. As the accuracy of the pedestrian flow prediction discussed in Sect. 4.1 is higher, power-control scheduling can achieve higher suitability. In particular, if a scheduling result is OFF despite the correct one being ON,

the performance of UEs such as bitrate might be affected by the wrong result because the UEs are obliged to access NTN, which may be unstable compared with TN. Therefore, the proposed power control scheduling optimizes two aspects: (1) obtaining as many correct scheduling results as possible with the limited capacity of NTN, and (2) reducing the frequency of the cases that incorrectly turn off the BSs.

5. Experimental System

5.1 Structure of Experimental System

In the experiment and performance evaluation of this study, we target at the case (b) as the proposed conceptual architecture illustrated in Fig. 2, but target at the case (a) to evaluate our proposed power control in the case that experiences the largest power consumption in O-RAN BSs. Note that our proposed power control can reduce the power consumption for O-RAN BSs in principle in other cases as well as in the case (a).

Figure 5 illustrates a schematic diagram of an experimental system using the mobile testbed called "Mobile Network Development Environment" [17], [18]. In this study, each piece of equipment was deployed inside the NICT Koganei headquarters in Tokyo, Japan. An O-RAN BS consists of a CU, DU, RU, and Near-RT RIC. We used one UE to communicate with an iperf server via the O-RAN BS and mobile core. The mobile core consists of 5G Network Functions (5GNFs), access and mobility management functions (AMFs), session management functions (SMFs), and user plane functions (UPFs). The UPF is connected to the DU in the O-RAN BS on the user plane and to the iperf server through the DN. The DN was constructed using an NICT high-speed R&D network testbed called the JGN [51]. The Non-RT RIC was placed near the mobile core equipment connected to the O-RAN BS in the control plane.

Table 2 lists the specifications of the components of the experimental system. The OS for each component was Ubuntu 22.04LTS. In addition, we used a FoxConn RHON-7901 for the RU and a RAKU Plus 5G Mobile Router for the UE.

Each function of the CU and DU was implemented on Kubernetes containers and the host OS, respectively. We adopted free5GC [52] as an open-source tool for the mobile core, and each function was deployed in Kubernetes containers. We used the software product of Radisys Corp. [53] to control the CU and DU for data communication between the UE and iperf server via the O-RAN and mobile core.

We implemented pedestrian flow analytics and power control scheduling in the Non-RT RIC and automatic power-on/off control of the CU and DU in the Near-RT RIC. In this study, we did not use the existing tools for Non-RT RIC and Near-RT RIC, which are rAPP and xAPP [3]. Instead of using the existing tools, we implemented scripts for the fundamental verification of pedestrian flow analytics and power control scheduling functions and installed them on general-purpose servers. We also implemented a gRPC protocol

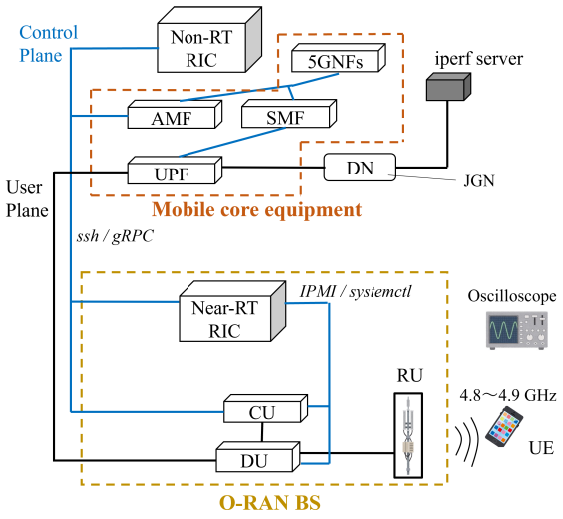


Fig. 5 A schematic diagram of experimental system.

Table 2 Specifications of components.

-	CU, DU	Mobile Core	RIC
Model	PowerEdge R740	Power Edge 750XA	HP ProLiant DL360 G6
CPU ×2	Xeon Gold 6248 2.5GHz (20C/40T)	Xeon Platinum 8358P 2.6GHz (32C/64T)	Xeon 5570 2.93GHz (4C/8T)
Memory	384GB	192GB	96GB
Storage ×4	RAID 8GB cache 800GB SSD SAS (RAID5+HS)	RAID 8GB cache 800GB SSD SAS (RAID5+HS)	146GB HDD (RAID0)

to send and receive power-control messages between the Non-RT RIC and Near-RT RIC. We used an Intelligent Platform Management Interface (IPMI) to turn on the CU and DU from the Near-RT RIC and used a *systemctl* command through SSH to turn off the CU and DU.

5.2 Control Sequence

Figure 6 illustrates the power-control sequence for an O-RAN BS. Prior to sending a control message, the Non-RT RIC executes pedestrian flow analytics and power control scheduling; consequently, the scheduling result of power-on/off at every specific time is obtained for each BS (i.e., each mesh area). At each prespecified time, the Non-RT RIC automatically sends a power control request message to the Near-RT RIC of every BS. After receiving the request message, the Near-RT RIC of each BS automatically executes power-on or power-off to the corresponding CU and DU, depending on the content of the message. In the power-on process, the CU and DU application processes start after their power is turned on. In the power-off process, the CU and DU are turned off after the application processes are stopped. It is also possible for Non-RT RIC and Near-RT RIC to confirm the power states of the CU and DU using the IPMI as necessary.

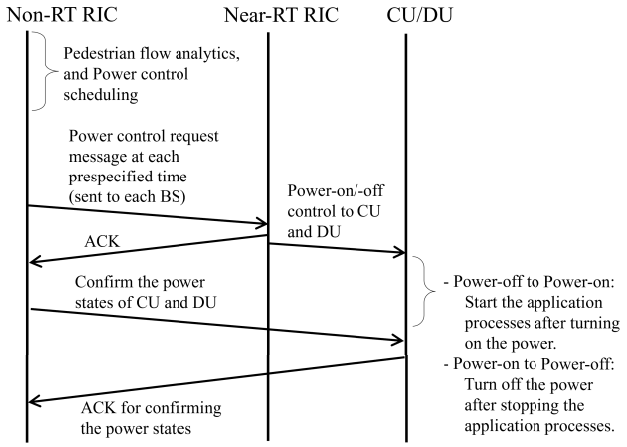


Fig. 6 Control sequence.

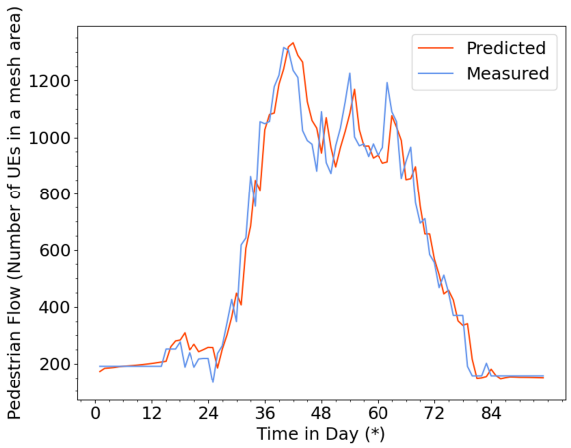


Fig. 7 Pedestrian flow prediction result for every 15 minutes in one mesh area of Koganei City.

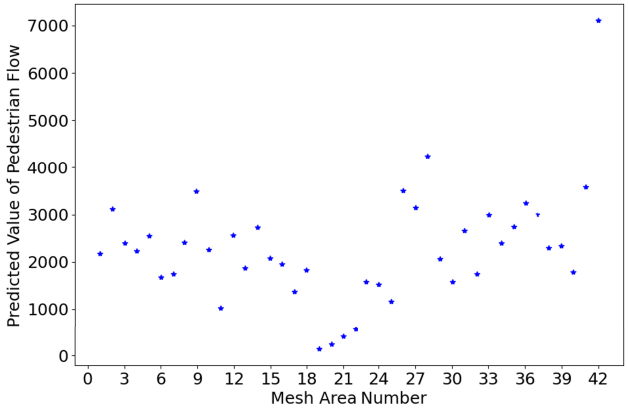


Fig. 8 Predicted values of pedestrian flow in 42 mesh areas of Koganei City at 14:00, September 30, 2022.

The power-on/-off control sequence in Fig. 6 is a straightforward procedure and does not have high novelty. Our objective here is to visually explain the control sequence. This is an essential procedure in our proposed power management architecture for O-RAN BSs. On the other hand, the following application processes and containers need to start and stop in CU/DU during the power-on/-off control: In the CU, several functions such as packet data convergence, radio resource control, connection to the mobile core, control of DU and RU, and O&M (operation and management) are consolidated into 2 containers (pods) created by Kubernetes. In the DU, three functions of (i) signal modulation/demodulation and layer-2 control, (ii) user plane's accelerator, and (iii) precision time protocol are implemented directly on the host OS to maintain the performance of the data transfer.

6. Performance Evaluation

6.1 Pedestrian Flow Prediction

In this study, we used KDDI Location Data (KLD) [54] as pedestrian flow datasets, which correspond to the number of UEs in each mesh area varying every 15 min (i.e., $T_{\text{pred}} = 0.25$). The size of each mesh area was 500 meters square. In this study, we focus on Koganei City and Shinjuku Ward, both in Tokyo, Japan, for pedestrian flow analytics and evaluations of power control scheduling. Koganei City and Shinjuku Ward consisted of 42 and 69 mesh areas, respectively. For each mesh area, we used the pedestrian flow dataset on weekdays within the period from July 8, 2022, to September 29, 2022, as the training dataset and the data from September 30, 2022, as the test data. Pedestrian flow on September 30, 2022, was predicted by using the dataset of the past 56 weekdays.

Figure 7 shows the LSTM-supported pedestrian flow prediction results for every 15 min in a single mesh area of Koganei City. The orange and blue lines represent the predicted and measured values, respectively. The analytics results indicated that LSTM predicts the pedestrian flow with high accuracy, and we observed comparable results in other mesh areas in Koganei City and Shinjuku Ward. It took approximately 3.5 min for Non-RT RIC to complete both the training and prediction for one mesh area. The time required for training and prediction does not depend on the number of mesh areas in the target city (or ward) because a pedestrian flow analytics engine is deployed for each mesh area inside the Non-RT RIC.

Figure 8 shows the pedestrian flow in 42 mesh areas of Koganei City at 14:00 on September 30, 2022. The horizontal scale is the mesh area number, whereas the vertical scale is the value of the pedestrian flow (i.e., the number of UEs) at that time. We can observe that the pedestrian flow values differ for the different mesh areas. Our proposed method optimizes power control scheduling by enabling UEs in mesh areas to access NTN in the ascending order of values of pedestrian flow under the limited capacity of NTN: This is equivalent to turning off as many O-RAN BSs as possible in TN and thus maximizes the energy-saving effect obtained by the proposed method.

Table 3 shows the root mean squared error (RMSE)

Table 3 RMSE in the three algorithms (Average values in all mesh areas).

Algorithm	Koganei City (N=42)	Shinjuku Ward (N=69)
AVS	366.30	1249.87
LS-SVR	366.01	1129.78
LSTM	221.18	639.26

values of AVS, LS-SVR, and LSTM for Koganei City and Shinjuku Ward. Each value represents the average of RMSE values for all mesh areas. AVS and LS-SVR obtain predicted values in less than 1 s. Generally, LS-SVR can create prediction models with a higher tolerance to noise (i.e., abnormal) data than AVS, but the RMSE values of both algorithms are almost the same. This is because the KLD used in this study includes a small amount of noise data that degrade the prediction accuracy of AVS. Meanwhile, LSTM reduced the RMSE values by 39.6 to 48.9% compared with the other two algorithms, owing to the high prediction accuracy of LSTM. Although LSTM takes approximately 3.5 min for the training process using the past dataset, it can achieve both a reduction in prediction errors and a high tolerance to noise data by effectively extracting the characteristics of trends in the fluctuations of pedestrian flow.

Our proposed power management system does not require high agility (e.g., control) in less than 1 s, and the prediction accuracy is much more crucial than the agility of training process using the past dataset. Thus, LSTM is a better solution than AVS and LS-SVR, and we applied LSTM to evaluate the relative power consumption discussed in Sect. 6.2.3. In addition, we compared the suitability of power control scheduling among the three algorithms, as discussed in Sect. 6.2.4.

6.2 Power Control Scheduling

6.2.1 Presupposition

i We measured the starting time of CU and DU activated simultaneously, and the result revealed that it took approximately 19 min to complete the starting processes, which is discussed in Sect. 7. Thereby, we set T_{sche} = 1 to maintain the stable operation of CU and DU in the testbed. We also set the target values of time t as 1:00, 2:00, 3:00, $\cdots$, 23:00. Specifically, we exclude 0:00 from the target values because the period from 0:00 to 0:59 needs to be used mainly for the two processes: (1) Pedestrian flow analytics for every mesh area and (2) Power control scheduling on the basis of the prediction results in all mesh areas.

ii The value of $b_{i,j}(t)$ in Table 1 fluctuates depending on each user and each time in the real world, and thus, the ratio of active UEs using communications services to all UEs fluctuates depending on the time. In this study, as T_{sche} was set to 1, we estimated the total value of all UEs' bitrates for each hour in each mesh area by the following calculations:

(1) Firstly, we utilized the real data of mobile traffic in June 2022 disclosed in [1], which include the average traffic for each hour in each day of the week.

(2) Secondly, we divided the traffic data in (1) by the number of subscriptions of mobile phones [1] in Japan in June 2022. In this way, we averaged all UEs' bitrates independent of i and j.

(3) Finally, we multiplied the averaged bitrate in (2) by the pedestrian flow data in each mesh area. In this way, we estimated the value of $b_{i,\text{total}}(t)$ in each mesh earea.

iii Because we used the estimated amount of traffic in each coverage area as explained above, the implementation and performance evaluation of the monitoring systems themselves are out of scope in this study although they are essential to the thorough implementation of the proposed conceptual architecture illustrated in Fig. 2.

iv We measured the power consumption of each component, and our measurement results revealed that the CU, the DU, and the RU consumed approximately 337 W, 514 W, and 84 W, respectively. These values of power consumption were uncorrelated with those of bitrates. The power consumption of RU remained substantially constant independent of the power-on/-off states of CU and DU, and besides, the amount of network traffic hardly affected the power consumption of RU (at most a few W) in our experimental system. Therefore, in this study, we assumed that the power consumption of RU is kept constant independent of the power states of CU and DU and the amount of network traffic. We reflect these measurement values of power consumption in performance evaluations. This means that $P_{\text{on}}(t)$=935 and $P_{\text{off}}(t)$=84, and the values do not depend on i.

v The results in Sect. 6.1 are the predicted values every 15 min. However, the power control scheduling is executed every 1 h (i.e., T_{sche} =1) at x:00 ($x \in (1, 2, \cdots, 23)$). For consistency with the power control scheduling, we adopted the maximum value of four kinds of predicted values at x:00, x:15, x:30, and x:45 as the predicted value at each time.

vi As we explained in Sect. 1, we targeted LEO, HAPS, and balloon as NTN lines and assumed that LEO and HAPS were equipped with the same type of O-RAN BSs as that in TN. In this study, we assume that UEs can access one LEO, one HAPS, five medium-sized balloons, and five small-sized balloons in each of Koganei City and Shinjuku Ward. We set the power consumption of the O-RAN BSs of LEO and HAPS at 935 W per BS, respectively, as in the case of TN. We also set the power consumption of BSs in medium-sized balloons and small-sized balloons as 120 W and 60 W per BS, respectively [16]. In this way, we obtained the total power consumption of NTN as P_{NTN}=2770 (W).

vii We set the communication capacity of the LEO, the HAPS, the medium-sized balloons, and the small-sized

balloons as 200 Mbps, 1.0 Gbps, 10 Mbps per balloon, and 10 Mbps per balloon, respectively [15], [16]. In this way, we obtained the total communication capacity of NTN as 1.3 Gbps. Note that it may be unrealistic to use all the NTN capacity of 1.3 Gbps for the scenario in this study. Therefore, we evaluate the relative power consumption achieved by our proposed power management by changing the value of B_{NTN} in the range from 100 Mbps to 1.3 Gbps.

viii We do not consider the power consumed by the ML platform by the following reason: To train the ML model and predict the pedestrian flow in one weekday, we used the past datasets of 56 weekdays. Their scale is much smaller than that of a famous example of big data analytics such as SNS advertisement, online shopping, and weather data, and it took around 3.5 min to train the ML model, which is sufficiently shorter than one day. This means the power consumption of our ML platform is negligibly much smaller than the total amount of power consumption of CU, DU, and RU.

6.2.2 A Visualized Power-On/Off Scheduling

Figure 9 shows an example of the power control scheduling results in 42 mesh areas of Koganei City at 14:00 on September 30, 2022. The horizontal scale is the mesh area number, while the vertical scale is the result of power control scheduling, where the values of "1" and "0" correspond to power-on and power-off, respectively. Figures 8 and 9 indicate that the proposed method turns off the BSs in the mesh areas where the values of the pedestrian flow are small.

6.2.3 Relative Power Consumption

Figures 10 and 11 show the relative power consumptions $R(t)$ achieved by the proposed power management system in Koganei City and Shinjuku Ward, respectively. $R(t) = 1$ corresponds to the conventional method, which always turns on all BSs in TN without utilizing NTN. The evaluation results show that our proposed power management method, utilizing the LSTM-supported pedestrian flow prediction results and NTN capacities, can reduce power consumption compared with the conventional method. Figures 10 and 11 indicate that the effect of power reduction increases as the NTN capacity increases. This is because the larger capacity of NTN can accommodate more UEs' traffic and, consequently, can turn off more BSs.

The effect of power reduction in the proposed method varies depending on the target city (or ward). As the values of pedestrian flow are larger over the entire city (or ward), the number of mesh areas that can turn off BSs is smaller. Figures 10 and 11 indicate that the effect of power reduction in Shinjuku Ward was smaller than that in Koganei City. For example, if the target value of relative power consumption is 0.6 (i.e., a power reduction of 40%), it can be achieved when

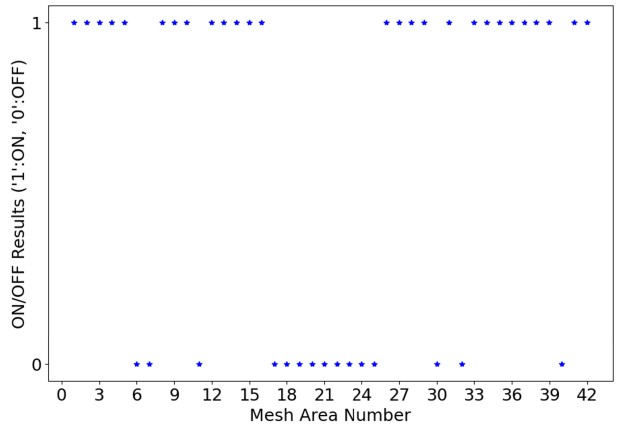


Fig. 9 power-on/off scheduling result in 42 mesh areas of Koganei City at 14:00, Sept. 30, 2022.

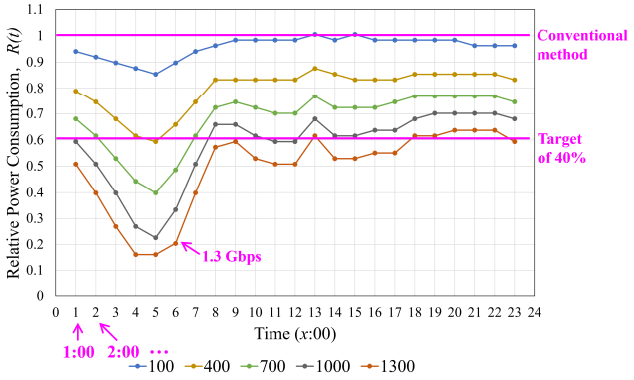


Fig. 10 Relative power consumption in Koganei City.

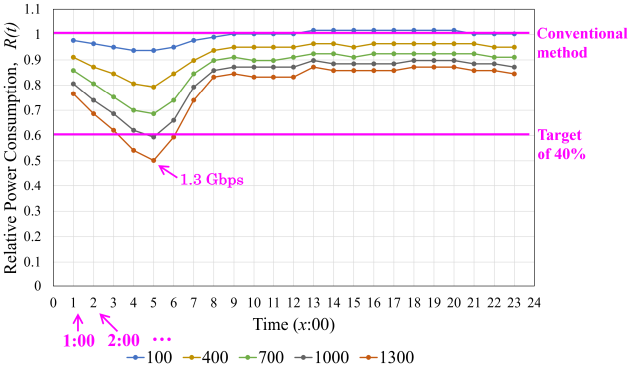


Fig. 11 Relative power consumption in Shinjuku Ward.

the NTN can accommodate UEs' traffic of 400 Mbps and 1 Gbps in Koganei City and Shinjuku Ward, respectively. In Shinjuku Ward, a heavily crowded place, the relative power consumption was more than 0.8 in the daytime, even when the NTN capacity was 1.3 Gbps.

Thus, for the B5G/6G era, it is necessary to increase the NTN capacity to further reduce the power consumption of O-RAN BSs in TN.

Table 4 Suitability of power control scheduling (Koganei City).

	(a) decided as ON, but OFF was correct	(b) decided as OFF, but ON was correct	(c) the decision was correct
AVS	40	51	875
LS-SVR	40	52	874
LSTM	14	15	937

Table 5 Suitability of power control scheduling (Shinjuku Ward).

	(a) decided as ON, but OFF was correct	(b) decided as OFF, but ON was correct	(c) the decision was correct
AVS	20	46	1521
LS-SVR	20	45	1522
LSTM	8	18	1561

6.2.4 Suitability of Power Control Scheduling

In this study, we set the NTN capacity to 500 Mbps in Koganei City and Shinjuku Ward. Tables 4 and 5 present the evaluation results of the suitability of the power control scheduling for the three methods: AVS, LS-SVR, and LSTM. Hereafter, we define "correct result" as the power control scheduling result using the measured values in the test data (i.e., the data on September 30, 2022). We evaluated the occurrence frequencies for three cases: (a) the correct result was OFF, although the prediction-based scheduling result was ON, (b) the correct result was ON, although the prediction-based scheduling result was OFF, and (c) the prediction-based scheduling result was the same as the correct result. Note that because times $t = 1{:}00, 2{:}00, \cdots, 23{:}00$, the sum of the frequencies in all cases is equal to $N \times 23$.

These results show that AVS and LS-SVR exhibit almost the same performance in all cases. This is because the pedestrian flow prediction accuracies of these two algorithms did not differ significantly, as shown in Table 3. These results show that LSTM outperformed the other two algorithms in terms of the suitability of power control scheduling. This is because LSTM enables more accurate power scheduling than the other two algorithms owing to the drastic reduction in prediction errors, as shown in Table 3.

In particular, if the frequency of (b) is high, the decision error affects the communication quality of UEs in the corresponding mesh areas. Because the NTN has limited capacity and may experience a higher delay and instability than TN, it is crucial to suppress the occurrence of case (b) as much as possible.

Table 6 shows the total number of UEs affected by case (b), which indicates the sum of the measured values of pedestrian flow in all combinations of the mesh area and time (on September 30, 2022) in which case (b) occurred. The results show that LSTM reduces the number of affected UEs drastically compared to the other two algorithms in both Koganei City and Shinjuku Ward. Thus, we conclude that LSTM is the optimal algorithm for pedestrian flow prediction to realize the correct scheduling of power-on/off control. We expect that a further reduction in pedestrian flow prediction will improve the suitability of power control scheduling.

Table 6 Total number of UEs affected by the case (b).

Algorithm	Koganei City	Shinjuku Ward
AVS	115,026	227,989
LS-SVR	118,413	222,484
LSTM	30,643	138,493

7. Field Experiments Using the NICT B5G Mobile Testbed

The experimental system illustrated in Fig. 5 did not include the monitoring systems described in Fig. 2. This is because, we used the estimated amount of traffic in each coverage area as explained in Sect. 6.2.1-ii, and thus, the implementation and performance evaluation of the monitoring systems themselves are out of scope in this study. Besides, the experimental system did not include the real NTN equipment and its link interfaces described in Fig. 2. This is because, as explained in Sect. 6.2.1-vi/vii, we used the estimated values of power consumption and communication capacities of NTN in this study to quantitatively evaluate the advantageous effect of our proposed method.

Figure 12 shows the captured images of the experiments when automatically executing the power control processes, in which the timing of power-on/off was prespecified. Cases (a) and (b) correspond to power-on and power-off states, respectively. Here, we define "ON process" as the process to start the radio-frequency radiation from the RU automatically by turning on the power of CU and DU and starting their application processes. We also define "OFF process" as the process to stop the radio-frequency radiation from the RU automatically by stopping the application processes of CU and DU and turning off their power. On the control plane, the Non-RT RIC sends a request message to the Near-RT RIC to execute the ON process at 11:00 as prespecified by the power control scheduling. Once the request message is received, the Near-RT RIC executes the ON process. One hour later, at 12:00, prespecified by the power-control scheduling, the Non-RT RIC sent a request message to the Near-RT RIC to execute the OFF process. When the request message is received, the Near-RT RIC executes an OFF process. As shown in Fig. 12, we confirmed that after the Non-RT RIC performed power-control scheduling based ON and OFF pedestrian flow prediction, both the ON and OFF processes were successfully demonstrated at the prespecified times in accordance with the scheduling results.

Figure 13 shows the power consumption of the CU and DU versus time, which indicates the time required for the ON process. We conducted this measurement at a different time from that shown in Fig. 12. The Near-RT RIC started the ON process at 13:49, and we confirmed that the radio-frequency radiation and data transfer restarted at 14:08. This implies that the ON process requires approximately 19 min on a

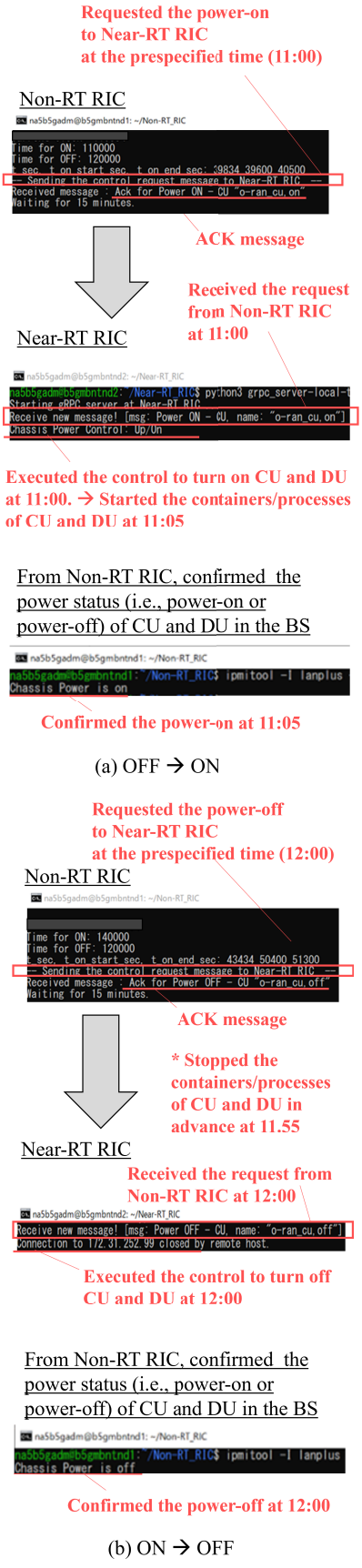


Fig. 12 Captured images of power control processes: (a) OFF to ON, (b) ON to OFF.

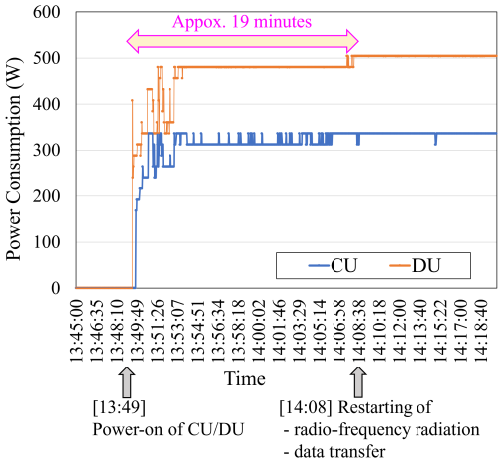


Fig. 13 Power consumption of CU and DU versus time, which indicates the time required for the ON process.

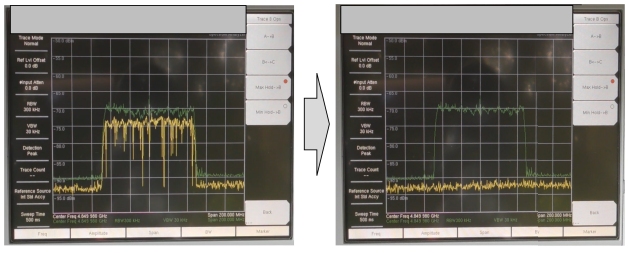


Fig. 14 Pictured images of frequency spectrum displayed on the oscilloscope when iperf UDP 1.0 Gbps data transfer was executed.

mobile testbed [17], [18]. This is acceptable in daily life because the power-on/off scheduling was executed every hour in this study. Meanwhile, it may be necessary to shorten the time required for the ON process in some emergency cases, such as sudden heavy rain, traffic accidents, and equipment failures. In such cases, we need to investigate how to reduce the time required for the ON process while maintaining the stability of the CU, DU, and RU equipment in the mobile testbed. However, it was confirmed that the OFF process was completed within a few seconds.

It took less than 1 s of overhead time to complete the message transfers between Non-RT RIC, Near-RT RIC and CU/DU in the control sequence shown in Fig. 6, which was negligibly much smaller than the time required for the ON process itself and did not adversely affect the proposed power management.

The Near-RT RIC executes the shell script command to activate each component in order of CU and DU. Since the CU and DU are deployed in different physical servers from each other, their activation processes are executed in parallel. The ON process of CU completed earlier by approximately 2 min than the ON process of DU.

Figure 14 shows photographs of the frequency spectrum with a 100 MHz width and a central frequency of 4.85 GHz, which were displayed on the oscilloscope. We confirmed that the radio wave signal power was increased by the ON pro-

cess, and then the iperf UDP 1.0 Gbps data transfer started. The throughput was 866 Mbps with full use of the 100 MHz frequency bandwidth. We also confirmed that the radio-wave signal power decreased and became almost zero by the OFF process, after which the iperf UDP 1.0 Gbps data transfer stopped.

8. Conclusion

We proposed, designed, and experimentally demonstrated energy-efficient power management for O-RAN BSs by utilizing LSTM-supported pedestrian flow prediction results and NTN capacities. For pedestrian flow prediction and power control scheduling, we evaluated the performance using real datasets of pedestrian flow (i.e., KLD) in Koganei City and Shinjuku Ward in Tokyo, Japan. We proved that LSTM outperforms AVS and LS-SVR in terms of prediction accuracy and consequently enhances the suitability of power control scheduling. In addition, we evaluated the relative power consumption with various values of the total capacity of NTN; for example, we proved that the power consumption can be reduced by up to 40% in Koganei City if the NTN can accommodate UEs' traffic of approximately 400 Mbps. In addition, we implemented the pedestrian flow analytics and power control scheduling in the Non-RT RIC, implemented the power-on/off control processes in the Near-RT RIC, and connected the controllers to the mobile testbed called "NICT B5G Development Environment." We experimentally demonstrated the on/off power control of the O-RAN BS and verified the feasibility of the control processes by monitoring the radio wave frequency spectrum displayed on an oscilloscope.

In future work, we plan to upgrade the functions of the pedestrian flow analytics and power control scheduling to further reduce the power consumption of O-RAN BSs, improve the suitability of power control scheduling, and extend the experimental system to enable larger-scale experiments using a mobile testbed. In addition, we plan to integrate our proposed management functions of Non-RT RIC with other technologies: (1) other data analytics such as weather prediction, equipment failure detection and prediction, and social networking services (SNS) information detection; and (2) networking technologies such as routing, traffic engineering, and resource allocation.

Acknowledgments

The authors would like to thank Yasunaga Kobari, Renri Nakano, and Kazuhiko Nakamura for their technical support and everyone who provided valuable comments on this work.

References

[1] Information & Communications Statistics Database, www.soumu.go.jp/johotsusintokei/english/
[2] NICT Beyond 5G/6G Whitepaper, beyond5g.nict.go.jp/en/download/index.html
[3] O-RAN Alliance, www.o-ran.org/
[4] 3GPP TS 28.310 v18.1.0 (March 2023), www.3gpp.org/ftp/Specs/archive/28_series/28.310/
[5] 3GPP TR 37.817 v17.0.0 (April 2022), www.3gpp.org/ftp/Specs/archive/37_series/37.817/
[6] KDDI Integrated Sustainability and Financial Report 2022, www.kddi.com/extlib/files/english/corporate/ir/ir-library/sustainability-integrated-report/pdf/kddi_sir2022_e.pdf
[7] K. Liolis, A. Geurtz, R. Sperber, D. Schulz, S. Watts, G. Poziopoulou, B. Evans, N. Wang, O. Vidal, B.T. Jou, M. Fitch, S.D. Sendra, P.S. Khodashenas, and N. Chuberre, "Use cases and scenarios of 5G integrated networks for enhanced mobile broadband: The SaT5G approach," International Journal of Satellite Communications and Networking, vol.21, no.4, pp.1–22, Nov. 2017.
[8] X. Zhu and C. Jiang, "Integrated satellite-terrestrial networks toward 6G: Architectures, applications and challenges," IEEE Internet Things J., vol.9, no.1, pp.437–461, Nov. 2022.
[9] V.P. Kafle, M. Sekiguchi, H. Asaeda, and H. Harai, "Integrated network control architecture for terrestrial and non-terrestrial network convergence in beyond 5G systems," 2022 ITU Kaleidoscope, Dec. 2022.
[10] Draft new Recommendation ITU-T Y.FMSC-INCA: "Fixed, mobile and satellite convergence - Integrated network control architecture framework for IMT-2020 networks and beyond," 20th March. 2023. www.itu.int/md/T22-SG13-230313-TD-WP1-0309/en
[11] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol.9, no.8, pp.1735–1780, Nov. 1997.
[12] S. Zeb, M.A. Rathore, A. Mahmood, S.A. Hassan, J. Kim, and M. Gidlund, "Edge intelligence in softwarized 6G: Deep learning-enabled network traffic predictions," Proc. IEEE GLOECOM Workshop (NETMAN6G), Madrid, Spain, Dec. 2021.
[13] N. Zhang, Y. Su, B. Wu, X. Tu, Y. Jin, and X. Bao, "Cloud resource prediction model based on LSTM and RBF," Proc. International Conference on Big Data and Artificial Intelligence and Software Engineering (ICBASE 2021), pp.189–194, Zhuhai, China, Sept. 2021.
[14] H. Tsuji, A. Miura, D.R. Kolev, and M. Toyoshima, "Future vision of non-terrestrial networks in beyond 5G/6G era," J. IEICE, vol.106, no.5, pp.358–363, May 2023.
[15] HAPS, www.softbank.jp/en/corp/philosophy/technology/special/ntn-solution/haps/
[16] ComBalloon. www.nittocom.co.jp/case/case-legasip/ (in Japanese).
[17] H. Harai, K. Nakamura, K. Terada, R. Nagareda, T. Kato, H. Nagano, and F. Kojima, "Deployment of circular evolutionary testbed toward beyond 5G mobile network," IEICE Technical Report, NS2022-63, Sept. 2022 (in Japanese).
[18] "Beyond 5G/IoT Testbed with High-reliability and High-elasticity, Beyond 5G Mobile Environment, a Mobile Network Development Environment," testbed.nict.go.jp/english/b5gm/
[19] N. Li, J. Martinez-Ortega, and V.H. Diaz, "Distributed power control for interference-aware multi-user mobile edge computing a game theory approach," IEEE Access, vol.6, pp.36105–36114, June 2018.
[20] Y. Zhang, Y. Zhu, W. Xia, F. Shen, X. Zuo, and F. Yan, "Game-based power control for downlink non-orthogonal multiple access in HetNets," Proc. IEEE Global Communications Conference (GLOBECOM) Dec. 2018.
[21] P. Gorla, A. Deshmukh, S. Joshi, V. Chamola, and M. Guizani, "A game theoretic analysis for power management and cost optimization of green base stations in 5G and beyond communication networks," IEEE Trans. Netw. Serv. Manag., vol.19, no.3, pp.2714–2725, Sept. 2022.
[22] L. Xiao, Y. Li, C. Dai, H. Dai, and H.V. Poor, "Reinforcement learning-based NOMA power allocation in the presence of smart jamming," IEEE Trans. Veh. Technol., vol.67, no.4, pp.3377–3389, April 2018.
[23] J. Ye and Y.A. Zhang, "DRAG: Deep reinforcement learning based base station activation in heterogeneous networks," IEEE Trans. Mobile Comput., vol.19, no.9, pp.2076–2087, Sept. 2020.
[24] Q. Wu, X. Chen, Z. Zhou, L. Chen, and J. Zhang, "Deep reinforce-

ment learning with spatio-temporal traffic forecasting for data-driven base station sleep control," IEEE/ACM Trans. Netw., vol.29, no.2, pp.935–948, April 2021.
[25] X. Tan, K. Xiong, B. Gao, P. Fan, and K.B. Letaief, "Energy-efficient base station switching-off with guaranteed cooperative profit gain of mobile network operators," IEEE Trans. Green Commun. Netw., vol.7, no.3, pp.1250–1266, Sept. 2023.
[26] O. Orhan, V.N. Swamy, T. Tetzlaff, M. Nassar, H. Nikopour, S. Talwar, "Connection management xAPP for O-RAN RIC: A graph neural network and reinforcement learning approach," Proc. IEEE International Conference on Machine Learning and Applications (ICMLA), Dec. 2021.
[27] Z. Ali, L. Giupponi, M. Miozzo, and P. Dini, "Multi-task learning for efficient management of beyond 5G radio access network architectures," IEEE Access, vol.9, pp.158892–158907, Nov. 2021.
[28] Y. Cao, S. Lien, Y. Liang, K. Chen, and X. Shen, "User access control in open radio access networks: A federated deep reinforcement learning approach," IEEE Trans. Wireless Commun., vol.21, no.6, pp.3721–3736, June 2022.
[29] S. D'Oro, L. Bonati, M. Polese, and T. Melodia, "OrchestRAN: Network automation through orchestrated intelligence in the open RAN," Proc. IEEE Conference on Computer Communications (INFOCOM), May 2022.
[30] P.H. Masur, J.H. Reed, and N.K. Tripathi, "Artificial intelligence in open-radio access network," IEEE Aerosp. Electron. Syst. Mag., vol.37, no.9, pp.6–15, Sept. 2022.
[31] A. Giannopoulos, S. Stantideas, N. Kapsalis, P. Gkonis, L. Sarakis, C. Capsalis, M. Vecchio, and P. Trakadas, "Supporting intelligence in disaggregated open radio access networks: Architectural principles, AI/ML workflow, and use cases," IEEE Access, vol.10, pp.39580–39595, April 2022.
[32] Plugfest, plugfestvirtualshowcase.o-ran.org/2023/SPRING
[33] O-RAN Alliance Specifications, "O-RAN Architecture Description 11.0 (Release R003)," Feb. 2024.
[34] O-RAN Alliance Specifications, "O-RAN Non-RT RIC Architecture 5.0 (Release R003)," Feb. 2024.
[35] O-RAN Alliance Specifications, "O-RAN Near-RT RIC Architecture 5.0 (Release R003)," Oct. 2023.
[36] O-RAN Alliance Specifications, "O-RAN A1 interface: General Aspects and Principles 3.02 (Release R003)," Feb. 2024.
[37] O-RAN Alliance Specifications, "O-RAN E2 General Aspects and Principles (E2GAP) 5.0 (Release R003)," Feb. 2024.
[38] O-RAN Alliance Specifications, "O-RAN Transport Specification 1.0," April 2020.
[39] O-RAN Alliance Specifications, "O-RAN O1 Interface Specification for Near Real Time RAN Intelligent Controller 1.0 (Release R003)," March 2023.
[40] O-RAN Alliance Specifications, "O-RAN Management Plane Specification 14.0 (Release R003)," Feb. 2024.
[41] G. Bathla, "Stock price prediction using LSTM and SVR," International Conference on Parallel, Distributed and Grid Computing (PDGC), pp.211–214, Nov. 2020.
[42] N. Nagar, P.K. Jatav, M. Gupta, and A. Limone, "Comparison of LSTM and SVR models in predicting stock prices," Journal of Harbin Engineering University, vol.44, no.7, pp.1–5, July 2023.
[43] H. Jiang, Y. Li, C. Zhou, H. Hong, T. Glade, and K. Yin, "Landslide displacement prediction combining LSTM and SVR algorithms: A case study of Shengjibao landslide from the three gorges reservoir area," Applied Sciences, vol.10, no.21, pp.1–21, Nov. 2020.
[44] G. Wang, Z. Li, G. Li, G. Dai, Q. Xiao, L. Bai, Y. He, Y. Liu, and S. Bai, "Real-time liver tracking algorithm based on LSTM and SVR networks for use in surface-guided radiation therapy," Radiat. Oncol., vol.16, no.13, pp.1–12, Jan. 2021.
[45] H. Wang and D. Hu, "Comparison of SVM and LS-SVM for regression," Proc. International Conference on Neural Networks and Brain (ICNNB), Beijing, China, pp.279–283, Oct. 2005.
[46] H. Drucker, C. Burges, L. Kaufman, A. Smola, and V. Vapnik, "Support vector regression machines," Advances in Neural Information Processing Systems 9, NIPS 1996, pp.155–161, MIT Press, Dec. 1996.
[47] X. Li, L. Yu, L. Tang, and W. Dai, "Coupling firefly algorithm and least squares support vector regression for crude oil price forecasting," Proc. IEEE BIFE, pp.80–83, Hangzhou, China, Nov. 2013.
[48] R. Andrae, "Least-squares support vector regression for GSP-Phot," Gaia Data processing and Analysis Consortium (DPAC), no.1, pp.1–13, Feb. 2016.
[49] T. Miyazawa, Y. Yokota, V.P. Kafle, Y. Naruse, and H. Asaeda, "Advanced data analytics using three-stage intelligent model pipelining for containerized microservices in 5G networks and beyond," ITU Journal on Future and Evolving Technologies, vol.4, no.2, pp.285–305, June 2023.
[50] J.P. Vert, K. Tsuda, and B. Scholkopf, "A primer on kernel methods," Kernel Methods in Computational Biology, pp.1–42, MIT Press July 2004.
[51] "JGN: High Speed R&D Network Testbed," testbed.nict.go.jp/jgn/english/index.html
[52] free5GC, free5gc.org/
[53] Radisys: www.radisys.com/solutions/openran
[54] KDDI Location Data, k-locationdata.kddi.com/

Takaya Miyazawa received an M.E. degree and a Ph.D. degree in information and computer science from Keio University, Yokohama, Japan, in 2004 and 2006, respectively. From April 2006 to March 2007, he was a visiting researcher at the University of California in Davis, CA, USA. He joined the National Institute of Information and Communications Technology (NICT) as a researcher in 2007, where he is currently a Research Manager. From April 2019 to July 2020, he worked at the Ministry of Internal Affairs and Communications, Japan, as a Deputy Director, and returned to NICT in August 2020. His research interests include network controls and management. He received the Hiroshi Ando Memorial Young Engineer Award in 2007, Funai Young Researcher Award in 2010, and Best Paper Award from the ITU Kaleidoscope Academic Conference in 2018. He is a member of IEEE.

Kentaro Ishizu received the Ph.D. degree in computer science from Kyushu University, Fukuoka, Japan, in 2005. Since then, he has been working with the National Institute of Information and Communications Technology (NICT) in Tokyo, Japan, and is dedicated to research on cognitive radio systems, TV white space systems, and local 5G. He has been involved in international standardizations since 2002, such as IEEE802, ETSI, 3GPP, and IETF. From April 2021. He is directing the Beyond 5G design initiative of NICT. He is a member of IEEE.

Hitoshi Asaeda is a Director of the Network Architecture Laboratory, National Institute of Information and Communications Technology (NICT) and also a Collaborative Professor with the Graduate School of Informatics and Engineering, the University of Electro-Communications (UEC). He holds a Ph.D. degree from Keio University. He was previously with IBM Japan, Ltd. and a Research Engineer Specialist at INRIA Sophia Antipolis, France. He was a Project Associate Professor at Keio University from 2005 to 2012. He was a Guest Editor-in-Chief of the special series of IEICE Trans. Commun. in 2016. He was a Chair of the IEICE Technical Committee on ICN from 2017 to 2019. He served as a General Chair of IEEE/ACM IWQoS 2021 and ACM ICN 2022 and has been a TPC member for premier conferences such as IEEE INFOCOM, WCNC, and ACM ICN. He was a Program Officer for several international projects and has been actively working in the IETF standards body. He received the IEICE Communications Society Outstanding Contributions Award in 2019. His research interests include ICN, network coding, high-quality streaming, and large-scale testbeds. He is a Senior Member of the IEEE and a Member of the ACM.

Hiroyuki Tsuji received the D.E. degree in electrical engineering from Keio University, Tokyo, Japan, in 1992. In 1992, he joined the Communications Research Laboratory (CRL), which is now part of the National Institute of Information and Communications Technology (NICT) in Tokyo. From 1999–2000, he was a Visiting Research Fellow of the University of Minnesota, Minneapolis, MN, USA. From 2002 to 2021, he was working at Yokohama National University in Yokohama, Japan as a visiting professor, which was an additional post. His research interests include array antennas for wireless communications, unmanned aircraft communication systems, and other satellite communication systems. He is currently the Director of the Space Communication Systems Laboratory of Wireless Networks Research Center, NICT, since April 2021. He is a member of IEEE.

Hiroaki Harai received his M.E. and Ph.D. degrees in information and computer science from Osaka University, Japan, in 1995 and 1998, respectively. He is currently the Director General of the Network Research Institute of NICT. His research interests include novel network architectures and optical networks. He received the Outstanding Young Researcher Award from the IEEE Communications Society Asia-Pacific Region in 2007 and the Young Researcher Award from the Ministry of Education, Culture, Sports, Science, and Technology, Japan, in 2009. He is a member of IEEE.

IEICE TRANS. COMMUN., VOL.E107–B, NO.11 NOVEMBER 2024

PAPER

DDMA-MIMO/Capon Observations Using the MU Radar: Beamwidth Verification Using the Moon's Reflection

Tomoya MATSUDA[†a)], ***Nonmember***, **Koji NISHIMURA**[††], ***and*** **Hiroyuki HASHIGUCHI**[††], ***Members***

SUMMARY Phased-array technology is primarily employed in atmospheric and wind profiling radars for meteorological remote sensing. As a novel avenue of advancement in phased-array technology, the Multiple-Input Multiple-Output (MIMO) technique, originally developed for communication systems, has been applied to radar systems. A MIMO radar system can be used to create a virtual receive antenna aperture plane with transmission freedom. The MIMO technique requires orthogonal waveforms on each transmitter to identify the transmit signals using multiple receivers; various methods have been developed to realize the orthogonality. In this study, we focus on the Doppler Division Multiple Access (DDMA) MIMO technique by using slightly different frequencies for the transmit waveforms, which can be separated by different receivers in the Doppler frequency domain. The Middle and Upper atmosphere (MU) radar is a VHF-band phased array atmospheric radar with multi-channel receivers. Additional configurations are necessary, requiring the inclusion of multi-channel transmitters to enable its operation as a MIMO radar. In this study, a comparison between the brightness distribution of the beamformer, utilizing echoes reflected from the moon, and the antenna pattern obtained through calculations revealed a high degree of consistency, which means that the MU radar functions effectively as a MIMO radar. Furthermore, it is demonstrated that the simultaneous application of MIMO and Capon techniques has a mutually enhancing effect.

key words: *atmospheric radar, Capon beamformer, Doppler Division Multiple Access, phased array, MIMO*

1. Introduction

The phased-array radar technology, originally developed as a defense radar, has been primarily utilized for atmospheric and wind profiling radars. It has also been employed as a weather radar for research purposes. As a further development of the phased-array technology, the Multiple-Input Multiple-Output (MIMO) technique, which was originally developed for communication systems, has been applied to radars, making a new contribution for radar signal processing [1].

More recently, various phased-array applications that perform Digital Beam Forming (DBF) with multiple receivers have been developed. DBF provides multiple receive beams in a single scan, which dramatically reduces the scan time. However, conventional phased array radars cannot distinguish transmit signals at the receivers. Therefore, they are categorized as Single-Input Multiple-Output (SIMO) radars.

Manuscript received November 18, 2023.
Manuscript revised March 9, 2024.
Manuscript publicized May 6, 2024.

[†]Mitsubishi Electric Corporation, Amagasaki-shi, 661-8661 Japan.
[††]Kyoto University, Uji-shi, 611-0011 Japan.
a) E-mail: matsuda.tomoya.b30@kyoto-u.jp

DOI: 10.23919/transcom.2023EBP3182

The MU radar [2], [3], which has been operational for four decades years, is one of the most advanced atmospheric radars with 475 transmitters, phase shifters, and corresponding antennas to orient the beam direction electrically. The MU radar has mostly been operated as a SIMO radar, although it can also be operated as a MIMO radar with additional settings.

The orthogonality of the transmit signals is the most definitive difference between SIMO and MIMO radars. Phased-array radars that separate transmit waves can be referred to as multiple-input transmitters, and understanding the methodology of separating transmit waves from the receive signals. The transmit waves can be separated by adopting orthogonal waveforms for each transmitter. Some methods to achieve their orthogonality were introduced in [4]. It is necessary to select an appropriate method according to the application and transmit frequency characteristics.

A MIMO approach for atmospheric radars was applied in [5], [6] for atmospheric and ionospheric synthesis radar imaging observation, attempting to use Time Division Multiple Access (TDMA), Code Division Multiple Access (CDMA), and polarization diversity. We used the Doppler Division Multiple Access (DDMA) method, which utilizes a slow-time direction to obtain the orthogonal transmit waveforms [4], [7], [8].

The effectiveness of DDMA-MIMO radar was introduced in [9], demonstrating the beam broadening effect during tropospheric observation. However, further analysis of improvements, including sidelobe suppression effects, is challenging owing to the nonuniform volume targets, necessitating the use of clear hard targets. In this study, beamwidth verification was performed using the moon compared with the calculated antenna pattern, which satisfies this condition. By utilizing the moon reflection echoes, we expect that various applications of the MIMO radar can be verified, and further combinations of multibeam and/or advanced beamforming techniques will be applied through this validation.

In general, few approaches exist to confirm the beamwidth directly. However, using the moon's reflection echo, which has been examined with the MU radar, could be one method to verify the beamwidth [10]. The observations with the moon reflection echoes and results are presented after discussions on the signal model of the MIMO radar and revised system of the MU radar. Furthermore, the combination of the MIMO virtual antenna and adaptive beamforming technique is expected to extract better performance, which was introduced in [11]. The Capon beamformer technique

was used with the MU radar for two-dimensional generalization of the brightness distribution in [12]; therefore, it is natural extension to confirm this combination effect.

This paper describes DDMA-MIMO observations using the MU radar by comparing it with other methods and discussing the signal model of the MIMO radar. The fundamental principle of the MIMO radar and adaptive beamforming methods are presented in Sect. 2. Four major methods to ensure the orthogonality of the transmit signals from the MIMO radar and the reasons for selecting the DDMA-MIMO for the MU radar are presented in Sect. 3. The system configuration of the MU radar as a MIMO radar and the observation results using the reflection echoes off the moon are presented in Sect. 4. Finally, the effectiveness of the MIMO radar is presented in Sect. 5.

2. Basic Theory

2.1 Basic Principle of the MIMO Radar

The MIMO technique has a long history. The wireless communications community has studied the characteristics and characterizations of the MIMO radar. In this section, the basic principle is introduced, as summarized in [1], [13]–[15].

Let there be M transmit signals. Let the m-th transmit signal be $x_m(t,\theta_0) = a_m(\theta_0)\phi_m(t)$, and let the n-th receive signal $y_n(t,\theta_0)$ be defined as

$$\begin{aligned} y_n(t,\theta_0) &= \alpha b_n(\theta_0)\sum_{m=1}^{M} a_m(\theta_0)\phi_m(t) + v_n(t) \\ &= \alpha b_n(\theta_0)\mathbf{a}(\theta_0)^{\mathrm{T}}\boldsymbol{\phi}(t) + v_n(t), \end{aligned} \tag{1}$$

where $v_n(t)$ represents the receive noise, $a_m(\theta_0)$ and $b_n(\theta_0)$ represent the transmit and receive phase shifts corresponding to the transmit target angle θ_0, $\boldsymbol{\phi}(t) \in \mathbb{C}^M$ is a normalized transmit waveform column vector composed of M transmitters, $\mathbf{a}(\theta_0) \in \mathbb{C}^M$ is a transmit steering column vector corresponding to the transmission angle θ_0, and α is a (complex-valued) backscatter coefficient.

To expand $y_n(t,\theta_0)$ to N receivers, the receive signal column vector $\mathbf{y}(t,\theta_0) \in \mathbb{C}^N$ is defined as

$$\mathbf{y}(t,\theta_0) = \alpha\mathbf{b}(\theta_0)\mathbf{a}(\theta_0)^{\mathrm{T}}\boldsymbol{\phi}(t) + \mathbf{v}(t), \tag{2}$$

where $\mathbf{b}(\theta_0) \in \mathbb{C}^N$ is the receive steering column vector that corresponds to the receive angle θ_0 (here, the transmit and receive angles are defined to be the same), and $\mathbf{v}(t) \in \mathbb{C}^N$ is the receive noise-column vector. Notably, $\mathbf{b}(\theta_0)\mathbf{a}(\theta_0)^{\mathrm{T}}$ represents an $N \times M$ matrix, that is, $\mathbf{b}(\theta_0)\mathbf{a}(\theta_0)^{\mathrm{T}} \in \mathbb{C}^{(N,M)}$.

Following range processing with time lag τ and matched filters $\boldsymbol{\phi}(t-\tau)^{\mathrm{H}}$ (the suffix H indicates a Hermitian transpose) to separate the transmit waveforms. The receive signal matrix $\mathbf{Z}(\tau,\theta_0) \in \mathbb{C}^{(N,M)}$ is expressed as

$$\begin{aligned} \mathbf{Z}(\tau,\theta_0) &\equiv \int_{-\infty}^{\infty} \mathbf{y}(t,\theta_0)\boldsymbol{\phi}(t-\tau)^{\mathrm{H}}\mathrm{d}t \\ &= \alpha\mathbf{b}(\theta_0)\mathbf{a}(\theta_0)^{\mathrm{T}}\int_{-\infty}^{\infty} \boldsymbol{\phi}(t)\boldsymbol{\phi}(t-\tau)^{\mathrm{H}}\mathrm{d}t \\ &\quad + \int_{-\infty}^{\infty} \mathbf{v}(t)\boldsymbol{\phi}(t-\tau)^{\mathrm{H}}\mathrm{d}t \\ &= \alpha\mathbf{b}(\theta_0)\mathbf{a}(\theta_0)^{\mathrm{T}}\mathbf{R}_{\phi}(\tau) + \mathbf{E}(\tau), \end{aligned} \tag{3}$$

where

$$\mathbf{R}_{\phi}(\tau) \equiv \int_{-\infty}^{\infty} \boldsymbol{\phi}(t)\boldsymbol{\phi}(t-\tau)^{\mathrm{H}}\mathrm{d}t \quad \in \mathbb{C}^{(M,M)} \tag{4}$$

represents the $M \times M$ MIMO signal correlation matrix that describes the correlation among the transmit waveforms, and

$$\mathbf{E}(\tau) \equiv \int_{-\infty}^{\infty} \mathbf{v}(t)\boldsymbol{\phi}(t-\tau)^{\mathrm{H}}\mathrm{d}t \quad \in \mathbb{C}^{(N,M)} \tag{5}$$

represents the filtered receive noise matrix.

The $N \times M$ data matrix expressed in (3) can be vectorized by stacking the columns of $\mathbf{Z}(\tau,\theta_0)$, and we define the receive MIMO signal as

$$\mathbf{z}(\tau,\theta_0) \equiv \mathrm{vec}[\mathbf{Z}(\tau,\theta_0)] \quad \in \mathbb{C}^{(NM,1)}. \tag{6}$$

Equation (6) is rewritten using the well-known relationships of the vectorization operator shown in [16]:

$$\begin{aligned} \mathrm{vec}\{\mathbf{XYZ}\} &= \{\mathbf{Z}^{\mathrm{T}} \otimes \mathbf{X}\}\mathrm{vec}\{\mathbf{Y}\}, \\ \mathrm{vec}\{\mathbf{XYZ}\} &= \{\mathbf{Z}^{\mathrm{T}} \otimes \mathbf{I}_N\}\mathrm{vec}\{\mathbf{XY}\}, \\ \mathrm{vec}\{\mathbf{XY}^{\mathrm{T}}\} &= \{\mathbf{Y} \otimes \mathbf{X}\}, \end{aligned} \tag{7}$$

where $\mathbf{X}$ represents an arbitrary $N \times K$ matrix, $\mathbf{Y}$ represents an arbitrary $K \times L$ matrix, $\mathbf{Z}$ represents an arbitrary $L \times N$ matrix, $\mathbf{I}_N$ represents a unit matrix of $N \times N$, and "$\otimes$" symbolizes the Kroneker product,

$$\begin{aligned} \mathbf{z}(\tau,\theta_0) &= \mathrm{vec}[\alpha\mathbf{b}(\theta_0)\mathbf{a}(\theta_0)^{\mathrm{T}}\mathbf{R}_{\phi}(\tau)] + \mathrm{vec}[\mathbf{E}(\tau)] \\ &= \alpha[\mathbf{R}_{\phi}^{\mathrm{T}}(\tau) \otimes \mathbf{I}_N]\mathrm{vec}[\mathbf{b}(\theta_0)\mathbf{a}(\theta_0)^{\mathrm{T}}] + \mathrm{vec}[\mathbf{E}(\tau)] \\ &= \alpha[\mathbf{R}_{\phi}^{\mathrm{T}}(\tau) \otimes \mathbf{I}_N][\mathbf{a}(\theta_0) \otimes \mathbf{b}(\theta_0)] + \mathrm{vec}[\mathbf{E}(\tau)] \\ &\equiv \alpha\mathbf{s}(\tau,\theta_0) + \mathbf{e}(\tau), \end{aligned} \tag{8}$$

where the MIMO steering vector $\mathbf{s}(\tau,\theta_0)$ for beam angle θ_0 and filtered noise columns vector $\mathbf{e}(\tau)$ are defined as

$$\mathbf{s}(\tau,\theta_0) \equiv [\mathbf{R}_{\phi}^{\mathrm{T}}(\tau) \otimes \mathbf{I}_N][\mathbf{a}(\theta_0) \otimes \mathbf{b}(\theta_0)] \tag{9}$$

$$\mathbf{e}(\tau) \equiv \mathrm{vec}[\mathbf{E}(\tau)]. \tag{10}$$

If it is assumed that the transmit signals $\boldsymbol{\phi}(t)$ are orthogonal with each other ($\phi_m(t)$ and $\phi_{m'}(t)$ $(m \neq m')$ are zero-correlation) and that each matched filtered range response is identical (each transmitter has the identical transmit waveform) and defined as $R_{\phi}(\tau)$, then each element of the MIMO signal correlation matrix can be expressed as

$$\begin{aligned} \mathbf{R}_{\phi}(\tau)_{m,m'} &\equiv \int_{-\infty}^{\infty} \phi_m(t)\phi_{m'}^{*}(t-\tau)\mathrm{d}t \\ &= \begin{cases} R_{\phi}(\tau), & (\text{for } m = m') \\ 0, & (\text{for } m \neq m') \end{cases} \end{aligned} \tag{11}$$

and (8) can be rewritten as

$$\mathbf{z}(\tau,\theta_0) = \alpha R_{\phi}(\tau)[\mathbf{a}(\theta_0) \otimes \mathbf{b}(\theta_0)] + \mathbf{e}(\tau). \tag{12}$$

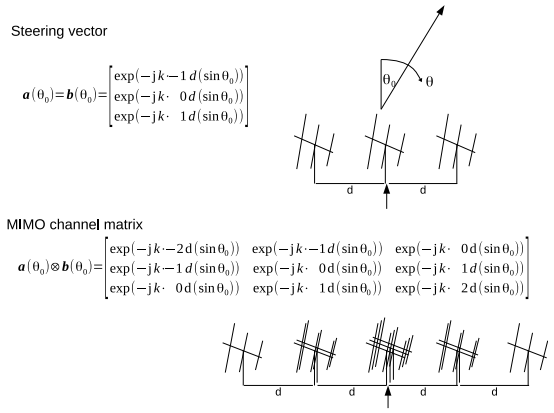


Fig. 1 Schematic of the MIMO radar for $M = 3$ transmitters and $N = 3$ receivers in one-dimension.

According to (12), the MIMO radar can be expanded to receive signal vectors using the MIMO channel matrix, which includes its transmit freedom, which can be expressed as $\mathbf{a}(\theta_0) \otimes \mathbf{b}(\theta_0) \in \mathbb{C}^{(NM,1)}$. Figure 1 shows a schematic of the MIMO radar with one-dimensional orthogonality of the transmit signals for $M = N = 3$.

2.2 SIMO and MIMO Adaptive Beamforming

The characteristics of the MIMO radar, which has a narrower receive and a wider transmit beam, are expected to work effectively for beamforming on a receiver referred to as the Capon beamformer, which generates angular brightness distribution. In this section, the Capon beamformer and its application to SIMO and MIMO radars are introduced.

In general, the output power $P_{BF}(\tau, \theta, \theta_0)$, obtained using the beamformer method at the range response time τ and the arrival angle θ, is expressed as [17], [18]

$$P_{BF}(\tau, \theta, \theta_0) = \frac{\mathbf{b}(\theta)^{\mathrm{H}} \mathbf{R}_z(\tau, \theta_0) \mathbf{b}(\theta)}{\mathbf{b}(\theta)^{\mathrm{H}} \mathbf{b}(\theta)} \tag{13}$$

and that obtained using the Capon beamformer, $P_{CP}(\tau, \theta, \theta_0)$, is [17]–[19]

$$P_{CP}(\tau, \theta, \theta_0) = \frac{1}{\mathbf{b}(\theta)^{\mathrm{H}} \mathbf{R}_z(\tau, \theta_0)^{-1} \mathbf{b}(\theta)}, \tag{14}$$

where $\mathbf{b}(\theta)$ represents the receive steering vector, and $\mathbf{R}_z(\tau, \theta_0)$ represents the covariance matrix of the receive signal vector $\mathbf{z}(\tau, \theta_0)$ [20], [21].

As shown earlier, MIMO radar processing is regarded as a virtual array of $M \times N$ elements. Therefore, the output power of the MIMO radar, obtained using the two methods, $P_{BF\text{–}MIMO}(\tau, \theta, \theta_0)$ and $P_{CP\text{–}MIMO}(\tau, \theta, \theta_0)$, are defined by replacing $\mathbf{b}(\theta)$ with $\mathbf{a}(\theta) \otimes \mathbf{b}(\theta)$ in (13) and (14), respectively [11],

$$P_{BF\text{–}MIMO}(\tau, \theta, \theta_0) = \frac{[\mathbf{a}(\theta) \otimes \mathbf{b}(\theta)]^{\mathrm{H}} \mathbf{R}_{z\text{–}MIMO}(\tau, \theta_0) [\mathbf{a}(\theta) \otimes \mathbf{b}(\theta)]}{[\mathbf{a}(\theta) \otimes \mathbf{b}(\theta)]^{\mathrm{H}} [\mathbf{a}(\theta) \otimes \mathbf{b}(\theta)]} \tag{15}$$

and

$$P_{CP\text{–}MIMO}(\tau, \theta, \theta_0) = \frac{1}{[\mathbf{a}(\theta) \otimes \mathbf{b}(\theta)]^{\mathrm{H}} \mathbf{R}_{z\text{–}MIMO}(\tau, \theta_0)^{-1} [\mathbf{a}(\theta) \otimes \mathbf{b}(\theta)]}, \tag{16}$$

where $\mathbf{R}_{z\text{–}MIMO}(\tau, \theta_0)$ represents the covariance matrix of the receive MIMO signal vector $\mathbf{z}(\tau, \theta_0)$ derived from (12) with the expression of a vector of expected values $E[\mathbf{x}]$ as

$$\mathbf{R}_{z\text{–}MIMO}(\tau, \theta_0) = E[\mathbf{z}(\tau, \theta_0) \mathbf{z}(\tau, \theta_0)^{\mathrm{H}}]. \tag{17}$$

3. Transmit Methods to Acquire Orthogonal Waveforms to be Determined for the MU Radar

Transmit/signal processing methods and system evaluation have attracted attention to ensure orthogonality of the transmit signal for the MIMO radar. In this section, four major methods introduced in [4] are discussed briefly, and the DDMA method is chosen for the MU radar.

3.1 Time Division Multiple Access (TDMA)

To guarantee transmit signal orthogonality by time separation, TDMA can be realized in a manner such that each transmit signal radiates at different times from different positions. The hardware and software systems used for this method are relatively simple, making the design of a radar system more easy. However, this method requires an adequate waiting time while other transmitters radiate; that is, it requires more dwell time. For the reasons mentioned earlier, the TDMA method requires a tolerance of the inter-pulse period times the number of MIMO transmitters (which also indicates duty ratio reduction), for the target identity, which is disadvantageous for atmospheric or weather radars. To overcome this effect, the staggered-TDMA method was introduced [4]. However, this method is restrictive, and it is only effective for low-frequency radars with continuous waves (CW).

3.2 Frequency Division Multiple Access (FDMA)

FDMA can be realized such that each transmit signal radiates at different frequencies in one time-series duration. The orthogonality of the transmitters guarantees that their signals can be extracted by a receiver using band-pass filters for each frequency, which would otherwise require certain frequency resources [22]. Therefore, the implementation cost of FDMA is relatively small. However, differences in the transmit frequencies can severely affect beamforming owing to the deterioration of the range sidelobes. To rectify this effect, FDMA using transmit frequencies that circulate toward slow-time is introduced, which also has certain limitations. To practically use atmospheric or weather radars, the influence of the range sidelobes should be reduced below an acceptable level.

Table 1 Comparison between orthogonal waveforms.

Method	Time division multiple access (TDMA)	Frequency division multiple access (FDMA)	Doppler division multiple access (DDMA)	Code division multiple access (CDMA)
Advantage	Good orthogonality	Good orthogonality	Good orthogonality	Approx. orthogonality Widely applicable
Judgment to be acceptable	Time loss Transmit power loss	High range sidelobe	Wide Doppler unambiguity	System costs, adequate correlation time
Add. approach for improvement	Staggered TDMA	Circulated FDMA	-	Slow-time CDMA CDMA with CCC

3.3 Doppler Division Multiple Access (DDMA)

DDMA can be realized using the principle of Doppler shift caused by the pulse-to-pulse phase difference. Each transmitter is set to its own initial phase per inter-pulse-period to generate a unique phase difference, which following transformation to the frequency division toward slow-time direction, divides the different frequency (Doppler) distribution. It has excellent transmit signal orthogonality, which makes it easier to configure the radar system.

DDMA can also achieve its objective using slightly different frequencies for the transmit waveforms to generate pulse-to-pulse phase differences [7], [23].

However, DDMA requires wide Doppler unambiguity to achieve transmit signal orthogonality. To satisfy this requirement, low transmit frequency and/or short inter-pulse period (short-range) radar systems are preferred. Therefore, a VHF radar such as the MU radar performs well, whereas a weather radar with a C-band or X-band must consider the trade-off between the observation range, maximum Nyquist velocity, and the number of orthogonal transmitters to apply this technique.

3.4 Code Division Multiple Access (CDMA)

CDMA can be realized such that orthogonal codes are used for the transmitters. The modulated transmit signals radiate to the target simultaneously and the returned signals are decoded by using the transmit codes for each signal. These codes are selected to be orthogonal to each other, such that the signals are completely separated by decoding their own codes in one receiver. This method has been widely employed, particularly in communication systems, for frequency efficiency, noise reduction, and high confidentiality.

Radar systems can apply CDMA to high range sidelobes but only in the fast-time direction. However, CDMA can easily overcome this disadvantage when slow-time direction is used. As one of the solutions, CDMA with complete complimentary codes (CCC) was proposed in [24] and [25], which has complementary codes to mitigate range sidelobes and eliminate all cross correlations in each code, such that their orthogonality to both fast-time and slow-time direction remains. Although CCC has development problems regarding the Doppler sidelobes for fast moving targets, atmospheric and weather radars can be applied owing to the relatively small the target velocity.

3.5 Optimal Method for the MU Radar

Four methods to realize the MIMO radar are introduced in previous subsections. Table 1 and Fig. 2 present comparisons between the orthogonal waveforms. Conventional SIMO radars are commonly used to identify a MIMO radar. The MU radar is one of the most multi-functional radars, which also functions as a MIMO radar with additional settings. From the previous discussion, DDMA and CDMA are suitable for the MU radar because it has lower frequencies and the Doppler speeds of the targets are relatively small. CDMA would be the best for higher frequency radars. However, it requires multiple transmitters with an independent pulse code setting. In contrast, DDMA can be modified to use multiple frequency sources to generate orthogonal waveforms. In this study, DDMA was chosen considering its easier application to the MU radar.

4. Beamwidth Verification Using the Moon's Reflection

4.1 MU Radar System Configuration

The MU radar [2], located in Shigaraki, Shiga, Japan, has been operational as an atmospheric radar since 1984. It consists of 475 elements, comprising 19-element antennas multiplied by 25 sub-array digital receivers, operating in the VHF band at 46.5 MHz. Upgraded to a digital modulator with frequency hopping functionality, it has 29 digital receivers for adaptive beamforming, including 25 primaries plus, an additional 4 receivers, as detailed in [20]. Although categorized as a SIMO radar, the MU radar can also function as a MIMO radar owing to its flexibility.

To use the MU radar as a DDMA-MIMO radar, we classified the transmit antennas into six parts, which is the same as the number of synchronized signal generators that have slightly different frequencies. In each receiver, the six orthogonal transmitter signals are separated using Doppler matched filters, which means that it consists of $6 \times 25 = 150$ receivers. Figure 3 shows images of the actual transmit and virtual receive antennas.

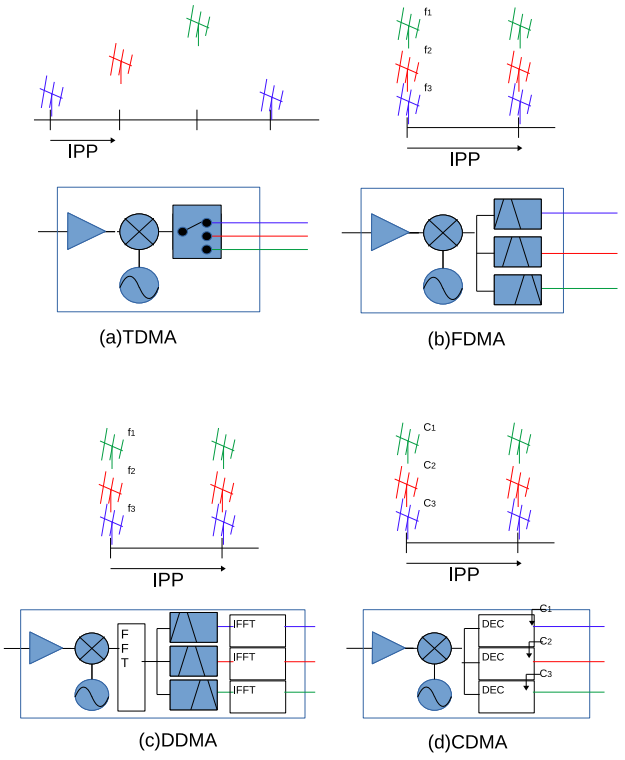


Fig. 2 Comparison between orthogonal waveforms (Configuration image).

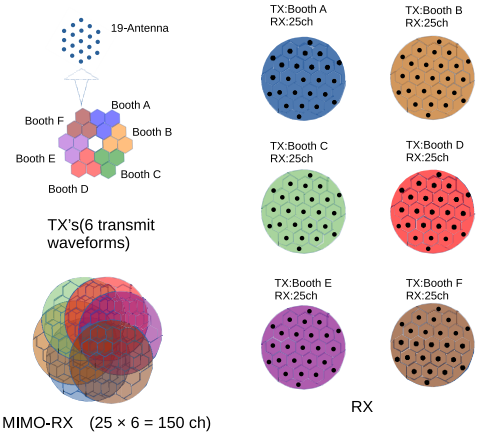


Fig. 3 Layout images of the actual transmit antenna (upper left), receive antenna (right), and virtual receive antenna (lower left).

4.1.1 Transmitter Configuration

The MU radar radiates a 46.5 MHz RF transmit signal by mixing a 5 MHz modulated IF-signal with a 41.5 MHz CW-local signal. In our study, the CW-local signal is used to replace the original signal with six separated local signals generated independently using signal generators but synchronized using a GPS-10 MHz oscillator. The frequencies of the signal generators are configured with intervals of f_{md} Hz between them to generate a pulse-to-pulse phase difference of $2\pi f_{md} T_{IPP}$ radians, where T_{IPP} is the inter-pulse period,

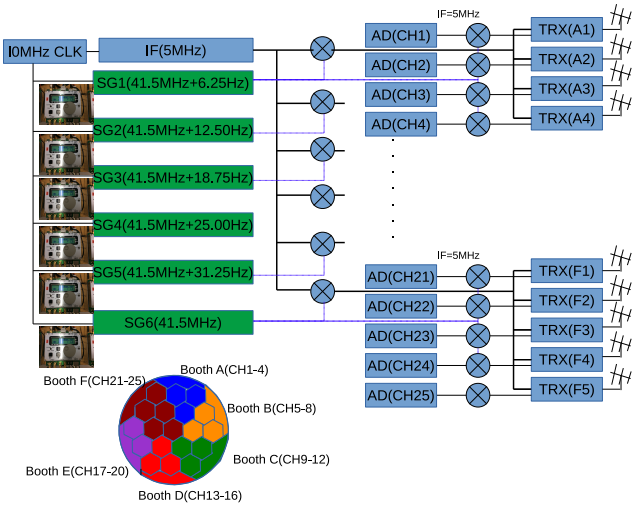


Fig. 4 DDMA-MIMO configuration of the MU radar. The green mark indicates additional settings, and the six orthogonal signals are distributed to the transmitters.

and each demodulator is processed by IF signals of 5 MHz, therefore down-converters by the same CW-local signals are equipped. The MIMO configuration of the MU radar is illustrated in Fig. 4.

4.1.2 Receiver Configuration

The receive signals that correspond to each transmit signal are mixed before the signal processing stage. A Doppler filter toward the slow-time direction is applied to separate the receive signals into orthogonal signals in the receiver.

Each receiver received six orthogonal waveforms, which were separated using Doppler filters. In this experiment, the interval of the Doppler offset velocity was selected to be $V_{Nyquist}/16 = 20.15\,\mathrm{ms}^{-1}$ because it is divisible by FFT points and enables us to consider the minimum frequency setting unit (0.01 Hz) of the SG, where $V_{Nyquist}$ is the Nyquist velocity, which is determined by the inter-pulse period, coherent integration number, and transmit frequency.

Because IF signals are down-converted by each transmit local signal, Doppler positions of the receive signal depend on the attributes of the receiver. Table 2 lists the Doppler offsets of the signal received from 25 receivers corresponding to the transmitters. In actual signal processing, amplitude and phase offset occurs because of the independent transmitters, which must be corrected on the receiver. The transmission phase adjustment process is presented in [9].

4.1.3 Antenna Position of the MU Radar

Figures 5 and 6 display the transmit and receive antenna positions of the MU radar, respectively. The transmit antennas are divided into six groups of sub-array antennas (composed of 57 antenna elements) that correspond to each antenna booth (depicted as booths A (A2/A3/A4 in blue); B (B2/B3/B4 in orange); C (C2/C3/C4 in green); D (D2/D3/D4

Table 2 DDMA-MIMO frequency settings (local signal frequency) and receive signal Doppler offset in each receiver.

Transmitter	Local signal frequency	Receive signal Doppler offset					
		A1-A4 CH1-CH4	B1-B4 CH5-CH8	C1-C4 CH9-CH12	D1-D4 CH13-CH16	E1-E4 CH17-CH20	F1-F5 CH21-CH25
TX1 (Booth A)	41.5 MHz + 6.25 Hz	0 ms^{-1}	20.15 ms^{-1}	40.29 ms^{-1}	60.44 ms^{-1}	80.59 ms^{-1}	-20.15 ms^{-1}
TX2 (Booth B)	41.5 MHz + 12.50 Hz	-20.15 ms^{-1}	0 ms^{-1}	20.15 ms^{-1}	40.29 ms^{-1}	60.44 ms^{-1}	-40.29 ms^{-1}
TX3 (Booth C)	41.5 MHz + 18.75 Hz	-40.29 ms^{-1}	-20.15 ms^{-1}	0 ms^{-1}	20.15 ms^{-1}	40.29 ms^{-1}	-60.44 ms^{-1}
TX4 (Booth D)	41.5 MHz + 25.00 Hz	-60.44 ms^{-1}	-40.29 ms^{-1}	-20.15 ms^{-1}	0 ms^{-1}	20.15 ms^{-1}	-80.29 ms^{-1}
TX5 (Booth E)	41.5 MHz + 31.25 Hz	-80.59 ms^{-1}	-60.44 ms^{-1}	-40.29 ms^{-1}	-20.15 ms^{-1}	0 ms^{-1}	-100.74 ms^{-1}
TX6 (Booth F)	41.5 MHz + 0 Hz	20.15 ms^{-1}	40.29 ms^{-1}	60.44 ms^{-1}	80.59 ms^{-1}	100.74 ms^{-1}	0 ms^{-1}

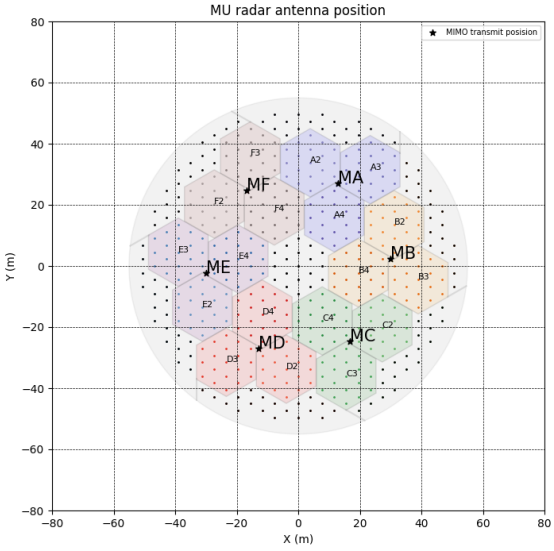


Fig. 5 Transmit antenna positions. Transmit antennas are divided into six groups of sub-array antennas (composed of 57 antenna elements) corresponding to the each booth (booth A (A2/A3/A4 in blue); B (B2/B3/B4 in orange); C (C2/C3/C4 in green); D (D2/D3/D4 in red); E (E2/E3/E4 in purple) and F (F2/F3/F4 in brown). Star markers denote the transmit antenna phase centers, and the black points are not used for transmission.

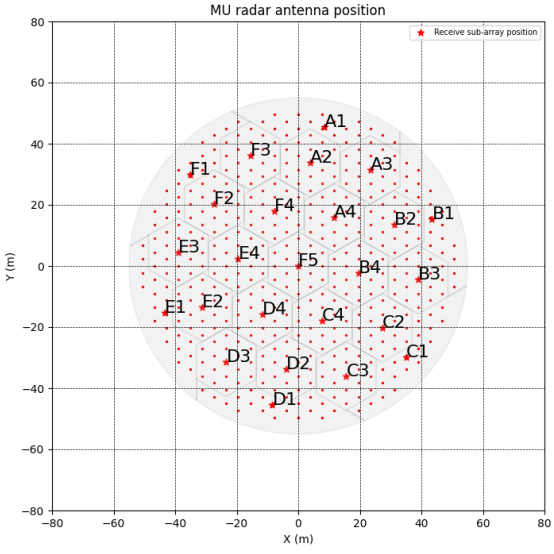


Fig. 6 Receive antenna positions. Star markers denote the receive antenna phase centers.

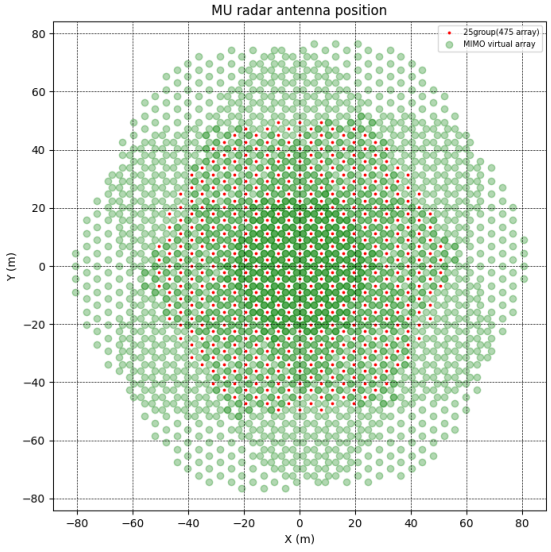


Fig. 7 Virtual receive antenna positions. The red points indicate the physical antenna layout, whereas the green dots represent virtual antennas. The green dots have transparency. The darker colors indicate overlapping antenna positions.

in red); E (E2/E3/E4 in purple); and F (F2/F3/F4 in brown). Star markers denote the transmit antenna phase centers, representing the transmit steering vector in Fig. 5, and highlight the receive steering vector in Fig. 6. These positions marked by stars correspond to the transmit steering vector $\mathbf{a}(\theta_0)$ and the receive steering vector $\mathbf{b}(\theta_0)$, representing the transmit and receive beam directions of θ_0, respectively.

4.1.4 Virtual Receive Antenna Position of the MU Radar for SIMO and MIMO

Figures 7 and 8 display the virtual receive and sub-array antenna positions for adaptive beamforming calculated using the MIMO channel matrix $\mathbf{a}(\theta_0) \otimes \mathbf{b}(\theta_0)$, respectively. The red points indicate the physical centers, and the green or blue dots indicate those of the virtual antennas. These dots have transparency, with darker colors indicating overlapping receiver positions. The receive MIMO antenna is larger than conventional physical ones, implying that MIMO-beamforming has narrower characteristics and sidelobe improvements compared with SIMO-beamforming.

4.2 Observation of the Moon's Reflection

As previously mentioned, MIMO radars can establish a virtual receive antenna aperture plane with transmission freedom. However, quantitatively confirming the enhancement of spatial resolution using observed atmospheric and ionospheric echoes is challenging owing to the nonuniform volume targets. To quantitatively assess the MIMO virtual array, we conducted observations of the beam pattern, as discussed by [10], derived from the reflection echo off the moon. We compared it with the calculated beam pattern from the virtual antenna layout.

4.2.1 Observation Parameters

The experiment was conducted for the 11th March, 2022 because the moon's position had a higher elevation angle (lower zenith angle), minimizing the radial velocity toward the radar, which made the analysis easier. The observation parameters are listed as experiment A presented in Table 3.

Figure 9 shows the time series of the estimated distance between the surface of the moon and observation point (Shigaraki, Japan), calculated using Skyfield (https://rhodesmill.org/skyfield/). From the estimation, the beam direction was determined for a zenith angle θ_0 of 8.59°, azimuth angle from the north ϕ_0 of 186.77° at 1849 JST 11th March, 2022, when the radial Doppler speed of the moon was expected to be zero when passing across the beam center.

For comparison, other experiments using conventional SIMO observations were conducted on the 27th June, 2022. The observation parameters and dates of these experiments are also listed in Table 3.

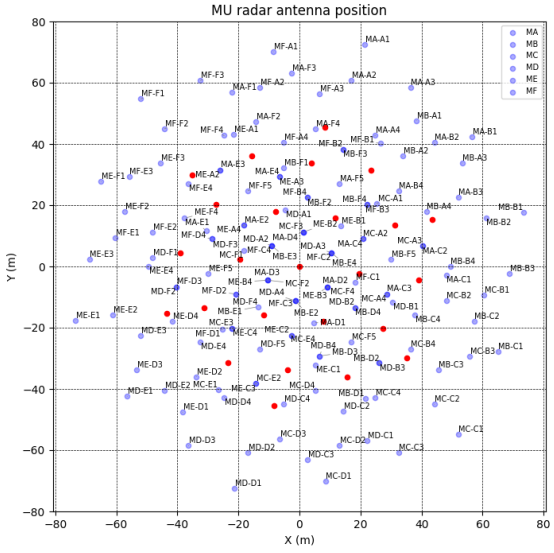


Fig. 8 Virtual receive sub-array positions for adaptive beamforming. The red points indicate the physical centers of the sub-array antennas, whereas the blue dots indicates those of the virtual sub-array antennas. The blue dots have transparency. The darker colors indicate overlapping antenna positions.

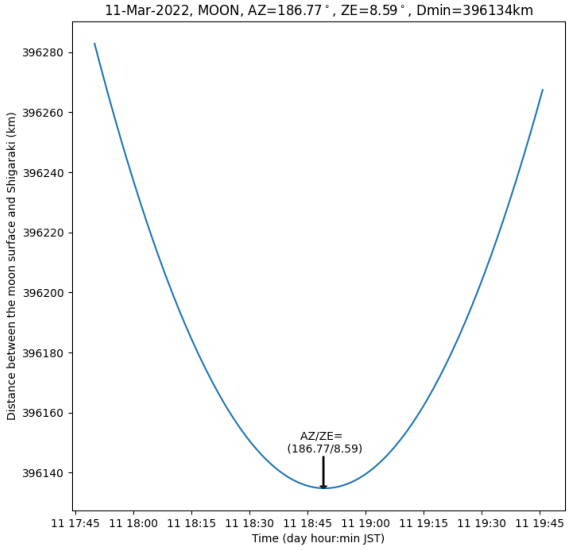


Fig. 9 Time series of estimated distance between the moon's surface and observation point (Shigaraki, Shiga, Japan) at 1750-1940 JST 11 Mar 2022 calculated using Skyfield (https://rhodesmill.org/skyfield/).

Table 3 Observation parameters and estimated distance from the MU radar to the moon's surface. Note this calculation does not consider the radius of curvature of the Earth or moon.

Item	Experiment A TX=57×6,RX=475×6(MIMO)	Experiment B TX=RX=475(SIMO)
Passing time of the beam center	1849 JST 11 Mar 2022	1008 JST 27 June 2022
Observation time	1750-1950 JST 11 Mar 2022	0920-1100 JST June 2022
Transmit antenna	57 (19 ×3ch) × 6	475 (19 × 25 ch)
Receive antenna	475 (19 ×25ch) × 6	475 (19 × 25 ch)
Inter-pulse period	10 000 μs	10 000 μs
Transmit beam direction (AZ/ZE)	186.77°/8.59°	166.24°/11.14°
Sub-pulse width	64 μs	64 μs
Pulse comp.	7bit Barker	7bit Barker
Coherent integration	1	1
FFT points	2 048	2 048
Estimated distance to the moon	396 134 km	396 903 km
N-th tripped echo	265	265
Estimated target range	412 km	1 180 km

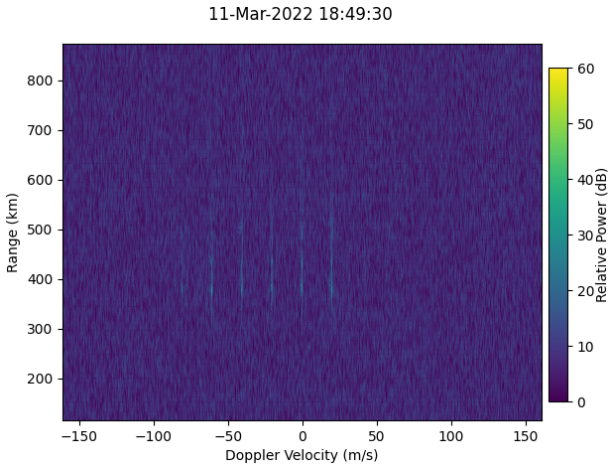


Fig. 10 Doppler spectra at A2 sub-array receivers observed at 1849 JST on 11 Mar 2022 before executing MIMO-processing. The horizontal and vertical axes show the target Doppler velocity and range with offset of 395 726 km due to echoes from 265 trips, respectively. Six receive signals derived from each orthogonal transmitter with DDMA-MIMO frequency offset (TX5, TX4, TX3, TX2, TX1, and TX6 from the left) were observed.

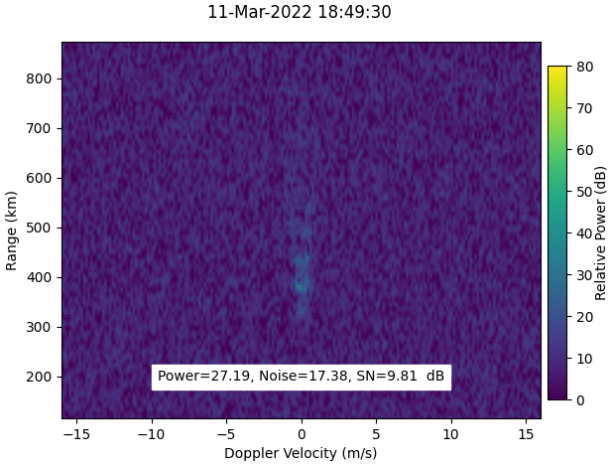


Fig. 11 Same as Fig. 10 except that the Doppler range was restricted from -16 to 16 ms^{-1}.

4.2.2 Observation Result: Doppler Spectrum of MIMO and SIMO Observation

Figure 10 illustrates the Doppler spectra obtained from the sub-array receivers A2 which is before executing MIMO-processing, where six separated signals received caused by transmit frequency offsets from the moon (265th trip echo) were confirmed.

In this figure, the estimated radial velocity of the moon was almost zero at the time of observation so that the observation results were consistent with those listed in Table 2, from which these signals could be identified as the signals received at TX5 ($-80.59\,\mathrm{ms}^{-1}$), TX4 ($-60.44\,\mathrm{ms}^{-1}$), TX3 ($-40.29\,\mathrm{ms}^{-1}$), TX2 ($-20.15\,\mathrm{ms}^{-1}$), TX1 ($0\,\mathrm{ms}^{-1}$), and TX6 ($20.15\,\mathrm{ms}^{-1}$)., respectively.

Figure 11 presents the same dataset as depicted in

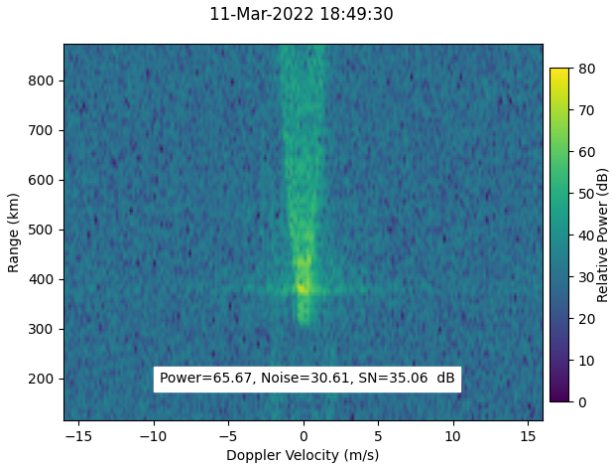


Fig. 12 Same as Fig. 11 except that the DDMA-MIMO processing was performed.

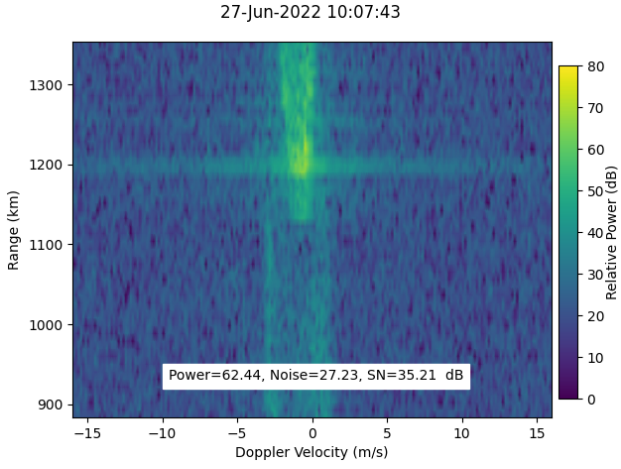


Fig. 13 Doppler spectra observed at 1007 JST on the 27th of June, 2022 as the conventional (SIMO) operation. The horizontal and vertical axes represent the target Doppler velocity and range with an offset of 395 726 km caused by echoes from 265 trips.

Fig. 10, with the Doppler range restricted from -16 to 16 ms^{-1}, where the signal-to-noise (S/N) ratio was calculated to be 9.81 dB. For a MIMO radar, the receive signals must be extracted using Doppler filters as independent IQ signals before combining them to obtain a MIMO receive signal by applying (12). Figure 12 illustrates the results after executing MIMO-processing, where the S/N ratio was calculated to be 35.06 dB. In this observation, the estimated summated power was the summation of 25×6 = 150 (21.76 dB) virtual receivers. Therefore, the summed S/N ratio was estimated to be 9.81+21.76=31.57 dB, where the difference in the observed data (Fig. 12) was considered to be caused by the individual difference (receive gain and the phase) in receivers.

Figure 13 shows the Doppler spectra observed at 1007 JST 27th June, 2022 as a conventional SIMO operation. This observation result was for a comparison between the MIMO and SIMO radars operated as experiment B, listed in Ta-

ble 3, where the S/N ratio was calculated to be 35.21 dB. From these results, the S/N ratio between the MIMO observation and the SIMO observation was also consistent from the point of the S/N ratio. Furthermore, the beam width appeared to be a wider distribution compared with that shown in Fig. 12. However, these differences are qualitative and cannot be definitively assessed.

4.2.3 Verification with the Beamformer and Capon

To verify the moon reflection echo quantitatively, we focus on the result whether the brightness distribution of the beamformer and the theoretical beam pattern comparing with the SIMO observation. Figures 14 and 15 show the two-dimensional angular imaging result of the moon's reflection echo. The MIMO observation data were the same as those used for the verification of the antenna pattern. As shown in Figs. 14 and 15, the MIMO radar had a narrower beamwidth than that of the SIMO radar.

Figure 16 illustrates the cut pattern at an azimuth angle of $\phi = 0^\circ$ to compare the MIMO and SIMO observation results with the receive antenna patterns. In this study, we compared the differences in power between the SIMO and MIMO beamformer and Capon methods. Specifically, the peak power was normalized for comparative analysis against the antenna pattern. The calculated peak differences between the beamformer and Capon brightness are indicated in Fig. 16, revealing differences of 2.65 dB for SIMO and 6.90 dB for MIMO. These differences are attributed to the influence of ionospheric scintillation on the observed data, which persists despite attempts to mitigate its effects through time-averaging. However, we confirmed that the Capon brightness with MIMO virtual arrays exhibited superior performance to that with SIMO physical arrays. Additionally, the dotted blue line (beamformer) and the red line (1-way antenna pattern) in the figure are consistent for both SIMO and MIMO observations, particularly in the mainlobe of both the SIMO and MIMO beamformer, indicating consistency

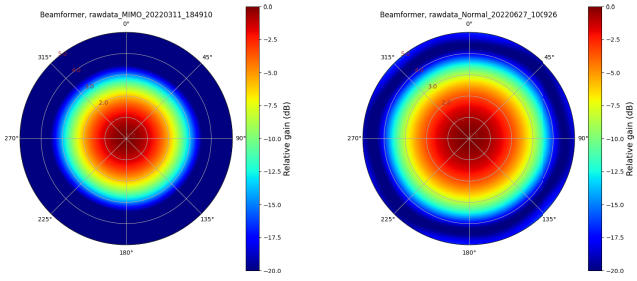


Fig. 14 Angular distribution of beamforming using the moon's reflection: Beamformer imaging result from MIMO (left) and SIMO (right) observations. Note that (azimuth, zenith) = (0,0) indicates the transmit beam direction.

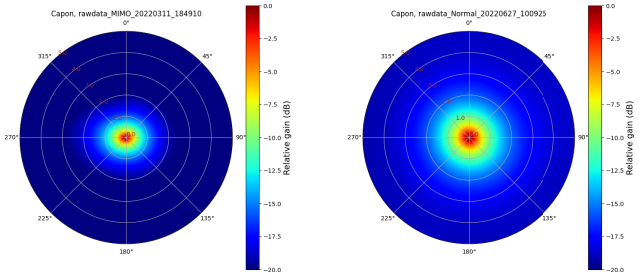


Fig. 15 Angular distribution of adaptive beamforming using the moon's reflection: Capon imaging result using MIMO (left) and SIMO (right) observations.

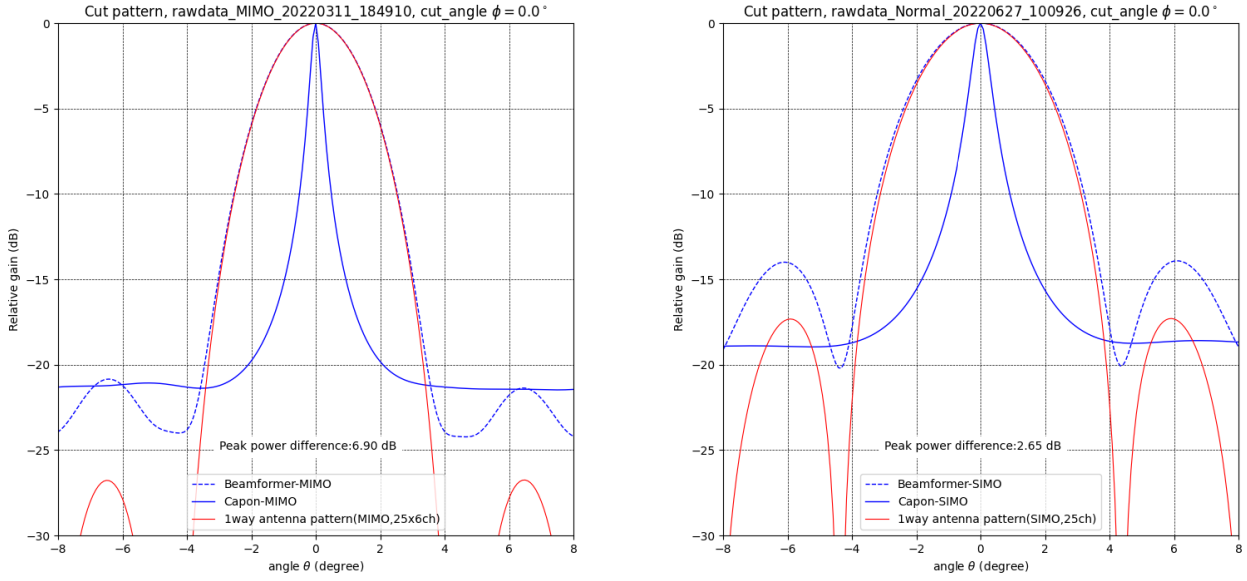


Fig. 16 Cut pattern at an azimuth angle of $\phi = 0^\circ$: Comparison between the beamformer, Capon, and calculated 1-way antenna pattern using MIMO (left) and SIMO (right) observations. The blue dotted and blue solid lines represent the angular distributions using the beamformer and Capon imaging, respectively, whereas the red line represents the calculated antenna pattern.

with theoretical expectations. Furthermore, Fig. 16 demonstrates the contribution of the virtual receive array to the suppression of antenna sidelobes, as predicted by theory.

The effectiveness of the Capon beamformer has already been demonstrated by [12], and the observation results were almost consistent. However, this experiment particularly showed that the Capon beamformer achieved even higher resolution when combined with MIMO radar, which the results clearly show.

From these results and considerations, the MU radar, in combination with the Capon beamformer, can operate as a MIMO radar with good performance and high angular resolutions.

5. Conclusion

In this study, an extension of the virtual receive array antenna that included the DDMA method was demonstrated through the experimental results using the MU radar, which was operated as a MIMO radar with additional settings. To achieve a DDMA-MIMO radar, local signals were replaced with six synchronized signal generators that generated different Doppler frequency offsets between the transmitters to realize transmit signal orthogonality. MIMO radar signal processing had a high compatibility with other processing methods, such as the Capon beamformer. It was confirmed through experimental results that a resolution finer than that of conventional methods can be obtained using a combination of the MIMO technique and Capon beamformer. Our findings are expected to contribute toward advancing the spatial super resolution technique intended for applications in atmospheric phased array radars.

Acknowledgments

The MU radar belongs to and is operated by the Research Institute for Sustainable Humanosphere (RISH), Kyoto University. This work was partially supported by ISHIZUE 2022 of Kyoto University and JSPS KAKENHI Grant Number JP23K17703.

References

[1] J. Li and P. Stoica, MIMO Radar Signal Processing, Wiley-IEEE Press, 2008.
[2] S. Fukao, T. Sato, T. Tsuda, S. Kato, K. Wakasugi, and T. Makihira, "The MU radar with an active phased array system: 1. Antenna and power amplifiers," Radio Sci., vol.20, no.6, pp.1155–1168, 1985.
[3] S. Kato, T. Tsuda, M. Yamamoto, T. Sato, and S. Fukao, "First results obtained with a middle and upper atmosphere (MU) radar," J. Atmos. Terr. Phys., vol.48, no.11-12, pp.1259–1267, 1986.
[4] H. Sun, F. Brigui, and M. Lesturgie, "Analysis and comparison of MIMO radar waveforms," 2014 Int. Radar Conf. Radar 2014, pp.1–6, 2014.
[5] J.M. Urco, J.L. Chau, M.A. Milla, J.P. Vierinen, and T. Weber, "Coherent MIMO to improve aperture synthesis radar imaging of field-aligned irregularities: First results at Jicamarca," IEEE Trans. Geosci. Remote Sens., vol.56, no.5, pp.2980–2990, 2018.
[6] J.M. Urco, J.L. Chau, T. Weber, and R. Latteck, "Enhancing the spatiotemporal features of polar mesosphere summer echoes using coherent MIMO and radar imaging at MAARSY," Atmos. Meas. Tech., vol.12, no.2, pp.955–969, 2019.
[7] D.J. Rabideau, "Doppler-offset waveforms for MIMO radar," IEEE Natl. Radar Conf. - Proc., pp.965–970, 2011.
[8] F. Yang, F. Xu, X. Yang, and Q. Liu, "DDMA MIMO radar system for low, slow, and small target detection," J. Eng., vol.2019, no.19, pp.5932–5935, 2019.
[9] T. Matsuda and H. Hashiguchi, "DDMA-MIMO observations with the MU radar: Validation by measuring a beam broadening effect," IEEE J. Sel. Topics Appl. Earth Observ., vol.16, pp.3083–3091, 2023.
[10] S. Fukao, T. Sato, T. Tsuda, M. Yamamoto, M.D. Yamanaka, and S. Kato, "MU radar: New capabilities and system calibrations," Radio Sci., vol.25, no.4, pp.477–485, 1990.
[11] H. Kimoto, N. Kikuma, and K. Sakakibara, "Target direction estimation characteristics of capon algorithm in MIMO radar," 2019 Int. Symp. Antennas Propagation, ISAP 2019 - Proc., 2019.
[12] R.D. Palmer, S. Gopalam, T.Y. Yu, and S. Fukao, "Coherent radar imaging using Capon's method," Radio Sci., vol.33, no.6, pp.1585–1598, 1998.
[13] M.S. Davis, Principles of Modern Radar: Advanced Techniques, Scitech publishing, 2013.
[14] M.S. Davis, G.A. Showman, and A.D. Lanterman, "Coherent MIMO radar: The phased array and orthogonal waveforms," IEEE Aerosp. Electron. Syst. Mag., vol.29, no.8, pp.76–91, 2014.
[15] M. Cattenoz, "MIMO radar processing methods for anticipating and compensating real world imperfections," Ph.D. dissertation, UniversitéParis Sud-Paris XI, Orsay, France, 2015.
[16] H.D. Macedo and J.N. Oliveira, "Typing linear algebra: A biproduct-oriented approach," Sci. Comput. Program., vol.78, no.11, pp.2160–2191, 2013.
[17] N. Kikuma, Adaptive Antenna Technology, Ohmsha, 2003 (in Japanese).
[18] H.L. Van Trees, Optimum Array Processing, Wiley-Interscience, New York, 2002.
[19] P. Stoica, Z. Wang, and J. Li, "Robust Capon beamforming," IEEE Signal Process. Lett., vol.10, no.6, pp.172–175, 2003.
[20] G. Hassenpflug, M. Yamamoto, H. Luce, and S. Fukao, "Description and demonstration of the new middle and upper atmosphere radar imaging system: 1-D, 2-D, and 3-D imaging of troposphere and stratosphere," Radio Sci., vol.43, no.2, 2008.
[21] J.S. Chen, C.Y. Wang, C.L. Su, and Y.H. Chu, "Meteor observations using radar imaging techniques and norm-constrained Capon method," Planet. Space Sci., vol.184, no.2020, p.104884, 2020.
[22] D. Cohen, D. Cohen, and Y.C. Eldar, "High resolution FDMA MIMO radar," IEEE Trans. Aerosp. Electron. Syst., vol.56, no.4, pp.2806–2822, 2020.
[23] D.J. Rabideau, "MIMO radar waveforms and cancellation ratio," IEEE Trans. Aerosp. Electron. Syst., vol.48, no.2, pp.1167–1178, 2012.
[24] J. Tang, N. Zhang, Z. Ma, and B. Tang, "Construction of Doppler resilient complete complementary code in MIMO radar," IEEE Trans. Signal Process., vol.62, no.18, pp.4704–4712, 2014.
[25] T. Kishigami, H. Yomo, N. Yosoku, A. Matsuoka, and J. Sato, "MIMO radar waveforms using orthogonal complementary codes with Doppler-offset," IEICE Trans. Commun., vol.E101-B, no.6, pp.1503–1512, June 2018.

Tomoya Matsuda received the B.S. and M.S. degrees in Department of electronic and communication engineering from Kyoto University, Japan, in 1997 and 1999, respectively. In 1999, he joined Mitsubishi Electric Corporation, Japan, where he has been engaged in development of active phased array radar system. His research interests include the development of phased array radar systems for atmospheric and weather radars. He was a recipient of the 2022 Gambo-Tatehira Award of the Meteorological Society of Japan on behalf of Mitsubishi Electric Corporation.

Koji Nishimura received his B.E. degree from Ritsumeikan University, Japan, in 1999, and his M.I. and Ph.D. degrees from Kyoto University, Japan, in 2001 and 2006, respectively. From 2001–2003, he worked for Sony Corporation. Since 2007, he was with the National Institute of Polar Research (NIPR), and the Research Organization for Information and Systems (ROIS), Tokyo, Japan. From 2021 to present, he has been with Kyoto University as an Associate Professor. His major research interests are radar signal processing, multi-channel and space-time signal processing, remote sensing for the atmosphere, and satellite communications. Dr. Nishimura was a recipient of the Commendation for Contributors to Promotion of an Oceanic State from the Prime Minister of Japan in 2015. He was a recipient of the Prize for Science and Technology from the Minister of Education, Culture, Sports, Science and Technology of Japan in 2014.

Hiroyuki Hashiguchi received Bachelor of Engineering from the Faculty of Engineering and Design, Kyoto Institute of Technology, Kyoto, Japan, in 1990. He completed the second half of his Ph.D. degree with the Department of Electrical Science and Engineering, Graduate School of Engineering, Kyoto University, Kyoto, in 1995. In 1997, he became a Research Associate with the Radio Atmospheric Science Center, Kyoto University (reorganized in 2000 as the Radio Science Center for Space and Atmosphere). In 2001, he became an Associate Professor at the Radio Science Center for Space and Atmosphere, Kyoto University [reorganized in 2004 as the Research Institute for Sustainable Humanosphere (RISH)]. In 2018, he became a Professor at RISH. His research interests include the development of atmospheric radars and research on observations using their radars. Dr. Hashiguchi is a member of the Meteorological Society of Japan, the Society of Geomagnetism and Earth, Planetary and Space Sciences, the American Meteorological Society, the American Geophysical Union, and the Institute of Electronics, Information and Communication Engineers, Japan. In 1992, he became Research Fellow of the Japan Society for the Promotion of Science. He was a recipient of the 2006 Minister of Education Science and Technology Award, and the 2008 Horiuchi Award and 2022 Gambo-Tatehira Award of the Meteorological Society of Japan.

PAPER

Cooperative Transmission of Energy-Constrained Wireless Devices in IRS-Assisted Wireless Powered Communication Networks*

Yun WU[†a)], **ZiHao CHEN**[†], **MengYao LI**[†], ***Nonmembers***, ***and*** **Han HAI**[†], ***Member***

SUMMARY Intelligent reflecting surface (IRS) is an effective technology to improve the energy and spectral efficiency of wireless powered communication network (WPCN). Under user cooperation, we propose an IRS-assisted WPCN system where the wireless devices (WDs) collect wireless energy in the downlink (DL) and then share data. The adjacent single-antenna WDs cooperate to form a virtual antenna array so that their information can be simultaneously transmitted to the multi-antenna common hybrid access point (HAP) through the uplink (UL) using multiple-input multiple-output (MIMO) technology. By jointly optimizing the passive beamforming at the IRS, the active beamforming in the DL and the UL, the energy consumed by data sharing, and the time allocation of each phase, we formulate an UL throughput maximization problem. However, this optimization problem is non-convex since the optimization variables are highly coupled. In this study, we apply the alternating optimization (AO) technology to decouple the optimization variables and propose an efficient algorithm to avoid the difficulty of directly solving the problem. Numerical results indicate that the joint optimization method significantly improves the UL throughput performance in multi-user WPCN compared with various baseline methods.

key words: *wireless powered communication networks, intelligent reflecting surface, virtual antenna array, beamforming*

1. Introduction

With the increasing requirement for the working time of the internet of things (IoT) devices [1], providing the wireless devices (WDs) with sufficient power continuously is a critical issue in practice. In the IoT, the working time of the WDs is generally extended by charging and replacing new batteries, but the cost of these two methods is high. In addition, an effective way to extend the working time of the WDs is to collect energy from the external environment, which can be water, wind, and solar energy. However, these energy sources are usually unstable or difficult to control [2]. Another practical method is to obtain energy through wireless power transfer (WPT) technology, that is, directly transmitting energy to the WDs through radio frequency signals [3], [4]. In a wireless powered communication network (WPCN) system, a transmitter transmits radio frequency (RF) signals, and a receiver obtains energy from the RF signals and then transmits information. Specifically, the WDs in WPCN first obtain energy from the RF signals transmitted by the hybrid access point (HAP) and then transmit their information to the HAP by this energy [5]–[7].

Recently, as the integration of intelligent reflecting surface (IRS) technology and wireless communication systems continues to deepen, IRS technology has gradually attracted widespread attention [8]. The IRS consists of many reconfigurable reflection elements, each of which can independently reflect signals and adjust a certain phase shift. Therefore, IRS can adjust the signal propagation environment to achieve signal enhancement or interference reduction. For IRS aided WPCN systems, the IRS can enhance channel gain, thereby improving energy transfer efficiency and information transmission rate.

Recent research about throughput maximization for IRS-assisted WPCN systems has attracted extensive attention [9]–[13]. Specifically, [9] addressed the incorporation of full-duplex transmission modes into IRS-assisted WPCN systems, where a model was developed to maximize system throughput by jointly optimizing time allocation, HAP transmission power, and IRS beamforming. In addition, [10] discussed an IRS-assisted WPCN system with multiple base stations (BS), aiming to maximize the throughput by jointly optimizing the time allocation, the energy beamforming in the downlink (DL) and receive beamforming in the uplink (UL). Furthering this study, [11] introduced dual IRSs to improve the energy and the information transmission, respectively. With multiple access schemes being crucial in multi-user IRS-assisted WPCN systems, [12] explored the effects of the non-orthogonal multiple access (NOMA) configuration on system throughput optimization. [13] further proposed a hybrid NOMA scheme, where users are partitioned into several clusters. The authors investigated optimizing the IRS beamforming and the time allocation among the power transfer of BS and the information transmission of different user clusters to maximize the throughput.

It is well known that multi-antenna techniques provide significant performance gains through spatial diversity and multiplexing. However, due to the limited physical volume and complexity of the device, it is difficult for the WDs to have multiple antennas, most of the existing works only focused on the individual information transmission of single-antenna users. How to incorporate multi-antenna technology into the user information transmission is still a challenge. For this problem, multiple adjacent single-antenna users can collaborate to form a virtual antenna array to achieve collaborative data transmission [14]. In [15]–[17], virtual MIMO was used in a wireless sensor network (WSN) to collect lots

Manuscript received January 9, 2024.
Manuscript revised April 20, 2024.
Manuscript publicized June 17, 2024.

[†]College of Information Science and Technology, Donghua University, Shanghai, 201620 China.

*This work was supported by the National Natural Science Foundation of China (No. 61801106).

a) E-mail: wuyun_hit@dhu.edu.cn

DOI: 10.23919/transcom.2024EBP3009

of information into the fusion center. In [15], the energy and delay efficiency of a virtual MIMO system based on space-time coding was analyzed. In [16], the author studied the transmit power allocation problem for different sensors in the virtual MIMO system to improve the communication between the sensors and the fusion center. In [17], a variety of optimal and sub-optimal fusion rules for virtual MIMO in WSN systems were studied.

To the best of our knowledge, the incorporation of the virtual MIMO in IRS-assisted WPCN system remains unexplored in the existing literature. Thus, in this paper, we propose an IRS-assisted WPCN system under virtual MIMO-based user cooperation, where the single-antenna WDs first obtain energy from the multi-antenna HAP and then share data with each other. The adjacent WDs are combined to form a virtual antenna array so that the shared data can be transmitted to the HAP simultaneously. The whole system consists of three stages. Specifically, in the first stage, the HAP transmits the energy signals to the WDs in the DL so that the WDs can obtain energy from them. In the second stage, the WDs share data with each other. In the final stage, the WDs form a virtual antenna array to transmit the shared data to the HAP cooperatively. Furthermore, the scheme proposed in this paper is suitable for scenarios where the WDs are in close proximity to each other, such as monitoring the working condition of each device in the factory and reporting the sensed information to the HAP.

To analyze the impact of virtual MIMO-based user-cooperative transmission in the IRS-assisted WPCN system, we formulate an UL throughput maximization problem and solve it by a joint optimization algorithm. Our main work is summarized as follows.

- We propose a virtual MIMO-based user cooperation scheme in the IRS-assisted WPCN system, in which the WDs first obtain energy from the signals transmitted by the HAP, then share data with each other and transmit information to the HAP cooperatively by forming a virtual antenna array. Furthermore, we deploy an IRS to enhance both the energy and the information transfer. Numerical simulation results indicate that the scheme proposed in this paper achieves higher UL throughput than other benchmark schemes.
- We propose a joint beamforming, power and time allocation (JO-BPT) algorithm to maximize the UL throughput. The algorithm uses an alternating optimization (AO) technology to decompose the problem into four sub-problems and optimizes them alternately. Specifically, we fix the passive beamforming for IRS and the time allocation, then solve the joint DL and UL active beamforming optimization subproblem using the Lagrange duality method.
- Given the active beamforming and the time allocation, the semi-definite relaxation (SDR) and the Gaussian randomization (GR) method are applied to optimize the passive beamforming for IRS. With the given active beamforming in the DL and the UL and passive beamforming for IRS, the time allocation optimization suproblem is convex and solved effectively by the CVX tool. Finally, the optimal cost of user data sharing E_S is found by a one-dimensional search.

The rest of the paper is organized as follows. Section 2 presents the IRS-assisted WPCN system based on virtual MIMO and describes the three phases of the system in detail. Section 3 formulates a problem of maximizing the UL throughput. Section 4 introduces the JO-BPT algorithm. In Sect. 5, we give the numerical simulation results and compare the different results. Finally, we make a conclusion of this article in Sect. 6.

Notation: Bold lowercase and uppercase letters indicate vectors and matrices, respectively. For the complex vector $\mathbf{a}$, $\|\mathbf{a}\|$ and $\mathbf{a}^*$ denote the Euclidean norm and conjugate of $\mathbf{a}$, respectively. diag($\mathbf{a}$) indicates a diagonal matrix created from the vector $\mathbf{a}$. Superscripts $(\cdot)^{\mathrm{T}}$ and $(\cdot)^{\mathrm{H}}$ indicate transposition and Hermitian transposition, respectively. $\mathbb{E}[\cdot]$ denotes the expectation. Tr($\mathbf{A}$) indicates the trace of matrix $\mathbf{A}$. $[x]^+$ indicates the $\max\{0, x\}$ of the real number x. $\mathbf{I}_n$ represents an $n \times n$ identity matrix. Vector $\mathbf{x}$ that follows a complex normal distribution with mean μ and variance σ^2 is denoted as $\mathbf{x} \sim \mathcal{CN}\left(\mu, \sigma^2\mathbf{I}\right)$. $(\cdot)^{(i)}$ represents the variable generated by the i-th iteration.

2. System Model

Figure 1 shows the proposed multi-user IRS-assisted WPCN system, which includes K single-antenna WDs, a HAP equipped with M antennas, and an IRS with N reflection elements to enhance the information and the energy transmission. During the energy and information transmission, the transmitter sends the reflective information to the IRS control unit to dynamically adjust the phase shift of each reflection element. We assume that each WD is equipped with a rechargeable battery and an energy harvesting (EH) circuit component, allowing it to store the harvested energy for powering its operation. Here, the set of WDs is defined as $\mathcal{K} \triangleq \{1, \ldots, K\}$. We use WD_k to denote the k-th WD. Let $\boldsymbol{\varphi}_d = \left[\varphi_{d,1}, \ldots, \varphi_{d,N}\right]^{\mathrm{T}}$ and

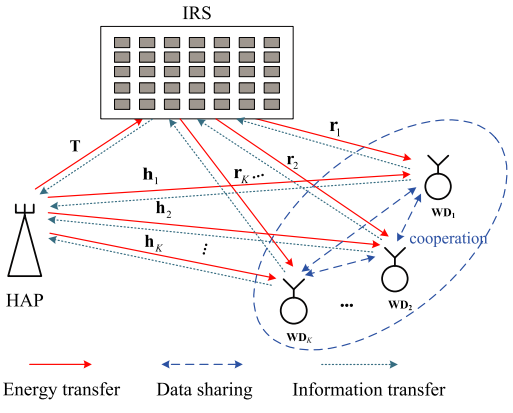


Fig. 1 IRS-assisted WPCN system based on virtual MIMO.

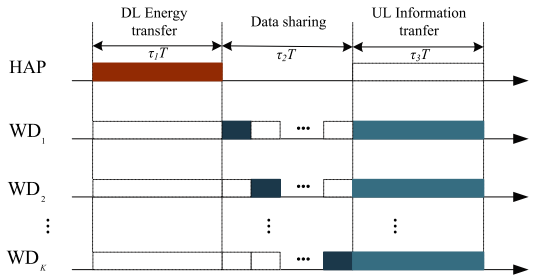


Fig. 2 The framework structure of the IRS-assisted multi-user WPCN system.

$\varphi_{d,n} = \beta_n e^{j\theta_{d,n}}, d \in \{1,2\}, \forall n \in \mathcal{N} \triangleq \{1,\cdots,N\}$, where $\theta_{d,n} \in [0,2\pi)$ and $\beta_n \in [0,1]$ denote phase shift and amplitude factor of the n-th reflected element on the IRS, respectively. Here, we assume that $\beta_n = 1, n = 1,\ldots,N$ to simplify the analysis. $\mathbf{\Theta}_d = \text{diag}\left(\boldsymbol{\varphi}_d\right), d \in \{1,2\}$ denotes the phase-shift matrix of the IRS, where $\mathbf{\Theta}_1$ and $\mathbf{\Theta}_2$ represent the phase-shift matrices of the IRS in the DL and the UL, respectively. It is assumed that the signal reflected more than once by the IRS is very weak and can be ignored due to the substantial path loss. Here, we assume that the transmission channels corresponding to different transceivers in the system are independent. The baseband equivalent channels of the HAP to the IRS, the IRS to the WD$_k$ and the HAP to the WD$_k$ are expressed as $\mathbf{T} \in \mathbb{C}^{N\times M}$, $\mathbf{r}_k \in \mathbb{C}^{N\times 1}$ and $\mathbf{h}_k \in \mathbb{C}^{M\times 1}, \forall k \in \mathcal{K}$, respectively. We consider that all WDs do not have a constant energy source and therefore require energy from the signal broadcasted by the HAP to transmit their data. Furthermore, we assume that the WDs are willing to share data with each other. Then, the WDs simultaneously transmit the shared data to the HAP through a virtual antenna array. The equivalent virtual MIMO channels of the WD to the HAP and the WD to the IRS links are expressed as $\mathbf{H} \in \mathbb{C}^{M\times K}$ and $\mathbf{R} \in \mathbb{C}^{N\times K}$, respectively.

Figure 2 shows the frame structure of the multi-user IRS-assisted WPCN system, in which the colored box with the solid line denotes the signal transmission, and the white box with the dotted line denotes the signal reception. Specifically, there are three phases: 1) DL energy transmission phase; 2) Data sharing phase; 3) UL information transmission phase. In Fig. 2, $\tau_1 T$, $\tau_2 T$ and $\tau_3 T$ represent the duration of each phase, respectively. $0 < \tau_i < 1, i = 1,2,3$. T is the total running time of the system. Generally, we assume $T = 1$ in this paper and $\tau_1 + \tau_2 + \tau_3 = 1$. In the following sections, we will explain each stage in detail.

2.1 Phase I: DL Energy Transmission

During the time τ_1 at Phase I, the WDs receive the energy signals transmitted by the HAP and obtain energy from them. The transmit signal at the HAP is expressed as

$$\begin{aligned}\mathbf{x}^{\text{DL}} &= \mathbf{V}\mathbf{s}^{\text{DL}} \\ &= \sum_{l=1}^{L} \mathbf{v}_l s_l^{\text{DL}},\end{aligned} \tag{1}$$

where L denotes the number of energy beams, $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_L] \in \mathbb{C}^{M\times L}$ is the active beamforming matrix at the HAP and $\mathbf{v}_l \in \mathbb{C}^{M\times 1}$ is the l-th energy beam, $\mathbf{s}^{\text{DL}} = \left[s_1^{\text{DL}}, s_2^{\text{DL}}, \ldots, s_L^{\text{DL}}\right]^{\text{T}}$, where s_l^{DL} is the l-th energy signal and $\mathbb{E}\left[\left|s_l^{\text{DL}}\right|^2\right] = 1$.

The transmit power is limited by

$$\sum_{l=1}^{L} \|\mathbf{v}_l\|^2 \le P_0, \tag{2}$$

where P_0 denotes the maximum transmit power of the HAP.

Then the received signal at the WD$_k$ is

$$y_k = \mathbf{b}_k^{\text{H}} \mathbf{x}^{\text{DL}} + z_k^{\text{DL}}, \forall k \in \mathcal{K}, \tag{3}$$

where $\mathbf{b}_k^{\text{H}} = \mathbf{r}_k^{\text{H}} \mathbf{\Theta}_1 \mathbf{T} + \mathbf{h}_k^{\text{H}}$ represents the composite channel of the HAP to the WD$_k$ in the DL, $z_k^{\text{DL}} \sim \mathcal{CN}\left(0, \sigma^2\right)$ represents the additive white gaussian noise (AWGN) at the WD$_k$.

Here, we assume that the energy collected by the WD increases with the power of the baseband signal received from the HAP. During the time of τ_1, the energy collected by the WD$_k$ is expressed as

$$\begin{aligned}E_k &= \tau_1 \eta \mathbb{E}\left\{|y_k|^2\right\} \\ &= \tau_1 \eta \left(\sum_{l=1}^{L} \left\|\mathbf{b}_k^{\text{H}} \mathbf{v}_l\right\|^2 + \sigma^2\right),\end{aligned} \tag{4}$$

where $0 < \eta \le 1$ represents the energy conversion efficiency.

2.2 Phase II: Data Sharing

During the time of τ_2 in Phase II, all WDs share information with each other so that they can transmit information to the HAP via virtual MIMO in the third stage. Each WD has data to transfer to the HAP. The adjacent WDs exchange data by continuously broadcasting signals to all other WDs in a time division multiple access (TDMA) manner. The data sharing time is divided into K periods equally, and each period can be expressed as τ_2/K. Let E_S denote the energy consumed by each WD for data exchange. We can consider that in the data sharing phase, the amount of information exchanged by the WDs will increase with the increase of E_S, but the throughput performance in the information transmission phase will decrease due to the insufficient transmit power of WDs. Therefore, it is necessary to optimize the E_S to maximize the UL throughput. Let the baseband equivalent channel between the k-th and the j-th WD be denoted by $b_{k,j}$. In order for all other WDs to successfully decode the information [18], the transmission rate of WD$_k$ is determined by the minimum channel gain $\left|b_{k,min}\right|^2$, where $\left|b_{k,\min}\right|^2 = \min_{j\neq k} \left|b_{k,j}\right|^2$. We assume that only the WDs close to each other can cooperate. Therefore, for simplicity, we assume $\left|b_{k,min}\right|^2 = \xi$, then in the data sharing phase, the sum rate of all WDs is expressed as

$$R^{\mathrm{DS}} = \sum_{k=1}^{K} \frac{\tau_2}{K} \log_2 \left(1 + \frac{KE_S\xi}{\tau_2\sigma^2}\right) = \tau_2 \log_2 \left(1 + \frac{KE_S\xi}{\tau_2\sigma^2}\right). \tag{5}$$

2.3 Phase III: UL Information Transmission

During the time of τ_3 at Phase III, K WDs cooperatively transmit a signal vector $\mathbf{x}^{\mathrm{UL}} \in \mathbb{C}^{K\times 1}$, which is defined by

$$\mathbf{x}^{\mathrm{UL}} = \mathbf{W}\mathbf{s}^{\mathrm{UL}}, \tag{6}$$

where $\mathbf{W} \in \mathbb{C}^{K\times K}$ is defined as the beamforming matrix in the UL, $\mathbf{s}^{\mathrm{UL}} = \left[s_1^{\mathrm{UL}}, s_2^{\mathrm{UL}}, \ldots, s_K^{\mathrm{UL}}\right]^{\mathrm{T}}$, and s_k^{UL} is the information signal of the WD_k shared during Phase II. We consider that $\mathbf{W}$ is calculated at the HAP and broadcasted to the WDs in the UL. The k-th element in $\mathbf{x}^{\mathrm{UL}}$ denotes the transmit signal at the WD_k which is expressed as

$$x_k^{\mathrm{UL}} = \sum_{i=1}^{K} [\mathbf{W}]_{k,i} s_i^{\mathrm{UL}}. \tag{7}$$

Within the time of the UL information transmission in Phase III, all WDs use the remaining energy given by $E_k - E_S - E_k^c$ to transmit the shared data to the HAP through the virtual MIMO channel, where $E_k^c \geq 0$ denotes the circuit energy consumption at the WD_k which is assumed to be constant [19]. The transmit power of the WD_k is given by

$$P_k = \frac{E_k - E_S - E_k^c}{\tau_3}, \forall k \in \mathcal{K}. \tag{8}$$

We use $\mathbf{Q} = \mathbb{E}\left[\mathbf{x}^{\mathrm{UL}} \left(\mathbf{x}^{\mathrm{UL}}\right)^{\mathrm{H}}\right]$ to represent the transmit covariance matrix of $\mathbf{x}^{\mathrm{UL}}$. So the constraint on each diagonal element of $\mathbf{Q}$ is given by

$$[\mathbf{Q}]_{k,k} \leq P_k, \forall k \in \mathcal{K}. \tag{9}$$

After data sharing, the received signal at the HAP in the UL is

$$\mathbf{y}^{\mathrm{UL}} = (\mathbf{T}\boldsymbol{\Theta}_2\mathbf{R} + \mathbf{H})\,\mathbf{x}^{\mathrm{UL}} + \mathbf{z}^{\mathrm{UL}}, \tag{10}$$

where $\mathbf{z}^{\mathrm{UL}} \sim \mathcal{CN}\left(0, \sigma^2\mathbf{I}_M\right)$ represents the AWGN at the HAP.

The throughput during Phase III can be expressed as

$$R^{\mathrm{UL}} = \tau_3 \log_2 \left|\mathbf{I}_M + \frac{(\mathbf{T}\boldsymbol{\Theta}_2\mathbf{R} + \mathbf{H})\,\mathbf{Q}\,(\mathbf{T}\boldsymbol{\Theta}_2\mathbf{R} + \mathbf{H})^{\mathrm{H}}}{\sigma^2}\right|. \tag{11}$$

3. Problem Formulation

According to the previous analysis, the throughput of Phase III is limited by the sum rate of data sharing in Phase II. With this constraint, we formulate an UL throughput maximization problem by jointly optimizing the DL and the UL active beamforming matrices $\{\mathbf{V}, \mathbf{Q}\}$, the DL/UL passive beamforming matrix for IRS (i.e., $\boldsymbol{\Theta}_1$ or $\boldsymbol{\Theta}_2$), the time allocation $\boldsymbol{\tau} = [\tau_1, \tau_2, \tau_3]$ and the energy E_S consumed during data sharing in this section. Mathematically, the optimization problem is expressed as

$$\begin{aligned}
\mathrm{P1}: \ & \max_{\boldsymbol{\tau},\mathbf{Q},\mathbf{V},\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2,E_S} \min\left\{R^{\mathrm{DS}}, R^{\mathrm{UL}}\right\} \\
\text{s.t.} \ & \sum_{l=1}^{L} \|\mathbf{v}_l\|^2 \leq P_0, \\
& [\mathbf{Q}]_{k,k} \leq P_k, \forall k \in \mathcal{K}, \\
& 0 < \tau_i < 1, i = 1,2,3, \\
& \tau_1 + \tau_2 + \tau_3 = 1, \\
& P_k \geq 0, \forall k \in \mathcal{K}, \\
& \left|\varphi_{d,n}\right| = 1, d \in \{1,2\}, \forall n \in \mathcal{N}.
\end{aligned} \tag{12}$$

Due to the coupled optimization variables and the constraint $\left|\varphi_{d,n}\right| = 1$ introduced by the IRS in the problem (P1), this problem is non-convex. Therefore, we decouple the optimization variables by the AO technique and then design a joint optimization algorithm to solve this optimization problem (P1) in the following.

4. Proposed Algorithm

Since R^{DS} increases with the increase of E_S, while R^{UL} decreases with the increase of E_S, it is easy to find the optimal E_S by one-dimensional search, such as the bisection method [20]. Then, we use the AO technique to divide the problem (P1) into four subproblems with the fixed E_S. Specifically, The problem (P1) is decomposed into the joint DL and UL active beamforming subproblem (optimizing $\{\mathbf{V}, \mathbf{Q}\}$), the DL passive beamforming of the IRS $\boldsymbol{\Theta}_1$ subproblem, the UL passive beamforming of the IRS $\boldsymbol{\Theta}_2$ subproblem, and the time allocation $\boldsymbol{\tau}$ subproblem, respectively. Then, we alternately optimize the four subproblems until the optimization problem converges.

4.1 Active Beamforming

First, we optimize $\{\mathbf{V}, \mathbf{Q}\}$ with given $\{\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2\}$ and $\boldsymbol{\tau}$. Define $\mathbf{G} = \mathbf{T}\boldsymbol{\Theta}_2\mathbf{R} + \mathbf{H}$, (11) can be reduced as

$$R^{\mathrm{UL}} = \tau_3 \log_2 \left|\mathbf{I}_M + \frac{\mathbf{G}\mathbf{Q}\mathbf{G}^{\mathrm{H}}}{\sigma^2}\right|. \tag{13}$$

For given E_S and $\boldsymbol{\tau}$, R^{DS} is fixed. Problem (P1) can be equivalently reformulated as

$$\begin{aligned}
\mathrm{P2}: \ & \max_{\mathbf{V},\mathbf{Q}} \tau_3 \log_2 \left|\mathbf{I}_M + \frac{\mathbf{G}\mathbf{Q}\mathbf{G}^{\mathrm{H}}}{\sigma^2}\right| \\
\text{s.t.} \ & \sum_{l=1}^{L} \|\mathbf{v}_l\|^2 \leq P_0, \\
& [\mathbf{Q}]_{k,k} \leq P_k, \forall k \in \mathcal{K}, \\
& P_k \geq 0, \forall k \in \mathcal{K}.
\end{aligned} \tag{14}$$

We use the Lagrange duality method to solve problem (P2) and formulate its Lagrangian function as

$$\mathcal{L}(\mathbf{V},\mathbf{Q},\mathbf{\Lambda}) = \tau_3 \log_2 \left| \mathbf{I}_M + \frac{\mathbf{G}\mathbf{Q}\mathbf{G}^{\mathrm{H}}}{\sigma^2} \right| + \sum_{k=1}^{K} \lambda_k \left(P_k - [\mathbf{Q}]_{k,k} \right), \tag{15}$$

where λ_k is the k-th Lagrange dual multiplier and $\mathbf{\Lambda} = \mathrm{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_K\}$. The Lagrange function in (15) can be reformulated as

$$\mathcal{L}(\mathbf{V},\mathbf{Q},\mathbf{\Lambda}) = \tau_3 \log_2 \left| \mathbf{I}_M + \frac{\mathbf{G}\mathbf{Q}\mathbf{G}^{\mathrm{H}}}{\sigma^2} \right| - \mathrm{tr}(\mathbf{\Lambda}\mathbf{Q}) + \sum_{k=1}^{K} \lambda_k \left(\frac{\tau_1 \eta \left(\sum_{l=1}^{L} \left\| \mathbf{b}_k^{\mathrm{H}} \mathbf{v}_l \right\|^2 + \sigma^2 \right) - E_S - E_k^c}{\tau_3} \right). \tag{16}$$

It can be seen from (16) that with a given $\mathbf{\Lambda}$, the problem is decomposed into two subproblems, i.e., the optimization subproblem of $\mathbf{V}$ and the optimization subproblem of $\mathbf{Q}$. When $\mathbf{Q}$ is given, the optimization subproblem of $\mathbf{V}$ is given by

$$\begin{aligned} \text{P2a:} \ \max_{\mathbf{V}} \ & \sum_{k=1}^{K} \lambda_k \left(\frac{\tau_1 \eta \left(\sum_{l=1}^{L} \left\| \mathbf{b}_k^{\mathrm{H}} \mathbf{v}_l \right\|^2 + \sigma^2 \right) - E_S - E_k^c}{\tau_3} \right) \\ \text{s.t.} \ & \sum_{l=1}^{L} \left\| \mathbf{v}_l \right\|^2 \le P_0. \end{aligned} \tag{17}$$

When $\mathbf{V}$ is given, the optimization subproblem of $\mathbf{Q}$ is given by

$$\begin{aligned} \text{P2b:} \ \max_{\mathbf{Q}} \ & \tau_3 \log_2 \left| \mathbf{I}_M + \frac{\mathbf{G}\mathbf{Q}\mathbf{G}^{\mathrm{H}}}{\sigma^2} \right| - \mathrm{tr}(\mathbf{\Lambda}\mathbf{Q}) \\ \text{s.t.} \ & [\mathbf{Q}]_{k,k} \le P_k, \forall k \in \mathcal{K}, \\ & P_k \ge 0, \forall k \in \mathcal{K}. \end{aligned} \tag{18}$$

Generally, we can apply the subgradient method to tackle the dual problem [21]. For our optimization problem, however, the subgradient method is infeasible because both $[\mathbf{Q}]_{k,k}$ and P_k change simultaneously, that is, P_k is not a fixed value. Therefore, another method is proposed.

4.1.1 DL Active Beamforming

Now, the optimal solution to problem (P2a) can be obtained, which is rewritten as

$$\begin{aligned} \text{P2}\bar{\text{a}}: \max_{\mathbf{V}} \ & \sum_{l=1}^{L} \mathbf{v}_l^{\mathrm{H}} \mathbf{\Omega} \mathbf{v}_l \\ \text{s.t.} \ & \sum_{l=1}^{L} \left\| \mathbf{v}_l \right\|^2 \le P_0, \end{aligned} \tag{19}$$

where $\mathbf{\Omega} = \sum_{k=1}^{K} \lambda_k \mathbf{b}_k \mathbf{b}_k^{\mathrm{H}}$, the weight of the channel $\mathbf{b}_k$ is denoted by the k-th Lagrangian dual variable λ_k. Since λ_k can be regarded as a marginal increment of the Lagrangian function for the k-th power constraint P_k, it is significant to optimize the beamforming vectors at the HAP to make the Lagrangian dual variable larger so that the device can receive more energy.

The optimal solution of problem P2$\bar{a}$ is expressed as

$$\begin{aligned} \mathbf{v}_1^{\star} &= \sqrt{P_0}\mathbf{u}_{\Omega}, \\ \mathbf{v}_l^{\star} &= \mathbf{0}, l = 2, 3, \ldots, L, \end{aligned} \tag{20}$$

where $\mathbf{u}_{\Omega}$ indicates the eigenvector that corresponds to the largest eigenvalue of $\mathbf{\Omega}$. According to (20), the transmit power of the WD$_k$ is given by

$$P_k = \frac{\tau_1 \eta \left(P_0 \left| \mathbf{b}_k^{\mathrm{H}} \mathbf{u}_{\Omega} \right|^2 + \sigma^2 \right) - E_S - E_k^c}{\tau_3}. \tag{21}$$

4.1.2 UL Active Beamforming

Then, this subsection will introduce the solution of the problem (P2b) in detail. If the beamforming vectors $\{\mathbf{v}_l\}$ are assumed to be fixed, P_k is fixed. When $\mathbf{V}$ is given, the problem (P2b) is rewritten as

$$\begin{aligned} \text{P2}\bar{\text{b}}: \max_{\mathbf{Q}} \ & \tau_3 \log_2 \left| \mathbf{I}_M + \frac{\mathbf{G}\mathbf{Q}\mathbf{G}^{\mathrm{H}}}{\sigma^2} \right| \\ \text{s.t.} \ & [\mathbf{Q}]_{k,k} \le P_k, \forall k \in \mathcal{K}. \end{aligned} \tag{22}$$

The above problem is to maximize the UL throughput with the power constraint of each antenna by optimizing the transmit covariance matrix $\mathbf{Q}$. The fixed-point iteration algorithm [22] can be applied to tackle it.

Note that the value of P_k keeps changing as the DL active beamforming $\mathbf{V}$. Therefore, if we directly combine the DL active beamforming $\mathbf{V}$ given by (20) and the UL active beamforming $\mathbf{Q}$ given by fixed-point iterative algorithm via an iterative approach, then this algorithm no longer works because it only converges under a fixed P_k. In order to solve this problem, a decreasing step size $\alpha = 1/i$ is introduced to adjust the transmit power vector smoothly at each iteration as

$$\mathbf{p}^{(i+1)} = \mathbf{p}^{(i)} + \alpha \left(\mathbf{p}_{\mathrm{new}}^{(i)} - \mathbf{p}^{(i)} \right), \tag{23}$$

where $\mathbf{p}^{(i)} = \left[P_1^{(i)}, P_2^{(i)}, \ldots, P_K^{(i)} \right]^{\mathrm{T}}$, $\mathbf{p}_{\mathrm{new}}^{(i)}$ and $P_k^{(i)}$ denote the transmit power vector, the transmit power vector calculated by (21), and the transmit power of the WD$_k$ respectively at the i-th iteration. By incorporating the DL active beamforming $\mathbf{V}$ calculated by (20) into the fixed-point algorithm and introducing the transmit power updated by (23) to ensure the convergence of each iterative solution, we design a joint DL and UL beamforming algorithm, which is interpreted in detail below.

First, we initialize $\mathbf{\Lambda}^{(0)} = \mathbf{I}_K$, where $\mathbf{\Lambda}^{(i)} = \mathrm{diag}\left\{ \lambda_1^{(i)}, \right.$

Algorithm 1 Joint DL and UL beamforming algorithm for problem (P2)

Input: $P_0, \eta, \mathbf{T}, \{\mathbf{h}_k, \mathbf{r}_k, \forall k \in \mathcal{K}\}, \mathbf{H}, \mathbf{R}$
Output: $\mathbf{V}^\star, \mathbf{Q}^\star$
1: Initialize $\mathbf{\Lambda}^{(0)} = \mathbf{I}_K$.
2: Set $\mathbf{p}^{(0)}$ to the solution in (21) with $\mathbf{\Lambda}^{(0)}$.
3: Set the iteration index $i = 1$.
4: Perform QR decomposition of $\mathbf{G}$: $\mathbf{G} = \mathbf{ZP}$, where $\mathbf{Z}$ and $\mathbf{P}$ are unitary matrix and upper triangular matrix, respectively.
5: **repeat**
6: Perform the singular value decomposition (SVD) of $\mathbf{P}\left(\mathbf{\Lambda}^{(i)}\right)^{-1/2}$: $\mathbf{P}\left(\mathbf{\Lambda}^{(i)}\right)^{-1/2} = \mathbf{U}^{(i)}\mathbf{\Sigma}^{(i)}\left(\mathbf{B}^{(i)}\right)^{\mathrm{H}}$, where $\mathbf{\Sigma}^{(i)} = \mathrm{diag}\{\varsigma_1, \varsigma_2, \ldots, \varsigma_r, 0, \ldots, 0\}$ and $r = \mathrm{rank}\left(\mathbf{P}\left(\mathbf{\Lambda}^{(i)}\right)^{-1/2}\right)$.
7: $\mathbf{\Phi}^{(i)} = \mathrm{diag}\left\{\left[1-\rho_1^{-1}\right]^+, \ldots, \left[1-\rho_r^{-1}\right]^+, 0, \ldots, 0\right\}$, where $\rho_j = \varsigma_j^2$.
8: $\mathbf{\Psi}^{(i)} = \mathbf{B}^{(i)}\left(\mathbf{I} - \mathbf{\Phi}^{(i)}\right)\left(\mathbf{B}^{(i)}\right)^{\mathrm{H}}$.
9: $\mathbf{Q}^{(i)} = \left(\mathbf{\Lambda}^{(i)}\right)^{-1} - \left(\mathbf{\Lambda}^{(i)}\right)^{-1/2}\mathbf{\Psi}^{(i)}\left(\mathbf{\Lambda}^{(i)}\right)^{-1/2}$.
10: $\tilde{\lambda}^{(i+1)} = \max\{\mathbf{p}^{(i)}, \mathbf{0}\} + \mathrm{diag}\{\mathbf{\Psi}^{(i)}\} \odot \tilde{\lambda}^{(i)}$.
11: $\mathbf{\Lambda}^{(i+1)} = \left(\mathrm{diag}\left(\tilde{\lambda}^{(i+1)}\right)\right)^{-1}$.
12: $\mathbf{\Omega}^{(i)} = \sum_{k=1}^{K}\lambda_k^{(i+1)}\mathbf{b}_k\mathbf{b}_k^{\mathrm{H}}$.
13: Calculate $\mathbf{v}_l^{(i)}, l = 1, 2, \ldots, L$ using (20), and obtain $\mathbf{V}^{(i)} = \left[\mathbf{v}_1^{(i)}, \mathbf{v}_2^{(i)}, \ldots, \mathbf{v}_L^{(i)}\right]$, where $\mathbf{v}_l^{(i)}$ is the l-th DL active beamforming vector at the i-th iteration.
14: Calculate the transmit power vector of each WD with $\mathbf{v}_l^{(i)}, l = 1, 2, \ldots, L$ using (21) and denote it by $\mathbf{p}_{\mathrm{new}}^{(i)}$.
15: $\mathbf{p}^{(i+1)} = \mathbf{p}^{(i)} + \alpha\left(\mathbf{p}_{\mathrm{new}}^{(i)} - \mathbf{p}^{(i)}\right)$.
16: $i = i + 1$.
17: **until** the objective value of (P2) converges.
18: **return** $\mathbf{V}^\star = \mathbf{V}^{(i)}$, $\mathbf{Q}^\star = \mathbf{Q}^{(i)}$

$\left\{\lambda_2^{(i)}, \ldots, \lambda_K^{(i)}\right\}$, $\lambda_k^{(i)}$ is the k-th Lagrangian dual variable at the i-th iteration. The initial transmit power vector $\mathbf{p}^{(0)}$ can be calculated by (21) when $\mathbf{\Lambda}^{(0)}$ is given. Similar to [22], the transmit covariance matrix $\mathbf{Q}^{(i)}$ with given $\mathbf{\Lambda}^{(i)}$ at the i-th iteration is given by

$$\mathbf{Q}^{(i)} = \left(\mathbf{\Lambda}^{(i)}\right)^{-1} - \left(\mathbf{\Lambda}^{(i)}\right)^{-1/2}\mathbf{\Psi}^{(i)}\left(\mathbf{\Lambda}^{(i)}\right)^{-1/2}. \tag{24}$$

The detailed explanation about $\mathbf{\Psi}^{(i)}$ is given in [22]. The Lagrangian dual variables are updated by

$$\tilde{\lambda}^{(i+1)} = \max\left\{\mathbf{p}^{(i)}, \mathbf{0}\right\} + \mathrm{diag}\left\{\mathbf{\Psi}^{(i)}\right\} \odot \tilde{\lambda}^{(i)}, \tag{25}$$

where $\tilde{\lambda}^{(i)} = \left[\left(\lambda_1^{(i)}\right)^{-1}, \left(\lambda_2^{(i)}\right)^{-1}, \ldots, \left(\lambda_K^{(i)}\right)^{-1}\right]^{\mathrm{T}}$. Under these updated Lagrangian dual variables, the DL active beamforming $\mathbf{V}$ is updated by (20), then the transmit power vector $\mathbf{p}_{\mathrm{new}}^{(i)}$ of the WDs in the UL is calculated by (21). Finally, the transmit power vector is updated by (23). This algorithm is summarized as Algorithm 1.

4.2 DL Passive IRS Beamforming

By analyzing the problem (P1), we can find that the effect of the IRS phase-shift matrix $\mathbf{\Theta}_1$ on the UL throughput is mainly related to the transmit power of the WDs due to the energy causality. Therefore, we can directly maximize the total energy collected by all WDs to find the optimal $\mathbf{\Theta}_1$. The final optimization problem of $\mathbf{\Theta}_1$ in the DL can be expressed as

$$\begin{aligned}\text{P3: } \max_{\mathbf{\Theta}_1} & \sum_{k=1}^{K}\tau_1\eta\left(\sum_{l=1}^{L}\left\|\left(\mathbf{r}_k^{\mathrm{H}}\mathbf{\Theta}_1\mathbf{T} + \mathbf{h}_k^{\mathrm{H}}\right)\mathbf{v}_l\right\|^2 + \sigma^2\right)\\ \text{s.t. } & [\mathbf{Q}]_{k,k} \le P_k, \forall k \in \mathcal{K},\\ & P_k \ge 0, \forall k \in \mathcal{K},\\ & \left|\varphi_{1,n}\right| = 1, \forall n \in \mathcal{N}.\end{aligned} \tag{26}$$

We optimize $\mathbf{\Theta}_1$ when $\{\mathbf{V}, \mathbf{Q}\}$ and $\boldsymbol{\tau}$ are given. Define $\boldsymbol{\zeta}_k = \mathrm{diag}\left(\mathbf{r}_k^{\mathrm{H}}\right)\mathbf{T} \in \mathbb{C}^{N\times M}, \forall k \in \mathcal{K}$, then

$$\begin{aligned}\mathbf{r}_k^{\mathrm{H}}\mathbf{\Theta}_1\mathbf{T} + \mathbf{h}_k^{\mathrm{H}} &= \boldsymbol{\varphi}_1^{\mathrm{H}}\,\mathrm{diag}\left(\mathbf{r}_k^{\mathrm{H}}\right)\mathbf{T} + \mathbf{h}_k^{\mathrm{H}}\\ &= \boldsymbol{\varphi}_1^{\mathrm{H}}\boldsymbol{\zeta}_k + \mathbf{h}_k^{\mathrm{H}}.\end{aligned} \tag{27}$$

In order to resolve the non-convex modulus constraint in the problem (P3), we define $\overline{\boldsymbol{\varphi}}_1^{\mathrm{H}} = \left[\boldsymbol{\varphi}_1^{\mathrm{H}}, 1\right] \in \mathbb{C}^{1\times(N+1)}$, $\bar{\boldsymbol{\zeta}}_k = \left[\boldsymbol{\zeta}_k; \mathbf{h}_k^{\mathrm{H}}\right] \in \mathbb{C}^{(N+1)\times M}$ and $\widetilde{\boldsymbol{\varphi}}_1 = \overline{\boldsymbol{\varphi}}_1\overline{\boldsymbol{\varphi}}_1^{\mathrm{H}} \in \mathbb{C}^{(N+1)\times(N+1)}$, then

$$\begin{aligned}\left\|\left(\mathbf{r}_k^{\mathrm{H}}\mathbf{\Theta}_1\mathbf{T} + \mathbf{h}_k^{\mathrm{H}}\right)\mathbf{v}_l\right\|^2 &= \left\|\left(\boldsymbol{\varphi}_1^{\mathrm{H}}\boldsymbol{\zeta}_k + \mathbf{h}_k^{\mathrm{H}}\right)\mathbf{v}_l\right\|^2\\ &= \left\|\overline{\boldsymbol{\varphi}}_1^{\mathrm{H}}\bar{\boldsymbol{\zeta}}_k\mathbf{v}_l\right\|^2\\ &= \mathrm{tr}\left(\bar{\boldsymbol{\zeta}}_k\mathbf{v}_l\mathbf{v}_l^{\mathrm{H}}\bar{\boldsymbol{\zeta}}_k^{\mathrm{H}}\widetilde{\boldsymbol{\varphi}}_1\right).\end{aligned} \tag{28}$$

Therefore, the problem (P3) can be reduced to

$$\begin{aligned}\text{P3a: } \max_{\boldsymbol{\varphi}_1} & \sum_{k=1}^{K}\tau_1\eta\left(\sum_{l=1}^{L}\mathrm{tr}\left(\bar{\boldsymbol{\zeta}}_k\mathbf{v}_l\mathbf{v}_l^{\mathrm{H}}\bar{\boldsymbol{\zeta}}_k^{\mathrm{H}}\widetilde{\boldsymbol{\varphi}}_1\right) + \sigma^2\right)\\ \text{s.t. } & [\mathbf{Q}]_{k,k} \le \frac{\tau_1\eta\left(\sum_{l=1}^{L}\mathrm{tr}\left(\bar{\boldsymbol{\zeta}}_k\mathbf{v}_l\mathbf{v}_l^{\mathrm{H}}\bar{\boldsymbol{\zeta}}_k^{\mathrm{H}}\widetilde{\boldsymbol{\varphi}}_1\right) + \sigma^2\right) - E_S - E_k^c}{\tau_3},\\ & P_k \ge 0, \forall k \in \mathcal{K},\\ & [\widetilde{\boldsymbol{\varphi}}_1]_{n,n} = 1, n = 1, \ldots, N+1,\\ & \mathrm{rank}\,(\widetilde{\boldsymbol{\varphi}}_1) = 1,\\ & \widetilde{\boldsymbol{\varphi}}_1 \ge 0,\end{aligned} \tag{29}$$

where $[\widetilde{\boldsymbol{\varphi}}_1]_{n,n} = 1, n = 1, \ldots, N+1$ holds from the modulus constraint of $\varphi_{1,n}$. Furthermore, due to $\widetilde{\boldsymbol{\varphi}}_1 = \overline{\boldsymbol{\varphi}}_1\overline{\boldsymbol{\varphi}}_1^{\mathrm{H}} \in \mathbb{C}^{(N+1)\times(N+1)}$, $\widetilde{\boldsymbol{\varphi}}_1$ must satisfy $\mathrm{rank}\,(\widetilde{\boldsymbol{\varphi}}_1) = 1, \widetilde{\boldsymbol{\varphi}}_1 \ge 0$. Due to the non-convex constraint $\mathrm{rank}\,(\widetilde{\boldsymbol{\varphi}}_1) = 1$, problem (P3a) is non-convex. The classical SDR technique coupled with the GR method can be used to tackle this problem with rank-1 constraint. The main principle of the GR method [23] is to randomly generate various possible solutions and then select the solution with the best performance as the approximate solution. Therefore, according to the problem (P3a), the

specific way to find the optimal $\boldsymbol{\Theta}_1$ is to remove the rank-1 constraint represented by rank $(\widetilde{\boldsymbol{\varphi}}_1) = 1$. Problem (P3a) is a standard semi-definite programming (SDP) problem and can be efficiently solved by CVX [24]. The relaxation problem (P3a) normally does not yield a rank-1 solution, and thus, the standard GR method is applied to generate high-quality rank-1 solutions. We first perform eigenvalue decomposition on $\widetilde{\boldsymbol{\varphi}}_1^{\star}$, that is, $\widetilde{\boldsymbol{\varphi}}_1^{\star} = \overline{\mathbf{U}}\,\overline{\boldsymbol{\Sigma}}\,\overline{\mathbf{U}}^{\mathrm{H}}$, where $\overline{\mathbf{U}} \in \mathbb{C}^{(N+1)\times(N+1)}$ and $\overline{\boldsymbol{\Sigma}} \in \mathbb{C}^{(N+1)\times(N+1)}$ represent unitary matrix and diagonal matrix, respectively. And then we get a suboptimal solution $\overline{\boldsymbol{\varphi}}_1 = \overline{\mathbf{U}}\,\overline{\boldsymbol{\Sigma}}^{1/2}\mathbf{q}$, where $\mathbf{q} \sim \mathcal{CN}(0, \mathbf{I}_{N+1}) \in \mathbb{C}^{(N+1)\times 1}$ is a Gaussian random vector. By constructing $\overline{\boldsymbol{\varphi}}_1$ randomly multiple times and then substituting the $\widetilde{\boldsymbol{\varphi}}_1 = \overline{\boldsymbol{\varphi}}_1\overline{\boldsymbol{\varphi}}_1^{\mathrm{H}}$ into the original optimization problem (P3a) so that the obtained solution satisfies the constraints, we take the set of $\overline{\boldsymbol{\varphi}}_1$ that maximizes the objective function as the suboptimal solution. Finally, we get $\boldsymbol{\varphi}_1^{\star} = e^{j\arg\left([\overline{\boldsymbol{\varphi}}_1]_{(1:N)}/\overline{\boldsymbol{\varphi}}_{1,N+1}\right)}$, and the optimal $\Theta_1^{\star} = \mathrm{diag}\left(\boldsymbol{\varphi}_1^{\star}\right)$.

4.3 UL Passive IRS Beamforming

Then we optimize $\boldsymbol{\Theta}_2$ when $\{\mathbf{V}, \mathbf{Q}\}$, $\boldsymbol{\Theta}_1$ and $\boldsymbol{\tau}$ are given. The element-by-element iterative algorithm in [23] is considered to solve this optimization problem. Define $\alpha_n = [\boldsymbol{\Theta}_2]_{n,n}, \forall n \in \mathcal{N}$, then with given the other variables, the problem (P1) can be re-represented as

$$\begin{aligned} \mathrm{P4}: &\max_{\alpha_n} \log_2 \left|\mathbf{B}_n + \alpha_n \mathbf{A}_n + \alpha_n^* \mathbf{A}_n^{\mathrm{H}}\right| \\ &\text{s.t. } |\alpha_n| = 1, \forall n \in \mathcal{N}. \end{aligned} \tag{30}$$

First, we express $\mathbf{T}$ and $\mathbf{R}$ as $\mathbf{T} = [\mathbf{t}_1, \mathbf{t}_2, \cdots \mathbf{t}_N]$ and $\mathbf{R} = [\mathbf{r}_1, \mathbf{r}_2, \cdots \mathbf{r}_N]^{\mathrm{H}}$, respectively. Introducing new variables $\widetilde{\mathbf{Q}} = \tau_3 \mathbf{Q}$, $\mathbf{B}_n$ and $\mathbf{A}_n$ in the optimization problem (P4) are respectively defined as

$$\begin{aligned} \mathbf{B}_n = \mathbf{I}_M + \frac{1}{\sigma^2 \tau_3} \Bigg(&\left(\mathbf{H} + \sum_{m=1, m\neq n}^{N} \alpha_m \mathbf{t}_m \mathbf{r}_m^{\mathrm{H}}\right) \widetilde{\mathbf{Q}} \\ &\left(\mathbf{H} + \sum_{m=1, m\neq n}^{N} \alpha_m \mathbf{t}_m \mathbf{r}_m^{\mathrm{H}}\right)^{\mathrm{H}} + \mathbf{t}_n \mathbf{r}_n^{\mathrm{H}} \widetilde{\mathbf{Q}} \mathbf{r}_n \mathbf{t}_n^{\mathrm{H}} \Bigg), \end{aligned} \tag{31}$$

$$\mathbf{A}_n = \mathbf{t}_n \mathbf{g}_n^{\mathrm{H}}, \tag{32}$$

$$\mathbf{g}_n = \frac{1}{\sigma^2 \tau_3} \left(\mathbf{H} + \sum_{m=1, m\neq n}^{N} \alpha_m \mathbf{t}_m \mathbf{r}_m^{\mathrm{H}}\right) \widetilde{\mathbf{Q}} \mathbf{r}_n. \tag{33}$$

Although the problem (P4) is non-convex due to the unit modulus constraint, we can still determine its closed-form solution by the rank-1 attribute and positive definiteness of the matrices $\mathbf{A}_n$ and $\mathbf{B}_n$. According to [25], the optimal solution of problem (P4) is summarized as

$$\alpha_n^{\star} = \begin{cases} e^{-\mathrm{j}\arg(\omega_n)} & \mathrm{tr}\left(\mathbf{B}_n^{-1}\mathbf{A}_n\right) \neq 0 \\ \forall \alpha \text{ with } |\alpha| = 1 & \mathrm{tr}\left(\mathbf{B}_n^{-1}\mathbf{A}_n\right) = 0 \end{cases}, \forall n \in N, \tag{34}$$

where $\omega_n = \mathrm{tr}\left(\mathbf{B}_n^{-1}\mathbf{A}_n\right), \forall n \in N$, then we can get the optimal $\boldsymbol{\Theta}_2$.

Algorithm 2 JO-BPT algorithm for problem (P1)

Input: $P_0, \eta, \mathbf{T}, \{\mathbf{h}_k, \mathbf{r}_k, \forall k \in \mathcal{K}\}, \mathbf{H}, \mathbf{R}$
Output: $\mathbf{V}^{\star}, \mathbf{Q}^{\star}, \boldsymbol{\Theta}_1^{\star}, \boldsymbol{\Theta}_2^{\star}, \boldsymbol{\tau}^{\star}$

1: Initialize $\boldsymbol{\Theta}_1^{(0)}, \boldsymbol{\Theta}_2^{(0)}, \boldsymbol{\tau}^{(0)}$.
2: Set the iteration index $i = 1$.
3: **repeat**.
4: Calculate $\mathbf{V}^{(i)}, \mathbf{Q}^{(i)}$ using algorithm 1 for given $\boldsymbol{\Theta}_1^{(i-1)}, \boldsymbol{\Theta}_2^{(i-1)}$, and $\boldsymbol{\tau}^{(i-1)}$.
5: Solve the problem (P3a) to get $\widetilde{\boldsymbol{\varphi}}_1^{(i)}$ for given $\mathbf{V}^{(i)}, \mathbf{Q}^{(i)}$ and $\boldsymbol{\tau}^{(i-1)}$, then obtain $\boldsymbol{\Theta}_1^{(i)}$ by using the SDR technique combined with GR method.
6: Solve the problem (P4) to get $\alpha_n^{(i)}, \forall n \in \mathcal{N}$ for given $\mathbf{Q}^{(i)}$ and $\boldsymbol{\tau}^{(i-1)}$, and then get $\boldsymbol{\Theta}_2^{(i)}$.
7: Solve the problem (P5) to get $\boldsymbol{\tau}^{(i)}$ for given $\mathbf{V}^{(i)}, \mathbf{Q}^{(i)}, \boldsymbol{\Theta}_1^{(i)}, \boldsymbol{\Theta}_2^{(i)}$.
8: Update $i = i + 1$.
9: **until** the objective value of (P1) converges.
10: **return** $\mathbf{V}^{\star} = \mathbf{V}^{(i)}, \mathbf{Q}^{\star} = \mathbf{Q}^{(i)}, \boldsymbol{\Theta}_1^{\star} = \boldsymbol{\Theta}_1^{(i)}, \boldsymbol{\Theta}_2^{\star} = \boldsymbol{\Theta}_2^{(i)}, \boldsymbol{\tau}^{\star} = \boldsymbol{\tau}^{(i)}$

4.4 Optimizing Time Allocation $\boldsymbol{\tau}$

Now, we optimize $\boldsymbol{\tau}$ for given $\mathbf{V}, \mathbf{Q}$ and $\{\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2\}$. The problem (P1) can be simplified to

$$\begin{aligned} \mathrm{P5}: &\max_{\boldsymbol{\tau}} \min\left\{R^{\mathrm{DS}}, R^{\mathrm{UL}}\right\} \\ &\text{s.t. } [\mathbf{Q}]_{k,k} \leq P_k, \forall k \in \mathcal{K}, \\ &\quad 0 < \tau_i < 1, i = 1, 2, 3, \\ &\quad \tau_1 + \tau_2 + \tau_3 = 1. \end{aligned} \tag{35}$$

The problem is convex and can be solved by CVX. Finally, it is easy to obtain the optimal $E_S^{\star}$ by one-dimensional search.

Considering the above description of each subproblem, we design the JO-BPT algorithm by jointly optimizing active beamforming, passive beamforming, transmit power and time allocation, which is summarized as Algorithm 2.

5. Simulation Results

In this section, we give the simulation results and analyses. In the simulation setting, we consider a scenario where multiple indoor WDs, located close to each other, monitor the process in production plants and report the sensed data to the HAP. We establish a three-dimensional coordinate system, where the HAP is positioned at (0 m, 0 m, 0 m), and the IRS is located at (1 m, 0 m, 1 m). K WDs are randomly distributed within a circle centered at (1 m, 0 m, 0 m) with a radius of 0.5 m. Moreover, the circuit energy consumption at the WD_k is set as $E_k^c = -5$ dBm, the energy conversion efficiency is set as $\eta = 0.8$ and the power of AWGN is set as $\sigma^2 = -50$ dBm. We assume that the space of antenna and IRS element is half wavelength. The distance path loss model is denoted by $L(s) = S_0\left(\frac{s}{1\mathrm{m}}\right)^{\epsilon}$, where $S_0 = -30$ dB is the path loss at the reference distance of 1 m, s represents the distance between two nodes and ϵ represents the path loss exponent. It is assumed that the antennas of the HAP and the array of IRS

are all employed in the uniform linear array (ULA).

For small-scale fading, we assume Rician fading model for all channels. As a result, the HAP-IRS channel $\mathbf{T}$ is modeled by

$$\mathbf{T} = \left(\sqrt{\frac{\kappa}{1+\kappa}}\mathbf{T}^{\mathrm{LoS}} + \sqrt{\frac{1}{1+\kappa}}\mathbf{T}^{\mathrm{NLoS}}\right), \tag{36}$$

where κ is the Rician factor, $\mathbf{T}^{\mathrm{LoS}}$ and $\mathbf{T}^{\mathrm{NLoS}}$ denote the line-of-sight (LoS) and the non-line-of-sight (NLoS) components, respectively. $\mathbf{T}^{\mathrm{LoS}}$ can be modeled as $\mathbf{T}^{\mathrm{LoS}} = \mathbf{a}_{\mathrm{HAP}}^{\mathrm{H}}(\theta_{\mathrm{HAP}})\,\mathbf{a}_{\mathrm{R}}(\theta_{\mathrm{R}})$, where $\mathbf{a}_{\mathrm{HAP}}(\theta_{\mathrm{HAP}})$ and $\mathbf{a}_{\mathrm{R}}(\theta_{\mathrm{R}})$ denote the array response vectors at the HAP and IRS, respectively. Moreover, $\mathbf{a}_{\mathrm{HAP}}(\theta_{\mathrm{HAP}}) = \left[1, e^{-j\pi\sin\theta_{\mathrm{HAP}}}, \cdots, e^{-j\pi(M-1)\sin\theta_{\mathrm{HAP}}}\right]$ and $\mathbf{a}_{\mathrm{R}}(\theta_{\mathrm{R}}) = \left[1, e^{-j\pi\sin\theta_{\mathrm{R}}}, \cdots, e^{-j\pi(N-1)\sin\theta_{\mathrm{R}}}\right]$, where θ_{HAP}, $\theta_{\mathrm{R}} \in [0, 2\pi)$ are the angle-of-departure (AoD) at the HAP and the angle-of-arrival (AoA) at the IRS, respectively. $\mathbf{T}^{\mathrm{NLoS}}$ is modeled by Rayleigh fading and each element $t_{n,m}^{\mathrm{NLoS}} \sim \mathcal{CN}(0,1), n = 1, \cdots N, m = 1, \cdots M$. The total channel gain is denoted by $\mathbf{T}\sqrt{L(s)}$. The channel matrices $\mathbf{r}_k$ and $\mathbf{h}_k$ are modeled similarly. For channels $\mathbf{T}$, $\mathbf{h}_k$ and $\mathbf{r}_k$, we set $\{\epsilon_{\mathrm{HAP,R}}, \epsilon_{\mathrm{HAP,U}}, \epsilon_{\mathrm{R,U}}\} = \{2.2, 2.3, 2.5\}$ and $\{\kappa_{\mathrm{HAP,R}}, \kappa_{\mathrm{HAP,U}}, \kappa_{\mathrm{R,U}}\} = \{\infty, 0, 0\}$.

Furthermore, we choose two representative benchmark schemes for performance comparison.

1. Random phase: In the two transmission phases of the system, the phase shift of each element on the IRS is randomly selected. Other parameters are optimized with the JO-BPT algorithm.
2. Without IRS: the WDs directly receive the energy transmitted by the HAP, then share data with each other, and finally transmit information signal directly to the HAP through a virtual MIMO channel without the help of IRS. Other parameters are optimized with the JO-BPT algorithm.

First, we consider the impact of the IRS passive beamforming matrix optimization on the UL throughput performance. Figure 3 displays the UL throughput versus the maximum transmit power (P_0) at the HAP, where $M = 2$, $K = 6$, $N = 20$. We observe that the UL throughput of these three schemes increases with increasing P_0. This is expected because the WDs can obtain more energy with the increase of the transmit power of the HAP. Furthermore, it is found that the proposed scheme considerably outperforms the other schemes, indicating that it utilizes the harvested energy efficiently. In addition, we can also note that even the scheme with randomly chosen phase shifts achieves higher UL throughput than the scheme without IRS because the IRS can provide additional array gain.

Figure 4 presents the UL throughput versus the number of reflection elements (N) contained in the IRS, where $M = 2$, $K = 6$, $P_0 = 30$ dBm. We observe that the UL throughput of the two IRS-assisted schemes increases significantly with the increase of N. This is expected because the increased IRS elements bring more cascaded channels, thus achieving higher

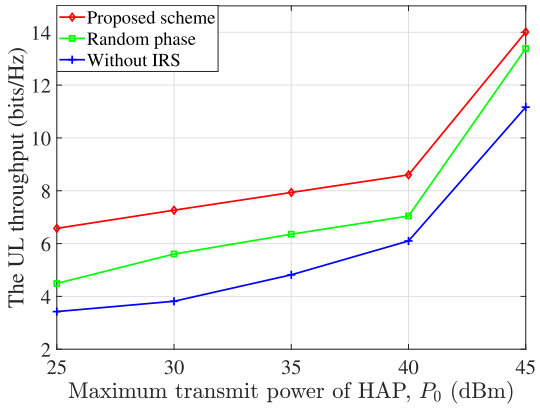


Fig. 3 Achievable UL throughput versus P_0.

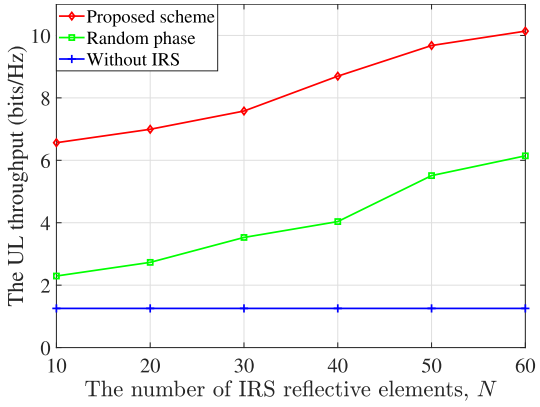


Fig. 4 Achievable UL throughput versus N.

channel gains in the IRS-assisted WPCN systems. Since we jointly optimize the passive beamforming for IRS and the active beamforming in the DL and the UL, the proposed scheme greatly improves the throughput performance compared with the other two benchmark schemes. The UL throughput of the proposed joint optimization scheme is 65.03% higher than that of the random phase scheme and 707.75% higher than that of the scheme without IRS, respectively, at $N = 60$.

Figure 5 shows the changes of the UL throughput in different schemes when the number of WDs (K) changes, where $M = 2$, $N = 20$, $P_0 = 30$ dBm. We can note that the UL throughputs under all schemes monotonically increase with the increasing WDs. Then, it can be found that the UL throughput performance of the proposed scheme is significantly better than the other two schemes. Moreover, we further observe that although the phase shifts are not optimized in the random phase scheme, the IRS can still provide a certain channel gain, making the UL throughput performance better than that of the scheme without IRS.

Furthermore, to verify the improvement of the HAP transmitter active beamforming design on the UL throughput performance in the proposed scheme, we introduce an equal gain beamforming scheme below and compare the proposed scheme with it. In the equal gain beamforming scheme, the HAP transmitter uses equal gain beamforming, and the

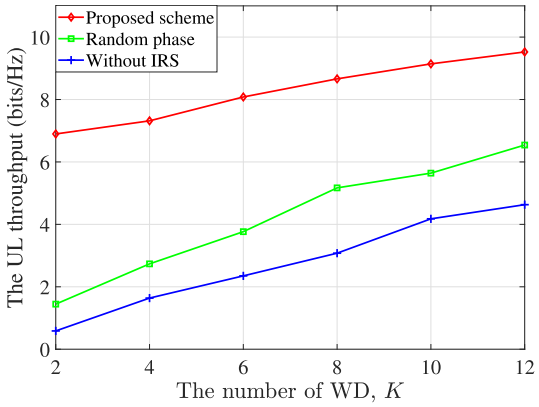


Fig. 5 Achievable UL throughput versus K.

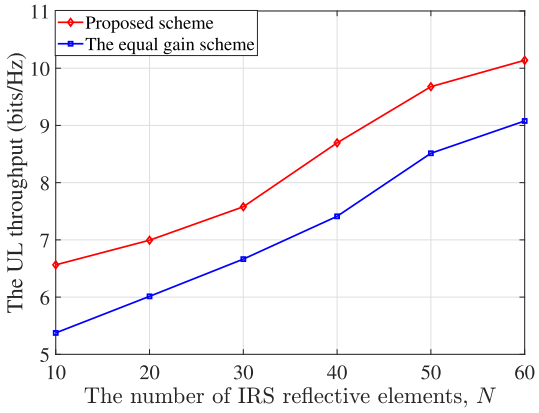


Fig. 6 Comparison of UL throughput performance of different beamforming schemes at the transmitter.

JO-BPT algorithm is used to optimize other variables, P_0 is equally allocated among K beamforming vectors. The equal gain beamforming scheme of the HAP can be specifically expressed as

$$\mathbf{v}_l = \sqrt{\frac{P_0}{K}} \frac{\mathbf{b}_l}{\|\mathbf{b}_l\|^2}, l = 1, 2, \ldots, K. \tag{37}$$

Figure 6 shows the UL throughput performance versus the number of IRS reflection elements (N) under the different beamforming schemes used at the HAP, where $M = 2$, $K = 6$, $P_0 = 30\,\text{dBm}$. Firstly, we can observe that the UL throughput performance is always proportional to N in all schemes. This is expected, with the increase of N, the reflection beamforming gain provided by IRS also increases gradually, and this can avoid signal attenuation in the DL and the UL. Moreover, we can see that the proposed HAP transmit beamforming gain design can achieve better throughput performance gain than the equal gain beamforming scheme.

In addition, the continuous phase shifts at IRS are difficult to implement in practical applications due to the hardware facilities and the cost. Therefore, the discrete IRS phase shift is discussed. Specifically, we compare the proposed scheme with three schemes where the quantized bit numbers of phase shifts of the IRS reflection elements are

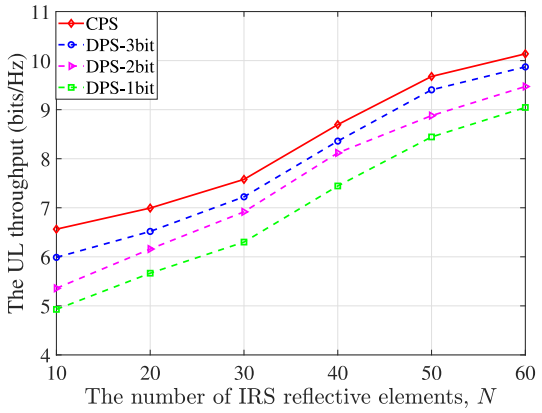


Fig. 7 Achievable UL throughput for continuous and discrete phase shifts versus N.

set to bit = 1, 2, 3, respectively. Actually, we consider that the discrete phase shift of each IRS reflection element is obtained by uniformly quantizing the interval $[0, 2\pi)$, and its value range is expressed as

$$\boldsymbol{\psi} = \{0, \Delta\theta, \ldots, (J-1)\Delta\theta\}, \tag{38}$$

where $\Delta\theta = 2\pi/J$, $J = 2^{\text{bit}}$.

The continuous phase shifting (CPS) scheme in Fig. 7 represents the optimization of the UL throughput by using the JO-BPT algorithm. The discrete phase shifting (DPS) scheme indicates that the optimal IRS reflection phase-shift vector $\boldsymbol{\varphi}^\star$ is first calculated using the JO-BPT algorithm proposed in this paper, and then we set $\boldsymbol{\psi}_n^\star = \arg\min_{m\in[1,J]} \left|\boldsymbol{\psi}_m - \boldsymbol{\varphi}_n^\star\right|, \forall n \in \mathcal{N}$, as the optimal discrete IRS reflection phase shift. The value range of $\boldsymbol{\psi}_m$ using a 3-bit phase shift is expressed as $\{0, \pi/4, \pi/2, 3\pi/4, \pi, 5\pi/4, 3\pi/2, 7\pi/4\}$.

Figure 7 shows the UL throughput performance trend versus the number of IRS reflection elements (N) under the CPS and the DPS scheme. As shown in Fig. 7, we can find that the DPS scheme suffers performance loss compared to the CPS scheme. However, as the number of phase-shift bits used in the system increases, performance improves. In addition, it is observed that the 3-bit phase shifting scheme achieves performance very close to the CPS scheme. Therefore, in practical scenarios, 3 bits of phase shift are enough. To reduce the cost of the system, we can further reduce the number of phase-shift bits, while causing performance loss. This is expected as lower phase-shift resolution will result in higher signal misalignment.

6. Conclusion

In this paper, we proposed an IRS-assisted WPCN system under user cooperation and formulated an UL throughput maximization problem by jointly optimizing the active beamforming in the DL and the UL, the passive beamforming of IRS, the time allocation and the energy consumed by data sharing. To tackle this non-convex optimization problem, we adopted the AO technique to decouple the optimization

variables and proposed an JO-BPT algorithm. By comparing with various baseline methods, we verified that the proposed JO-BPT algorithm has significant performance advantages under different real network settings, effectively improving the throughput performance of multi-user IRS-assisted WPCN. In further work, the strategy of grouping the devices will be considered when the difference between the channels of the WDs is large. Moreover, we will extend our method to systems with multiple IRSs and apply other more effective optimization strategies to reduce the complexity.

References

[1] J.A. Stankovic, "Research directions for the internet of things," IEEE Internet Things J., vol.1, no.1, pp.3–9, 2014.
[2] A. Tarighati, J. Gross, and J. Jaldén, "Decentralized hypothesis testing in energy harvesting wireless sensor networks," IEEE Trans. Signal Process., vol.65, no.18, pp.4862–4873, 2017.
[3] M. Hua, Q. Wu, and H.V. Poor, "Power-efficient passive beamforming and resource allocation for IRS-aided WPCNs," IEEE Trans. Commun., vol.70, no.5, pp.3250–3265, 2022.
[4] P. Chen, B. Lyu, Y. Liu, H. Guo, and Z. Yang, "Multi-irs assisted wireless-powered mobile edge computing for internet of things," IEEE IEEE Trans. Green Commun. Netw., vol.7, no.1, pp.130–144, 2023.
[5] M. Hua and Q. Wu, "Throughput maximization for IRS-aided MIMO FD-WPCN with non-linear EH model," IEEE J. Sel. Topics Signal Process., vol.16, no.5, pp.918–932, 2022.
[6] P. Zeng, D. Qiao, Q. Wu, and Y. Wu, "Throughput maximization for active intelligent reflecting surface-aided wireless powered communications," IEEE Wireless Commun. Lett., vol.11, no.5, pp.992–996, 2022.
[7] W. Wang, Y. Gong, L. Yang, Y. Zhan, and D.W.K. Ng, "Robust resource allocation design for secure IRS-aided WPCN," IEEE Trans. Wireless Commun., vol.22, no.4, pp.2715–2729, 2023.
[8] Q. Wu and R. Zhang, "Towards smart and reconfigurable environment: Intelligent reflecting surface aided wireless network," IEEE Commun. Mag., vol.58, no.1, pp.106–112, 2020.
[9] M. Hua and Q. Wu, "Joint dynamic passive beamforming and resource allocation for IRS-aided full-duplex WPCN," IEEE Trans. Wireless Commun., vol.21, no.7, pp.4829–4843, 2022.
[10] X. Li, C. Zhang, C. He, G. Chen, and J.A. Chambers, "Sum-rate maximization in IRS-assisted wireless power communication networks," IEEE Internet Things J., vol.8, no.19, pp.14959–14970, 2021.
[11] H. Cao, Z. Li, and W. Chen, "Resource allocation for IRS-assisted wireless powered communication networks," IEEE Wireless Commun. Lett., vol.10, no.11, pp.2450–2454, 2021.
[12] D. Song, W. Shin, and J. Lee, "A maximum throughput design for wireless powered communication networks with IRS-NOMA," IEEE Wireless Commun. Lett., vol.10, no.4, pp.849–853, 2021.
[13] D. Zhang, Q. Wu, M. Cui, G. Zhang, and D. Niyato, "Throughput maximization for IRS-assisted wireless powered hybrid NOMA and TDMA," IEEE Wireless Commun. Lett., vol.10, no.9, pp.1944–1948, 2021.
[14] L. Liu, R. Zhang, and K.C. Chua, "Multi-antenna wireless powered communication with energy beamforming," IEEE Trans. Commun., vol.62, no.12, pp.4349–4361, 2014.
[15] S. Jayaweera, "Virtual MIMO-based cooperative communication for energy-constrained wireless sensor networks," IEEE Trans. Wireless Commun., vol.5, no.5, pp.984–989, 2006.
[16] X. Zhang, H.V. Poor, and M. Chiang, "Optimal power allocation for distributed detection over MIMO channels in wireless sensor networks," IEEE Trans. Signal Process., vol.56, no.9, pp.4124–4140, 2008.
[17] D. Ciuonzo, G. Romano, and P.S. Rossi, "Channel-aware decision fusion in distributed MIMO wireless sensor networks: Decode-and-fuse vs. decode-then-fuse," IEEE Trans. Wireless Commun., vol.11, no.8, pp.2976–2985, 2012.
[18] M.N. Soorki, M.H. Manshaei, B. Maham, and H. Saidi, "On uplink virtual MIMO with device relaying cooperation enforcement in 5G networks," IEEE Trans. Mobile Comput., vol.17, no.1, pp.155–168, 2018.
[19] L. Liu, R. Zhang, and K.C. Chua, "Multi-antenna wireless powered communication with energy beamforming," IEEE Trans. Commun., vol.62, no.12, pp.4349–4361, 2014.
[20] D.V. Griffiths and I.M. Smith, Numerical Methods for Engineers, CRC Press, 2006.
[21] L. Xiao, M. Johansson, and S. Boyd, "Simultaneous routing and resource allocation via dual decomposition," IEEE Trans. Commun., vol.52, no.7, pp.1136–1144, 2004.
[22] T.M. Pham, R. Farrell, and L.N. Tran, "Revisiting the MIMO capacity with per-antenna power constraint: Fixed-point iteration and alternating optimization," IEEE Trans. Wireless Commun., vol.18, no.1, pp.388–401, 2019.
[23] S.P. Boyd and L. Vandenberghe, Convex Optimization, Cambridge University Press, 2004.
[24] Z.q. Luo, W.k. Ma, A.M.c. So, Y. Ye, and S. Zhang, "Semidefinite relaxation of quadratic optimization problems," IEEE Signal Process. Mag., vol.27, no.3, pp.20–34, 2010.
[25] S. Zhang and R. Zhang, "Capacity characterization for intelligent reflecting surface aided mimo communication," IEEE J. Sel. Areas Commun., vol.38, no.8, pp.1823–1838, 2020.

Yun Wu received her B.S. and M.S. degrees in electronic engineering from Harbin Institute of Technology, Harbin, China in 1999 and 2001, respectively, and Ph.D. degree in communication and information system from Shanghai Jiao Tong University, Shanghai, China in 2006. In 2014, She was a visiting scholar at University of Texas, Dallas, USA. She is currently an associate professor at College of Information Science and Technology, Donghua University, Shanghai, China. Her research interests include wireless communications, channel estimation and spatial modulation.

ZiHao Chen received his B.S. degree in electronic information science and technology from Guangzhou University, and is currently pursuing the M.S. degree at College of Information Science and Technology, Donghua University, Shanghai, China. His current research interests focus on mobile edge computing, reconfigurable smart surfaces.

MengYao Li received her B.S.degree in electronic information engineering from Jiangsu Normal University in 2020 and the M.S.degree in information and communication engineering from Donghua University in 2023. Her research interests include wireless communications, intelligent metasurfaces, and convex optimization.

Han Hai received his B.S. degree in information and computing sciences from South Central University for Nationalities, China, and the M.S. and the Ph.D. degrees in electronics engineering from Chonbuk National University, Korea, in 2013 and 2018, respectively. He is currently an associate professor at College of Information Science and Technology, Donghua University, Shanghai, China. His research interests include wireless communications, MIMO and spatial modulation.

IEICE TRANS. COMMUN., VOL.E107–B, NO.11 NOVEMBER 2024

PAPER

Analysis of Beamforming for OAM Communication Using Loop Antenna Arrays and Paraboloids

Akira SAITOU[†a)], **Kaito UCHIDA**[†], **Kanki KITAYAMA**[†], **Ryo ISHIKAWA**[†], ***Members***, ***and*** **Kazuhiko HONJO**[†], ***Fellow***

SUMMARY Analytical expression of transmission for the orbital angular momentum (OAM) communication using loop antenna arrays and paraboloids is derived to achieve a communication distance of 100 m. With the field distribution of the single "transformed OAM mode" radiated by a loop antenna, the collimated field by the transmitting paraboloid and its diffracted field are analytically derived. Effects of frequencies, sizes of paraboloids, and shifts of transmitting and receiving arrays from the focal planes are included. With the diffracted field distribution on the focal plane of the receiving paraboloid, transmission between the transmitting and receiving loop antennas is analytically estimated. It is shown that the transmission between the antennas with different OAM modes is null, but the transmission between the antennas with the same mode can be reduced. To clarify the mechanism of the reduction, factors of the reduction are quantitatively defined, and the explicit formulae are derived. Based on the analytical results, numerical estimation for a communication distance of 100 m is demonstrated, where the frequency, the focal length, and the size of the paraboloid are 150 GHz, 50 cm and 100 cm, respectively. Where both arrays are located on each focal plane, the transmission for the signal is more than −7.78 dB for eight kinds of OAM modes. The transmission is the least for the highest-order mode. The transmission loss is shown to be mitigated by optimizing the shifts of transmitting and receiving arrays from their focal planes. The loss is made almost even by exploiting the tradeoff of the improvement for the mode orders. The transmission is improved by 5.98 dB, to be more than −1.80 dB, by optimizing the shifts of the arrays.

key words: *OAM wave, diffraction, eigenmode, paraboloid, numerical calculation*

1. Introduction

A variety of multiplexing methods have been investigated to improve frequency usage efficiency, because frequency bands are precious and finite resources. Whereas the efficiency can be improved within the scope of the Shannon-Hartley theorem, the methods have been exploited only by sacrificing the exponentially increased power. While the MIMO multiplexing communication scheme relaxes the severe demand for the power, it requires complex signal processing to obtain spatially independent channels. On the other hand, the orbital angular momentum (OAM) communication method relaxes the complexity by exploiting the spatial orthogonality of the electromagnetic fields [1]–[7]. An important feature of the OAM eigenmode is its spatial distribution of $\exp(jm\phi)$ in the spherical or cylindrical coordinate system, where m denotes the index that is called the phase mode number or the magnetic quantum number. When different signal sequences are overlaid on n kinds of OAM eigenmodes, n-channel spatial multiplexing becomes possible due to the spatial orthogonality of $\exp(jm\phi)$. Thus, multiplexing in the same frequency band is enabled. In addition, the feature of the eigenmode transmission relaxes the demands for the complexity of signal-processing. As the OAM communication exploits the spatial distribution, the direction of the receiver is strictly specified. Thus, the OAM communication may be anticipated to provide a novel physical dimension of a multiplexing method for the fixed wireless communication.

To obtain pure OAM single-modes, various schemes have been proposed. In the frequency range of light, a spiral phase plate or computer-generated holograms have been used to realize the phase distribution of $\exp(jm\phi)$ [8]–[14], where the plane wave is transformed into the OAM wave.

In the microwave frequency range, a uniform circular array has often been used, where each antenna is stimulated for the phase to be shifted by 2π divided by integers related with the element number of the array [5], [7], [16]–[19]. Whereas the phase distribution of the OAM wave results only at discreet points physically, communication is enabled owing to the sampling theorem. The number of the discreet points can be increased by increasing the number of elements, where feasible modes are also increased. As its element radiation pattern may be arbitrary, flexible configurations of the array can be exploited. Signals are artfully distributed with the Butler Matrix. In addition, dual polarizations can be easily exploited. Thus, the data rate is more than 100 Gbit/s for a communication distance of 100 m, where 15 data streams are exploited at 40 GHz [20].

On the other hand, a continuous field distribution of $\exp(jm\phi)$ has been obtained with a circular travelling antenna in the microwave frequency range, where a cylindrical slot and a stimulation phase-shifted by 90 degrees are exploited [21], [22]. In addition, an OAM communication scheme using loop antennas and paraboloids has been proposed, where the antenna radiates an almost single "transformed OAM mode" directly by controlling the radius of the loop antenna [23], [24]. The transformed mode is a linear combination of $+m^{th}$ and $-m^{th}$ order OAM modes. This mode may be understood as a state in which two kinds of photons, one with an OAM of $m\hbar$ and another with an OAM of $-m\hbar$, are entangled. In addition, its independent mode can be radiated by rotating the port azimuth of the antenna

Manuscript received February 7, 2024.
Manuscript revised June 17, 2024.
Manuscript publicized August 1, 2024.

[†]The University of Electro-Communications, Chofu-shi, 182-8585 Japan.
a) E-mail: asaitou@uec.ac.jp

DOI: 10.23919/transcom.2024EBP3028

by $\pi/(2m)$, where the radiated field distribution is an independent linear combination of $\pm m^{th}$ order modes [25], [26]. Thus, the number of feasible modes is identical to that for the other OAM communication schemes. The mode unity is further improved by a reflector plane and port-azimuth control [27], [28]. An array of concentric loop antenna elements radiates plural single-mode waves concurrently, where each wave propagates independently in favor of the spatial orthogonality. For a receiver, the antenna receives only the mode. Thus, an ideal communication is enabled without the troublesome signal-processing, if the diffraction can be neglected for the collimated field.

Whereas the OAM communication has been shown to be effective in the lower frequency bands for short-range communication [23]–[28], higher frequency bands are desirable for long-range communication, such as 100 m in terms of mitigating the diffraction effect. Whereas the diffraction effect has been analyzed also for the uniform circular array scheme [29], [30] to increase the distance, the effect is more complex for the loop antenna scheme, because the received field should be identical to the radiated field distribution for ρ as well as ϕ. This requirement may be said to be a harmful influence of communication without the troublesome signal processing. Thus, to estimate the performance of this scheme, a more-detailed investigation of the diffraction is required. For long-range communication such as 100 m, a large paraboloid and a high frequency would be required, where a prohibitively long time is concerned for simulations.

Thus, the analytical expression of the diffraction of the transformed OAM wave is investigated in this paper. In Sect. 2, the field distribution is analytically derived for the collimation of the transformed OAM waves with a paraboloid. With the obtained distribution, the diffracted field distribution is analytically derived. In Sect. 3, the analytical expression of the transmission between the transmitting and receiving loop antennas is obtained. In addition, factors of the transmission loss are investigated to clarify the mechanism. With the obtained analytical expression, the field distributions and transmissions are numerically estimated for a simple case in which the transmitting and receiving arrays are on their focal planes in Sect. 4. In Sect. 5, improvement of the transmission is investigated by optimizing the shifts of the arrays away from their focal planes.

2. Collimation and Diffraction of Transformed OAM Wave

The m^{th}-order transformed OAM eigenmode radiated by the loop antenna with the planar reflector is given by (1) and (2), where the circumference of the antenna is adjusted to be the integral multiple of the wavelength [23], [27].

The spherical coordinates are expressed by r', θ' and ϕ', and d denotes the distance between the loop antenna and the planar reflector. The assumed time convention is $\exp(-j\omega t)$. The radiated power is normalized as follows, where η and R denote the wave impedance in free space and the radius of the sphere for the integration, respectively:

$$\mathrm{P_{rad}} = \frac{1}{2\eta}\int_0^{2\pi} d\phi' \int_0^{\frac{\pi}{2}} d\theta' \{R^2 \mathbf{E}_i^{(m)} \cdot \mathbf{E}_i^{(m')^*}\} = \delta_{mm'} \tag{3}$$

Where the center of the loop antenna is shifted by z_{0t} along the Z-axis, the far field $\boldsymbol{E}_1$ is estimated with the Fraunhofer approximation as follows:

$$\mathbf{E}_1 = \exp\{-jkz_{0t}\cos\theta'\}\mathbf{E}_i^{(m)} \tag{4}$$

The far field, shown in (1) and (4), may be interpreted to be radiated from the origin, because it is proportional to $e^{jkr'}/r'$. In this case, the dependence of the magnitude and phase on θ' and ϕ' is given by the equations.

The collimated field of the OAM wave by the paraboloid is analyzed as shown in Fig. 1. The focal point is located at the origin, and the focal length is denoted by f. To collimate all the radiated power, the radius is selected to be $2f$.

Where the collimated field on the focal plane (XY plane) is expressed with a cylindrical coordinate (ρ, ϕ, z), the transformation by the paraboloid is expressed as follows:

$$\phi' = \phi \tag{5}$$

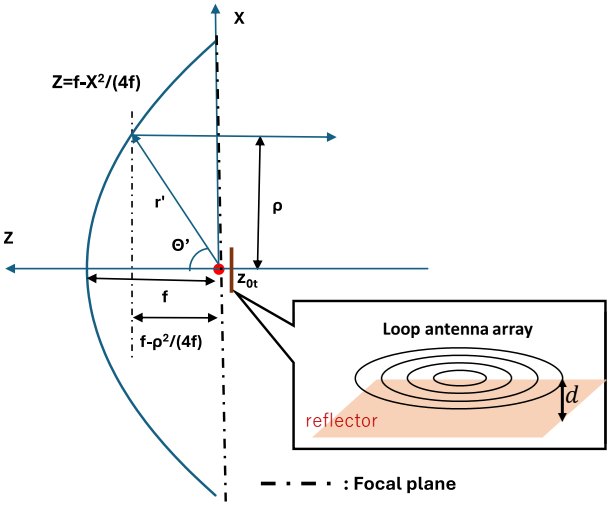


Fig. 1 Configuration of collimation with transmitting paraboloid.

$$\mathbf{E}_i^{(m)} = \frac{(-j)^{m-1}\sqrt{\frac{2\eta}{\pi}}\sin(kd\cos\theta')}{A(m)}\frac{e^{jkr'}}{r'}\begin{bmatrix}\approx 0\\ \cot\theta' J_m(m\sin\theta')\sin m\phi'\\ \frac{dJ_m(x)}{dx}|_{x=m\sin\theta'}\cos m\phi'\end{bmatrix} \quad m \geq 0,\ \ 0 \leq \theta' \leq \frac{\pi}{2} \tag{1}$$

$$A(m) \equiv \sqrt{\int_0^{\frac{\pi}{2}}\left[\sin^2(kd\cos\theta')\{\cot^2\theta'(J_m(m\sin\theta'))^2 + (\frac{dJ_m(m\sin\theta')}{dx})^2\}\right]\sin\theta' d\theta'} \tag{2}$$

$$\sin\theta' = \frac{\rho}{\sqrt{\rho^2 + \left(f - \frac{\rho^2}{4f}\right)^2}} \qquad 0 \le \theta' \le \frac{\pi}{2} \tag{6}$$

The collimated field $\boldsymbol{E}_{\rm ap}$ has been given as follows [31], [32], where the source is at the focal point:

$$\begin{aligned}\mathbf{E}_{\rm ap} &= \begin{bmatrix} \left[\mathbf{E}_{ap}(\rho,\phi,0)\right]_\rho \\ \left[\mathbf{E}_{ap}(\rho,\phi,0)\right]_\phi \\ \left[\mathbf{E}_{ap}(\rho,\phi,0)\right]_z \end{bmatrix} \\ &= -\left(\sqrt{\frac{\varepsilon}{\mu}}\frac{\mathrm{P}_t}{2\pi}\right)^{1/2} \frac{e^{-2jkf}}{\sqrt{\rho^2+f^2}} \begin{bmatrix} \mathrm{G}_\theta(\theta',\phi')^{1/2} \\ \mathrm{G}_\phi(\theta',\phi')^{1/2} \\ 0 \end{bmatrix} \end{aligned} \tag{7}$$

Whereas the loop antenna array is shifted along the Z-axis in the case of (4), the source may be considered to be at the focal point. Thus, the collimated fields $\boldsymbol{E}_{ap}$ of the OAM wave can be estimated with (1), (4) and (7). The gains for the cylindrical coordinate are estimated with (1), (4), (5), and (6). Thus, the collimated field is given by (8).

The collimated field expands by the diffraction, and the diffracted field $\boldsymbol{E}(\rho,\phi,z,m)$ is given by (9) [33], where (ρ,ϕ,z) and (ρ',ϕ',z') denote the coordinates of the observation point and the aperture points on the transmitting focal plane for the cylindrical coordinate system, respectively.

The diffracted field on the receiving focal plane is obtained by substituting (8) for the electric field of (9). In addition, the integration for ϕ' can be analytically obtained, as shown in (10). Thus, the diffracted field is given by the integration only for ρ', as shown in (11).

$$\left.\begin{aligned} &\int_0^{2\pi} \frac{\exp\left(ik\sqrt{\rho^2+\rho'^2-2\rho\rho'\cos(\phi-\phi')+z^2}\right)}{k\sqrt{\rho^2+\rho'^2-2\rho\rho'\cos(\phi-\phi')+z^2}} \sin m\phi' d\phi' \\ &\qquad = g(\rho',\rho,z)\sin m\phi \\ &\int_0^{2\pi} \frac{\exp\left(ik\sqrt{\rho^2+\rho'^2-2\rho\rho'\cos(\phi-\phi')+z^2}\right)}{k\sqrt{\rho^2+\rho'^2-2\rho\rho'\cos(\phi-\phi')+z^2}} \cos m\phi' d\phi' \\ &\qquad = g(\rho',\rho,z)\cos m\phi \\ &g(\rho',\rho,z) \equiv \int_0^{2\pi} \cos m\phi' \\ &\qquad \times \frac{\exp\left(ik\sqrt{\rho^2+\rho'^2-2\rho\rho'\cos(\phi-\phi')+z^2}\right)}{k\sqrt{\rho^2+\rho'^2-2\rho\rho'\cos(\phi-\phi')+z^2}} d\phi' \end{aligned}\right\} \tag{10}$$

$$\begin{aligned}\mathbf{E}(\rho,\phi,z,m) = \frac{2jk^2e^{2jkf}\sqrt{\eta}}{(2\pi)^{3/2}A(m)} \times \\ \begin{bmatrix} \sin m\phi \int_0^{2f} \exp\left[\frac{-jkz_0\left(f-\frac{\rho'^2}{4f}\right)}{\sqrt{\rho'^2+\left(f-\frac{\rho'^2}{4f}\right)^2}}\right] f_1(\rho')g(\rho',\rho,z)d\rho' \\ \cos m\phi \int_0^{2f} \exp\left[\frac{-jkz_0\left(f-\frac{\rho'^2}{4f}\right)}{\sqrt{\rho'^2+\left(f-\frac{\rho'^2}{4f}\right)^2}}\right] f_2(\rho')g(\rho',\rho,z)d\rho' \\ 0 \end{bmatrix}\end{aligned} \tag{11}$$

3. Transmission between Loop Antennas

The transmission is estimated for a communication distance of z_c, where z_c is defined as the distance between the focal planes of the transmitting and receiving paraboloids. For the receiver, the center of the loop antenna is shifted by z_{0r} along the Z-axis, as shown in Fig. 2. The structures of both the loop antenna array and the paraboloid are identical to those for the transmitter.

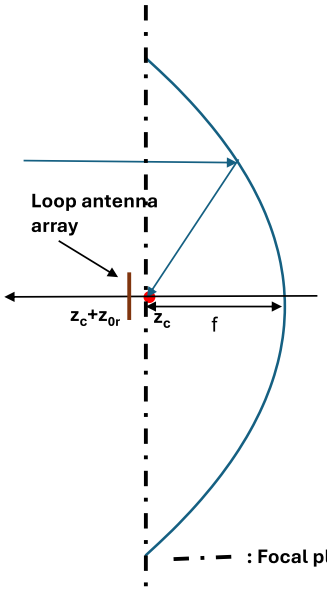


Fig. 2 Configuration of receiver consisting of array and paraboloid.

$$\mathbf{E}_{\rm ap}(\rho,\phi,m) = -\frac{\exp\left(2jkf\sqrt{\frac{2\eta}{\pi}}\right)}{A(m)} \exp\left[\frac{jkz_{0r}\left(f-\frac{\rho^2}{4f}\right)}{f+\frac{\rho^2}{4f}}\right] \begin{bmatrix} f_1(\rho)\sin m\phi \\ f_2(\rho)\cos m\phi \\ 0 \end{bmatrix} \qquad 0 \le \rho \le 2f \tag{8}$$

$$f_1(\rho) = \sin\left[\frac{kd\left(f-\frac{\rho^2}{4f}\right)}{f+\frac{\rho^2}{4f}}\right] \frac{f-\frac{\rho^2}{4f}}{\rho\left(f+\frac{\rho^2}{4f}\right)} J_m\left(\frac{m\rho}{f+\frac{\rho^2}{4f}}\right), \quad f_2(\rho) = \sin\left[\frac{kd\left(f-\frac{\rho^2}{4f}\right)}{f+\frac{\rho^2}{4f}}\right] \frac{1}{f+\frac{\rho^2}{4f}} \frac{J_{m-1}\left(\frac{m\rho}{f+\frac{\rho^2}{4f}}\right) - J_{m+1}\left(\frac{m\rho}{f+\frac{\rho^2}{4f}}\right)}{2}$$

$$\mathbf{E}(\rho,\phi,z,m) = \frac{k^2}{j2\pi}\int_0^{2f}\int_0^{2\pi} \begin{bmatrix} (\mathrm{E}_{\rm ap})_{\rho'} \\ (\mathrm{E}_{\rm ap})_{\phi'} \\ 0 \end{bmatrix} \frac{\exp\left[ik\sqrt{\rho^2+\rho'^2-2\rho\rho'\cos(\phi-\phi')+z^2}\right]}{k\sqrt{\rho^2+\rho'^2-2\rho\rho'\cos(\phi-\phi')+z^2}} \rho' d\rho' d\phi' \tag{9}$$

Considering the reciprocal behavior of the field and the current on the antenna, the received field distribution $\boldsymbol{E}_r$ on the receiving focal plane is proportional to (8), except for the phase caused by the shift of the array, as follows:

$$\mathbf{E_r}(\rho,\phi,m) = \sqrt{\frac{2\eta}{\pi}}\frac{1}{A(m)} \times \exp\left[\frac{jkz_{0r}\left(f-\frac{\rho^2}{4f}\right)}{f+\frac{\rho^2}{4f}}\right]\begin{bmatrix} f_1(\rho)\sin m\phi \\ f_2(\rho)\cos m\phi \\ 0 \end{bmatrix} \quad 0 \le \rho \le 2f \tag{12}$$

The constant phase over the focal plane is neglected. In a hypothetical case in which the positions for the transmitting and receiving focal planes are coincident, the magnitude of the transmission would be 0 dB for z_{0r} to be z_{0t}, which implies that the arrays should be shifted toward the same direction along the Z-axis.

As the field distribution on the receiving focal plane is given by (11), the ratio $T^{\text{total}}_{m'm}$ for the transmitted and received power is given by (13), where the OAM mode orders of the transmitting and receiving antennas are m and m', respectively.

The integration terms of the denominator are the total powers identical to the transmitting and receiving focal planes, which are identical to the radiated power shown in (3). Thus, the denominator of (13) is 1.

From (13), the differences of both the magnitude and phase for $\boldsymbol{E}_r$ and $\boldsymbol{E}$ affect the transmission. The numerator of (14) is proportional to the received power. Where the phase distributions of $\boldsymbol{E}$ and $\boldsymbol{E}_r$ are identical, it becomes the denominator. It should be noted that constant phase difference does not affect the value of (14). Thus, the effect of the phase distribution difference, T^{phase}, can be estimated as follows:

$$T^{\text{phase}} = \left|\frac{\int_0^{2\pi}\left[\int_0^{2f}(\mathbf{E}_r(\rho,\phi,m)^* \cdot \mathbf{E}(\rho,\phi,z_c,m))\rho d\rho\right]d\phi}{\int_0^{2\pi}\left[\int_0^{2f}|\mathbf{E}_r(\rho,\phi,m)^* \cdot \mathbf{E}(\rho,\phi,z_c,m)|\rho d\rho\right]d\phi}\right|^2 \tag{14}$$

The collection ratio by the receiving paraboloid, T^c, is given by the power portion inside the radius of the paraboloid ($2f$) for the diffracted field as follows:

$$T^c = \frac{\int_0^{2\pi}[\int_0^{2f}|\mathbf{E}(\rho,\phi,z_c)|^2\rho d\rho]d\phi}{\int_0^{2\pi}[\int_0^{\infty}|\mathbf{E}(\rho,\phi,z_c)|^2\rho d\rho]d\phi} \tag{15}$$

The residual loss of transmission, T^{mag}, is assumed to be the effect of the magnitude difference, as follows:

$$T^{\text{mag}} = T^{\text{phase}} \cdot T^c / T^{\text{total}} \tag{16}$$

By substituting (11) and (12) in (13)–(15), definite equations are obtained, as shown in (17)–(19).

$$T^{\text{total}}_{mm'} = \frac{k^4\pi^2}{8\eta}\delta_{mm'} \times \left|\int_0^{2f}\exp\left[\frac{-jkz_{0r}(f-\frac{\rho^2}{4f})}{f+\frac{\rho^2}{4f}}\right]\{f_1(\rho)f_3(\rho)+f_2(\rho)f_4(\rho)\}\rho d\rho\right|^2 \tag{17}$$

$$T^{\text{phase}} = \left|\frac{\int_0^{2f}\exp\left[\frac{-jkz_{0r}(f-\frac{\rho^2}{4f})}{f+\frac{\rho^2}{4f}}\right]\{f_1(\rho)f_3(\rho)+f_2(\rho)f_4(\rho)\}\rho d\rho}{\{\int_0^{2f}|f_1(\rho)f_3(\rho)+f_2(\rho)f_4(\rho)|\rho d\rho\}^2}\right|^2 \tag{18}$$

$$T^c = \frac{k^4}{8\pi\eta}\int_0^{2f}\left\{|f_3(\rho)|^2+|f_4(\rho)|^2\right\}\rho d\rho \tag{19}$$

$$\left.\begin{aligned} f_3(\rho) &= \int_0^{2f}\exp\left[\frac{-jkz_{0t}\left(f-\frac{\rho'^2}{4f}\right)}{f+\frac{\rho'^2}{4f}}\right]f_1(\rho')g(\rho',\rho,z_c)\rho' d\rho' \\ f_4(\rho) &= \int_0^{2f}\exp\left[\frac{-jkz_{0t}\left(f-\frac{\rho'^2}{4f}\right)}{f+\frac{\rho'^2}{4f}}\right]f_2(\rho')g(\rho',\rho,z_c)\rho' d\rho' \end{aligned}\right\} \tag{20}$$

Kronecker delta is denoted by $\delta_{mm'}$. From (17), the loop antennas of the other OAM modes receive no power, because the field distribution along ϕ of each OAM mode is maintained after the diffraction. The transmission loss is caused by the different field distribution along ρ. Thus, whereas interference waves are not generated, the signal wave can be attenuated for an OAM communication scheme using loop antennas. From (18), the effect of the phase difference can be partly mitigated by controlling the value of

$$\begin{aligned} T^{\text{total}}_{m'm} &= \left|\frac{\frac{1}{2\eta}\int_0^{2\pi}\left[\int_0^{2f}(\mathbf{E}_r(\rho,\phi,m')^* \cdot \mathbf{E}(\rho,\phi,z_c,m))\rho d\rho\right]d\phi}{\frac{1}{2\eta}\sqrt{\int_0^{2\pi}\left[\int_0^{2f}|\mathbf{E}_r(\rho,\phi,m')|^2\rho d\rho\right]d\phi \cdot \int_0^{2\pi}\left[\int_0^{\infty}|\mathbf{E}(\rho,\phi,z_c,m)|^2\rho d\rho\right]d\phi}}\right|^2 \\ &= \left|\frac{\int_0^{2\pi}\left[\int_0^{2f}(\mathbf{E}_r(\rho,\phi,m)^* \cdot \mathbf{E}(\rho,\phi,z_c,m))\rho d\rho\right]d\phi}{2\eta}\right|^2 \end{aligned} \tag{13}$$

z_{0r}.

4. Numerical Calculation for Transmitting and Receiving Arrays Located on Each Focal Plane

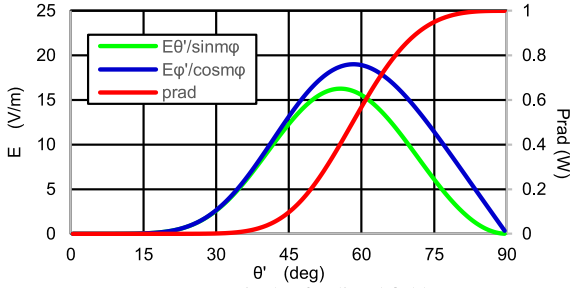
To estimate the proper values for the frequency and the focal length, simple configurations are numerically calculated with Mathematica™, where the transmitting and receiving arrays are located on each focal plane. To achieve a communication distance of 100 m for OAM modes up to the 8^{th} order, the values were selected to be 150 GHz and 50 cm, respectively, after preliminary calculations. The radius of the paraboloid is $2f$, as explained in Sect. 2.

As the diffraction effect is larger for higher-order modes, the calculated field distributions for the 8^{th}-order OAM mode are shown in Fig. 3 for a communication distance, z_c, of 100 m. Fig. 3 (a) shows the magnitudes of θ'- and ϕ'-polarized radiated fields and their accumulated power. Fig. 3 (b) shows the magnitudes of collimated ρ- and ϕ-polarized fields and their accumulated power up to ρ. Whereas the fields are dependent on $\sin m\phi$ or $\cos m\phi$, as shown in (8), the dependence is removed. Thus, they correspond to the magnitudes where $\sin m\phi$ or $\cos m\phi$ is ± 1. For arbitrary ϕ, the magnitudes are obtained by multiplying $|\sin m\phi|$ or $|\cos m\phi|$. As the radius of the paraboloid is 100 cm ($2f$), the accumulated power becomes identical to the radiated power at a ρ of 100 cm. The phase distribution is not shown, because it is constant over ρ, as shown in (8). Fig. 3(c) and (d) show the magnitudes and phases of the diffracted fields on the receiving focal plane, respectively. The positions of the maximum magnitudes move about 20 cm outward by the diffraction, and the accumulated power up to the radius of the receiving paraboloid becomes slightly less than 1 W. Whereas the phase is constant at the transmitting focal plane, the phases increase with ρ by the diffraction. From (8) and (12), the magnitude of the received field distribution is identical to the collimated field shown in Fig. 3(b). In addition, as the receiving array is on its focal plane, the phase distribution of the received field is constant. Thus, the differences of the phase distributions for T^{phase} are identical to those shown in Fig. 3(d).

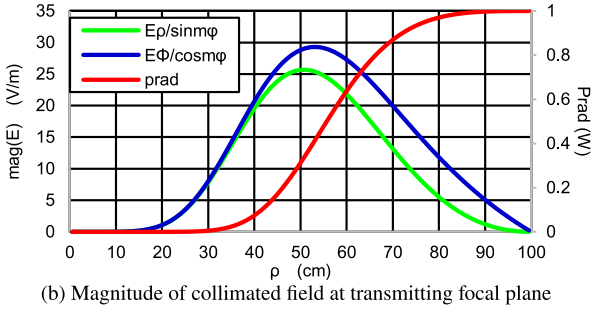
Table 1 and Fig. 4 show the transmission characteristics explained in (13)–(16). To clarify the effect of the mode order, the characteristics for the 1^{st} to 8^{th} orders are compared.

Table 1 Transmission characteristics for m.

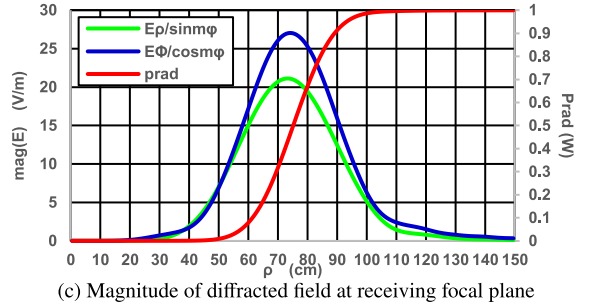
m	T^c (dB)	T^{phase} (dB)	T^{mag} (dB)	T^{total}_{mm} (dB)
1	−0.05	−0.89	−0.73	−1.66
2	−0.01	−1.05	−0.47	−1.54
3	−0.01	−1.44	−0.56	−2.01
4	−0.02	−1.96	−0.79	−2.76
5	−0.02	−2.57	−1.14	−3.74
6	−0.03	−3.24	−1.64	−4.91
7	−0.04	−3.94	−2.28	−6.26
8	−0.06	−4.62	−3.10	−7.78



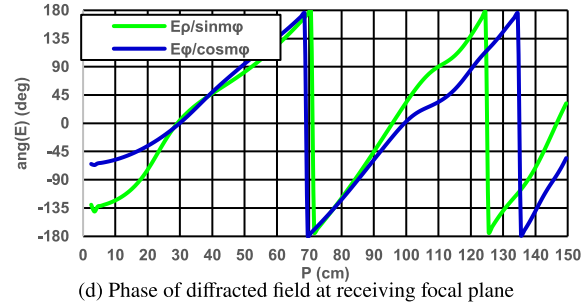
(a) Magnitude of radiated field



(b) Magnitude of collimated field at transmitting focal plane



(c) Magnitude of diffracted field at receiving focal plane



(d) Phase of diffracted field at receiving focal plane

Fig. 3 Calculated field distributions.
freq. = 150 GHz, f = 50 cm, z_c = 100 m, m = 8.

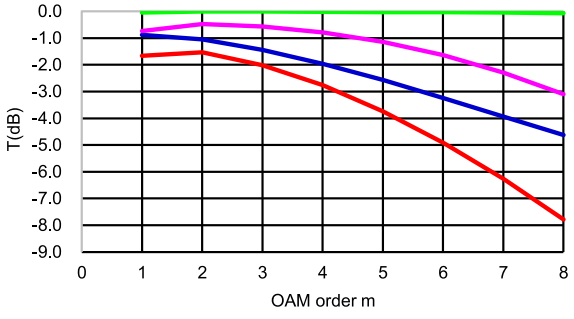


Fig. 4 Effect of m on transmission characteristics
freq. = 150 GHz, f = 50 cm, z_c = 100 m.

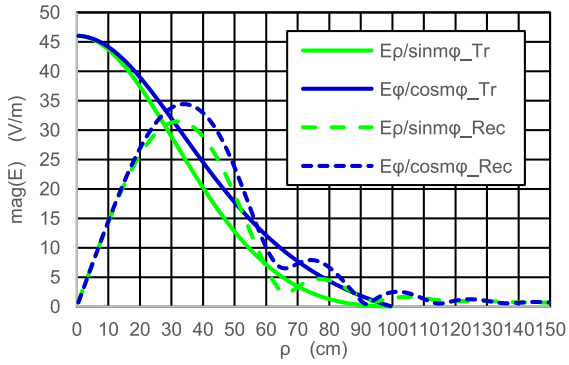


Fig. 5 Magnitudes of fields on transmitting and receiving focal planes $m = 1$, freq. = 150 GHz, $f = 50$ cm, $z_c = 100$ m.

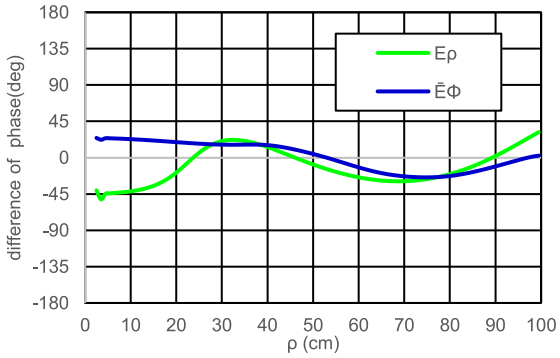


Fig. 6 Difference in phase for $\boldsymbol{E}(\rho, \phi, z, m)$ and $\boldsymbol{E}_r(\rho, \phi, m)$ $m = 8$, freq. =150 GHz, f =50 cm, $z_c = 100$ m, $z_{0r} = 0.005f$.

The loss for T^c is less than 0.1 dB, because the radius of the receiving paraboloid is as large as 100 cm and the frequency is as high as 150 GHz. T^{phase} is the largest factor to decrease the transmission. It is −4.62 dB for the 8^{th} order, which is explained by the phase difference shown in Fig. 3(d). T^{mag} is −3.10 dB, which is explained by the magnitude difference shown in Figs. 3(b) and (c). The total transmission, $T^{\text{total}}_{8,8}$, is −7.78 dB. The overall tendency for the modes is that both the total transmission and the factors decrease with the orders. However, only the transmissions for the 1^{st} order are slightly reduced, especially for T^{mag}, because the magnitudes of the fields change drastically during the propagation. Indeed, whereas the magnitude of the field is the maximum at ρ of 0 on the transmitting focal plane, it is null on the receiving focal plane, as shown in Fig. 5. Nevertheless, the OAM mode is maintained, because the difference is related to the distribution along ρ. The OAM mode is totally determined by the distribution along ϕ. Thus, even where the signal is attenuated, the interference does not appear, as already shown in (17).

5. Improvement of Transmission by Optimizing Shifts

For the simple configuration shown in Sect. 4, whereas the power is almost wholly collected by the receiving paraboloid, the received power ratio was only −7.8 dB for the 8^{th} order. To improve transmission, we tried to reduce the largest factor of the phase distribution difference. It can be reduced by controlling the shift of the receiving array, as explained in Sect. 3. In this case, whereas the phase of the received field, $\boldsymbol{E}_r(\rho, \phi, m)$, changes, the magnitude remains identical, as shown in (12). In addition, the field distribution on the receiving focal plane, $\boldsymbol{E}(\rho, \phi, z, m)$, is identical. Thus, both T^c and T^{mag} remain identical, and only T^{phase} changes. Figure 6 shows the differences in the phases for $\boldsymbol{E}_r(\rho, \phi, m)$ and $\boldsymbol{E}(\rho, \phi, z_c, m)$, where z_{0r} is $0.005f$. The phase differences are almost constant when compared with those shown in Fig. 3(d). As a result, the calculated T^{phase} is reduced to −0.13 dB, where it is improved by 4.49 dB. $T^{\text{total}}_{8,8}$ becomes −3.29 dB, as shown in Table 2. This implies that phase distribution caused by the diffraction is almost compensated for by the receiving array shift of z_{0r}.

Table 2 Transmission characteristics for m freq. = 150 GHz, $f = 50$ cm, $z_c = 100$ m, $z_{0r} = 0.005f$.

m	T^c (dB)	T^{phase} (dB)	T^{mag} (dB)	T^{total}_{mm} (dB)
1	−0.05	−2.31	−0.73	−3.09
2	−0.01	−2.63	−0.47	−3.12
3	−0.01	−2.13	−0.56	−2.70
4	−0.02	−1.46	−0.79	−2.27
5	−0.02	−0.90	−1.14	−2.06
6	−0.03	−0.49	−1.64	−2.16
7	−0.04	−0.24	−2.28	−2.57
8	−0.06	−0.13	−3.10	−3.29

However, the phase distribution caused by the diffraction depends on the mode order, and the same phase compensation by z_{0r} overcancels the phase for the lower-order modes. Table 2 also shows the T^{phase} for the other orders.

Whereas T^{phase} is improved remarkably for the 8 th-order mode compared with that in Table 1, it becomes worse for the 1^{st}-, 2^{nd}-, and 3^{rd}-order modes. The obtained total transmissions are more even, and may be more convenient for communication. However, where the transmissions for the lower-order modes are more important, the value of z_{0r} should be decreased.

Next, the improvement of T^{mag} is examined. As the difference in the magnitude distributions is mainly caused by the outward shift of the maximum field position, a slight focusing of the radiated field with the transmitting paraboloid was examined. In this case, the focused beam expands less by the diffraction, and the maximum field position on the receiving focal plane, can be expected to be reduced. To compare the obtained values of T^{mag} with those in Table 1, the receiving array is kept on the focal plane.

Figure 7 shows the magnitudes of the diffracted fields on the receiving focal plane for varied z_{0t}. To compare the distribution, the received field is also shown. Where the magnitude of the diffracted field is identical to that of the received field, T^{mag} becomes 0 dB. The position of the peak is reduced when z_{0r} changes from 0 to $-0.01f$; however, it increases thereafter, and the distribution tends to broaden remarkably. As a result, the tail of the distribution exceeds

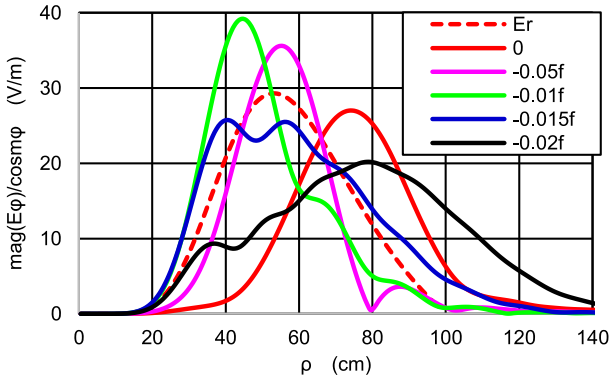


Fig. 7 Magnitude of $\boldsymbol{E}(\rho, \phi, z_c, m)$ for different z_{0t} and $\boldsymbol{E}_r(\rho, \phi, m)$ $m = 8$, freq. = 150 GHz, $f = 50$ cm, $z_c = 100$ m, $z_{0r} = 0$ cm. Numerical values in the legend show the values of z_{0t}.

Table 3 Effect of z_{0t} on transmission characteristics freq. = 150 GHz, $f = 50$ cm, $z_c = 100$ m, $z_{0r} = 0.005f$.

z_{0t}/f	T^c (dB)	T^{phase} (dB)	T^{mag} (dB)	T_{88}^{total} (dB)
0	−0.06	−4.62	−3.10	−7.78
−0.005	0.00	−2.94	−0.35	−3.29
−0.01	0.00	−9.02	−0.51	−9.53
−0.015	−0.06	−26.56	−0.41	−27.02
−0.02	−0.92	−40.13	−3.08	−44.13

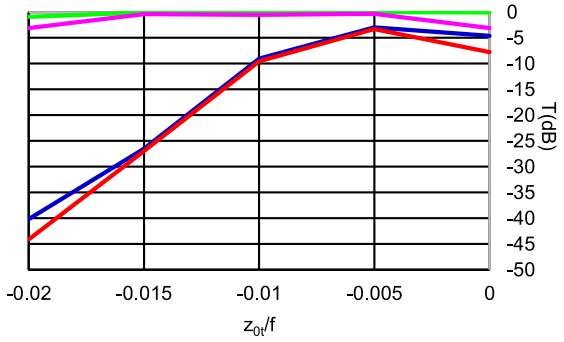


Fig. 8 Effect of z_{0t} on transmission characteristics.

the end of the receiving paraboloid, and the collection ratio of T^c increases, as shown in Table 3 and Fig. 8. In the cases of $-0.005f$ and $-0.015f$ for z_{0t}, the distribution is similar to that of $\boldsymbol{E}_r$, and each T^{mag} is reduced. Thus, T^{mag} can be reduced by reducing the difference of the magnitude distribution. However, in the case of $-0.015f$, the T^{phase} is much larger. Thus, the shift of $-0.005f$ is concluded to be optimal for the 8^{th}-order mode. In this case, T^{mag} is improved by 2.75 dB, and T^{phase} is additionally improved by 1.68 dB, which results in a 4.49 dB improvement to T_{88}^{total}.

The contribution of T^{phase} has been shown to be improved by controlling z_{0r}, where z_{0t} is null. However, the improvement is complicated in the case of nonzero z_{0t}, because the phase difference depends both on z_{0t} and z_{0r} as well as the mode order. By iterating calculations, one of the best characteristics was obtained for the eight kinds of orders, as shown in Table 4. The values of z_{0t} and z_{0r} are $-0.0025f$ and $0.0025f$, respectively. The value of z_{0t} is considerably smaller than the optimum value for the 8^{th}-order mode to improve the transmissions for the lower-order modes. The worst transmission for the eight kinds of modes is −1.8 dB, which is better by 5.98 dB than that shown in Table 1.

Table 4 Transmission characteristics for appropriate z_{0t} and z_{0r} freq. = 150 GHz, $f = 50$ cm, $z_c = 100$ m, $m_t = 8$, $z_{0t} = 0.0025f$, $z_{0r} = 0.0025f$.

m	T^c (dB)	T^{phase} (dB)	T^{mag} (dB)	T_{mm}^{total} (dB)
1	−0.04	−1.22	−0.53	−1.80
2	−0.00	−1.11	−0.38	−1.50
3	−0.00	−0.70	−0.30	−1.01
4	−0.00	−0.32	−0.24	−0.56
5	−0.00	−0.11	−0.22	−0.34
6	−0.01	−0.10	−0.30	−0.41
7	−0.01	−0.26	−0.50	−0.77
8	−0.01	−0.57	−0.85	−1.44

6. Conclusion

An analytical expression of transmission for OAM communication using loop antenna arrays and paraboloids was derived to achieve a communication distance of 100 m. With the field distribution of the single "transformed OAM mode" radiated by a loop antenna, the collimated field by the transmitting paraboloid and its diffracted field were analytically derived. The effects of frequencies, sizes of paraboloids, and the shifts in transmitting and receiving arrays away from their focal planes were included. With the diffracted field distribution on the focal plane of the receiving paraboloid, the transmission between the transmitting and receiving loop antennas was analytically estimated. It was shown that the transmission between antennas with different OAM modes was null, but the transmission between antennas with the same mode could be reduced. To clarify the mechanism of the reduced transmission, the factors of the reduction were quantitatively defined, and the explicit formulae were derived. Based on the analytical results, numerical estimation for a communication distance of 100 m was demonstrated, where the frequency, the focal length, and the size of the paraboloid were 150 GHz, 50 cm and 100 cm, respectively. Where both arrays were located on each focal plane, the transmission for the signal was more than −7.78 dB for eight kinds of OAM modes, where the transmission was the least for the highest order mode. The transmission loss was shown to be reduced by optimizing the shifts of the transmitting and receiving arrays away from their focal planes. The loss was made almost even by exploiting the tradeoff of the improvement for the mode orders. The transmission was improved by 5.98 dB, to more than −1.80 dB, by optimizing the shifts of the arrays.

Acknowledgments

This work is partly funded by Ministry of Internal Affairs and Communications, Japan, for SCOPE Programme (#JP225003001).

References

[1] L. Allen, M.W. Beijersbergen, R.J.C. Spreeuw, J.P. Woerdman, "Orbital angular momentum of light and the transformation of Laguerre Gaussian laser modes," Phys. Rev. A., vol.45, no.11, pp.8185–8189, June 1992.
[2] H. He, M.E.J. Friese, N.R. Heckenberg, and H. Rubinsztein-Dunlop, "Direct observation of transfer of angular momentum to absorptive particles from a laser beam with a phase singularity," Phys. Rev. Lett., vol.75, no.5, pp.826–829, July 1995.
[3] S. Franke-Arnold, L. Allen, and M. Padgett, "Advances in optical angular momentum," Laser Photon. Rev., vol.2, no.4, pp.299–313, July 2008.
[4] J. Wang, Jeng-Yuan Yang, I.M. Fazal, N. Ahmed, Y. Yan, H. Huang, Y. Ren, Y. Yue, S. Dolinar, M. Tur, and A.E. Willner, "Terabit free-space data transmission employing orbital angular momentum multiplexing," Nature Photon., vol.6, pp.488–496, July 2012.
[5] S.M. Mohammadi, L.K.S. Daldorff, J.E.S. Bergman, R.L. Karlsson, B. Thidé, K. Forozesh, T.D. Carozzi, and B. Isham, "Orbital Angular momentum in Radio — A system study," IEEE Trans. Antennas Propag., vol.58, no.2, pp.565–572, Feb. 2010.
[6] F. Tamburini, E. Mari, A. Sponselli, B. Thidé, A. Bianchini, and F. Romanato, "Encoding many channels on the same frequency through radio vorticity: First experimental test," New J. Phys., vol.14, pp.1–17, March 2012.
[7] D. Lee, H. Sasaki, H. Fukumoto, K. Hiraga, and T. Nakagawa, "Orbital angular momentum (OAM) multiplexing: An enabler of a new era of wireless communications," IEICE Trans. Commun., vol.E100-B, no.7, pp.1044–1063, July 2017.
[8] M.W. Beijersbergen, R.P.C. Coerwinkel, M, Kristensen, and J.P. Woerdman, "Helical-wavefront laser beams produced with a spiral phaseplate" Opt. Commun., vol.112, no.5-6, pp.321–327, 1994.
[9] G.A. Turnbull, D.A. Robertson, G.M. Smith, L. Allen, and M.J. Padgett, "The generation of free-space Laguerre-Gaussian modes at millmetre-wave frequencies by use of a spiral phaseplate," Opt. Commun., vol.127, no.4-6, pp.183–188, 1996.
[10] S. Tao, X. Yuan, and J. Lin, "Sequence of focused optical vortices generated by a spiral fractal zone plate," Appl. Phys. Lett., vol.89, no.3, pp.1758–1767, 2006.
[11] A. Ostrovsky, C. Rickenstorff, and V. Arrizon, "Generation of the perfect optical vortex using a liquid crystal spatial light modulator," Opt. Lett., vol.38, no.4, pp.534–536, 2013.
[12] X. Su, H. Zhou, K. Zou, A. Minoofar, H. Song, R. Zhang, K. Pang, H. Song, N. Hu, Z. Zhao, A. Almaiman, S. Zach, M. Tur, A.F. Molisch, H. Sasaki, D. Lee, and A.E. Willner, "Demonstration of 8-channel 32-Gbit/s QPSK wireless communications at 0.28–0.33 THz using 2 frequency, 2 polarization, and 2 mode multiplexing," Opt. Fiber Commu. Conf., 2021.
[13] N.R. Heckenberg, R. McDuff, C.P. Smith, H. Rubinsztein-Dunlop, and M.J. Wegener, "Laser beams with phase singularities," Opt. Quant Electron, vol.24, pp.951–962, 1992.
[14] Z. Guo, S. Qu, and S. Liu, "Generating optical vortex with computer-generated hologram fabricated inside glass by femtosecond laser pulses," Opt. Commun., vol.273, no.1, pp.286–289, May 2007.
[15] A. Carpentier, H. Michinel, and J. Salgueiro, "Makling optical vortices with computer-generated holograms," Amer. J. Physics, vol.76, no.10, pp.916–921, 2008.
[16] B. Thidé, H. Then, J. Sjoholm, K. Palmer, J, Bergman, T.D. Carozzi, and Y.N. Istomin, "Utilization of photon orbital angular momentum in the low-frequency radio domain," Phys. Rev. Lett., vol.99, 087701, 2007.
[17] Q. Bai, A. Tennant and B. Allen, "Experimental circular phased array for generating OAM radio beams," Electron. Lett., vol.50, no.20, pp.1414–1415, 2014.
[18] D. Lee H. Sasaki, H. Fukumoto, Y. Yagi, T Kaho, H. Shiba, and T. Shimizu, "An experimental demonstration of 28 GHz band wireless OAM-MIMO (orbital angular momentum multi-input and multi-output) multiplexing," Proc. IEEE 87th Veh. Technol. Conf., pp.1–5, 2018.
[19] Y. Yagi, H. Sasaki, T. Yamada, and D. Lee, "200 Gbit/s wireless transmission using dual-polarized OAM-MIMO multiplexing on 28 GHz band," Proc. IEEE GLOBECOM Workshop, pp.1–4, 2019.
[20] D. Lee, H. Sasaki, Y. Yagi, and H. Shiba, "Orbital angular momentum multiplexing using radio wave and its extension to multishape radio," J. Lightw. Technol., vol.41, no.7, pp.1985–1996, 2023.
[21] S. Zheng, X. Hui, X. Jin, H. Chi, and X. Zhang, "Transmission characteristics of a twisted radio wave based on circular traveling-wave antenna," IEEE Trans. Antenna Propag., vol.63, no.4, pp.1530–1536, 2015.
[22] W. Zhang, S. Zheng, X. Hui, Y. Chen, X. Jin, H. Chi, and X. Zhang, "Four-OAM-mode antenna with traveling-wave ring-slot structure," IEEE Antennas Wireless Propag. Lett., vol.16, pp.194–197, 2017.
[23] A. Saitou, R. Ishikawa, and K. Honjo, "4-value multiplexing orbital-angular-momentum communication scheme using loop-antenna arrays," APMC, Dec. 2016.
[24] H. Otsuka, R. Yamagishi, A. Saitou, R. Ishikawa, and K. Honjo, "Analytical and measured estimation for 4-value multiplexing OAM communication using loop array antennas," European Microw. Conf., Oct. 2017.
[25] A. Saitou, H. Otsuka, Y. Yamagishi, R. Ishikawa, H. Suzuki, and K. Honjo, "Double multiplicity exploiting orthogonal polarizations of OAM-wave for OAM communication with loop arrays," Proc. APMC, pp.494–496, Nov. 2018.
[26] H. Miyake, A. Saitou, H. Suzuki, R. Ishikawa, and K. Honjo," Improved performance for 8-channel multiplexing OAM communication by suppressing interference," Proc. 2020 Asia Pacific Mic. Conf., pp.161–163, Dec. 2020.
[27] Y. Yamagishi, H. Otsuka, A. Saitou, R. Ishikawa, and K. Honjo, "Improvement of mode uniqueness for OAM communication using loop array with reflector plane," Proc. APMC, 2017.
[28] H. Otsuka, Y. Yamagishi, A. Saitou, H. Suzuki, R. Ishikawa, and K. Honjo, "High performance OAM communication exploiting port-azimuth effect of loop antennas," IEICE Trans. Commun., vol.E102-B, no12, pp.2267–2275, Dec. 2019.
[29] T. Nguyen, R. Zenkyu, M. Hirabe, T. Maru, and E. Sasaki, "A study of orbital angular momentum generated by parabolic reflector with circular array feed," Proc. ISAP, pp.708–709, 2016.
[30] S. Saito, Y. Yagi, D. Lee, and F, Maehara, "Link distance and carrier frequency dependence of propagation attenuation in OAM multiplexing using parabolic antenna," IEEE, PIMRC, Oct. 2023.
[31] C.A. Balanis, Antenna Theory, 3rd ed., pp.883–937, John Wiley & Sons, U.S., 1999.
[32] S. Silver, Microwae Antenna Theory and Design, pp.413–464, McGraw-Hill, 2008.
[33] M. Born and E. Wolf, Principles of Optics, 7th ed., pp.436–443, Cambridge University Press, U.K., 2002.

Akira Saitou received his B.E. and M.E. degrees in applied physics from the University of Tokyo in 1975 and 1977, respectively, and his D.E. degree from the University of Electro-Communications in 2008. From 1977 to 2002, he was employed at NEC Corporation to develop GaAs FETs and MMICs for microwave and millimeter-wavecommunication. From 2002–2009, he worked for YKC Corporation to develop microwave circuits and antennas. In 2009, he joined the University of Electro-Communications as a guest professor to develop wideband, dual-band, and OAM antennas.

Kaito Uchida received the B.E. degree in chemical engineering from the Gunma University, Kiryu, Japan in 2022 and is currently working toward the M.E. degree at the University of Electro-communications, Tokyo, Japan.

Kanki Kitayama received the B.E. degree in infomation and commuication engineering from the University of Electro-commuications, Tokyo, Japan in 2023 and is currently working toward the M.E. degree at the University of Electro-communications.

Ryo Ishikawa received the B.E., M.E., and D.E. degrees in electronic engineering from Tohoku University, Sendai, Japan, in 1996, 1998, and 2001, respectively. In 2001, he joined the Research Institute of Electrical Communication, Tohoku University, Sendai, Japan. In 2003, he joined the University of Electro-Communications, Tokyo, Japan. His research interest is the development of microwave compound semiconductor devices and related techniques. Dr. Ishikawa is a member of the Japan Society of Applied Physics. He was the recipient of the 1999 Young Scientist Award for the Presentation of an Excellent Paper of the Tohoku Chapter, Japan Society of Applied Physics.

Kazuhiko Honjo received the B.E. degree from the University of Electro-Communications, Tokyo, in 1974, and the M.E. and D.E. degrees in electronic engineering from the Tokyo Institute of Technology, Tokyo, in 1976 and 1983, respectively. From 1976 to 2001, he worked for NEC Corporation, Kawasaki, Japan. In 2001, he joined the University of Electro-Communications as a professor in the Information and Communication Engineering Department. He has been involved in research and development of high-power/broadband/low-distortion microwave amplifiers, MMICs, HBT device and processing technology, miniature broadband microwave antennas and FDTD electro-magnetic wave and device co-analysis. Prof. Honjo received both the 1983 Microwave Prize and the 1988 Microwave Prize granted by the IEEE Microwave Theory and Techniques Society. He also received the 1980 Young Engineer Award, and the 1999 Electronics Award both presented by the Institute of Electrical, Information and Communication Engineers (IEICE), Japan. He is Life Fellow of IEEE.

PAPER

Enhanced Spatial Modulation Based Orthogonal Time Frequency Space System

Anoop A[†a)], ***Member***, **Christo K. THOMAS**[††], ***and*** **Kala S**[†], ***Nonmembers***

SUMMARY In this paper, a novel Enhanced Spatial Modulation-based Orthogonal Time Frequency Space (ESM-OTFS) is proposed to maximize the benefits of enhanced spatial modulation (ESM) and orthogonal time frequency space (OTFS) transmission. The primary objective of this novel modulation is to enhance transmission reliability, meeting the demanding requirements of high transmission rates and rapid data transfer in future wireless communication systems. The paper initially outlines the system model and specific signal processing techniques employed in ESM-OTFS. Furthermore, a novel detector based on sparse signal estimation is presented specifically for ESM-OTFS. The sparse signal estimation is performed using a fully factorized posterior approximation using Variational Bayesian Inference that leads to a low complexity solution without any matrix inversions. Simulation results indicate that ESM-OTFS surpasses traditional spatial modulation-based OTFS, and the newly introduced detection algorithm outperforms other linear detection methods.

key words: *enhanced spatial modulation (ESM), orthogonal time frequency space (OTFS), variational Bayesian inference (VBI), delay-Doppler (DD), bit error rate (BER)*

1. Introduction

The primary role of modern multicarrier modulation techniques in enabling future wireless communication with rigorous requirements on reliability and spectral efficiency has been extensively recognized. Future multicarrier modulation schemes are expected to handle highly fading wireless scenarios such as vehicle-to-vehicle communication, communications in high-speed drones and fast moving bullet trains, vehicle-to-infrastructure communication, etc. The recently proposed orthogonal time-frequency (OTFS) is a feasible option that can be used to combat high mobility conditions in multicarrier modulation schemes [1]. The information symbols are distributed in the delay-Doppler (DD) domain instead of in the time-frequency (TF) domain in the case of OFDM. The DD domain channel characteristics have made the design of equalizers and channel estimation relatively easy as proposed in [2]. These characteristics make OTFS a highly desirable candidate for future practical wireless communications.

The OTFS proposed in [1] is very effective in representing the time-fluctuating nature of the high-doppler channel. OTFS spreads information in the DD domain using a two-dimensional (2D) basis function [4]. So we can say that OTFS modulation has the capacity to transform a channel that varies over time in the TF domain into a channel with time-invariant properties in the DD domain. OTFS is very effective in doubly selective wireless channels because these basis functions span the entire time-frequency plane.

The performance evaluation of OTFS is done in the millimetre wave system, and it is estimated that the bit error rate (BER) performance of OTFS is much better than orthogonal frequency division multiplexing (OFDM) [3]. OTFS modulation can be put into practice by adding inverse sympletic finite Fourier transform (ISFFT) at the transmitter side and sympletic finite Fourier transform (SFFT) at the receiver [5]. So we can easily integrate OTFS into conventional OFDM systems. A simplified discrete input-output relationship for OTFS was proposed in [6]. In [7], authors investigated that OTFS exhibits excellent peak-to-average power ratio (PAPR) endurance in high mobility channels. An new method for channel estimation of OTFS system utilising embedded pilot signals was proposed in [8]. An iterative detector for OTFS was proposed in [9].

Multiple-input multiple-output (MIMO) technology has been part of fourth and fifth-generation wireless networks due to its ability to provide high spectral efficiency. But MIMO networks encounter many more challenges in environments characterised by a high degree of mobility compared to fixed MIMO networks. So conventional MIMO channels experience many significant drawbacks as hilighted in [10]. The DD domain OTFS channel has a sparse nature, and this property helps higher-order MIMO networks to mitigate channel equalization and channel estimation in high-doppler environments. MIMO-OTFS systems have been well investigated since their inception, and their detailed signal processing, detection methods, and channel estimation techniques are studied in [11]–[13].

Spatial modulation (SM) represents a cutting-edge technique in wireless communication, particularly enhancing Multiple-Input Multiple-Output (MIMO) systems by utilising the spatial dimension to transmit additional information [14]. This method efficiently combines antenna selection along with symbol modulation, significantly reducing inter-channel interference and system complexity. This means that only one transmit antenna is operational at the same instant in the case of SM, and other antennas remain inactive. This makes SM a modulation scheme with high spectral efficiency, low complexity, and good BER performance.

Manuscript received December 14, 2023.
Manuscript revised March 26, 2024.
Manuscript publicized May 31, 2024.

[†]Department of Electronics and Communication Engineering, Indian Institute of Information Technology, Kottayam, 686635 India.
[††]Wireless@VT, Bradley Department of Electrical and Computer Engineering, Virginia Tech., Arlington, VA, USA.
a) E-mail: anoop.phd2206@iiitkottayam.ac.in
DOI: 10.23919/transcom.2023EBP3206

In order to enhance the spectral efficiency further, quadrature-spatial modulation (QSM) was proposed [22]. QSM, a derivative of SM, distinguishes itself by dividing the modulated symbol into in-phase and quadrature components. These components are then assigned to two different transmit antennas, each activated by its respective group of index bits to transmit the in-phase and quadrature components. Subsequently, these modulated signal components are transmitted using carriers that are mutually orthogonal. This method effectively prevents channel interference and enhances diversity gain. Generalized Spatial Modulation (GSM) is another innovative spatial modulation technique that selectively activates a subset of antennas for transmission at any given time [23]. In comparison to SM, GSM requires more complex signal processing and antenna selection algorithms, leading to increased computational complexity and potentially higher power consumption [24].

Enhanced spatial modulation (ESM) was first proposed in [16]. ESM was formulated by combining different ideas. ESM involves the use of primary and secondary constellations. In the case of ESM, when one transmit antenna is active, information symbols used for modulation are selected from a primary constellation, and when two transmit antennas are active, information symbols are modulated using the secondary constellation, and other antennas remain silent. The total count of information bits that can be sent in a single transmission of ESM depends on the size of the combinations of transmit antennas and constellation symbols. So this property increases the effective throughput of ESM compared to conventional SM. Furthermore, the secondary constellation is designed using geometric interpolation of the primary constellation, which optimize the minimum Euclidean distance among the transmitted signal vectors. This approach marks a significant divergence between ESM from traditional SM and GSM. When we select signal constellations that maintain the operational spectral efficiency of ESM on par with conventional SM, ESM exhibits superior performance. This concept of augmenting combination numbers to boost spectral efficiency is similarly observed in QSM. Nevertheless, QSM experiences a diminished minimum squared Euclidean distance and yields inferior performance compared to ESM.

SM demonstrates remarkable flexibility and adaptability, allowing it to be seamlessly integrated with various transmission technologies. When combined with OFDM to form SM-OFDM, this approach is capable of combating frequency-selective fading, obviating the need for complex equalization methods, and thereby significantly boosting the spectral efficiency of the system [25]. While OFDM performs well in static or low-mobility environments with multipath propagation, its performance can degrade significantly in high mobility scenarios due to Doppler shifts affecting the orthogonality of the subcarriers. To address this issue, spatial modulation based on OTFS (SM-OTFS) is proposed [15] to excel in environments with high Doppler shifts and significant delay spreads. The authors introduce SM-OTFS based on MIMO to enhance spectral efficiency and reduce detection complexity in [15]. It demonstrates through simulations that SM-OTFS provides significant performance gains over space-time-coded OTFS (STC-OTFS), especially in high-mobility scenarios.

The system model of the SM-OTFS scheme and its related signal processing techniques are elaborated in [26], and this paper also provides closed-form expressions for the average symbol error rate (ASER) and average bit error rate (ABER) over delay-Doppler channels. The authors also highlighted the superior performance of SM-OFDM over SM-OTFS under high Dopple environments in [26]. The system model and its associated signal processing techniques for genaralised spatial modulation on the OTFS (GSM-OTFS) system are outlined in [27], including the use of the union bound technique and moment generating function (MGF) for theoretical analysis of average BER performance. In the paper [27], the authors demonstrate through theoretical and simulation results that GSM-OTFS offers better BER performance and spectral efficiency compared to conventional SM-OTFS. Even if the GSM-OTFS system enhances spectral efficiency and BER performance compared to SM-OTFS, it introduces challenges such as increased system complexity, hardware demands, power consumption, and advanced signal processing requirements.

Quadrature spatial modulation based on OTFS (QSM-OTFS) proposed in [20] is another OTFS-based index modulation utilising QSM. In the paper [20], the authors detail the system model, signal processing steps, and performance analysis, including theoretical ABER analysis and an innovative detection technique named enhanced minimum mean square error (EMMSE) for reduced complexity in detection. Furthermore, it compares the proposed QSM-OTFS system with the traditional SM-OTFS system, highlighting the advantages in terms of ABER performance. Motivated by the characteristics of ESM and OTFS, we propose a novel ESM-based OTFS scheme called enhanced modulation-based OTFS (ESM-OTFS) that performs well in high mobility scenarios and provides high spectral efficiency compared to SM-OTFS. QSM-OTFS has the same spectral efficiency as ESM-OTFS but has inferior performance under the same channel conditions.

Enhanced spatial modulation combined with orthogonal time frequency space (ESM-OTFS) modulation presents a versatile and efficient solution for modern wireless communication challenges, particularly in environments requiring high mobility and robustness. ESM-OTFS is particularly suited for environments with high mobility, such as high-speed trains, vehicular networks (vehicle-to-vehicle and vehicle-to-infrastructure communication), and drones, due to its resilience to Doppler shifts and ability to maintain reliable communication at high speeds. This innovative approach is also well-suited for a wide range of applications, including 5G and beyond wireless systems, Internet of Things (IoT) networks, satellite and deep space communication, and underwater acoustic communication, as well as enhancing capacity and reliability in urban cellular networks. By offering improved spectral efficiency and resilience to Doppler

shifts and multipath propagation, ESM-OTFS stands out as a promising technology for ensuring reliable, high-speed communication across various challenging environments and applications, driving advancements in both current and future wireless communication landscapes.

The transmit vectors of SM-OTFS and ESM-OTFS include significant numbers of zero entries rather than non-zero values, and this property makes the data transmit vector a sparse vector in most of the system configurations. So it is a good option that we can utilise sparse signal estimation methods for the detection of ESM-OTFS. Sparse Bayesian Learning (SBL) is one of the popular techniques used in sparse signal estimation. But it involves the use of matrix inversion in each iteration, so that it is computationally intensive even for datasets of intermediate size. Instead of using SBL, we can use the Variational Bayesian Inference method [17], [18] for sparse signal estimation.

Variational Bayesian Inference (VBI), also known as Variational Bayes, is a method used in Bayesian statistics and machine learning for approximating the posterior distribution of latent variables in a probabilistic model. This method is beneficial in situations involving complicated models where the use of exact inference or sampling techniques is computationally intractable. The main idea behind VBI is to approximate the true posterior distribution with a simpler, parameterized variational distribution that is easier to work with. These variational distributions are normally selected from a family of distributions, such as Gaussian distributions. The goal is to find the parameters of this simpler distribution that best approximate the true posterior. The SAVE (Space Alternating Variational Estimation for Sparse Bayesian Learning) algorithm proposed in [19] is an excellent algorithm for sparse signal estimation. Motivated by the SAVE algorithm, we are able to come up with a novel sparse signal estimation-based detector for the newly proposed ESM-OTFS.

2. System Model

2.1 ESM Modulation

Spatial modulation (SM) is a wireless communication technique based on Mutiple Input Mutiple Output (MIMO) that uses transmit antenna index for transmitting information symbols. SM involves the activation of a single transmit antenna at any given moment in time and that selected antenna is used to transmit a signal from a chosen constellation. If the N_T is the number of transmit antennas and $s_a = 2^{n_a}$ is the size of size of signal constellation, the number of bits transmitted in SM is $n_a + \log_2(N_T)$.

Enhanced spatial modulation improves the efficiency and robustness of SM by transmitting different symbols from different constellations depending on the number of active antennas as established in [16]. This is done by first transmitting symbols from a primary constellation during single transmit antenna activation just like the conventional SM. When two transmit antennas are enabled, symbols are transmitted from a secondary constellation. The size of the secondary constellation is taken half of the primary constellation so that the same amount of information bits are transmitted during single antenna activation and double antenna activation periods. This allows for a higher diversity gain as the signal is transmitted from two different antennas, which makes it more robust to interference. In order to maximise the minimal Euclidean distance between transmitted signal vectors, secondary constellations are created via geometric interpolation. This ensures that the symbols in the secondary constellations are as far apart as possible, which makes it more difficult for the receiver to make errors.

Table 1 ESM, 2 TX - 4 BPCU.

	TX1	TX2
$C1$	QPSK	0
$C2$	0	QPSK
$C3$	BPSK0	BPSK0
$C4$	BPSK1	BPSK1

In conventional SM, the number of information bits transmitted depend on the number of transmit antennas and the size of symbol constellation used for modulation. In ESM, it is decided by the antenna and constellation symbol combinations. An illustration of ESM system with 2 numbers of transmit antennas, QPSK as the primary constellation and BPSK as the secondary constellation is depicted in Table 1. Here the number of transmitted bits or bits per channel use (bpcu) is 4.

QPSK $= \pm 1 \pm j$, BPSK0 and BPSK1 are respectively given by BPSK0 $= \pm 1$ & BPSK1 $= \pm j$. The first two combinations $C1$ and $C2$ appear like the transmission from one of the antenna and convey one of symbols from QPSK constellation. The last combinations $C3$ and $C4$ corresponds to transmissions of symbols from secondary constellations BPSK0 or BPSK1 coming out of both antennas.

2.2 OTFS Signal Modulation and Demodulation

OTFS modulation is a way of transmitting information symbols that is robust to Doppler shift. It works by multiplexing the data symbols in the DD domain, which is another way of expressing the signal that is invariant to Doppler shift. This makes OTFS ideal for high-mobility applications. Let $\{u[l,k], l = 0,1,\ldots,M-1, k = 0,1,\ldots,N-1\}$ be a two dimensional information signal in the DD domain where M denotes the number of sub carriers and N denotes the number of OFDM symbols or time slots. The time-frequency (TF) domain signal $U[m,n]$ is obtained from $u[l,k]$ through the inverse symplectic finite Fourier transform (ISFFT), i.e.,

$$U[m,n] = \frac{1}{\sqrt{MN}} \sum_{l=0}^{M-1} \sum_{k=0}^{N-1} u[l,k]\, e^{j2\pi\left(\frac{nk}{N} - \frac{ml}{M}\right)} \tag{1}$$

Using Heisenberg transform along with transmitter side pulse shaping signal $p_{tx}(t)$, $U[m,n]$ is converted to a time domain (TD) signal $s(t)$

$$s(t) = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} U[m,\ n] p_{\mathrm{tx}}(t - mT) e^{j2\pi n\Delta f(t-mT)} \quad (2)$$

where Δf and T represent the sub carrier spacing and OFDM symbol period respectively, and $\Delta f = \frac{1}{T}$.

At the receiver, the received signal $r(t)$ is matched with receiver side pulse shaping signal $p_{rx}(t)$ and is converted to TF domain using Wigner transform and the resulting cross-ambiguity $\Lambda_{p_{rx},r}(t,f)$ is given as

$$\Lambda_{p_{rx},r}(t,f) = \int r(t^*) p_{\mathrm{rx}}(t^* - t) e^{j2\pi f(t^*-t)} \mathrm{d}t^* \quad (3)$$

The resultant signal is sampled with sub carrier spacing Δf and the frame duration T. The received TF domain signal is given as

$$V[m,n] = \Lambda_{p_{\mathrm{rx,r}}(t,f)|_{t=nT,f=m\Delta f}} \quad (4)$$

The DD domain signal $v[l,k]$ is obtained from $V[m,n]$ by applying symplectic finite Fourier transform (SFFT) and is given as

$$v\,[l,k] = \frac{1}{\sqrt{MN}} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} V\,[m,n]\, e^{-j2\pi\left(\frac{nk}{N} - \frac{ml}{M}\right)} \quad (5)$$

2.3 OTFS System Model

Consider a DD domain channel with P channel taps, each with a delay of τ_i, a doppler of ν_i and a fading channel gain of κ_i. The channel impulse response in the DD domain as in [1] can be represented as

$$\kappa(\tau,\nu) = \sum_{i=1}^{P} \kappa_i \delta(\tau - \tau_i)\delta(\nu - \nu_i) \quad (6)$$

where $\delta(.)$ denotes the dirac-delta function. The delay and doppler of the i^{th} can be formulated as

$$\tau_i = \frac{l_i}{M\Delta f}, \nu_i = \frac{k_i}{NT} \quad (7)$$

Given the TD input signal $s(t)$, the received signal $r(t)$ is given as

$$r(t) = \int_{\nu} \int_{\tau} \kappa(\tau,\nu) s(t-\tau) e^{j2\pi\nu(t-\tau)} \mathrm{d}\tau \mathrm{d}\nu \quad (8)$$

Substituting equation (6) in equation (8) yields as established in [5].

$$r(t) = \int_{\nu} \int_{\tau} \left(\sum_{i=1}^{P} \kappa_i \delta(\tau - \tau_i)\delta(\nu - \nu_i) \right) s(t-\tau) e^{j2\pi\nu(t-\tau)} \mathrm{d}\tau \mathrm{d}\nu \quad (9)$$

and can be simplified as

$$r(t) = \sum_{i=1}^{P} \kappa_i s(t - \tau_i) e^{j2\pi\nu_i(t-\tau_i)} + n(t) \quad (10)$$

where $n(t)$ is the noise signal in TD. The sampled TD signal $r(p)$ is given as

$$r(p) = \sum_{i=1}^{P} \kappa_i s(p - \tau_i) e^{j2\pi\nu_i(p-\tau_i)} + n(p) \quad (11)$$

where $p = 0, 1, 2 \ldots, NM - 1$.

It can be depicted in the matrix form as follows

$$r = \begin{bmatrix} r_0 \\ r_1 \\ \vdots \\ r_{MN-1} \end{bmatrix} = \left(\sum_{i=1}^{P} \kappa_i \Pi^{l_i} \Delta^{k_i} \right) s + w = \breve{H} s + n \quad (12)$$

where $\Pi = \begin{bmatrix} 0 & \cdots & 0 & 1 \\ 1 & \ddots & 0 & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 1 & 0 \end{bmatrix}$ is the permutation matrix with dimension $MN \times MN$, Δ is the $MN \times MN$ dimensional diagonal matrix as $\Delta = diag\left\{e^{j2\pi\frac{p}{MN}}\right\}_{p=0}^{MN-1}$, $\breve{H}$ is the TD channel matrix with dimension $MN \times MN$, meanwhile without the loss of generality l_i and k_i are assumed to be integers.

At the receiver, the received signal vector r gets converted to DD domain and the input-output relationship in DD domain is given as follows

$$v = Hu + w \quad (13)$$

where w is the noise vector in the DD domain and H is the DD domain channel matrix given as $H = (F_N \otimes I_M)\breve{H}(F_N^H \otimes I_M)$ assuming the use of rectangular pulse shaping filter at the transmitter and receiver, F_N is the N point discrete Fourier transform (DFT) matrix, F_N^H is the N point inverse discrete Fourier transform(IDFT) matrix, $\otimes$ is the operator of the Kronecker product and I_M is the M dimensional identity matrix.

2.4 ESM-OTFS

The system model of the proposed ESM-OTFS is shown in Fig. 1. From the Fig. 1, it is observed that the ESM-OTFS system is equipped with N_T transmit antennas and N_R receive antennas. A detail description of the signal processing used in ESM-OTFS is given as the following. A random bit sequence $b = [b_0\ \ b_1 \cdots b_{totbits}]$ of an ESM-OTFS frame in the DD domain enter the ESM-OTFS system where $totbits = MNlog_2(\mathbb{C})$ and $\mathbb{C}$ is the size of ESM constellation. The mapping rule of ESM modulation is given in Table 2. For each of $\mathbb{C}$ incoming bits, an ESM transmit vector is selected as per the presented mapping rule. For each of MN time slots of an ESM-OTFS frame, a constellation symbol is assigned to each of N_T transmit antennas. These MN symbols enter

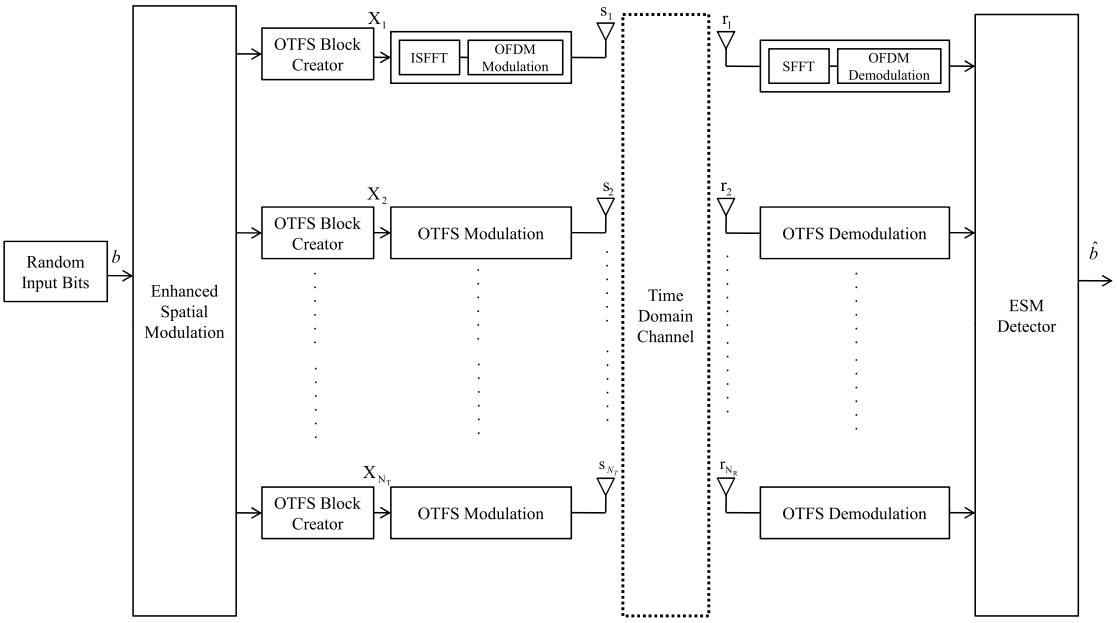


Fig. 1 System model of ESM-OTFS system.

Table 2 ESM mapping rule for $N_T = 2$ and QSPK as primary constellation.

$\mathbb{C}$ bits	Transmit Vector
0000	[0+0j, -1+1j]
0001	[0+0j, -1-1j]
0010	[0+0j, 1+1j]
0011	[0+0j, 1-1j]
0100	[-1+1j, 0+0j]
0101	[-1-1j, 0+0j]
0110	[1+1j, 0+0j]
0111	[1-1j, 0+0j]
1000	[-1+0j, -1+0j]
1001	[-1+0j, 1+0j]
1010	[1+0j, -1+0j]
1011	[1+0j, 1+0j]
1100	[0-1j, 0-1j]
1101	[0-1j, 0+1j]
1110	[0+1j, 0-1j]
1111	[0+1j, 0+1j]

the OTFS Block Creator unit of the corresponding transmit antenna and form the transmit DD domain vector for the respective antenna.

The transmission rate of ESM-OTFS can be stated as $R_{ESM-OTFS} = MNlog_2(\mathbb{C})$ whereas under the same setup, SM-OTFS has a transmission rate of $R_{SM-OTFS} = MNlog_2(s_a)$ where s_a is the order of the modulation used or the size of the modulation constellation used. Since the size of the ESM constellation formed by the combination of transmit antennas and primary and secondary constellation symbols is higher than the size of the modulation order used in SM, the transmission rate of ESM-OTFS is much better than that of SM-OTFS under the same MIMO environment. The dimension of the data transmit matrix of an ESM-OTFS frame is $MN \times N_T$ and is given as follows

$$\mathbf{X_{ESM}} = \begin{bmatrix} \mathbf{x}^1_{0,0} & \cdots & \mathbf{x}^i_{0,0} & \cdots & \mathbf{x}^{N_T}_{0,0} \\ \mathbf{x}^1_{0,1} & \cdots & \mathbf{x}^i_{0,1} & \cdots & \mathbf{x}^{N_T}_{0,1} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \mathbf{x}^1_{0,N-1} & \cdots & \mathbf{x}^i_{0,N-1} & \cdots & \mathbf{x}^{N_T}_{0,N-1} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \mathbf{x}^1_{k,l} & \cdots & \mathbf{x}^i_{k,l} & \cdots & \mathbf{x}^{N_t}_{k,l} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \mathbf{x}^1_{M-1,N-1} & \cdots & \mathbf{x}^i_{M-1,N-1} & \cdots & \mathbf{x}^{N_T}_{M-1,N-1} \end{bmatrix} \tag{14}$$

X_i is the DD domain information matrix of dimension $M \times N$ transmitted from the i^{th} transmit antenna. X_i is formed by collecting all the elements of i^{th} column of $\mathbf{X_{ESM}}$.

The TD signal that is being transmitted from the i^{th} transmit antenna s_i is derived from X_i by passing through ISSFT unit and OFDM modulator which simulates the Heisenberg transform and is given as follows

$$s_i = vec\left(F_M^H\left(F_M X_i F_N^H\right)\right) = \left(F_N^H \otimes I_M\right) x_i \tag{15}$$

where s_i is the $MN \times 1$ dimensional column vector and $vec(.)$ is the column wise vector operation.

The TD signal s_i is transmitted from the i^{th} transmit antenna and travels through the multi-path wireless channel. The received signal at the j^{th} receive antenna from the i^{th} transmit antenna is given as

$$r_j = \breve{\mathbf{H}}_{ji} s_i + n_j \tag{16}$$

In the receiver, the TD signal received at each receive antenna is converted to DD domain through SFFT and OFDM demodulator which simulates the Wigner transform. The

DD domain signal received at j^{th} receive antenna is given as

$$y_j = \mathbf{H}_{j1}x_1 + \mathbf{H}_{j2}x_2 + \cdots + \mathbf{H}_{ji}x_i + \cdots + \mathbf{H}_{jN_R}x_{N_R} + w_j \tag{17}$$

The received signal in the DD domain considering all the receive antennas is shown as follows

$$\mathbf{y}_{ESM} = \mathbf{H}_{eff}\mathbf{x}_{ESM} + \mathbf{w}_{eff} \tag{18}$$

where $\mathbf{y}_{ESM} = [y_0, y_1, \cdots, y_{N_R}]$, $\mathbf{x}_{ESM} = [x_0, x_1, \cdots, x_{N_T}]$ and w_{eff} is the effective noise vector. At the same time, we have $\mathbf{y}_{ESM}, \mathbf{w}_{eff} \in \mathbb{C}^{N_R MN \times 1}$, $\mathbf{x}_{ESM} \in \mathbb{C}^{N_T MN \times 1}$ and $\mathbf{H}_{eff} \in \mathbb{C}^{N_R MN \times N_T MN}$. The effective channel matrix $\mathbf{H}_{eff}$ of ESM-OTFS in the DD domain is given as

$$\mathbf{H}_{eff} = \begin{bmatrix} \mathbf{H}_{11} & \mathbf{H}_{12} & \dots & \mathbf{H}_{1N_T} \\ \mathbf{H}_{21} & \mathbf{H}_{22} & \dots & \mathbf{H}_{2N_T} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{H}_{N_R1} & \mathbf{H}_{N_R2} & \cdots & \mathbf{H}_{N_RN_T} \end{bmatrix} \tag{19}$$

3. Detection of ESM-OTFS

3.1 MMSE Detector

In order to bring down the complexity of detection, a detector combining MMSE equalization and minimum Euclidean distance detection is proposed for the detection of ESM-OTFS. Firstly an estimate of the transmitted signal $\hat{\mathbf{x}}_{ESM}^{MMSE}$ is obtained using a MMSE based equalizer. The MMSE equalization is given as follows

$$\hat{\mathbf{x}}_{ESM}^{MMSE} = \left(\mathbf{H}^{\mathbf{H}} + \frac{I_{N_T MN}}{\rho_{snr}}\right)^{-1} \mathbf{H}\mathbf{y}_{ESM} \tag{20}$$

where $I_{N_T MN}$ is the identity matrix of order $N_T MN \times N_T MN$ and ρ_{snr} is the average SNR in the DD domain. Now a minimum euclidean distance detector is applied on each row of $\hat{\mathbf{X}}_{ESM}^{MMSE}$ against all the possible combination of ESM constellation vectors C_{ESM}. This can be formulated as follows

$$\left\{\hat{\mathbf{X}}_{ESM}^{\eta}\right\} = \underset{\mathbf{C_{ESM}} \in C_{\mathbf{ESM}}}{\arg\min} \left|\hat{\mathbf{X}}_{ESM}^{MMSE}(\eta) - \mathbf{C_{ESM}}\right|^2, \quad 1 \le \eta \le MN, \tag{21}$$

where $\hat{\mathbf{X}}_{ESM}^{MMSE}(\eta)$ is the η^{th} row of the matrix $\hat{\mathbf{X}}_{ESM}^{MMSE}$. Then the ESM constellation vector with minimum euclidean distance is taken as the detected ESM-OTFS transmit vector for each time slot of ESM-OTFS frame. After the equalization and detection, the demodulated ESM-OTFS frame $\hat{\mathbf{X}}_{ESM} \in \mathbb{C}^{MN \times N_T}$ can be expressed in the matrix form as

Algorithm 1 MMSE Detection Algorithm

Given: $\mathbf{y}_{ESM}$, $\mathbf{H}$, C_{ESM}.

1. Compute $\hat{\mathbf{X}}_{ESM}^{MMSE}$ using (20).
2. Estimate the transmitted ESM vectors using (21) and form $\hat{\mathbf{X}}_{ESM}$.
3. Retrieve the transmitted bits using ESM mapping Rule.

$$\hat{\mathbf{X}}_{ESM} = \begin{bmatrix} \hat{\mathbf{x}}_{0,0}^1 & \cdots & \hat{\mathbf{x}}_{0,0}^i & \cdots & \hat{\mathbf{x}}_{0,0}^{N_T} \\ \hat{\mathbf{x}}_{0,1}^1 & \cdots & \hat{\mathbf{x}}_{0,1}^i & \cdots & \hat{\mathbf{x}}_{0,1}^{N_T} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \hat{\mathbf{x}}_{0,N-1}^1 & \cdots & \hat{\mathbf{x}}_{0,N-1}^i & \cdots & \hat{\mathbf{x}}_{0,N-1}^{N_T} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \hat{\mathbf{x}}_{k,l}^1 & \cdots & \hat{\mathbf{x}}_{k,l}^i & \cdots & \hat{\mathbf{x}}_{k,l}^{N_t} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \hat{\mathbf{x}}_{M-1,N-1}^1 & \cdots & \hat{\mathbf{x}}_{M-1,N-1}^i & \cdots & \hat{\mathbf{x}}_{M-1,N-1}^{N_T} \end{bmatrix} \tag{22}$$

Each row of $\hat{\mathbf{X}}_{ESM}$ represents the estimate of ESM-OTFS transmit vector for each of the MN time slots of ESM-OTFS frame. Now an estimate of the original transmitted bits $\hat{b}$ can be recovered from comparing each row of $\hat{\mathbf{X}}_{ESM}$ using a look-up table as shown in Table 2. The steps involved in detection of ESM-OTFS using MMSE detector is summarized in Algorithm 1.

3.2 Variational Bayesian Inference Based ESM-OTFS Detector

Since the computational complexity of the maximum likelihood detector increases exponentially with the increase in the number of transmit N_T and receive N_R antennas, we propose a new detection algorithm based on the sparsity of the ESM-OTFS transmit frame. In ESM-OTFS, only one or two transmit antennas radiate at the same time and others remain silent. This results in the formation of data transmit matrix with most of the entries are zero making the matrix sparse. The key idea behind the proposed detector is the sparse signal estimation of the received signal using Variational Bayesian Learning. So the problem of ESM-OTFS signal detection can be restated as

$$y = \mathbf{H}x + w, \tag{23}$$

where x and y are transmit and receive signals of dimensions $N_T MN \times 1$ and $N_R MN \times 1$ respectively in the DD domain, $\mathbf{H}$ is the effective DD domain channel matrix of dimension $N_R MN \times N_T MN$ and w is the Gaussian distributed white noise signal with zero mean and variance γ.

It is assumed to have a two-layer hierarchical prior for the transmit data signal x as proposed in [19] so that it motivates the sparsity property of x. It is assumed that x follows a Gaussian distribution parameterized by $\alpha = [\alpha_1, \alpha_2, \cdots, \alpha_{N_T MN}]$ where α_i is the inverse variance parameter of x_i and x is given as

$$p(x/\alpha) = \prod_{i=1}^{N_T MN} p(x_i/\alpha_i) = \prod_{i=1}^{N_T MN} \mathcal{N}(0, \alpha_i^{-1}). \quad (24)$$

Moreover a Gamma prior distribution is assumed for α

$$p(\alpha) = \prod_{i=1}^{N_T MN} p(\alpha_i/a, b) = \prod_{i=1}^{N_T MN} \Gamma^{-1}(a) b^a \alpha_i^{a-1} e^{-b\alpha_i}. \quad (25)$$

It is presumed that the variance γ of white noise signal w is already known and a full DD domain Channel State Information (CSI) is accessible at the receiver. The likelihood distribution of the received signal y is given as follows

$$p(y/x) = (2\pi)^{-N_R MN/2} \gamma^{N_R MN/2} e^{\frac{-\gamma ||y - \mathbf{H}x||^2}{2}}. \quad (26)$$

The estimation of the posterior distribution of the received signal y is very cumbersome. To cope with this problem, variational Bayesian technique is applied and the posterior distribution $p(x/y, \alpha)$ is estimated by a varitional distribution $q(x, \alpha)$. It can be represented as

$$q(x, \alpha) = \prod_{i=1}^{N_T MN} q_{x_i}(x_i) \prod_{i=1}^{N_T MN} q_{\alpha_i}(\alpha_i). \quad (27)$$

Variational Bayesian technique calculates the factors of $q(x, \alpha)$ by minimizing the Kullback-Leibler (KL) distance between the variational distribution $q(x, \alpha)$ and the true posterior distribution $p(x, \alpha/y)$. The KL distance between $q(x, \alpha)$ and $p(x, \alpha/y)$ is denoted as KLD_{VBI}.

$$KLD_{VBI} = KL\left(p(x, \alpha/y) || q(x, \alpha)\right). \quad (28)$$

Minimization of KL distance is equivalent to maximizing the evidence lower Bound (ELBO). To discuss this pont further, KLD_{VBI} can be expressed as

$$\begin{aligned} KLD_{VBI} &= -\int q(\theta) \ln \frac{p(\theta/y)}{q(\theta)} d\theta \\ &= -\int q(\theta) \ln \frac{p(y, \theta)}{p(y) q(\theta)} d\theta. \end{aligned} \quad (29)$$

This further simplifies to

$$KLD_{VBI} = \ln p(y) - \int q(\theta) \ln \frac{p(y, \theta)}{q(\theta)} d\theta, \quad (30)$$

where $\theta = \{x, \alpha\}$ and can be rearranged as

$$\ln p(y) = KLD_{VBI} + L(q). \quad (31)$$

Since KLD_{VBI} is a distance, $KLD_{VBI} \geq 0$. It means that the ELBO $L(q)$ is the lower bound on $\ln p(y)$. As we know, $\ln y$ is independent of $q(\theta)$ and minimization of KLD_{VBI} is analogous to maximization of the lower bound $L(q)$. The ELBO maximization results in the following expression

$$\ln(q_i(\theta_i)) = < \ln p(y, \theta) >_{k \neq i} + c_i, \quad (32)$$

where $\theta = \{x, \alpha\}$ and θ_i denotes each scalar in θ. Here $<>_{k \neq i}$ symbolizes the operator of expectation across the distributions q_k for all $k \neq i$. The joint probability distribution $p(y, \theta)$ can be represented as

$$p(y, \theta) = p(y/x, \alpha) p(x/\alpha) p(\alpha) \quad (33)$$

Now the focus is to find an iterative solution. For this purpose, $\ln p(y, \theta)$ can be expanded as follows

$$\begin{aligned} \ln p(y, \theta) = &\frac{N_R MN}{2} \ln \gamma - \frac{\gamma}{2} \|y - \mathbf{H}x\|^2 \\ &+ \sum_{i=1}^{N_T MN} \left(\frac{1}{2} \ln \alpha_i - \frac{\alpha_i}{2} x_i^2 \right) \\ &+ \sum_{i=1}^{N_T MN} \left((a-1) \ln \alpha_i + a \ln b - b\alpha_i \right) \\ &+ \text{constants}. \end{aligned} \quad (34)$$

Now using (32) and (34), we need to find the update expressions for $\ln q_{x_i}(x_i)$ and $\ln q_{\alpha_i}(\alpha_i)$.

$$\begin{aligned} &\ln q_{x_i}(x_i) = \\ &-\frac{\gamma}{2} \Big\{ < \left| y - \mathbf{H}_{\bar{i}} x_{\bar{i}} \right|^2 > - (y - \mathbf{H}_{\bar{i}} < x_{\bar{i}} >)^H \mathbf{H}_i x_i - \\ &x_i \mathbf{H}_i^H (y - \mathbf{H}_{\bar{i}} < x_{\bar{i}} >) + ||\mathbf{H}_i||^2 x_i^2 \Big\}, - \tfrac{<\alpha_i>}{2} x_i^2 + c_{x_i} \\ &= -\tfrac{1}{2\sigma_i^2} (x_i - \mu_i)^2 + c'_{x_i}, \end{aligned} \quad (35)$$

where we represent $\mathbf{H}x$ as $\mathbf{H}x = \mathbf{H}_i x_i + \mathbf{H}_{\bar{i}} x_{\bar{i}}$ where where $\mathbf{H}_i$ denotes the i^{th} column of $\mathbf{H}$,$\mathbf{H}_{\bar{i}}$ is formed by removing i^{th} column of $\mathbf{H}$,x_i is the i^{th} element of x and $x_{\bar{i}}$ is the vector formed by removing i^{th} element of x. c_{x_i} and c'_{x_i} are the normalization constants. From (35), we can understand that the expansion of $\ln q_{x_i}(x_i)$ is quadratic in nature and can be expressed as a Gaussian distributed random variable. The mean and variance of the subsequent Gaussian distribution is given as follows

$$\begin{aligned} &\sigma_i^2 = \tfrac{1}{\gamma ||\mathbf{H}||^2 + \alpha_i}, \\ &< x_i > = \mu_i = \sigma_i^2 \mathbf{H}^H \left(y - \mathbf{H}_{\bar{i}} < x_{\bar{i}} > \right) \gamma, \end{aligned} \quad (36)$$

where μ_i is the point estimate of i^{th} element of the transmitted signal x. In a similar way $\ln q_{\alpha_i}(\alpha_i)$ can be expressed as follows

$$\begin{aligned} &\ln q_{\alpha_i}(\alpha_i) = (a - 1 + \tfrac{1}{2}) \ln \alpha_i - \alpha_i \left(\tfrac{<x_i^2>}{2} + b \right) + c_{\alpha_i}, \\ &q_{\alpha_i}(\alpha_i) \propto \alpha_i^{a+\frac{1}{2}-1} e^{-\alpha_i \left(\frac{<x_i^2>}{2} + b \right)}, \end{aligned} \quad (37)$$

where c_{α_i} is the normalization constant. From (37), we can come to a conclusion that variational approximation of $q_{\alpha_i}(\alpha_i)$ is following a Gamma distribution. The mean of the resulting Gamma pdf is given as follows

$$< \alpha_i > = \frac{a + \frac{1}{2}}{\left(\frac{<x_i^2>}{2} + b \right)}, \text{ where } < x_i^2 > = \mu_i^2 + \sigma_i^2. \quad (38)$$

Algorithm 2 VBI based ESM-OTFS Detector

Given: $\mathbf{y}_{ESM}, \mathbf{H}, C_{ESM}, \gamma$.
Initialization: a, b are considered to be significantly low, on the order of 10^{-10}. $\alpha_i^0 = \frac{a}{b}, \forall i$ and $\sigma_i^{2,0} = \frac{1}{\|\mathbf{H}_i\|^2\gamma+\alpha_i^0}$, $x^0 = \mathbf{0}$.
At iteration $l+1$,

1. Update $\sigma_i^{2,l+1}$, $x_i^{l+1} = \mu_i, \forall i$ from (36) using x_{i-}^{l+1} and x_{i+}^{l}.
2. Calculate $< x_i^{2,l+1} >$ from (38) and update α_i^l.
3. Repeat steps $1-3$ until the algorithm reaches convergence. Convergence condition is $\frac{\left\|x^{(l+1)}-x^{(l)}\right\|^2}{\left\|x^{(l)}\right\|^2} \leq 10^{-18}$.
4. Compute $\hat{\mathbf{X}}_{ESM}^{VBI}$ using all the point estimates μ_i from step-3.
5. Estimate the transmitted ESM vectors using (21) and form $\hat{\mathbf{X}}_{ESM}$.
6. Retrieve the transmitted bits using ESM mapping Rule.

Now the equalized version of the transmitted signal matrix $\hat{\mathbf{X}}_{ESM}^{VBI}$ in the DD domain is formed by collecting all the point estimates μ_i where $i = 1, 2 \cdots N_T MN$. Again, the minimum Euclidean distance detector is applied on each row of $\hat{\mathbf{X}}_{ESM}^{VBI}$ using (21) to obtain $\hat{\mathbf{X}}_{ESM}$ as depicted in (22). Now each row of $\hat{\mathbf{X}}_{ESM}$ is decoded using the ESM mapping rule as illustrated in Table 2 to retrieve an estimated of the original transmitted bits. The detection of the entire ESM-OTFS frame is accomplished in this manner. The steps involved in detection of ESM-OTFS using VBI detector is summarized in Algorithm 2.

4. The Computational Complexity of the Detector

The signal detection algorithm at the receiver is composed of two distinct components that contribute to its overall complexity. First part is the computational complexity of the algorithm used for equalisation and second part is the complexity of the demodulation algorithm. MMSE detector utilises (20) for equalisation which has a computational complexity of $O(M^3N^3N_T^3)$ [21]. The minimum euclidean distance detector is used for demodulation purpose which has a computational load of $O(MN\mathbb{C})$, where $\mathbb{C}$ represents the size of ESM constellation as depicted in Table 2. There is an additional search complexity of $O(\mathbb{C})$ for decoding the bits.

The VBI based detector does not require any matrix inversion operation for the equalisation operation. It has a computational complexity of $O(M^2N^2N_T^2L)$ where L is the number of iteration required for convergence of the algorithm. Since VBI based detector also uses euclidean distance based detector for the demodulation part, it has the same computational complexity in comparison to MMSE detector for demodulation.

5. Spectral Efficiency

Predominantly spectral efficiency (SE) has been characterised as the capacity of data transmission within a given bandwidth and can be expressed as a ratio between the information rate and the total bandwidth occupied. The total time span of an ESM-OTFS frame is NT and the total bandwidth engaged by a frame is $M\Delta f$. The number of bits transmitted by an ESM-OTFS frame is $MNlog_2(\mathbb{C})$. So we can express the SE of OTFS-ESM as follows

$$\begin{aligned} SE_{ESM-OTFS} &= \frac{MN \times log_2(\mathbb{C})}{NT \times M\Delta f} \\ &= log_2(\mathbb{C}), \end{aligned} \tag{39}$$

where $T\Delta f = 1$. The SE of the OTFS-SM and OTFS-QSM [20] systems can be represented using the same parameters as

$$\begin{aligned} SE_{SM-OTFS} &= log_2(M_{mod}) + log_2(N_T) \\ SE_{QSM-OTFS} &= log_2(M_{mod}) + 2log_2(N_T), \end{aligned} \tag{40}$$

where M_{mod} is size of the modulation alphabet used. As an example, consider an ESM-OTFS system with QPSK as primary modulation and $N_T = 2$. The ESM constellation corresponding to this system configuration is depicted in Table 2 and from this, we can see that there are 16 distinct constellation vectors. So the spectral efficiency of ESM-OTFS is 4 bits/s/Hz and this is also referred to as bit per channel use (bpcu). Under the same settings, SE of SM-OTFS and QSM-OTFS are 3 and 4 bits/s/Hz respectively.

6. Simulation Results and Discussions

In this section, we discuss the findings from the simulations of the bit error rate (BER) performance of ESM-OTFS compared to SM-OTFS and QSM-OTFS under different system configurations. It is assumed that the receiver has the perfect channel state information (CSI) and all the channels have the Rayleigh fading. The main simulation parameters are given in Table 3.

The BER performance comparison of ESM-OTFS, QSM-OTFS, SM-OTFS and SIMO-OTFS for 4 bpcu is given in Fig. 2. The constellation used in each scheme is indicated in the legend of the figure and in case of ESM-OTFS, primary constellation is indicated. A MIMO setup of 2×4 is used for this simulation. It can be seen from Fig. 2 that ESM-OTFS outperforms other schemes in BER performance. At the BER value of 10^{-4}, ESM-OTFS outperforms QSM-OTFS by 2.1 dB, SM-OTFS by 3 dB and SIMO-OTFS by 5.3 dB.

The BER performance comparison of of ESM-OTFS, QSM-OTFS, SM-OTFS and SIMO-OTFS for 6 bpcu is given in Fig. 3. A MIMO setup of 4×4 is used for this simulation. At the BER value of 10^{-3}, ESM-OTFS outperforms QSM-OTFS by 2 **dB** and SM-OTFS by 3.8 **dB**.

Figure 4 depicts the BER performance of ESM-OTFS for 8 bpcu using various detectors. A MIMO setup of $4 \times$

Table 3 Simulation parameters.

System Parameters	Value
Mobile user velocity (v)	120 km/h
Carrier Frequency (f_c)	4 Ghz
Number of Subcarriers (M)	4
Number of OFDM symbols (N)	4
Subcarrier Spacing (Δf)	15 KHz

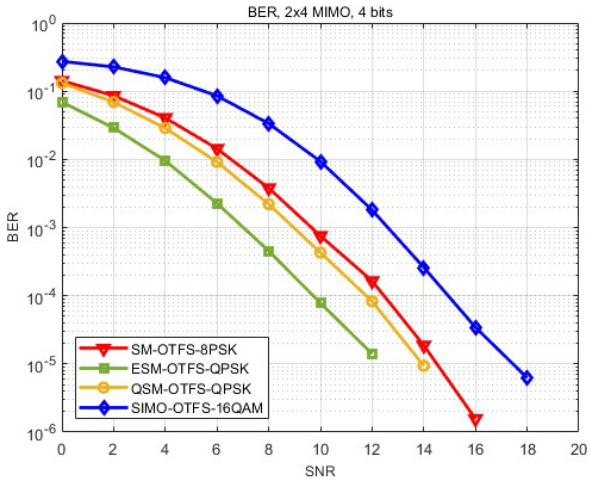


Fig. 2 BER performance of ESM-OTFS, QSM-OTFS, SM-OTFS and SIMO-OTFS systems based on MMSE detector with $M = 4$, $N = 4$ for 4 bits/s/Hz.

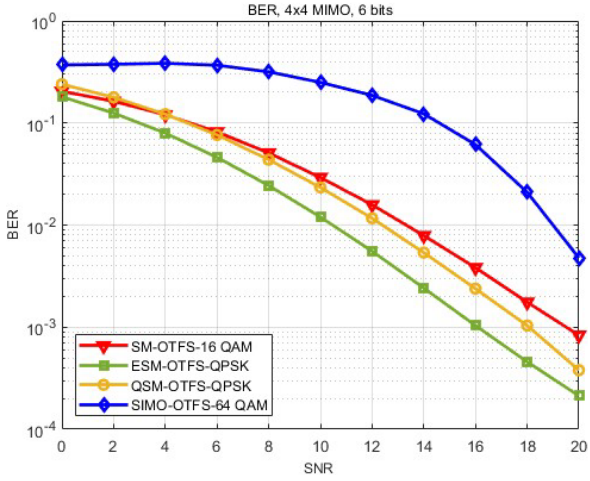


Fig. 3 BER performance of ESM-OTFS, QSM-OTFS, SM-OTFS and SIMO-OTFS systems based on MMSE detector with $M = 4$, $N = 4$ for 6 bits/s/Hz

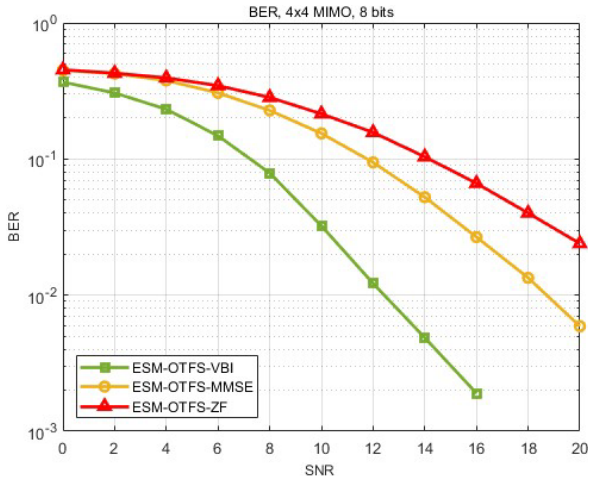


Fig. 4 BER performance of ESM-OTFS using ZF, MMSE and VBI detectors with $M = 4$, $N = 4$ for 8 bits/s/Hz.

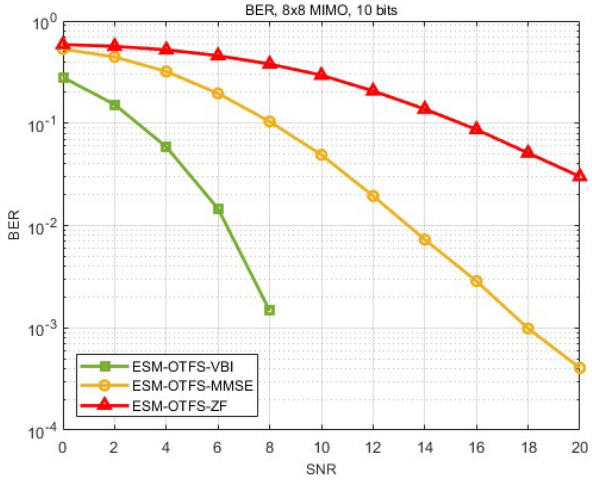


Fig. 5 BER performance of ESM-OTFS using ZF, MMSE and VBI detectors with $M = 4$, $N = 4$ for 10 bits/s/Hz

4 is used for this simulation and 16-QAM is used as the primary modulation. It can be shown that VBI based detector outperforms the linear detectors. ESM-OTFS detector has a gain of 6.1 dB at the value of 10^{-2} compared to detection using MMSE detector.

The BER performance of ESM-OTFS for 10 bpcu using various detectors is shown in the Fig. 5. A MIMO setup of 8×8 is used for this simulation and 16-QAM is used as the primary modulation. VBI based detector has a gain of 7.8 dB over MMSE based detector. It is observed from Fig. 4 and Fig. 5 that the performance of VBI based detector improves as the transmit data matrix becomes more sparse.

ESM-OTFS is highly robust to Doppler shifts, which are prevalent in high mobility environments. Figures 6 and 7 demonstrates the performance of ESM-OTFS at different velocities. Figures 6 and 7 showcases the BER performance of ESM-OTFS for 4 bpcu and 6 bpcu respectively. From the figures, we can understand that the BER performance of ESM-OTFS remains almost same from normal velocity of 30 kmph to very high velocity of 500 kmph. There is only a slight performance degradation at higher velocities and we can estimate that ESM-OTFS is highly immune to doppler shift resulting from velocity variations.

The computational complexity of the proposed ESM-OTFS detectors was discussed in Sect. 4. To illustrate further, consider an ESM-OTFS system $M = 4$, $N = 4$ and $N_T = 2$. In order to compare the receiver complexity at different bpcu, QPSK,16-QAM,64-QAM,256-QAM are used as the primary constellation to have bpcu of 4, 6, 8 and 10 respectively. Since the computational complexity of the equalisation algorithms depend on M, N and N_T, the computational complexity remains same in all four cases of bpcu. Since the demodulation algorithm has a computational complexity of $O(MN\mathbb{C})$, the computational load in terms of multiplication increases with increase in size of $\mathbb{C}$. As we know, the bpcu of the ESM-OTFS system is $log_2(\mathbb{C})$, and the

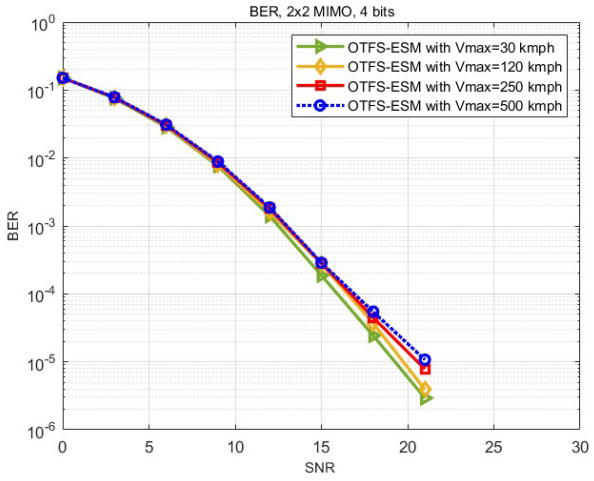


Fig. 6 BER performance of ESM-OTFS at different velocities $M = 16$, $N = 16$ for 4 bits/s/Hz.

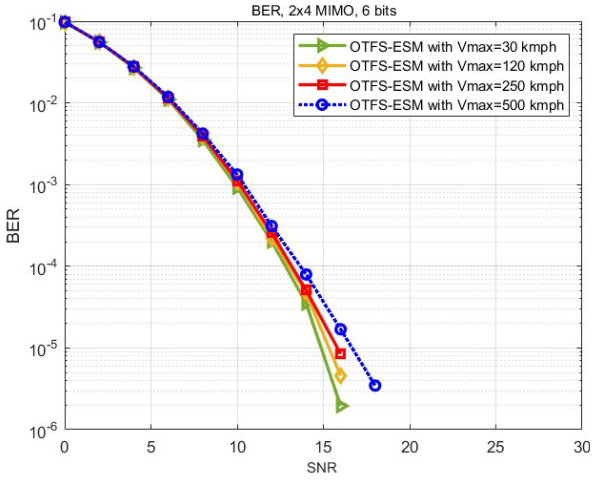


Fig. 7 BER performance of ESM-OTFS at different velocities $M = 16$, $N = 16$ for 6 bits/s/Hz.

computational complexity of the detector increases with an increase in bpcu. This is highlighted in Fig. 8.

As the modulation order increases, the receiver must perform more sophisticated signal processing to distinguish between closely spaced constellation points, thereby increasing the computational burden. The graph plotting computational complexity against bpcu (Fig. 8) effectively captures this relationship. As bpcu increases — moving from QPSK to 256-QAM — the computational complexity of the receiver escalates due to the need for more refined signal processing techniques to accurately decode the higher-order modulated signals. This trade-off is critical in designing and optimizing communication systems for specific application scenarios, balancing the need for high data rates (and thus higher bpcu) against the constraints of receiver complexity, power consumption, and real-time processing capabilities.

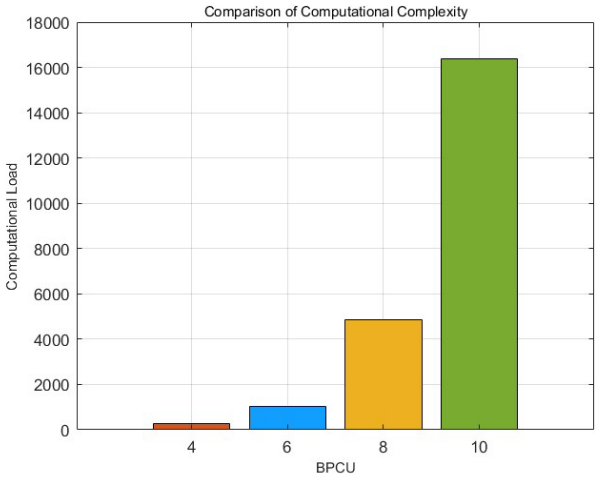


Fig. 8 Trade-off between detector complexity and bpcu in ESM-OTFS systems.

7. Conclusion

In this paper, we propose an ESM-OTFS scheme which is suitable for high doppler shift wireless communication environments and enhances the system reliability and spectral efficiency. The system model and signal processing of ESM-OTFS have been discussed. A novel detector based Variational Bayesian Inference has been proposed for the detection of ESM-OTFS. It is demonstrated through the simulation that ESM-OTFS has a better BER performance compared to SM-OTFS and QSM-OTFS. It is also verified that VBI based ESM-OTFS detector outperforms the linear detectors.

References

[1] R. Hadani, S. Rakib, M. Tsatsanis, A. Monk, A.J. Goldsmith, A.F. Molisch, and R. Calderbank, "Orthogonal time frequency space modulation," 2017 IEEE Wireless Communications and Networking Conference (WCNC), pp.1–6, 2017. doi: 10.1109/WCNC.2017.7925924.

[2] R. Hadani and A. Monk, "OTFS: A new generation of modulation addressing the challenges of 5G," CoRR, vol.abs/1802.02623, 2018, [Online]. Available: http://arxiv.org/abs/1802.02623

[3] R. Hadani, S. Rakib, A.F. Molisch, C. Ibars, A. Monk, M. Tsatsanis, J. Delfeld, A. Goldsmith, and R. Calderbank, "Orthogonal time frequency space (OTFS) modulation for millimeter-wave communications systems," 2017 IEEE MTT-S International Microwave Symposium (IMS), pp.681–683, 2017. doi: 10.1109/MWSYM.2017.8058662.

[4] M. Kollengode Ramachandran and A. Chockalingam, "MIMO-OTFS in high-Doppler fading channels: Signal detection and channel estimation," 2018 IEEE Global Communications Conference (GLOBECOM), pp.206–212, 2018. doi: 10.1109/GLOCOM.2018.8647394.

[5] P. Raviteja, K.T. Phan, Q. Jin, Y. Hong, and E. Viterbo, "Low-complexity iterative detection for orthogonal time frequency space modulation," 2018 IEEE Wireless Communications and Networking Conference (WCNC), pp.1–6, 2018. doi: 10.1109/WCNC.2018.8377159.

[6] A. Farhang, A. RezazadehReyhani, L.E. Doyle, and B. Farhang-Boroujeny, "Low complexity modem structure for OFDM-based

orthogonal time frequency space modulation," IEEE Wireless Commun. Lett., vol.7, no.3, pp.344–347, 2018, doi: 10.1109/LWC.2017.2776942.
[7] G.D. Surabhi, R.M. Augustine, and A. Chockalingam, "Peak-to-average power ratio of OTFS modulation," IEEE Commun. Lett., vol.23, no.6, pp.999–1002, 2019, doi: 10.1109/LCOMM.2019.2914042.
[8] S.S. Das, V. Rangamgari, S. Tiwari, and S.C. Mondal, "Time domain channel estimation and equalization of CP-OTFS under multiple fractional Dopplers and residual synchronization errors," IEEE Access, vol.9, pp.10561–10576, 2021, doi: 10.1109/ACCESS.2020.3046487.
[9] P. Raviteja, K.T. Phan, Y. Hong, and E. Viterbo, "Interference cancellation and iterative detection for orthogonal time frequency space modulation," IEEE Trans. Wireless Commun., vol.17, no.10, pp.6501–6515, 2018, doi: 10.1109/TWC.2018.2860011.
[10] P. Yang, Y. Xiao, Y.L. Guan, M. Di Renzo, S. Li, and L. Hanzo, "Multidomain index modulation for vehicular and railway communications: A survey of novel techniques," IEEE Veh. Technol. Mag., vol.13, no.3, pp.124–134, 2018, doi: 10.1109/MVT.2018.2814023.
[11] W. Shen, L. Dai, J. An, P. Fan, and R.W. Heath, "Channel estimation for orthogonal time frequency space (OTFS) massive MIMO," IEEE Trans. Signal Process., vol.67, no.16, pp.4204–4217, 2019, doi: 10.1109/TSP.2019.2919411.
[12] Y. Liu, S. Zhang, F. Gao, J. Ma, and X. Wang, "Uplink-aided high mobility downlink channel estimation over massive MIMO-OTFS system," IEEE J. Sel. Areas Commun., vol.38, no.9, pp.1994–2009, 2020, doi: 10.1109/JSAC.2020.3000884.
[13] Y. Shan and F. Wang, "Low-complexity and low-overhead receiver for OTFS via large-scale antenna array," IEEE Trans. Veh. Technol., vol.70, no.6, pp.5703–5718, 2021, doi: 10.1109/TVT.2021.3072667.
[14] R.Y. Mesleh, H. Haas, S. Sinanovic, C.W. Ahn, and S. Yun, "Spatial modulation," IEEE Trans. Veh. Technol., vol.57, no.4, pp.2228–2241, 2008, doi: 10.1109/TVT.2007.912136.
[15] C. Zhang, D. Feng, M. Liu, and B. Bai, "Spatial modulation based MIMO-OTFS transmissions," 2021 IEEE/CIC International Conference on Communications in China (ICCC Workshops), pp.427–432, 2021. doi: 10.1109/ICCCWorkshops52231.2021.9538858.
[16] C.-C. Cheng, H. Sari, S. Sezginer, and Y.T. Su, "Enhanced spatial modulation with multiple signal constellations," IEEE Trans. Commun., vol.63, no.6, pp.2237–2248, 2015, doi: 10.1109/TCOMM.2015.2422306.
[17] M. Beal, "Variational algorithms for approximate Bayesian inference," Ph.D. Thesis, University of London, Jan. 2003.
[18] D.G. Tzikas, A.C. Likas, and N.P. Galatsanos, "The variational approximation for Bayesian inference," IEEE Signal Process. Mag., vol.25, no.6, pp.131–146, 2008, doi: 10.1109/MSP.2008.929620.
[19] C.K. Thomas and D. Slock, "SAVE - SPACE alternating variational estimation for sparse Bayesian learning," 2018 IEEE Data Science Workshop (DSW), pp.11–15, 2018. doi: 10.1109/DSW.2018.8439891.
[20] Y. Cai, Z. Bai, H. Liu, Y. Yang, K. Pang, X. Hao, and Y. Cao, "Quadrature spatial modulation based orthogonal time frequency spatial system," AEU - International Journal of Electronics and Communications, vol.160, p.154495, 2023, doi: 10.1016/j.aeue.2022.154495.
[21] P. Singh, A. Gupta, H.B. Mishra, and R. Budhiraja, "Low-complexity ZF/MMSE MIMO-OTFS receivers for high-speed vehicular communication," IEEE Open J. Commun. Soc., vol.3, pp.209–227, 2022, doi: 10.1109/OJCOMS.2022.3147569.
[22] R. Mesleh, S.S. Ikki and H.M. Aggoune, "Quadrature spatial modulation," IEEE Trans. Veh. Technol., vol.64, no.6, pp.2738–2742, June 2015, doi: 10.1109/TVT.2014.2344036.
[23] A. Younis, N. Serafimovski, R. Mesleh, and H. Haas, "Generalised spatial modulation," 2010 Conference Record of the Forty Fourth Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, USA, pp.1498–1502, 2010, doi: 10.1109/ACSSC.2010.5757786.
[24] M. Di Renzo, H. Haas, A. Ghrayeb, S. Sugiura, and L. Hanzo, "Spatial modulation for generalized MIMO: Challenges, opportunities, and implementation," Proc. IEEE, vol.102, no.1, pp.56–103, Jan. 2014, doi: 10.1109/JPROC.2013.2287851.
[25] S. Ganesan, R. Mesleh, H. Ho, C.W. Ahn, and S. Yun, "On the performance of spatial modulation OFDM," 2006 Fortieth Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, USA, pp.1825–1829, 2006, doi: 10.1109/ACSSC.2006.355077.
[26] Y. Yang, Z. Bai, K. Pang, P. Ma, H. Zhang, X. Yang, and D. Yuan, "Design and analysis of spatial modulation based orthogonal time frequency space system," China Commun., vol.18, no.8, pp.209–223, Aug. 2021, doi: 10.23919/JCC.2021.08.015.
[27] B. Li, Z. Bai, J. Guo, Y. Yang, M. Yan, and X. Hao, "Generalized spatial modulation based orthogonal time frequency space system," 2021 IEEE 94th Vehicular Technology Conference (VTC2021-Fall), Norman, OK, USA, pp.1–5, 2021, doi: 10.1109/VTC2021-Fall52928.2021.9625452.

Anoop A received BTech degree in electronics and communication engineering from National Institute of Technology, Calicut, India in 2010. He is currently an MS-PhD student in the Department of ECE at the Indian Institute of Information Technology (IIIT) Kottayam, Kerala, India. His research interests include OTFS modulation, OTFS-based index and spatial modulation schemes.

Christo K. Thomas received his BS in Electronics and Communication Engineering from National Institute of Technology, Calicut, India in year 2010, his MS in Telecommunication Engineering from Indian Institute of Science, Bangalore, India in year 2012, and his PhD from EURECOM, France in year 2020. He is currently a postdoctoral fellow at the Electrical and Computer Engineering Department at Virginia Tech. His research interests include semantic communications, statistical signal processing, and machine learning for wireless communications. From 2012 to 2014, he was a staff design engineer on 4G LTE with Broadcom communications, Bangalore, and from 2014 to 2017, he was a design engineer with Intel corporation, Bangalore. During November 2020 till June 2022, he was a staff engineer on 5G modems with wireless research and development division of Qualcomm Inc., Espoo, Finland. He was a recipient of the best student paper award at IEEE SPAWC 2018, Kalamata, Greece, and also received third prize for his team titled "Learned Chester" ML5G-PHY channel estimation challenge, as part of the ITU AI/ML in 5G challenge, conducted at NCSU, US, 2020.

Kala S (Senior Member, IEEE) received BTech degree in electronics and communication engineering from MG University, India in 2006 and MS (Engg) from CeNSE, Indian Institute of Science Bangalore (IISc), India in 2013. She received her PhD degree in Electronics Engineering from Cochin University of Science and Technology, India in 2020. She is currently an Assistant Professor in the Department of ECE at the Indian Institute of Information Technology (IIIT) Kottayam, Kerala, India. Her research interests include FPGA based system design, Wireless Communications, hardware acceleration of deep learning algorithms and DSP algorithms, neuromorphic architectures and hardware security.

PAPER

Secrecy Analysis of Transmit Antenna Selection with Hybrid GSC/SC in AF-MIMO Relay Systems

Donghun LEE[†a)], ***Member***

SUMMARY This paper studies the secrecy outage probability of transmit antenna selection (TAS) with hybrid generalized selection combining (GSC)/selection combining (SC) in amplify and forward (AF)-multiple input multiple output (MIMO) relay system. This paper derives the exact cumulative distribution (CDF) expression of the received signal to noise ratio (SNR) for TAS with hybrid GSC/SC system. Using derived CDF, this work derives the lower bound and asymptotic forms for the hybrid combining system for the secrecy outage probability. Asymptotic results shows that the proposed hybrid system provides the secrecy diversity of product of the number of antennas in the relay node and the number with the smaller number of antennas among the source node and user node. An interesting result is that the secrecy diversity order is independent of the number of combining signals and the number of eavesdroppers.

key words: *transmit antenna selection (TAS), generalized selection combining (GSC), amplify and forward (AF)*

1. Introduction

In physical layer security, transmit antenna selection (TAS) is a promising technique to improve secrecy performance in [1], [2]. The previous works in [3]–[5] studied the secrecy analysis of TAS in multiple-input multiple-output (MIMO) system. The work in [3] investigated the secrecy performance of the TAS-based system where maximal-ratio combining (MRC) and selection combining (SC) were considered. The work in [4] derived the exact and asymptotic secrecy performance of TAS with generalized selection combining (GSC) over Rayleigh fading channels. The work in [5] investigated the secrecy performance of TAS-based system where GSC was considered.

Amplify and forward (AF)-relay system in physical layer security has received great attention, which can improve the throughput, reliability and service coverage of wireless networks in [6], [7]. For this reason, the TAS-based AF-relay system has been researched in [8]–[10]. The study in [8] assessed how correlated fading channels affect the secrecy performance of TAS at source node. The work in [9] studied the secrecy performance of TAS and user selection over Rayleigh fading channels. The work in [10] investigated the TAS-based AF relay system where relay and user selection was considered.

Even if GSC offers many performance and implementation benefits in [11]–[13], previous studies have generally given limited consideration to investigating the secrecy performance of the TAS-based AF relay system employing GSC. Since GSC requires a considerable amount of power consumption to keep all antennas always active in [14], this work adapts GSC at relay node and SC at user node. Thus, this paper investigates secrecy analysis of TAS with hybrid GSC/SC in AF-MIMO relay system.

The first contribution of this work is to derive a new exact cumulative distribution function (CDF) representation of the received SNR for TAS with hybrid GSC/SC. Another contribution of this research is the derivation of lower bound and asymptotic expressions for the secrecy outage probability in TAS with hybrid GSC/SC. Through asymptotic analysis, this study quantifies the secrecy diversity order associated with the secrecy outage probability.

2. System Overview

This work studies the secrecy outage probability of TAS with hybrid combining in AF-MIMO relay system over Rayleigh fading channels. In the proposed system, the relay node and user node adapt GSC and SC, respectively. The source, relay and user nodes in the proposed hybrid combining (i.e., GSC/SC) have M, N, and P antennas and an eavesdropper has E antennas. The number of antennas on source-side and user-side are the same at the relay node. TAS is applied to the source node and relay node maximizing the receive SNR at each node. Thus, the source node selects the transmission antenna for the source node and the relay node using feedback information. Feedback information from the relay node and user node to the source node contains the largest SNR value and the corresponding antenna index. In this work, γ_r and γ_u denote the average SNR at the relay node and user node, respectively, and γ_e denotes the average SNR at eavesdropper. This work considers relay link (i.e., source-relay link and relay-user link), not direct link (i.e., source-user link).

CDF and probability density function (PDF) of the single received antenna output SNR is given by [15]

$$F_{\mathrm{sa}}(x) = 1 - e^{-\frac{x}{\gamma}} \tag{1}$$

$$f_{\mathrm{sa}}(x) = \frac{1}{\gamma} e^{-\frac{x}{\gamma}} \tag{2}$$

and the CDF for the GSC's output SNR is written by [11]

$$F_{\mathrm{gsc}}(x) = c_0 + \sum_{l=1}^{\mathrm{L}} \frac{c_l x^{l-1}}{(l-1)!} e^{-\frac{x}{\gamma}} + \sum_{l=\mathrm{L}+1}^{\mathrm{N}} c_l e^{-\frac{lx}{\mathrm{L}\gamma}} \tag{3}$$

Manuscript received September 25, 2023.
Manuscript revised April 2, 2024.
Manuscript publicized May 14, 2024.

[†]Division of Information & Communication Engineering, Kongju National University, Kongju, 31080, Rep. of Korea.
a) E-mail: mmdang@kongju.ac.kr

DOI: 10.23919/transcom.2023EBP3154

where N is received antennas, L denotes the number of combining signals from the L largest received antennas among N. γ denotes the average receive SNR. The coefficient of c_l is given by [11]

$$c_l = \begin{cases} 1, & \text{l=0} \\ \left[\sum_{k=\mathrm{L}+1}^{\mathrm{N}} \frac{(-1)^{k-l}\binom{\mathrm{N}}{\mathrm{N}-k}\binom{k-1}{k-\mathrm{L}-1}}{(\frac{k}{\mathrm{L}}-1)^{\mathrm{L}-l+1}} - 1\right]\gamma^{-l+1}, \\ 1 \le l < \mathrm{L} \\ -\gamma^{-\mathrm{L}+1}\binom{\mathrm{N}}{\mathrm{N}-\mathrm{L}}, \quad l = \mathrm{L} \\ (-1)^l(\frac{l}{\mathrm{L}}-1)^{-\mathrm{L}}\binom{\mathrm{N}}{\mathrm{N}-l}\binom{l-1}{l-\mathrm{L}-1}, \quad \mathrm{L} < l \le \mathrm{N} \end{cases}. \tag{4}$$

3. Secrecy Outage Probability

3.1 Distribution

CDF of the received SNR for TAS with hybrid GSC/SC can be obtained by using [16]

$$F_{\mathrm{hyb}}(x) = 1 - \int_0^\infty \left(1 - F_{1,\mathrm{gsc}}(x + \frac{x^2}{t})\right) f_{2,\mathrm{sc}}(t+x)dt. \tag{5}$$

CDF of the GSC output SNR in Eq. (3) can be represented to make more compact form given by

$$F_{\mathrm{gsc}}(x) = \sum_{l_r=0}^{\mathrm{N}} c_{l_r}\left(\frac{x}{\gamma}\right)^{a_{l_r}} e^{-b_{l_r}\frac{x}{\gamma}} \tag{6}$$

where the coefficients a_{l_r}, b_{l_r} and c_{l_r} are given by

$$a_{l_r} = \begin{cases} 0, & l_r = 0, \ \mathrm{L} < l_r \le \mathrm{N} \\ l_r - 1, & 1 \le l_r \le \mathrm{L}, \end{cases} \tag{7}$$

$$b_{l_r} = \begin{cases} 0, & l_r = 0 \\ 1, & 1 \le l_r \le \mathrm{L} \\ \frac{l_r}{\mathrm{L}}, & \mathrm{L} < l_r \le \mathrm{N}, \end{cases} \tag{8}$$

and

$$c_{l_r} = \begin{cases} 1, & l_r = 0 \\ \sum_{k=\mathrm{L}+1}^{\mathrm{N}} \frac{(-1)^{k-l_r}\binom{\mathrm{N}}{\mathrm{N}-k}\binom{k-1}{k-\mathrm{L}-1}}{(l_r-1)!(\frac{k}{\mathrm{L}}-1)^{\mathrm{L}-l_r+1}} - \frac{1}{(l_r-1)!}, & 1 \le l_r < \mathrm{L} \\ -\frac{1}{(l_r-1)!}\binom{\mathrm{N}}{\mathrm{N}-\mathrm{L}}, & l_r = \mathrm{L} \\ (-1)^{l_r}(\frac{l_r}{\mathrm{L}}-1)^{-\mathrm{L}}\binom{\mathrm{N}}{\mathrm{N}-l_r}\binom{l_r-1}{l_r-\mathrm{L}-1}, & \mathrm{L} < l_r \le \mathrm{N}. \end{cases} \tag{9}$$

Using order statistics and Eq. (6), CDF of the received SNR for the relay node with GSC is given by

$$F_{1,\mathrm{gsc}}(x) = \left[F_{\mathrm{gsc}}(x)\right]^{\mathrm{M}} \tag{10}$$

and using the binomial expansion, Eq. (10) can be derived by

$$\begin{aligned} F_{1,\mathrm{gsc}}(x) &= \sum_{\mathrm{M}}\left[\prod_{r=1}^{\mathrm{M}} c_{l_r}\right]\left(\frac{x}{\gamma_r}\right)^{d_1} e^{-e_1\frac{x}{\gamma_r}} \\ &= 1 + \sum_{l_1=l_2=\cdots=l_{\mathrm{M}}\neq 0}\left[\prod_{r=1}^{\mathrm{M}} c_{l_r}\right]\left(\frac{x}{\gamma_r}\right)^{d_1} e^{-e_1\frac{x}{\gamma_r}} \end{aligned} \tag{11}$$

where d_1, e_1 and $\sum_{\mathrm{M}}$ are denoted by,

$$d_1 = \sum_{r=1}^{\mathrm{M}} a_{l_r}, \qquad e_1 = \sum_{r=1}^{\mathrm{M}} b_{l_r} \tag{12}$$

and

$$\sum_{\mathrm{M}} = \sum_{l_1=0}^{\mathrm{N}} \cdots \sum_{l_r=0}^{\mathrm{N}} \cdots \sum_{l_{\mathrm{M}}=0}^{\mathrm{N}}. \tag{13}$$

Using order statistics and Eq. (1), Eq. (2), the received SNR's PDF for the SC-based user node is given by

$$f_{2,\mathrm{sc}}(x) = \mathrm{NP}\left[F_{\mathrm{sa}}(x)\right]^{\mathrm{NP}-1} f_{\mathrm{sa}}(x) \tag{14}$$

and using the binomial expansion, Eq. (14) can be obtained by

$$f_{2,\mathrm{sc}}(x) = \mathrm{NP}\sum_{\mathrm{NP}-1}\frac{(-1)^{d_2}}{\gamma_u} e^{-(d_2+1)\frac{x}{\gamma_u}} \tag{15}$$

where d_2 and $\sum_{\mathrm{NP}-1}$ are denoted by,

$$d_2 = \sum_{r=1}^{\mathrm{NP}-1} q_r, \qquad \sum_{\mathrm{NP}-1} = \sum_{q_1=0}^{1}\cdots\sum_{q_r=0}^{1}\cdots\sum_{q_{\mathrm{NP}-1}=0}^{1}. \tag{16}$$

Using the binomial expansion, Eq. (11) and Eq. (15) can be rewritten by

$$\begin{aligned} F_{1,\mathrm{gsc}}(x + \frac{x^2}{t}) &= 1 + \sum_{l_1=l_2=\cdots=l_{\mathrm{M}}\neq 0}\sum_{i=0}^{d_1}\left[\prod_{r=1}^{\mathrm{M}} c_{l_r}\right] \\ &\quad \times \binom{d_1}{i}\left(\frac{1}{\gamma_r}\right)^{d_1} x^{d_1+i} t^{-i} e^{-\frac{e_1}{\gamma_r}(x+\frac{x^2}{t})} \end{aligned} \tag{17}$$

and

$$f_{2,\mathrm{sc}}(x+t) = \mathrm{NP}\sum_{\mathrm{NP}-1}\frac{(-1)^{d_2}}{\gamma_u} e^{-\frac{(d_2+1)}{\gamma_u}(x+t)}. \tag{18}$$

Substituting Eq. (17) and Eq. (18) into Eq. (5), the exact CDF of the received SNR for TAS with hybrid GSC/SC is obtained by using Eq. (3.471.9) in [17]

$$\begin{aligned} F_{\mathrm{hyb}}(x) &= 1 - 2\mathrm{NP}\sum_{\mathrm{NP}-1}\sum_{l_1=l_2=\cdots=l_{\mathrm{M}}\neq 0}\sum_{i=0}^{d_1}\left[\prod_{r=1}^{\mathrm{M}} c_{l_r}\right] \\ &\quad \times \binom{d_1}{i}\frac{(-1)^{d_2}}{\sigma}\left(\frac{\sigma e_1}{d_2+1}\right)^{\frac{-i+1}{2}}\left(\frac{x}{\gamma_r}\right)^{d_1+1} \\ &\quad \times e^{-(e_1+\frac{(d_2+1)}{\sigma})\frac{x}{\gamma_r}} K_{-i+1}\left(\frac{2x}{\gamma_r}\sqrt{e_1\frac{(d_2+1)}{\sigma}}\right) \end{aligned} \tag{19}$$

where $\gamma_u = \sigma\gamma_r$ and $K(\cdot)$ denotes the modified Bessel function of second kind.

Similar to Eq. (15), using order statistics and Eq. (1), Eq. (2), PDF of the received SNR for the eavesdropper with SC is given by

$$f_{e,\mathrm{sc}}(x) = \mathrm{E}\sum_{\mathrm{E}-1}\frac{(-1)^{d_e}}{\gamma_e} e^{-(d_e+1)\frac{x}{\gamma_e}} \tag{20}$$

where d_e and $\sum_{\mathrm{E-1}}$ are denoted by,

$$d_e = \sum_{j=1}^{\mathrm{E-1}} q_j, \quad \sum_{\mathrm{E-1}} = \sum_{q_1=0}^{1} \cdots \sum_{q_j=0}^{1} \cdots \sum_{q_{\mathrm{E-1}}=0}^{1}. \tag{21}$$

3.2 Lower Bound Analysis

Because the secrecy outage probability of TAS with hybrid GSC/SC in AF-MIMO relay system is given by [3]

$$P_{\mathrm{out}} = \int_0^{\infty} F_{\mathrm{hyb}}(2^{C_0}(1+x)-1) f_{e,\mathrm{sc}}(x)dx, \tag{22}$$

the lower bound secrecy outage probability can be evaluated by [18]

$$P_{\mathrm{lb}} = \int_0^{\infty} F_{\mathrm{hyb}}(2^{C_0}x) f_{e,\mathrm{sc}}(x)dx, \tag{23}$$

where C_0 is target capacity.

Substituting Eq. (19) and Eq. (20) into Eq. (23), the secrecy outage probability of TAS with hybrid GSC/SC can be lower bounded by using integral table [19]

$$\begin{aligned} P_{\mathrm{lb}} &= 1 - 2\mathrm{NPE} \sum_{\mathrm{NP-1}} \sum_{\mathrm{E-1}} \sum_{l_1=l_2=\cdots=l_{\mathrm{M}}\neq 0} \sum_{i=0}^{d_1} \Big[\prod_{r=1}^{\mathrm{M}} c_{l_r}\Big] \\ &\quad \times \binom{d_1}{i} \frac{(-1)^{d_2+d_e}}{\sigma} \Big(\frac{\sigma e_1}{d_2+1}\Big)^{\frac{-i+1}{2}} 2^{C_0(d_1+1)} \frac{1}{\gamma_e} \frac{1}{\gamma_r^{d_1+1}} \\ &\quad \times \frac{h^{\nu-f} g^{-\nu} \sqrt{\pi}}{2^f} \frac{\Gamma(f-\nu)\Gamma(f+\nu)}{\Gamma(f+\frac{1}{2})} \\ &\quad \times {}_2F_1\Big(\frac{f-\nu}{2}, \frac{f-\nu+1}{2}; f+\frac{1}{2}; 1-\frac{g^2}{h^2}\Big), \end{aligned} \tag{24}$$

where $f = d_1 + 2$, $\nu = -i + 1$, $h = 2^{C_0}(e_1 + \frac{(d_2+1)}{\sigma})\frac{1}{\gamma_r} + \frac{d_e+1}{\gamma_e}$ and $g = \frac{2^{C_0+1}}{\gamma_r}\sqrt{e_1 \frac{(d_2+1)}{\sigma}}$.

3.3 Asymptotic Analysis

Moment generating function (MGF) for the GSC output SNR is given by [20]

$$\begin{aligned} \Psi_\gamma(s) &= \mathrm{L}\binom{\mathrm{N}}{\mathrm{L}} \int_0^{\infty} e^{-sx} f_{\mathrm{sa}}(x) [\Psi(s,x)]^{\mathrm{L-1}} \\ &\quad \times [F_{\mathrm{sa}}(x)]^{\mathrm{N-L}} dx, \end{aligned} \tag{25}$$

where $\Psi(s,x)$ is complementary incomplete MGF given by [20]

$$\Psi(s,x) = \frac{1}{(1+s\gamma)} e^{-(s+\frac{1}{\gamma})x}. \tag{26}$$

As $\gamma \to \infty$, $F_{\mathrm{sa}}(x)$ can be approximated by

$$F_{\mathrm{sa}}(x) \approx \frac{x}{\gamma}. \tag{27}$$

Substituting Eq. (2), Eq. (26) and Eq. (27) into Eq. (25), MGF for the GSC output SNR can be approximated by

$$\Psi_\gamma(s) \approx \frac{(\mathrm{N-L})!}{\mathrm{L}^{\mathrm{N-L}}} \binom{\mathrm{N}}{\mathrm{L}} \frac{1}{[\gamma(s+\frac{1}{\gamma})]^{\mathrm{N}}}. \tag{28}$$

Because the Laplace transform of $\mathcal{L}[F_{\mathrm{gsc}}(x)] = \frac{\Psi_\gamma(s)}{s}$,

$$\mathcal{L}[F_{\mathrm{gsc}}(x)] \approx \frac{(\mathrm{N-L})!}{\mathrm{L}^{\mathrm{N-L}}} \binom{\mathrm{N}}{\mathrm{L}} \frac{1}{s[\gamma(s+\frac{1}{\gamma})]^{\mathrm{N}}}, \tag{29}$$

then, by taking the inverse Laplace transform, $\mathcal{L}^{-1}[F_{\mathrm{gsc}}(x)]$, CDF of the GSC output SNR is derived by

$$F_{\mathrm{gsc}}(x) \approx \frac{(\mathrm{N-L})!}{\mathrm{N!L}^{\mathrm{N-L}}} \binom{\mathrm{N}}{\mathrm{L}} \Big(\frac{x}{\gamma}\Big)^{\mathrm{N}}. \tag{30}$$

Using order statistics and Eq. (27), the SC output SNR's CDF for P received antennas is written by

$$F_{\mathrm{sc}}(x) \approx \Big(\frac{x}{\gamma}\Big)^{\mathrm{P}}. \tag{31}$$

By employing Eq. (30) and Eq. (31), CDF of the hybrid GSC/SC output SNR in AF-MIMO relay is approximated by

$$\begin{aligned} F_{\mathrm{hyb}}(x) &\approx \Big[F_{\mathrm{gsc}}(x)\Big]^{\mathrm{M}} + \Big[F_{\mathrm{sc}}(x)\Big]^{\mathrm{N}} \\ &= \left[\frac{(\mathrm{N-L})!}{\mathrm{N!L}^{\mathrm{N-L}}} \binom{\mathrm{N}}{\mathrm{L}}\right]^{\mathrm{M}} \Big(\frac{x}{\gamma_r}\Big)^{\mathrm{MN}} + \Big(\frac{x}{\gamma_u}\Big)^{\mathrm{PN}}. \end{aligned} \tag{32}$$

Consequently, using Eq. (20), Eq. (23) and Eq. (32), the asymptotic secrecy performance of the hybrid GSC/SC is written by

$$\begin{aligned} P_{\mathrm{asy}} &= \mathrm{E}\left[\frac{(\mathrm{N-L})!}{\mathrm{N!L}^{\mathrm{N-L}}} \binom{\mathrm{N}}{\mathrm{L}}\right]^{\mathrm{M}} \sum_{\mathrm{E-1}} (-1)^{d_e} \frac{2^{C_0\mathrm{MN}}(\mathrm{MN})!}{(\mathrm{d_e}+1)^{\mathrm{MN}+1}} \\ &\quad \times \Big(\frac{\gamma_e}{\gamma_r}\Big)^{\mathrm{MN}} + \mathrm{E} \sum_{\mathrm{E-1}} (-1)^{d_e} \frac{2^{C_0\mathrm{PN}}(\mathrm{PN})!}{(\mathrm{d_e}+1)^{\mathrm{PN}+1}} \Big(\frac{\gamma_e}{\sigma\gamma_r}\Big)^{\mathrm{PN}}. \end{aligned} \tag{33}$$

When the secrecy performance is in form of $P_{\mathrm{out}} = (\frac{S}{\gamma})^D$ where S and D denote the secrecy SNR gain and the secrecy diversity order, we can quantify two important parameters using Eq. (33). From above asymptotic analysis, we find that the secrecy diversity order of TAS with hybrid GSC/SC is quantified by product of minimum value of M, P and N (i.e., min[M, P]N). Further, it can be observed that L and E do not affect the secrecy diversity order of MIMO wiretap channels, but affect the secrecy SNR gain.

4. Numerical Results and Discussions

This section presents the TAS-based AF relay system's secrecy performance over Rayleigh fading channels. The secrecy analysis such as lower bound analysis and asymptotic analysis is verified by the Monte-Carlo simulation. Here, γ denotes the average receive SNR of the relay node.

Figure 1 and Fig. 2 plot the secrecy outage probability of TAS with hybrid GSC/SC as a function of L. From Fig. 1, the secrecy outage probability with L = 3 provides the best

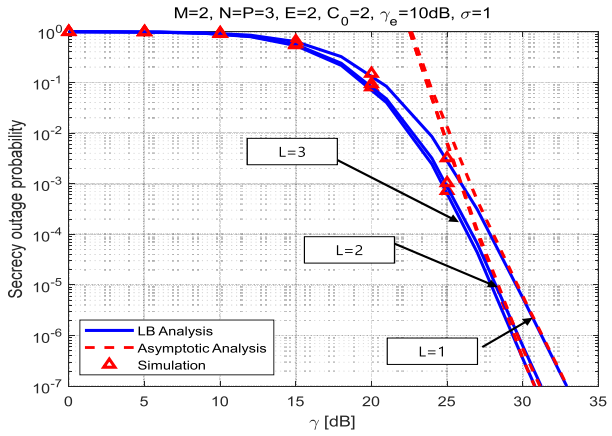


Fig. 1 Secrecy outage probability of TAS with hybrid GSC/SC as a function of L.

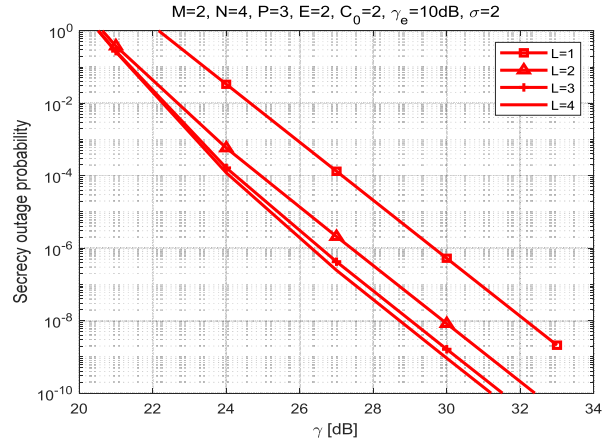


Fig. 2 Asymptotic secrecy outage probability of TAS with hybrid GSC/SC as a function of L.

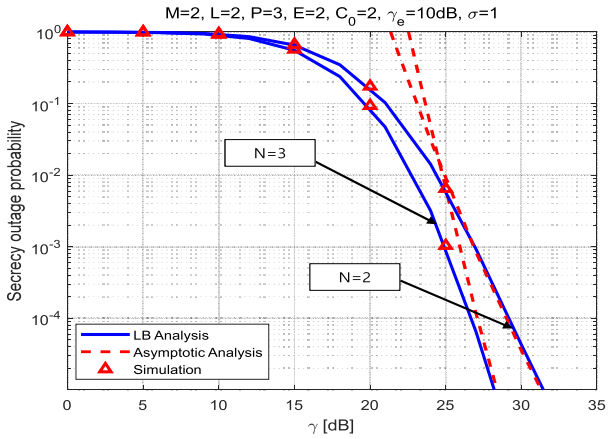


Fig. 3 Secrecy outage probability of TAS with hybrid GSC/SC as a function of N.

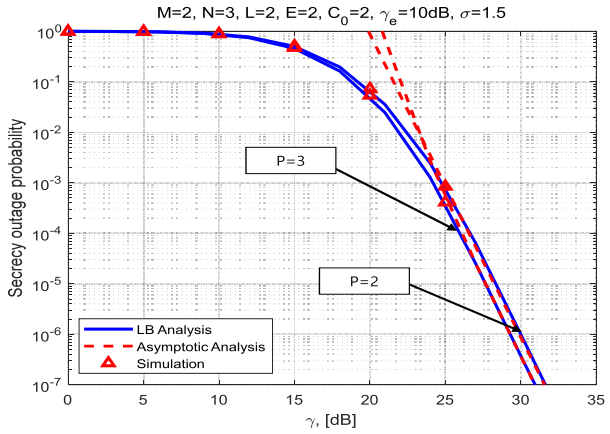


Fig. 4 Secrecy outage probability of TAS with hybrid GSC/SC as a function of P.

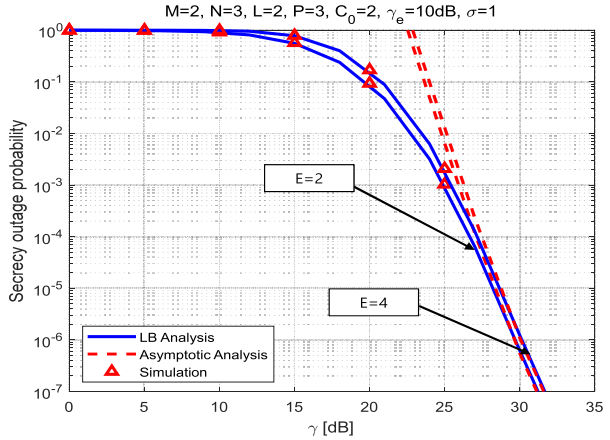


Fig. 5 Secrecy outage probability of the with hybrid GSC/SC as a function of E.

performance and the case with L = 1 shows the worst performance. For this reason, when the required performance is 10^{-5}, the L = 3 case has the SNR gain of 2 dB compared to the L = 1 case. L = 2, 3 cases have the almost same performance. Figure 2 shows asymptotic performance curves of TAS with hybrid GSC/SC for L = 1, 2, 3 and 4. Similarly to Fig. 1, L = 3, 4 cases have the almost same performance. It means that in case of L = N − 1, the secrecy outage probability is almost the same as that of L = N. Further, because of the same secrecy diversity order for all cases in Fig. 2, the SNR gain is constant at high SNR regardless of the requirement of secrecy performance.

Figure 3 is the secrecy outage probability of TAS with hybrid GSC/SC as a function of N. The secrecy outage probability with N = 3 has the better performance than that with N = 2. This is due to that the scenario with N = 3 has a higher secrecy diversity order compared to the case with N = 2, which is quantified in Eq. (33). For this reason, when the required performance is 10^{-5}, the N = 3 scenario has the SNR gain of 3 dB compared to the case with N = 2. In addition, since the secrecy diversity order is proportional to N, the SNR gain increases as the required secrecy performance increases.

Figure 4 and Fig. 5 present the secrecy outage probability of TAS with hybrid GSC/SC as a function of the number of antennas at the user and eavesdropper. From Fig. 4, we know that both cases of P = 2, 3 have the same secrecy di-

versity order as verified in Eq. (33). Since the secrecy diversity order of the proposed hybrid system is independent of E, it can be seen from Fig. 5 that the SNR gain degradation is not large by increasing E. When the required performance is 10^{-5}, the E = 4 case has the SNR gain degradation of 0.5 dB compared to the E = 2 case. Since the secrecy diversity order is proportional to min[M, P]N, the SNR gain in Fig. 4 is consistent at high SNR. In addition, because E is independent of the secrecy diversity order, the SNR gain is constant at high SNR regardless of the required secrecy performance.

5. Conclusions

This work analyzed the secrecy performance of TAS using hybrid GSC/SC in AF-MIMO relay system. This research obtained lower bound and asymptotic expressions for the secrecy performance in the hybrid combining system. From asymptotic results, we can be observed that the secrecy diversity order is proportional to product of the number of antennas at relay node and minimum number of antennas at source and user nodes. It means that the secrecy diversity order is independent of L and E. From numerical results, we found that in case of L = N − 1 and L = N, the secrecy outage probability is almost the same. Further, since the secrecy diversity order is proportional to N, the SNR gain increases as the desired secrecy performance rises.

Acknowledgments

This work was supported by the research grant of Kongju National University in 2024.

References

[1] H. Alves, R.D. Souza, M. Debbah, and M. Bennis, "Performance of transmit antenna selection physical layer security schemes," IEEE Signal Process. Lett., vol.19, no.6, pp.372–375, June 2012.
[2] N. Yang, H.A. Suraweera, I.B. Collings, and C. Yuen, "Physical layer security of TAS/MRC with antenna correlation," IEEE Trans. Inf. Forensics Security, vol.8, no.1, pp.254–259, Jan. 2013.
[3] N. Yang, P. Yeoh, M. Elkashlan, R. Scholber, and I. Collings, "Transmit antenna selection for security enhancement in MIMO wiretap channels," IEEE Trans. Commun., vol.61, no.1, pp.144–154, Jan. 2013.
[4] N. Yang, P.L. Yeoh, M. Elkashlan, R. Schober, and J. Yuan, "MIMO wiretap channels: Secure transmission using transmit antenna selection and receive generalized selection combining," IEEE Commun. Lett., vol.17, no.9, pp.1754–1757, Sept. 2013.
[5] L. Wang, M. Elkashlan, J. Huang, R. Schober, and R. Mallik, "Secure transmission with antenna selection in MIMO Nakagami-*m* fading channels," IEEE Trans. Wireless Commun., vol.13, no.11, pp.6054–6067, Nov. 2014.
[6] A. Sendonaris, E. Erkip, and B. Aazhang, "User cooperation diversity-part I: System description," IEEE Trans. Commun., vol.51, no.11, pp.1927–1938, Nov. 2003.
[7] J.N. Laneman, D.N.C. Tse, and G.W. Wornell, "User cooperation diversity-part I: System description," IEEE Trans. Inf. Theory, vol.50, no.12, pp.3062–3080, Dec. 2004.
[8] L. Fan, R. Zhao, F. Gong, N. Yang, and G. Karagiannidis, "Secure multiple amplify-and-forward relaying over correlated fading channels," IEEE Trans. Commun., vol.65, no.7, pp.2811–2820, July 2017.
[9] M. Shukla, S. Yadav, and N. Purohit, "Secure transmission in cellular multiuser two-way amplify-and-forward relay networks," IEEE Trans. Veh. Technol., vol.67, no.12, pp.11886–11899, Dec. 2018.
[10] D. Lee, "Secrecy analysis of relay-user selection in AS-AF systems over Nakagami fading channels," IEEE Trans. Veh. Technol., vol.70, no.3, pp.2378–2388, March 2021.
[11] X. Cai and G.B. Giannakis, "Performance analysis of combined transmit selection diversity and receive generalized selection combining in Rayleigh fading channels," IEEE Trans. Wireless Commun., vol.3, no.6, pp.1980–1983, Nov. 2004.
[12] I. Ahmed, A. Nasri, R. Schober, and R.K. Mallik, "Asymptotic performance of generalized selection combining in generic noise and fading," IEEE Trans. Commun., vol.60, no.4, pp.916–922, April 2012.
[13] M. Alouini and M.K. Simon, "An MGF-based performance analysis of generalized selection combining over Rayleigh fading channels," IEEE Trans. Commun., vol.48, no.3, pp.401–415, March 2000.
[14] H. Yang, "New results on ordered statistics and anlaysis of minimum-selection gereralized selection combining," IEEE Trans. Wireless Commun., vol.5, no.7, pp.1876–1885, July 2006.
[15] D. Costa and S. Aissa, "Cooperative dual-hop relaying systems with beamforming over Nakagami-m fading channels," IEEE Trans. Wireless Commun., vol.8, no.8, pp.3950–3954, Aug. 2009.
[16] R. Louie, Y. Li, H. Suraweera, and B. Vucetic, "Performance analysis of beamforming in two-hop amplify and forward relay networks with antenna correlation," IEEE Trans. Wireless Commun., vol.8, no.6, pp.3132–3141, June 2009.
[17] I. Gradshteyn, and I. Ryzhik, Table of Integrals, Series, and Products, 7th ed., Academic Press, 2007.
[18] K. Peppas, N. Sagias, and A. Maras, "Physical layer security for multiple-antenna systems: A unified approach," IEEE Trans. Commun., vol.64, no.1, pp.314–328, Jan. 2016.
[19] A. Prudnikov, Y. Brychkov, and O. Marichev, Integrals and Series vol.2: Special Functions, 1st ed., Gordon and Breach Science Publishers, 1986.
[20] Y. Ma, and S. Pasupathy, "Efficient performance evaluation for generalized selection combining on gneralized fading channels," IEEE Trans. Wireless Commun., vol.3, no.1, pp.29–34, Jan. 2004.

Donghun Lee received the B.Eng. and M. Eng. degrees, in electronics engineering, from Chungbuk National University, Korea, in 2001 and 2003. He received Ph.D. degree in information and communications engineering at Gwangju Institute of Science and Technology (GIST) in 2010. From 2010 to 2015, he was with Cognitive Radio Research Section, Electronics and Telecommunications Research Institute, as a senior member of research staff. From 2016 to 2020, he was a Researcher with the Certification Industry Promotion Division, Korean Agency for Technology and Standards. From 2021 to 2023, he was an associate professor in the Department of Electronics Engineering at Chosun University. He is currently working as an associate professor in the Division of Information & Communication Engineering at Kongju National University. His primary research interests include the reliability and security improvement of cognitive radio systems and multiple-input multiple-output systems.

IEICE TRANS. COMMUN., VOL.E107–B, NO.11 NOVEMBER 2024

PAPER

Simulation of Radar Sea Clutter in Correlated Generalized Compound Distribution Based on Improved ZMNL

Yi CHENG[†,††a)], **Kexin LI**[†,††b)], **Chunbo XIU**[†c)], ***and*** **Jiaxin LIU**[†d)], ***Nonmembers***

SUMMARY In modern radar systems, the Generalized compound distribution model is more suitable for describing the amplitude distribution characteristics of radar sea clutter. Accurately and efficiently simulating sea clutter has important practical significance for radar signal processing and sea surface target detection. However, in traditional zero memory nonlinearity (ZMNL) method, the correlated Generalized compound distribution model cannot deal with non-integral or non-semi-integral parameter. In order to overcome this shortcoming, a new method of generating correlated Generalized compound distributed clutter is proposed, which changes the generation method of Generalized Gamma distributed random sequences in traditional Generalized compound distribution models. Firstly, by combining with the Gamma distribution and using the additivity of the Gamma distribution, the Probability Density Function (PDF) of Gamma function is transformed into a second-order nonlinear ordinary differential equation, and the Gamma distributed sequence under arbitrary parameter is solved. Then the Generalized Gamma distributed sequence with arbitrary parameter can be obtained through the nonlinear transformation relationship between the Generalized Gamma distribution and the Gamma distribution, so that the shape parameters of the Generalized compound distributed sea clutter are extended to general real numbers. Simulation results show that the proposed method is not only suitable for clutter simulation with non-integral or non-semi-integral shape parameter values, but also further improves the fitting degree.

key words: *clutter simulation, Generalized compound distribution model, Generalized Gamma distribution, Gamma distribution*

1. Introduction

Radar sea clutter modeling and simulation technology has become a research topic that attracts broad attention. Accurately and quickly simulating sea clutter is of great significance for radar signal processing and sea surface target detection [1], [2]. A lot of analysis has been conducted on the measured radar sea clutter data both domestically and internationally, and the traditional radar clutter models include Rayleigh distribution, Log-normal distribution, Weibull distribution, and K-distribution [3]. The Rayleigh distribution can better fit the amplitude distribution of radar sea clutter at low resolution. However, at higher sea conditions and higher radar resolution, the clutter has a longer trailing tail and the Log-normal distribution fits the measured data better [4]. The dynamic range of the Weibull distribution is between the Log-normal and Rayleigh distributions, and it can more accurately describe the amplitude changes in the actual marine environment [5]. The K-distribution is a compound model composed of speckle and texture components, which can more accurately simulate sea clutter. However, with the increase of radar broadband, the speckle component of radar clutter deviates from the Gaussian distribution, the texture component deviates from the Gamma distribution, so that the K-distribution is no longer applicable [6], [7]. In response to this phenomenon, Anastassopoulos et al. proposed the Generalized compound distribution model suitable for broadband radar clutter [8]–[10], and this model treats the distributions of both speckle component and texture component of broadband radar clutter as Generalized Gamma (GΓ) distribution.

On this basis, Hou et al. [11] proposed a simulation method for correlated Generalized compound distributed radar clutter based on zero memory non-linear (ZMNL) method. This method is suitable for incoherent clutter model, the drawback is that the shape parameter of the clutter model must be an integer or semi-integer, which will cause clutter simulation bias, and it cannot independently control the power spectrum and amplitude distribution. To satisfy the requirements of power spectrum characteristics, the calculation of the correlation coefficients is relatively complex. A new method for simulation of Generalized compound distributed clutter is proposed in Ref. [12], which allows the generation of clutter with an arbitrarily specified power spectrum and relatively simple operation compared to the traditional ZMNL method, but an approximate approach is made for the simulation when the shape parameter is non-integer or non-semi-integer. Conte et al. [13] and Zhu and Tang [14] used the additivity of the Gamma distribution to simulate K-distributed clutter, dividing the shape parameter into the sum of the integral or semi-integral part named v_1 and the non-integral or non-semi-integral part named v_2, solving the problem of approximating the shape parameter v, but the specific Gamma distribution with $v_2$is generated by the product of Beta distributed random numbers and exponential distributed random numbers. Although the approximation problem of shape parameters has been solved, the simulation of Beta distribution may deviate when the parameters are small, which will lead to a certain error between the final clutter simulation curve and the theoretical Probability Density Function (PDF). The GΓ distribution does not have the

Manuscript received February 16, 2024.
Manuscript revised April 1, 2024.
Manuscript publicized May 31, 2024.

[†]Tianjin Key Laboratory of Intelligent Control of Electrical Equipment, Tiangong University, Tianjin, 300387, China.
[††]School of Control Science and Engineering, Tiangong University, Tianjin, 300387, China.
a) E-mail: chengyi@tiangong.edu.cn
b) E-mail: likexin20000429@163.com (Corresponding author)
c) E-mail: xiuchunbo@tiangong.edu.cn
d) E-mail: ljx1793561421@163.com
DOI: 10.23919/transcom.2024EBP3032

additivity and cannot solve the problem of approximation when the shape parameters are non-integers or non-semi-integers by adding branches of the GΓ distribution.

In order to further improve the effect of Generalized compound distributed sea clutter simulation, a new correlated Generalized compound distributed clutter simulation method is proposed to change the generation approach of GΓ distributed random sequences in the traditional correlated GΓ distribution model. Gamma distributed random sequences transformed by specific nonlinear transformation to obtain GΓ distributed random sequences. Using the additivity of the Gamma distribution, the PDF of the Gamma function is transformed into a second-order nonlinear ordinary differential equation by adding the Gamma distribution generating branches before performing the nonlinear transformation, and solving the Gamma distributed random numbers with the scale parameter of 1 and the shape parameter of arbitrary value. Then, nonlinear transformation is performed to obtain the GΓ distributed random numbers under arbitrary parameter. The proposed method not only extends the shape parameters of Generalized compound Gaussian distributed clutter to general real numbers, but also fits better with theoretical distribution.

2. Correlated Generalized Compound Distributed Clutter Simulation Data Model

2.1 Correlated Generalized Compound Distributed Clutter Mathematical Model

The compound model of radar sea clutter can be described using Eq. (1) [15]:

$$Z = XY \tag{1}$$

where X and Y are independent short-term correlated fast changing component (speckle component) and long-term correlated slow changing component (texture component), respectively. The PDF of Z describes the distributions of the distribution of the results obtained by multiplying the speckle component and texture component. For Generalized compound distribution, X and Y are the independent GΓ distributions, which can be obtained from Gaussian distributed random sequences through nonlinear transformation:

$$X = \left(\sum_{m=1}^{M} \varepsilon_m^2\right)^{\frac{1}{b_1}}, \quad Y = \left(\sum_{n=1}^{N} \eta_n^2\right)^{\frac{1}{b_2}} \tag{2}$$

where b_1 and b_2 are positive parameters ($b_1 > 0$, $b_2 > 0$), and ε_m and η_n are independent and identically distributed random sequences which follow the Gaussian distributions $\varepsilon_m \sim N\left(0, \frac{1}{2}\right)$ and $\eta_n \sim N\left(0, \frac{a^{b_2}}{2}\right)$ with a scale parameter a, respectively.

If new parameters v_1 and v_2 defined by $M = 2v_1$ and $N = 2v_2$ are introduced, it can be seen that the PDF of X and Y are giveb by Eq. (3) [16], [17]:

$$\begin{aligned} f(X) &= \frac{b_1}{\Gamma(v_1)} X^{b_1 v_1 - 1} \exp\left(-X^{b_1}\right) \\ g(Y) &= \frac{b_2}{a \cdot \Gamma(v_2)} \left(\frac{Y}{a}\right)^{b_2 v_2 - 1} \exp\left(-\left(\frac{Y}{a}\right)^{b_2}\right) \end{aligned} \tag{3}$$

where $\Gamma(\cdot)$ is the Gamma function, the definition of which is given later (in Eq. (5)).

Based on Eqs. (1) to (3), the expression for the PDF of Generalized compound distribution clutter can be written as:

$$\begin{aligned} & f_{GC}(Z; a, b_1, b_2, v_1, v_2) \\ &= \int_0^{+\infty} \frac{1}{s} f(Z/s) g(s) ds \\ &= \frac{b_1 b_2}{\Gamma(v_1)\Gamma(v_2)} \cdot \frac{Z^{b_1 v_1 - 1}}{a^{b_2 v_2}} \\ &\quad \cdot \int_0^{+\infty} s^{b_2 v_2 - b_1 v_1 - 1} \exp\left(-\left(\frac{s}{a}\right)^{b_2} - \left(\frac{Z}{s}\right)^{b_1}\right) ds \end{aligned} \tag{4}$$

where v_1 and v_2 are the shape parameter; b_1 and b_2 are the power parameter; a is the scale parameter; $\Gamma(\cdot)$ is the Gamma function, and its expression is:

$$\Gamma(x) = \int_0^{+\infty} t^{x-1} \exp(-t) dt \tag{5}$$

When the Generalized compound distribution takes a special value, it degenerates into other common distributions [18]: 1) When $b_1 = b_2$, it corresponds to the Generalized K-distribution; 2) When $b_1 = b_2$, $v_1 = 1$ and $v_2 = 0.5$, it corresponds to the Weibull distribution; 3) When $b_1 = b_2 = 2$ and $v_1 = 1$, it corresponds to the K-distribution. According to the different values of different parameters, the Generalized compound distribution can also evolve into other distributions, as shown in Fig. 1.

2.2 Power Spectrum Model

The power spectrum characteristic of clutter is another important parameter to describe radar clutter, and the power

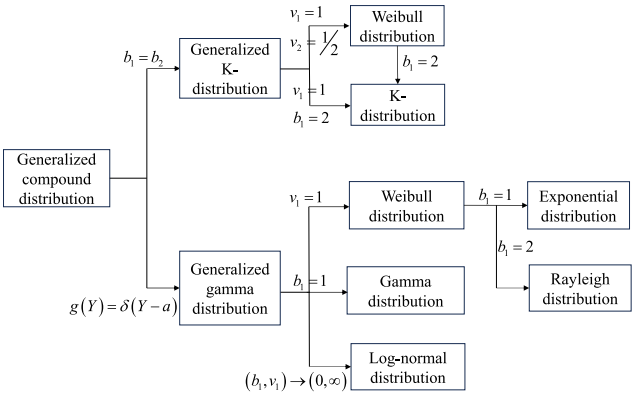


Fig. 1 Relationship between generalized compound distribution and other distributions ($\delta(\cdot)$ is the Dirac delta function).

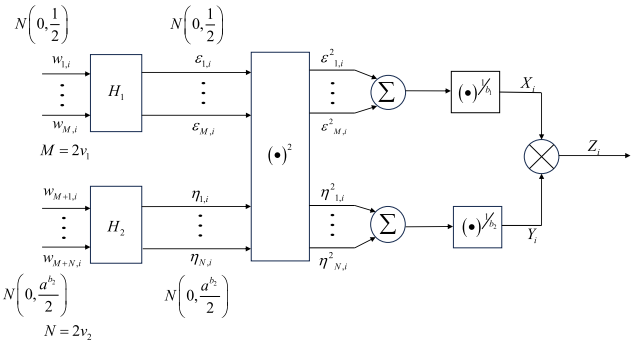


Fig. 2 Simulation flow diagram of correlated generalized compound distributed clutter (in the figure, H_1 and H_2 are filters).

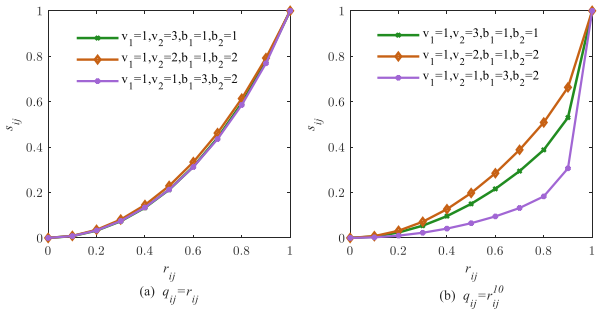


Fig. 3 Relationship curve between the autocorrelation coefficient s_{ij} of the Generalized compound distributed sequences r_{ij} and q_{ij}.

spectrum distribution of clutter is directly related to the design of filter. The Gaussian power spectrum model is expressed as follows [19]:

$$S(f) = S_0 \exp\left[-\left(\frac{(f-f_d)^2}{2\sigma_f}\right)\right] \tag{6}$$

where S_0 is the average clutter power, f_d is the doppler frequency and σ_f is the broadening of the power spectrum reflected by the standard deviation of the power spectrum.

2.3 Simulation of Generalized Compound Distributed Clutter

The ZMNL method is used to generate correlated Generalized compound distributed sea clutter sequences, the flowchart is shown in Fig. 2. The Generalized compound Gaussian distributed clutter simulation of ZMNL method consists of two branches, one of which is to generate X_i, which is generated by the sum of squares of M independent Gaussian distributed random sequences $(w_{1,i}, w_{2,i}, \ldots, w_{M,i})$ to the power of $1/b_1$. The other branch is used to generate Y_i, which generates the Y_i through the sum of squares of N independent Gaussian distributed random sequences $(w_{M+1,i}, w_{M+2,i}, \ldots, w_{M+N,i})$ to the power of $1/b_2$. Among them, $M = 2v_1$, $N = 2v_2$, M and N are integers determined by the shape parameters v_1 and v_2. It can be seen that this simulation method can only simulate Generalized compound distributed sea clutter with shape parameters v_1 and v_2 being integers or semi-integers.

Xie et al. [15] derived the relationship between the autocorrelation coefficients r_{ij} and q_{ij} of the correlated Gaussian distributed random sequences $\{\varepsilon_i\}$ and $\{\eta_i\}$ under the condition of Generalized compound distribution and the autocorrelation coefficient s_{ij} of the output sequence $\{Z_i\}$, as shown in Eq. (7):

$$s_{ij} = \left\{\left[\left(1-r_{ij}^2\right)^{v_1+\frac{2}{b_1}}\left(1-q_{ij}^2\right)^{v_2+\frac{2}{b_2}} {}_2F_1\left(v_1+\frac{k}{b_1}, v_1+\frac{l}{b_1}, v_1, r_{ij}^2\right) {}_2F_1\left(v_2+\frac{k}{b_2}, v_2+\frac{l}{b_2}, v_2, q_{ij}^2\right) - 1\right]\right\} \cdot \frac{1}{\Lambda - 1} \tag{7}$$

where k and l are the order, Λ is given by the next Eq. (8) and ${}_2F_1$ is the hypergeometric function defined by Eq. (9).

$$\Lambda = \frac{\Gamma(v_1)\Gamma(v_2)\Gamma\left(\frac{2}{b_1}+v_1\right)\Gamma\left(\frac{2}{b_2}+v_2\right)}{\Gamma^2\left(\frac{1}{b_1}+v_1\right)\Gamma^2\left(\frac{1}{b_2}+v_2\right)} \tag{8}$$

$${}_2F_1(a,b,c,d) = \sum_{n=0}^{+\infty} \frac{(a)_n (b)_n}{(c)_n} \cdot \frac{d^n}{n!} \tag{9}$$

where $(a)_n$ etc. means Pochhammer's symbol defined by $(a)_n = \Gamma(a+n)/\Gamma(a)$. The right side of Eq. (9) converges only when $|d| < 1$.

Due to the autocorrelation coefficients r_{ij} and q_{ij} cannot be determined directly according to s_{ij}, it needs to be determined based on different shape parameters v_1, v_2 and power parameters b_1, b_2. In actual simulation process, it is often assumed that $q_{ij} = r_{ij}$ or $q_{ij} = r_{ij}^{10}$, and the relationship curves between the autocorrelation coefficient s_{ij} and r_{ij}, q_{ij} are shown in Fig. 3. Observing the Fig. 3, it can be observed that when $q_{ij} = r_{ij}$, the changes of the shape parameters and power parameters have little effect on the values of r_{ij} and q_{ij}, which can simplify the calculation to a certain extent. When $q_{ij} = r_{ij}^{10}$, taking different values of the shape parameters and power parameters results in a significant change in the values of r_{ij} and q_{ij}. Considering reducing the computational complexity of clutter simulation, this paper selects $q_{ij} = r_{ij}$ in the clutter simulation experiment.

3. Improved Correlated Generalized Compound Distributed Clutter Simulation Method

The traditional ZMNL method shown in Fig. 2 requires the shape parameters to be integer or semi-integer, which cannot effectively simulate Generalized compound distributed clutter with non-integral or non-semi-integral shape parameters. Due to the fact that the shape parameters of the simulated clutter are determined by the intermediate $G\Gamma$ distribution variables, in order to extend the shape parameters of the simulated GΓ distributed clutter from integers or semi-integers to general real numbers, it is necessary to improve the generation method of the GΓ distribution variables in Fig. 2. We

would like to propose to use the specific nonlinear transformation relationship between the GΓ distribution and the Gamma distribution to generate the GΓ distributed random sequences. The specific process is as follows:

Assuming that V follows the GΓ distribution, i.e., $V\sim G\Gamma(a,v,b,u=0)$, if T follows the scale parameter of 1 and the shape parameter of v, i.e., $T\sim\Gamma(1,v)$, then we have Eq. (10) [20]:

$$V = aT^{\frac{1}{b}} \sim G\Gamma(a,v,b,0) \tag{10}$$

For the Generalized compound distribution, X and Y are mutually independent GΓ distributions, where $X\sim G\Gamma(X_i;a_1,v_1,b_1,0)$ and $Y\sim G\Gamma(Y_i;a_2,v_2,b_2,0)$. Combining with Eq. (10), it can be obtained that:

$$X = a_1 x^{\frac{1}{b_1}} \tag{11}$$

$$Y = a_2 y^{\frac{1}{b_2}} \tag{12}$$

where $a_1=1$, $x\sim\Gamma(x;1,v_1)$, $y\sim\Gamma(y;1,v_2)$, due to the fact that the GΓ distribution does not have the additivity, but through the nonlinear transformation relationship between the GΓ distribution and the Gamma distribution in Eq. (10), it can be obtained that the shape parameters of the GΓ distributed random sequences X and Y are determined by the shape parameters of the Gamma distributed random sequences x and y.

From this conclusion, by combining with the Gamma distributed random sequences generation method and using the additivity of the Gamma distribution, v_1 is separated into the integral or semi-integral part v_{11} and the non-integral or non-semi-integral part v_{12}, and v_2 is separated into the integral or semi-integral part v_{21} and the non-integral or non-semi-integral part v_{22}, i.e., $v_1=v_{11}+v_{12}$, $v_2=v_{21}+v_{22}$. The non-integral or non-semi-integral part v_{12} and v_{22} are generated by increasing the Gamma distribution generation branches. By converting the PDF of Gamma function into a second-order nonlinear ordinary differential equation [21], [22], the Gamma distribution with arbitrary parameters is generated by power series expansion method. Then the GΓ distributed random sequences with arbitrary parameters can be obtained by the nonlinear transformation Eq. (10).

We are considering the Γ distribution with the scale parameter 1 whose PDF is shown in Eq. (13):

$$f(y) = \frac{y^{v-1}}{\Gamma(v)}\exp(-y) \tag{13}$$

where $\Gamma(v)$ is the Gamma function, v is the shape parameter.

Convert the PDF of the Eq. (13) to the following form:

$$\frac{1}{f(y)} = \frac{dy}{dt} = \Gamma(v)y(t)^{1-v}e^{y(t)} \tag{14}$$

where y is a function of t, and t is a random number with distributions of $[0,1]$, differentiating Eq. (14) with respect to t yields:

$$\frac{d^2y}{dt^2} = \Gamma(v)\left[y^{1-v}(t)e^{y(t)} + (1-v)y^{-v}(t)e^{y(t)}\right]\frac{dy}{dt} \tag{15}$$

Substituting of Eq. (14) into Eq. (15) yields, after simplification:

$$\frac{d^2y}{dt^2} = \frac{1-v}{y}\left(\frac{dy}{dt}\right)^2 \tag{16}$$

$$y\frac{d^2y}{dt^2} - [y+1-v]\left(\frac{dy}{dt}\right)^2 = 0 \tag{17}$$

and the following boundary conditions are assumed:

$$y(0)=0, \quad y(t)\sim[t\Gamma(v+1)]^{\frac{1}{v}} \ \ as \ \ t\to 0 \tag{18}$$

Applying the transformation:

$$z = [t\Gamma(v+1)]^{\frac{1}{v}} \tag{19}$$

Substituting Eq. (19) into Eq. (17), we obtain:

$$y\left(\frac{d^2y}{dz^2} + \frac{1-v}{z}\frac{dy}{dz}\right) - (y+1-v)\left(\frac{dy}{dz}\right)^2 = 0 \tag{20}$$

Assume that the solution of Eq. (20) is given by the infinite power series:

$$y(z) = \sum_{n=1}^{\infty} c_n z^n \ \text{ with } \ c_1 = 1 \tag{21}$$

Substituting the series solution into (20), we find:

$$n(n+v)c_{n+1} = \sum_{k=1}^{n}\sum_{l=1}^{n-k+1} c_k c_l c_{n-k-l+2} l(n-k-l+2)$$
$$-\Delta(n)\sum_{k=2}^{n} c_k c_{n-k+2} k[k-v-(1-v)(n+2-v)] \tag{22}$$

where $\Delta(n)=0$ if $n<2$ and $\Delta(n)=1$ if $n\geq 2$. In the following we list some c_n:

$$\begin{aligned} c_1 &= 1 \\ c_2 &= \frac{1}{1+v} \\ c_3 &= \frac{1}{2}\frac{5+3v}{(1+v)^2(2+v)} \\ c_4 &= \frac{1}{3}\frac{31+33v+8v^2}{(1+v)^3(2+v)(3+v)} \end{aligned} \tag{23}$$

These solutions are reported in [23]. In summary, the approximate solution of the differential equation function can be obtained as follows:

$$y(t) = \sum_{n=1}^{\infty} c_n \left([t\Gamma(v+1)]^{\frac{1}{v}}\right)^n \tag{24}$$

where the coefficients c_n are given by Eq. (23). By using the additivity of the Gamma distribution, we add the Gamma distribution branches with shape parameters v_{12} and v_{22} in the flow diagram of the traditional Generalized compound

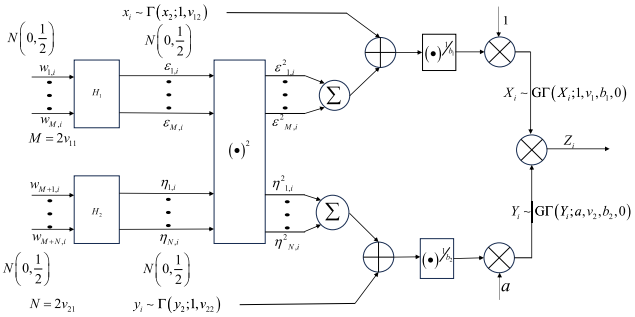


Fig. 4 Flow diagram for improved generation of generalized compound distributed clutter by ZMNL method.

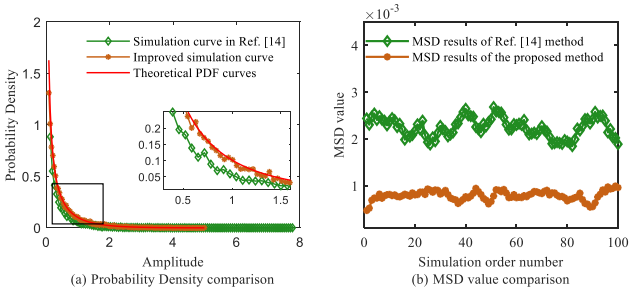


Fig. 5 Improved Gamma distributed clutter model.

distributed clutter simulation. The improved Generalized compound Gaussian distributed clutter generation ZMNL process is shown in Fig. 4:

4. Clutter Simulation Experiment and Analysis

In Ref. [14], according to the additive property of the Gamma distributed shape parameter, Zhu and Tang generated the specific Gamma distribution with non-integral or non-semi-integral shape parameter by the product of Beta distributed random numbers and exponential distributed random numbers, which overcame the shortcoming of the traditional ZMNL. Now, we compare the proposed method with Zhu and Tang's method. In the experiment, we set the shape parameter of the Gamma distribution is $v = 0.15$, $a = 1$, Sampling frequency is set 1,000 Hz and the total simulation number is 20,000. The Gamma branches whose shape parameters v_{12} and v_{22} ($v_{12} = v_{22} = 0.15$) are generated by the method of in Ref. [14] and the proposed method respectively. The comparison of results is shown in Fig. 5(a). Obviously, it can be seen from the locally enlarged curve that the fitting degree between the average histogram of simulated data and the theoretical PDF curve of the proposed method is higher than that of Zhu's method.

To illustrate the performance of the proposed method, the mean squared difference (MSD) technique is used to test the fitting degree of the simulated data. Simulation was done 100 times and a MSD value was obtained in each time. The MSD value comparison is shown in Fig. 5(b). The proposed method has improved the fitting degree on the generation of the GΓ distributed random sequence.

Table 1 Simulation parameters for amplitude distribution.

a	υ_1	υ_2	b_1	b_2	Distribution type
2.5	1	0.5	2.5	2.5	Weibull Distribution
2.5	1	1.65	2	2	K-Distribution
2.5	1.65	1.65	2	2	Generalized K-Distribution
2	1.65	1.65	2	3	Generalized compound Distribution

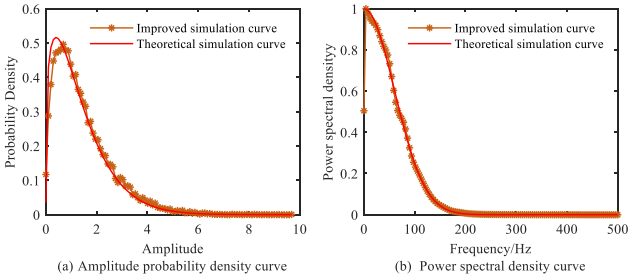


Fig. 6 Probability density function and power spectral density function of Weibull distributed sequence.

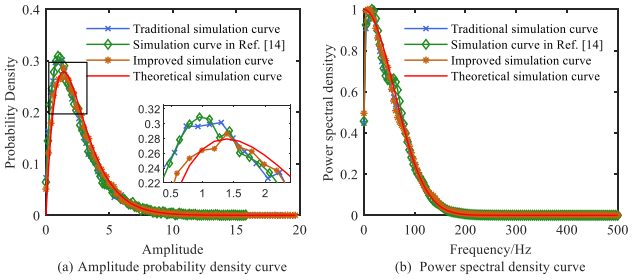


Fig. 7 Probability density function and power spectral density function of K-distributed sequence.

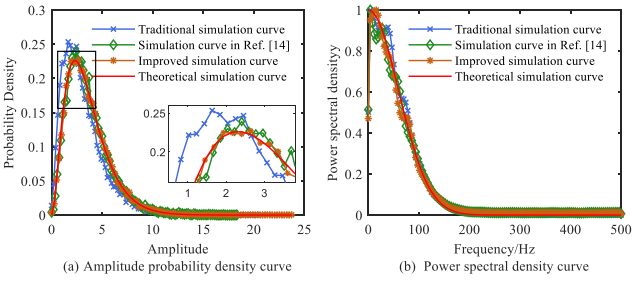
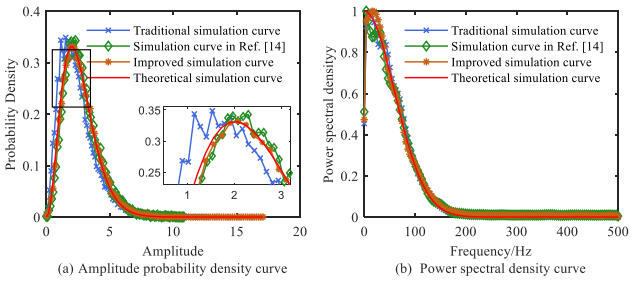


Fig. 8 Probability density function and power spectral density function of generalized K-distributed sequence.

To further verify the effect of non-integral shape parameters, four sets of Generalized compound distributed sea clutter sequences with the data length of 20,000 and the radar pulse repetition frequency of 1,000 Hz were generated using the method proposed in Sect. 3. The simulated power spectrum is Gaussian spectrum with a bandwidth of 60 Hz. The specific simulation parameters are shown in Table 1. Figures 6, 7, and 8 show the simulation results of the Generalized compound distribution degenerated into Weibull distribution, K-distribution, and Generalized K-distribution, respectively, and Fig. 9 shows the simulation results of the ordinary Generalized compound distribution.



(a) Amplitude probability density curve (b) Power spectral density curve

Fig. 9 Probability density function and power spectral density function of generalized compound distributed sequence.

Among them, Fig. 6(a) shows the comparison between the simulated Weibull distributed clutter and the theoretical amplitude probability density curves. When degenerating to the Weibull distribution, the shape parameter can only be fixed values, i.e., $v_1 = 1$, $v_2 = 0.5$, which does not involve the problem of approximating shape parameters. The Weibull distributed clutter sequence simulated by this Generalized compound distribution model fits well with the theoretical values. Figures 7(a)–9(a) show the comparison of amplitude probability density curves for the K-distribution, Generalized K-distribution, and Generalized compound distribution of traditional and improved methods, where the shape parameters of the traditional method are taken down to integers or semi-integers, i.e., $v = 1.5$. In Figs. 7(a)–9(a), the probability density curves of the K-distributed, Generalized K-distributed, or Generalized compound distributed clutter simulated by the traditional ZMNL method have some deviations from the theoretical values. Compared with the improved Generalized compound distributed clutter simulation method, especially when the amplitude value is small, the proposed method has a higher fitting degree than the traditional method. Figures 7(b)–9(b) show the comparison of the power spectral density curves between traditional and improved methods. In Figs. 7(b)–9(b), the power spectrum characteristics fit well with the ideal Gaussian spectrum curve within the effective bandwidth, and the newly added branch has a negligible impact on the power spectral density of the Generalized compound clutter simulation.

5. Conclusion

Traditional Weibull distribution, K-distribution, Generalized K-distribution, etc. are special cases of Generalized compound distribution, and the Generalized compound distribution is more universal, which is the core of radar sea clutter modeling and simulation. This paper analyzes the Generalized compound distributed sea clutter model. Focusing on the problem that the traditional ZMNL method can only simulate random sequences with shape parameters of integers or semi-integers in the GΓ distribution, it is proposed to combine the GΓ distribution with Gamma distribution through the proprietary nonlinear transformation relationship. By using the additivity of the Gamma distribution, the shape parameters of the Generalized compound distributed sea clutter are extended to general real numbers by adding the Gamma distribution generating branches before the nonlinear transformation. The simulation results show that the PDF simulation curves obtained from the simulation of the improved ZMNL method fit the theoretical PDF curves better and improves the performance of the amplitude characteristic simulation of the Generalized compound distributed sea clutter.

References

[1] H. Ding, Y.L. Dong, N.B. Liu, G.Q. Wang, and J. Guan, "Overview and prospects of research on sea clutter property cognition," Journal of Radars, vol.5, no.5, pp.499–516, 2016.

[2] J. Carretero-Moya, J. Gismero-Menoyo, Á. Blanco-del-Campo, and A. Asensio-Lopez, "Statistical analysis of a high-resolution sea-clutter database," IEEE Trans. Geosci. Remote Sens., vol.48, no.4, pp.2024–2037, April 2010.

[3] S. Bocquet, "Analysis and simulation of low grazing angle X-Band coherent radar sea clutter using memoryless nonlinear transformations," IEEE Trans. Geosci. Remote Sens., vol.60, no.5111113, pp.1–13, 2022.

[4] Y. Chen, Y.Y. Li, L.B. Yang, X.H. Ni, B.S. Yang, and Y.H. Wang, "An overview of statistical characteristics of ground-sea clutter," AIR&SPACE DEFENSE, vol.3, no.4, pp.44–51, 2020.

[5] L. Zhou, H.X. Kuang,Y.T. Zhang, C. Ding, and L.L. Wang, "Correction probability density distribution function of sea clutter based on logarithmic normal distribution," RADAR &ECM, vol.41, no.01, pp.18–22, 2021.

[6] K. Zhang, P.L. Shui, and G.H. Wang, "Non-coherent integration constant false alarm rate detectors against K-distributed sea clutter for coherent radar systems," Journal of Electronics & Information Technology, vol.42, no.7, pp.1627–1635, 2022.

[7] S. Watts, "Modeling and simulation of coherent sea clutter," IEEE Trans. Aerosp. Electron. Syst., vol.48, no.4, pp.3303–3317, 2012.

[8] V. Anastassopoulos and G.A. Lampropoulos, "A generalized compound model for radar clutter," IEEE National Radar Conference, pp.41–45, 1994.

[9] V. Anastassopoulos and G.A. Lampropoulos, "High resolution radar clutter classification," Proc. International Radar Conference, pp.662–667, 1995.

[10] V. Anastassopoulos, G.A. Lampropoulos, A. Drosopoulos, and N. Rey, "High resolution radar clutter statistics," IEEE Trans. Aerosp. Electron. Syst., vol.35, no.1, pp.43–60, 1999.

[11] X.M. Hou, H. Ma, H. Ruan, K. Jiang, and S.J. Wei, "Simulation of correlated generalized compound distribution radar sea clutter," Journal of Telemetry, Tracking and Command, vol.35, no.1, pp.43–60, 2015.

[12] C. Zhao, C.S. Jiang, J. Ma, and Y. Zhao, "Simulation of correlated generalized-compound-distributed radar sea clutter based on ZMNL transformation," Modern Radar, vol.37, no.2, pp.75–78, 2015.

[13] E. Conte, M. Longo, and M. Lops, "Modelling and simulation of non-Rayleigh radar clutter," IEE Proceedings-F: Radar & Signal Processing, vol.138, no.2, pp.121–13, 1991.

[14] J.L. Zhu and J. Tang, "K-distribution clutter simulation methods based on improved ZMNL and SIRP," J. Radar, vol.3, no.5, pp.533–540, 2014.

[15] L.Q. Xie, Z.M. Chen, C.S. Jiang, and Y. Zhou, "Simulation of the correlated generalized compound distribution wide band radar clutter using ZMNL method," Signal Processing, vol.25, no.09, pp.1463–1468, 2009.

[16] F.J. Marques and F. Loingeville, "Improved near-exact distributions for the product of independent generalized gamma random variables," Computational Statistics & Data Analysis, vol.102, pp.55–66, 2016.

[17] R.K. Yang, J.X. H, F.P. Wu, F. Lu, and Y. Zhou, "Research on simulating method of optical intensity scintillation series with double generalized gamma distribution in atmospheric turbulence," Acta Photonica Sinica, vol.50, no.10, pp.332–337, 2021.
[18] H.C. Li, W. H, Y.R. Wu, and P.Z. Fan, "On the empirical-statistical modeling of SAR images with generalized gamma distribution," IEEE J. Sel. Topics Signal Process., vol.5, no.3, pp.386–397, 2011.
[19] L. Rosenberg and S. Watts, "Modeling the statistics of microwave radar sea clutter," IEEE Aerosp. Electron. Syst. Mag., vol.34, no.10, pp.44–75, 2019.
[20] C. Combes and H.K.T. Ng, "On parameter estimation for Amoroso family of distributions," Mathematics and Computers in Simulation, vol.191, pp.309–327, 2022.
[21] H.I. Okagbue, M.O. Adamu, and T.A. Anake, "Closed form expressions for the quantile function of the Chi square distribution using the hybrid of quantile mechanics and spline interpolation," Wireless Pers. Commun., vol.115, no.3, pp.2093–2112, 2020.
[22] H.I. Okagbue, M.O. Adamu, and T.A. Anake, "Approximations for the inverse cumulative distribution function of the gamma distribution used in wireless communication," Heliyon, vol.6, no.11, pp.240–884, 2022.
[23] A. Kleefeld and V. Brazauskas, "A statistical application of the quantile mechanics approach: MTM estimators for the parameters of *t* and gamma distributions," European Journal of Applied Mathematics, vol.23, no.5, pp.593–610, 2012.

Yi Cheng received the M.S. degree and the Ph.D. degree in Navigation, guidance and control specialty from Harbin Engineering University, Heilongjiang, China, in 2004 and 2008, respectively. She is currently an associate professor at Tiangong University, and engaged in radar signal processing research.

Kexin Li received the B.S. degrees in Automation from Zhongyuan University of Technology of Science and Technology, Henan, China, in 2022. She is the master's student majoring in Control Engineering at Tiangong University, research direction is sea clutter modeling and simulation.

Chunbo Xiu in Automatic Control and the Ph.D. degree in Navigation, Guidance and Control from Beijing Institute of Technology, China, in 2000 and 2004, respectively. He is currently a professor at Tiangong University, China, and is engaged in the research of intelligent control.

Jiaxin Liu received the B.S. degrees in Automation from Shenyang Jianzhu University of Electrical Engineering and Automation, Liaoning, China, in 2023. He is the master's student majoring in Control Engineering at Tiangong University, research direction is clutter suppression.

PAPER

A Mobile Reception Experiment of Galileo Improved I/NAV Navigation Messages

Satoshi TAKAHASHI †a), ***Senior Member***

SUMMARY In satellite positioning, both the reception of ranging signals and the acquisition of navigation messages are necessary. In general, the acquisition of navigation messages does not always require the reception of radiowaves; however, when radiowaves are used for acquisition, a period of continuous reception significantly longer than one second is required. The European satellite positioning system, Galileo, started broadcasting new navigation messages from August 2022. The improvement is based on a secondary synchronization pattern, secondary forward error correction, and reduced ephemeris to aid in the rapid recovery from interruptions in message acquisition caused by temporary deterioration in radio reception. This paper evaluates the recovery characteristics from interruptions in navigation message acquisition by moving reception of this improved I/NAV navigation message.

key words: *GNSS, Galileo, navigation message, I/NAV, software defined radio, Pocket SDR*

1. Introduction

When GNSS (Global Navigation Satellite System), such as GPS (Global Positioning System), is used, the user's location can be determined solely by receiving radiowaves. GNSS receivers are also installed in devices such as smartphones, car navigation systems, and smart watches, and have permeated our daily lives as a means of self-positioning outdoors.

A GNSS receiver estimates the distance from a satellite to the user by using the ranging signal embedded in the satellite's radiowaves. It then calculates the satellite's position from the navigation message included in these radiowaves, determining the user's location based on these results. As the latitude, longitude, ellipsoid height, and time at the user's location are unknown, the GNSS receiver performs distance estimation and satellite position calculation for at least four different satellites. Each satellite transmits a unique navigation message that conveys information such as the current time, satellite position, and ephemeris, which is a satellite internal time correction parameter. The GNSS receiver calculates the user's position based on the known time, synchronized with an error of less than 1 second, and the acquired information [1].

The ephemeris describes orbit parameters based on the motion model rather than the satellite position and is not updated frequently. With a parameter validity period of 2 hours, the acquisition of navigation messages does not necessarily require radiowave reception. Smartphones can also obtain navigation messages through cellular networks. For instance, intermittent reception is feasible by alternating between distance measurement using radiowave reception for 100 milliseconds and receiver sleep for 900 milliseconds [2]. This method of intermittent reception is anticipated to reduce the power consumption required for positioning to approximately one-tenth of that needed for continuous reception.

On the other hand, obtaining navigation messages solely through the reception of satellite radiowaves has the advantage of incurring no communication costs and not being dependent on communication infrastructure. However, a drawback is the need for continuous reception for approximately 30 seconds to acquire all the basic information. This requirement is due to the satellite's location approximately 20,000 kilometers from the Earth's surface and the deliberate limitation of the effective message transmission rate to between 50 and 120 bits per second. This rate restriction facilitates the positioning of high-speed moving objects such as aircraft. Signal interruption due to changes in the user's reception environment is one of the primary challenges in satellite positioning.

The European positioning satellite system, Galileo, commenced broadcasting the Improved I/NAV (Integrity Navigation) message in August 2022. This enhancement introduces three types of information to the conventional I/NAV message, aimed at supporting quick recovery from interruptions in message acquisition [3]. Conventional receivers remain unaffected by this improvement. However, receivers capable of interpreting this enhanced message can acquire ephemeris more quickly, thereby mitigating the increase in the time required to initiate positioning following a signal interruption.

Currently, there are no commercially available receivers capable of utilizing the Improved I/NAV, and its efficacy in real-world environments remains unverified. This report conducts a post-analysis of the outputs from a commercially available receiver and a software-defined radio, recorded during movement, to evaluate the recovery characteristics of the Improved I/NAV navigation messages following interruptions in their acquisition.

2. Navigation Messages Sent from Galileo

Galileo's positioning signals comprise the E1 signal, broadcast in the 1.5 GHz band, and the E5 signal, broadcast in the

Manuscript received December 20, 2023.
Manuscript revised March 28, 2024.
Manuscript publicized May 31, 2024.
†Hiroshima City University, Hiroshima-shi, 731-3194 Japan.
a) E-mail: s.takahashi@m.ieice.org
DOI: 10.23919/transcom.2023EBP3212

Table 1 Navigation message format comparison between GPS LNAV and Galileo I/NAV.

GPS LNAV	Galileo I/NAV
Frame period of 30 seconds	Frame period of 30 seconds
Transmission rate of 50 symbol/s	Transmission rate of 250 symbol/s
Error correction of Hamming (32, 26) code	Error correction of 1/2-rate convolutional code
A 300-bit message in a 6-second cycle	A 240-bit message in a 2-second cycle

1.2 GHz band. The E5 signal is further subdivided into the E5a and E5b signals, each broadcast on a distinct frequency. The F/NAV (Free Navigation) message is transmitted on the E5a signal, while the I/NAV message is transmitted on both the E1 and E5b signals [3]. However, the I/NAV messages in the E5b signals have not been included in the recent improvement. The focus of this discussion is on the I/NAV messages transmitted on the E1 signal. Notably, the E1 signal shares the same frequency as the GPS L1 signal and is therefore receivable by many GNSS receivers.

Navigation messages broadcast by positioning satellites encompass ephemeris, almanacs representing approximate orbital information of other satellites, time, and other relevant information. Compared to GPS, Galileo represents a more recent satellite system that achieved Full Operational Capability (FOC) in August 2014. Consequently, the method for transmitting navigation messages has also been modernized. These updates include error correction through a combination of deinterleaving and Viterbi decoding, along with expedited information transmission. A brief comparison between the GPS LNAV message format and the Galileo I/NAV message format is presented in Table 1. Such enhancements allow I/NAV messages to accommodate the transmission of additional information.

In GPS LNAV (Legacy Navigation) messages, information is transmitted in a cycle of 30 seconds, comprising units of 300-bit data segments, each transmitted over 6 seconds. At the beginning of each unit, there is a special 8-bit bit pattern, known as the preamble, which indicates the start of the message [4].

Contrastingly, Galileo I/NAV messages consist of units with a 2-second duration and adhere to a similar 30-second period as GPS LNAV messages. The I/NAV message is divided into two parts, with each part undergoing independent error correction and featuring a 10-symbol preamble. Each page part is 114 bits in length. At the beginning of each page part, there is a 1-bit page type identifier and a 1-bit validity (alert) identifier. Consequently, the I/NAV can effectively transmit a total of 228 bits of information within a 2-second timeframe.

Out of the 228 bits in the I/NAV message, the segment dedicated to transmitting navigation information is only 128 bits long, and the CRC (Cyclic Redundancy Check) code, used for error detection, occupies 24 bits. This leaves a space of 76 bits within the I/NAV page for transmitting additional information.

In the I/NAV message transmitted on the E5b signal, the aforementioned 76 bits are allocated as reserved space. Additionally, on the E5b signal, 7 out of the 15 word slots that are transmitted over the total message period are designated as spare words (word number 0).

For the message on the E1 signal, 40 bits of the 76-bit space are utilized for navigation message authentication OSNMA (Open Signal Navigation Message Authentication) [5], and 22 bits are dedicated to transmitting the SAR (Search and Rescue) return link message [3] of Cospas-Sarsat, the search and rescue mission satellite [6]. In January 2021, the Galileo signal specification, OS SIS ICD (Open Signal Signal-in-Space Interface Control Document), was updated to Issue 2.0, introducing Improved I/NAV on the E1 signal [7], [8]. Subsequently, in Issue 2.1 revised in November 2023, one of the three spare words was modified to the ARAIM (Advanced Receiver Autonomous Integrity Monitoring) word (word number is 22). Moreover, the disaster information DCX (Satellite Report for Disaster and Crisis Management—Extended Information) message, slated for new broadcast on the L1S signal of the Quasi-Zenith Satellite System, Michibiki, will be shared with Galileo as a common emergency message, EWS (Emergency Warning Service). This message is expected to be transmitted through the I/NAV word [9].

3. Analysis of Qualitative Properties

3.1 SSP

The initial enhancement made to the I/NAV message involved the allocation of 8 bits from the remaining 10 bits of reserved space for a Secondary Synchronization Pattern (SSP). This involves the assignment of three types of fixed bit patterns (`0x04`, `0x2b`, `0x2f` in hexadecimal notation) outside the scope of CRC error detection. By executing pattern matching between the deinterleaved symbol string and the converted SSP symbol string in the receiver, it is anticipated that time synchronization errors within a 3-second range can be corrected, even under erroneous radio propagation conditions [3], [8].

Time within the Galileo system is measured in GST (Galileo System Time). GST is composed of the number of weeks (WN: Week Number) and the number of seconds (TOW: Time of Week) elapsed since the start of the week. The reference epoch for GST is set at 13 seconds before Sunday, August 22, 1999, at 00:00:00 UTC, with the 13-second adjustment accounting for leap seconds. I/NAV messages are broadcast in alignment with GST, and those on E1 signals are specifically broadcast at TOW odd times. The content of each word and the SSP corresponding to the remainder of TOW divided by 30 (GST mod 30) are both consistently assigned [3].

From the integrated information presented in Table 2, it is evident that GST mod 30 is uniquely determined by the combination of the word type number and SSP.

For instance, in the case of a message that successfully passes the CRC test, if its word type number is 0 and its SSP is 3, then GST mod 30 is uniquely determined to be 17

Table 2 GST mod 30, word type, word content, and SSP.

GST mod 30	word type	word content	SSP
1	2	ephemeris 2	1
3	4	ephemeris 4	2
5	6	time conversion	3
7	7 or 9	almanac 1 or 3	1
9	8 or 10	almanac 2 or 4	2
11	17 or 18	FEC2	3
13	19 or 20	FEC2	1
15	16	Reduced CED	2
17	0	spare	3
19	22	ARAIM	1
21	1	ephemeris 1	2
23	3	ephemeris 3	3
25	5	time & health	1
27	0	spare	2
29	16	Reduced CED	3

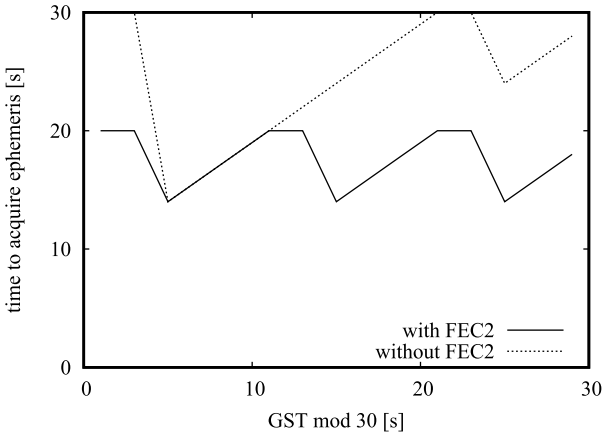


Fig. 1 GST mod 30 vs. ephemeris acquisition time.

seconds[†]. If both the word type and the SSP are verified to be correct, the time synchronization range can be extended from 3 seconds to 30 seconds.

3.2 FEC2

The second improvement to I/NAV is the implementation of a secondary FEC (Forward Error Correction), named FEC2, for words, supplementing the existing FEC that utilizes Viterbi decoding. Previously, for ephemeris acquisition, it was essential to acquire all four ephemeris words from 1 to 4. However, with the Improved I/NAV, Galileo satellites broadcast these four ephemeris words along with four additional Reed-Solomon parity words. At the receiver, when at least four of these words are received, the content of any unreceived words is set to all zeros, and Reed-Solomon decoding is then forcibly applied using erasure-code error correction. In [10], proposals are not limited to FEC2 using Reed-Solomon codes; FEC2 utilizing LDPC (Low-Density Parity-Check) codes and LD-MDS (Low-Density Maximum Distance Separable) codes has also been suggested.

Here, the time required for ephemeris acquisition as a function of GST mod 30 is calculated and summarized in Fig. 1. For instance, at a time when GST mod 30 is 25 seconds, a receiver incapable of utilizing FEC2 would receive the following ephemeris words:

- Ephemeris 3, which started transmitting 2 seconds ago,
- Ephemeris 1, which started 4 seconds ago,
- Ephemeris 4, which started 22 seconds ago, and
- Ephemeris 2, which started 24 seconds ago.

In this scenario, the time to obtain all ephemeris words is estimated to be 24 seconds, as the receiver utilizes the word transmitted 24 seconds ago. Conversely, with FEC2 capability, the receiver needs only to receive:

- Ephemeris 3 from 2 seconds ago,
- Ephemeris 1 from 4 seconds ago,
- FEC2 (word 19 or 20) from 12 seconds ago, and
- FEC2 (word 17 or 18) from 14 seconds ago.

This reduces the time required to acquire all ephemeris words to 14 seconds.

Additionally, during the GST mod 30 range of 7–11 seconds, all ephemeris words are received before any FEC2 word is transmitted. Consequently, in this specific time range, the use of FEC2 does not reduce the ephemeris acquisition time. However, in other time ranges, the time required to acquire ephemeris is indeed shortened by utilizing FEC2.

3.3 Reduced CED

The third improvement in the I/NAV system is the implementation of Reduced CED (Clock and Ephemeris Data). While the acquisition of normal precision ephemeris typically requires 4 words, Reduced CED needs only 1 word due to data compression. This approach allows for the use of Reduced CED before all words of normal precision ephemeris are received, enabling quicker positioning [11]. However, while waiting for the acquisition of normal precision ephemeris and utilizing Reduced CED, the estimated distance measurement error at the 95% worst-case scenario ranges from 0.25 meters to 4.5 meters, a consequence of information compression [8]. In the Improved I/NAV, 2 words of Reduced CED are broadcast within each 30-second period.

The time required to acquire either normal ephemeris using FEC2 or Reduced CED is calculated using the method described in Sect. 3.2. This acquisition time ranges between 2 to 14 seconds.

4. Mobile Reception Experiment

4.1 Signal Recording

For the acquisition of Improved I/NAV navigation messages, a commercial receiver, the u-blox ZED-F9P-02B (FW version 1.00 HPG 1.32), and an open-source software radio,

[†]As of December 2023, ARAIM (Advanced Receiver Autonomous Integrity Monitoring) is not being broadcast, and instead, a spare word is being transmitted. Even under these circumstances, GST mod 30 can be uniquely determined from the word type number and SSP.

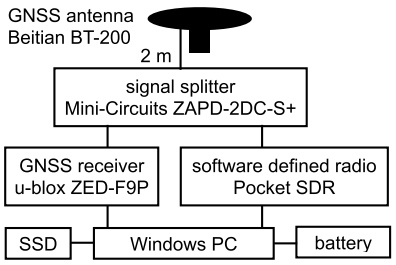


Fig. 2 Equipment connection diagram.

Pocket SDR version 0.8 [12], were utilized. The diagram illustrating the connections of the equipment is depicted in Fig. 2. Note that in this measurement system, there is a deterioration from the ideal signal reception state. The specific amount of degradation at each receiver input terminal includes a 1.8 dB noise figure (catalog value) due to the LNA (low noise amplifier) in the antenna, a 0.5 dB loss (actual measurement) due to the 2 meter antenna cable, and the signal splitter causing a 3.1 dB loss (measured), for a total of 5.4 dB.

A loop road with a total length of approximately 2.4 kilometers was chosen for the mobile reception experiment. This course features varying elevations with a height difference of 80 meters, and visibility is often limited by slopes on one or both sides of the road. The receivers were switched on 30 minutes before the start of the measurement to warm up the internal oscillator and to allow the ZED-F9P receiver to acquire the almanac in advance.

For the ZED-F9P receiver, the navigation message output `UBX-RXM-SFRBX` was enabled. The receiver was set to a reception frequency of 1575.42 MHz. For the Pocket SDR receiver, settings included a third-order polyphase filter with a bandwidth of 4.2 MHz, a sampling frequency of 24 MHz, and synchronous detection with 2 bits each for I (In-phase) and Q (Quadrature) components.

Subsequently, data acquisition was initiated with both the ZED-F9P and Pocket SDR, and signals were recorded while completing two laps of the course. The start time of the movement, as determined from the I/NAV navigation message, was 2023-11-23 23:19:30 UTC (WN=1265, TOW=343183), with a recording duration of 10.5 minutes. The trajectory of the driving course is depicted in Fig. 3. The driving course trajectory is plotted using the ZED-F9P's raw data and the RTKLIB [13] single point positioning mode. Each black dot that makes up the driving course trajectory represents coordinates determined at a 5-second interval. In the figure, the sky-view photos and the sky view factors for some locations are also shown.

The sky plot is shown in Fig. 4. In this plot, positions of satellites available for reception are shown with circles, and positions of other satellites are shown with cross marks. At the signal recording location (34.23 degrees north latitude and 132.27 degrees east longitude), it was estimated that 11 Galileo satellites were visible at the time, each with an elevation angle of 0 degrees or more. However, throughout the course, only 3–4 satellites with elevation angles of 50 degrees or more were consistently usable. The temporal changes in the number of received satellite signals and cumulative distribution of signal interruption time duration during this observation are plotted in Figs. 5 and 6. Signals from satellites E03 and E07 were often lost and signals from E15 were sometimes lost during this experiment. This situation presented a challenging environment for the reception of navigation messages.

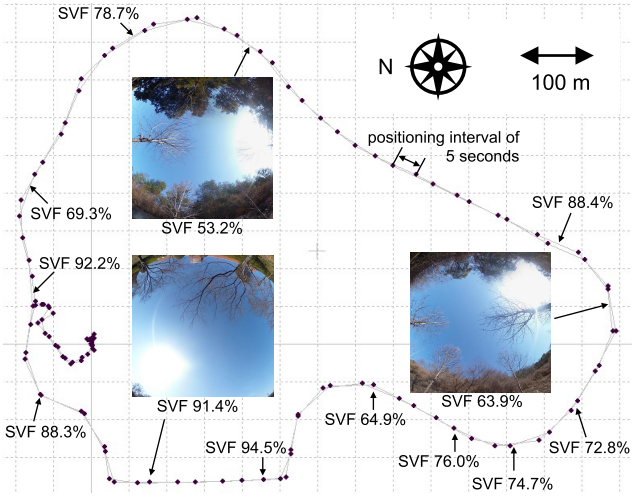


Fig. 3 The driving course trajectory, sky-view photos, and sky view factor (SVF).

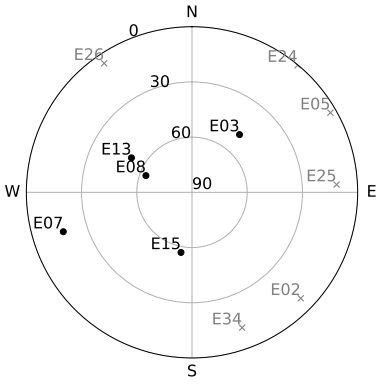


Fig. 4 The sky plot.

After completing the signal recording, the I/NAV message on the E1 signal was specifically extracted from the `UBX-RXM-SFRBX` data output by the ZED-F9P receiver[†].

Conversely, the source code of Pocket SDR was altered to enable the output of I/NAV messages to the log file, in-

[†]It is important to note that the ZED-F9P firmware manual does not provide details on how to store the two pages comprising the I/NAV message in `UBX-RXM-SFRBX`, particularly regarding the read position for the second half of the page. In this analysis, the position was identified based on the description found in lines 801–804 of the source code `src/rcv/ublox.c` from RTKLIB version 2.4.3 b34 [13], allowing for the correct extraction and interpretation of the data.

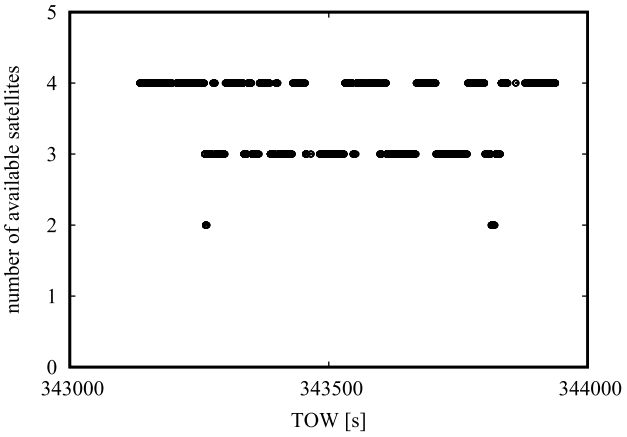


Fig. 5 The number of received satellite signals during this experiment.

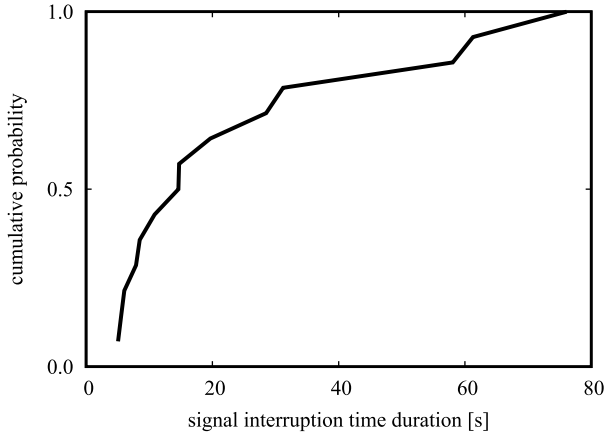


Fig. 6 Cumulative distribution of signal interruption time duration.

cluding those that fail the CRC test, from the recorded data[†].

Subsequently, the time required to obtain ephemeris from an arbitrary point in time was calculated. It's important to note that only words 0, 5, and 6 of the I/NAV messages contain the TOW, which is essential for determining the GST mod 30 time. To ascertain the GST mod 30 time for every word, the method outlined in Sect. 3.1 was employed.

During the recording period, a total of 1420 messages were obtained from the ZED-F9P receiver, while 1135 messages were captured by the Pocket SDR receiver. The higher number of satellites captured by the ZED-F9P receiver is likely attributable to undisclosed optimizations that leverage almanac and Doppler frequency information from the captured signals. In contrast, the signal reception code of Pocket SDR is open-source and is written in a straightforward manner, adhering closely to the specifications. This allows it to output messages that might otherwise be discarded due to CRC test failures. After processing the messages from both receivers, the messages from the Pocket SDR receiver are chosen for the subsequent analysis.

[†]Specifically, a modification was made in the Python code `python/sdr_nav.py`. The line 553 was changed from `data = pack_bits(bits)` to `data = pack_bits(np.hstack([bits, bits2[106:106+8]]))`, which added SSP to the I/NAV message output. Additionally, `if test_CRC(bits):` on line 550 was modified to `if True:` to bypass the CRC test, allowing the output of erroneous I/NAV messages as well.

4.2 SSP

Messages that pass the CRC test can have their GST mod 30 time estimated accurately. During the recording period, two messages failed the CRC test. Examples of some decoded messages are as follows:

```
E03 SSP2 Word 16 (15)
E13 SSP2 Word 16 (15)
E07 SSP2 Word 16 (15)
E08 SSP3 Word 0 (17) 2023-11-22 23:21:34 (WN=1265 TOW=343307)
E03 SSP? (c6) Word 0 CRC error:  2c4baa != 37f4c0
E13 SSP3 Word 0 (17) 2023-11-22 23:21:34 (WN=1265 TOW=343307)
E07 SSP3 Word 0 CRC error:  0be731 != c64ecc
E08 SSP1 Word 0 (19) 2023-11-22 23:21:36 (WN=1265 TOW=343309)
E03 SSP1 Word 0 (19) 2023-11-22 23:21:36 (WN=1265 TOW=343309)
E13 SSP1 Word 0 (19) 2023-11-22 23:21:36 (WN=1265 TOW=343309)
E08 SSP2 Word 1 (21)
```

In these results, each message displays various pieces of information including the satellite number (e.g., E03), SSP value, word number, estimated GST mod 30 (determined by the method described in Sect. 3.1), and the GST itself, if it contains GST information.

The first message that failed the CRC test displayed an SSP value with the hexadecimal representation `0xc6`, which is not defined in the specifications. This anomaly is likely attributable to a bit error, suggesting that the SSP bit string was also received incorrectly.

The subsequent message that failed the CRC test had an SSP value of 3 and a word number of 0. According to Sect. 3.1, the GST mod 30 for this combination is 17 seconds. The correctness of this SSP value and word number is supported by their consistency with the messages that passed the CRC test immediately before and after this instance.

Based on these observations, it can be concluded that when the CRC test following Viterbi decoding fails, the reliability of the SSP value becomes questionable. Therefore, it is improbable that SSP can be effectively utilized for messages that contain a high number of bit errors.

4.3 FEC2

Figure 7 illustrates the time required for ephemeris acquisition when FEC2 is not available. In the figure, the overlap of the circles and the straight line indicates that the minima of time to acquire ephemeris are consistent with the ideal values. In contrast, Fig. 8 depicts the ephemeris acquisition time when FEC2 is available, providing a comparative view of the impact of FEC2 on the efficiency of ephemeris acquisition. This calculation operates under the assumption that if the receiver is unable to receive a required word, it will continue its reception efforts until the next word is successfully acquired. Under this framework, the maximum reception period is set at 60 seconds, which corresponds to two complete cycles of transmission.

When comparing the data presented in Figs. 7 and 8,

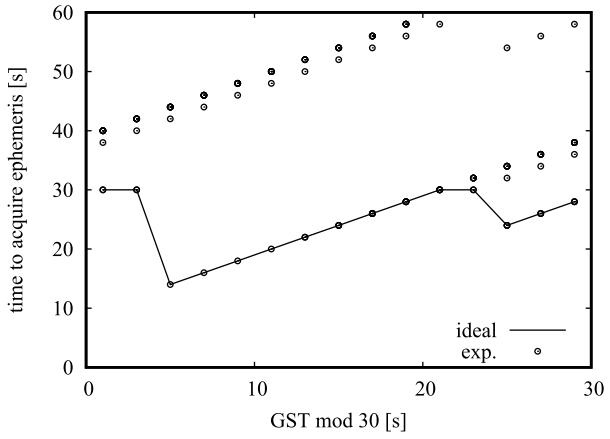


Fig. 7 Ephemeris acquisition time when FEC2 is unavailable.

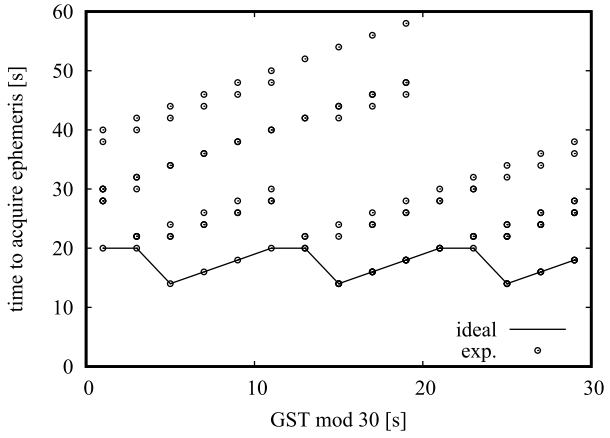


Fig. 8 Ephemeris acquisition time when FEC2 is available.

it is evident that the introduction of FEC2 contributes to a reduction in the ephemeris acquisition time. Additionally, the increased density of plots at specific GST mod 30 times indicates that message retrieval is frequently retried when a necessary message is lost, ultimately contributing to a shorter overall retrieval time.

Moreover, even in the 7–11 second range of GST mod 30, where, as discussed in Sect. 3.2, FEC2 does not significantly reduce the ephemeris acquisition time under normal circumstances, it is observed that FEC2 helps mitigate the increase in acquisition time that typically occurs due to message loss in a mobile environment. This highlights the utility of FEC2 in enhancing the robustness and efficiency of ephemeris acquisition, particularly in scenarios where signal reception is challenging.

Indeed, in both scenarios – with and without FEC2 – the plots align along a straight line in relation to the GST mod 30 time. This pattern arises because the sequence of message transmissions is pre-established, and consequently, the waiting period incurred when a message is missed remains constant. The occurrence of multiple plots at the same GST mod 30 time suggests that multiple messages were consecutively lost. This repetitive loss pattern underscores the

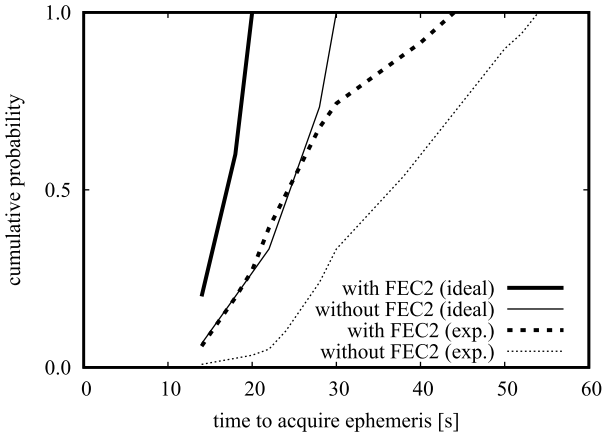


Fig. 9 Cumulative distribution of time to obtain ephemeris.

importance of the message transmission order and the fixed waiting intervals in determining the overall efficiency and reliability of message retrieval in these scenarios.

Next, a cumulative distribution of the ephemeris acquisition times at arbitrary points is calculated and plotted. This distribution is presented in Fig. 9. This visualization will provide a clearer understanding of the distribution and frequency of ephemeris acquisition times under the conditions of this experiment, offering valuable insights into the overall performance and efficiency of the system in different scenarios. In the figure, the solid line represents the probability of acquiring all messages, as detailed in Sect. 3.2, under ideal conditions where all messages can be successfully obtained. The broken line, on the other hand, depicts the probability when taking into account the loss of messages in a real-world environment. This distinction provides a comparative view of the system's performance in theoretical versus actual operational conditions, highlighting the impact of message losses on the reliability and efficiency of ephemeris acquisition.

Focusing on the 50% value in Fig. 9, in an environment where all messages are successfully received, the implementation of FEC2 reduces the ephemeris acquisition time by approximately 30%, from 24.1 seconds to 16.8 seconds. In a mobile reception environment, FEC2 reduces the acquisition time by approximately 34%, from 36.4 seconds to 24.1 seconds.

However, it is noteworthy that in a mobile reception environment, the effectiveness of FEC2 in reducing acquisition time diminishes when the value exceeds 75 percent. This trend is likely attributed to certain locations where the reception of required messages was consistently problematic for extended periods. Such locations can significantly impact the overall acquisition time, reducing the efficacy of FEC2 in these particular circumstances.

4.4 Reduced CED

Figure 10 illustrates the time required to obtain either normal-precision ephemeris with the aid of FEC2 or

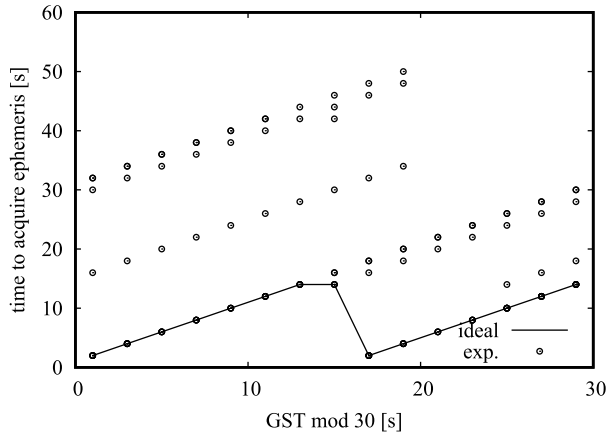


Fig. 10 Time required to obtain ephemeris when both FEC2 and Reduced CED are available.

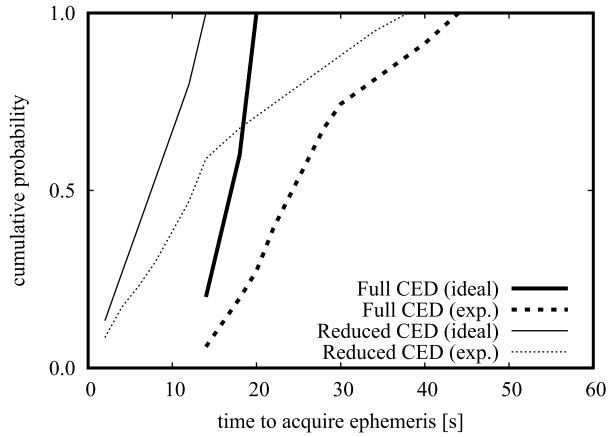


Fig. 11 Cumulative distribution of time to acquire Full CED and Reduced CED.

reduced-precision ephemeris using Reduced CED. This comparison provides insights into the efficiency of acquiring ephemeris data under different conditions and with the application of varying error correction and data reduction strategies. Since Reduced CED enables the acquisition of ephemeris in just one word, it significantly shortens the acquisition time. In Fig. 10, the number of plots is smaller compared to the scenario with only FEC2 in use. This is because ephemeris is often obtained more quickly through Reduced CED before FEC2 can complete its acquisition process.

The cumulative distribution that reflects this observation is presented in Fig. 11. This distribution offers a visual representation of the time efficiency gains achieved through the use of Reduced CED instead of losing Full CED accuracy, especially in comparison to the acquisition times when using the conventional ephemeris and FEC2. When comparing Full CED and Reduced CED results, it is evident that the introduction of Reduced CED significantly impacts ephemeris acquisition times in a mobile environment. Specifically, the time required to acquire ephemeris at the 50 percent probability level is reduced from 26.5 seconds to 12.3 seconds with the implementation of Reduced CED, marking a decrease of 54%.

Furthermore, the comparison of the two plots in Fig. 11 highlights the substantial role Reduced CED plays in mitigating the increase in ephemeris acquisition time caused by message losses in a mobile environment. The 50 percent values for ephemeris acquisition time are 7.3 seconds when Reduced CED is available, as opposed to 12.3 seconds in its absence. This underscores the effectiveness of Reduced CED in enhancing the efficiency of ephemeris acquisition, particularly in challenging reception conditions.

5. Conclusion

The conducted mobile reception experiment focused on the Improved I/NAV navigation message, which the European positioning satellite system Galileo began broadcasting in August 2022. In the context of receiving navigation messages, continuous reception over a certain period is essential; however, in mobile environments, maintaining continuous reception is not always feasible. The Improved I/NAV system is designed to facilitate rapid recovery from reception interruptions through the implementation of various features, including SSP, FEC2, and Reduced CED.

This experiment has provided valuable insights into the effectiveness of these improvements in real-world, mobile conditions, underscoring the potential for Enhanced I/NAV to significantly enhance the reliability and efficiency of satellite-based positioning systems.

In this study, I/NAV messages recorded in an environment with frequent radiowave interruptions were analyzed. The analysis demonstrated that SSP could extend the time synchronization range from 3 seconds to 30 seconds, within the limits where error correction is effective. Typically, the absence of continuous reception leads to an increase in the time required to acquire ephemeris; however, this increase was mitigated by the use of FEC2. Furthermore, the introduction of Reduced CED, with its compressed expression and frequent transmission, significantly shortened the ephemeris acquisition time.

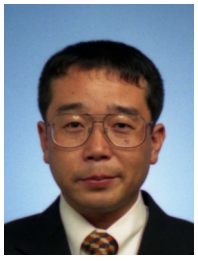
The mobile reception experiments validated the effectiveness of Improved I/NAV in rapidly recovering from signal interruptions. However, it was also observed that continuous signal interruptions could sometimes prevent a significant reduction in the time required to acquire ephemeris. This underscores the importance of continuous reception for optimal performance, even with advanced technologies like Improved I/NAV. The findings from these experiments provide valuable insights into the performance of satellite navigation systems in challenging reception environments and highlight areas for potential improvement in future system designs.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 20K04470, 23K03858.

References

[1] P. Misra and P. Enge, Global Positioning System: Signals, Measurements, and Performance, Ganga-Jamuna Press, Massachusetts, 2001.
[2] European Union Agency for the Space Programme (EUSPA), Using GNSS raw measurements on Android devices, 2017. https://www.euspa.europa.eu/system/files/reports/gnss_raw_measurement_web_0.pdf
[3] European Union Agency for the Space Programme (EUSPA), European GNSS open service signal-in-space interface control document (OS SIS ICD), issue 2.1, Nov. 2023. https://www.gsc-europa.eu/sites/default/files/sites/all/files/Galileo_OS_SIS_ICD_v2.1.pdf
[4] U.S. Space Force, Navstar GPS space segment/navigation user segment interfaces, Aug. 2022. https://www.gps.gov/technical/icwg/IS-GPS-200N.pdf
[5] European Union Agency for the Space Programme (EUSPA), Galileo open service navigation message authentication (OSNMA) signal-in-space interface control document (SIS ICD), Issue 1.0, Dec. 2022.
[6] Cospas-Sarsat, Specification for Cospas-Sarsat 406 MHz distress beacons, C/S T.001, Issue 4, Revision 10, Nov. 2022. http://www.cospas-sarsat.int/images/stories/SystemDocs/Current/T001-NOV-29-2022.pdf
[7] M. Paonni, M. Anghileri, T. Burger, L. Ries, S. Schlötzer, B.E. Schotsch, M. Ouedraogo, S. Damy, E. Chatre, M. Jeannot, J. Godet, and D. Hayes, "Improving the performance of Galileo E1-OS by optimizing the I/NAV navigation message," Proc. 32nd International Technical Meeting of the Satellite Division of The Institute of Navigation (ION GNSS+ 2019), pp.1134–1146, Miami, Sept. 2019. https://www.gsc-europa.eu/sites/default/files/sites/all/files/IonGnss19_InavOptimisation_FullPaper_GSC.pdf
[8] European Union Agency for the Space Programme (EUSPA), I/NAV Navigation Message Improvements Info Note, March 2023. https://www.gsc-europa.eu/sites/default/files/sites/all/files/INAV_Navigation_Message_Improvements_Info_Note.pdf
[9] K. Shimazu, K. Makabe, N. Nishii, R. Mori, N. Harada, H. Sashida, and K. Oikawa, "Emergency warning services via GNSS signals," 2020 IEEE Aerospace Conference, pp.1–16, March 2020. DOI: 10.1109/AERO47225.2020.9172377
[10] L.O. Espluga, C. Poulliat, M.-L. Boucheret, M. Aubault, and H.A. Bitar, "New solutions to reduce the time-to-CED and to improve the CED robustness of the Galileo I/NAV message," 2018 IEEE/ION Position, Location and Navigation Symposium (PLANS), pp.1399–1408, Monterey, April 2018. DOI: 10.1109/PLANS.2018.8373532
[11] M. Anghileri, M. Paonni, E. Gkougkas, and B. Eissfeller, "Reduced navigation data for a fast first fix," 6th ESA Workshop on Satellite Navigation Technologies (Navitec 2012) & European Workshop on GNSS Signals and Signal Processing, pp.1–7, Noordwijk, Dec. 2012. DOI: 10.1109/NAVITEC.2012.6423105
[12] T. Takasu, "Pocket SDR - An Open-Source GNSS SDR," v.0.8, Oct. 2021. https://github.com/tomojitakasu/PocketSDR
[13] T. Takasu, "RTKLIB: An open source program package for GNSS positioning," v.2.4.3 b34, Dec. 2020. https://github.com/tomojitakasu/RTKLIB



Satoshi Takahashi received M.E. and Ph.D. degrees from Tokyo Denki University, Japan, in 1992 and 2001. In 1992, he joined Hitachi, Ltd., where he engaged in research on radio propagation for indoors. During 1996–1999, he was a research engineer at YRP Key Tech Labs, where he engaged in research on radio propagation for the future 4th generation mobile radio communication systems. In 2002, he joined Communications Research Laboratory (now National Institute of Information and Communications Technology), where he engaged in future mobile radio communication systems. Since 2005, he is an associate professor at Hiroshima City University, Japan. Dr. Takahashi is a member of IPNTJ and IEEE.

PAPER

Long-Term Adaptive Bitrate Control Mechanism

Pierre LEBRETON[†], *Nonmember and* **Kazuhisa YAMAGISHI**[†a)], *Member*

SUMMARY Adaptive bitrate (ABR) video streaming is an important application on the Internet. To ensure that users enjoy high-quality services, ABR control mechanisms need to be designed that select chunks wisely on the basis of the available network throughput. To address the chunk selection problem, this paper describes an adaptive bitrate control mechanism that leverages long-term throughput information in the chunk selection process. While previous work has considered how quality should be requested on a per-chunk basis, the proposed method increases the timeframe of the analysis and allows higher quality of experience (QoE) to be reached. This is done by appropriately selecting a sequence of consecutive chunks' quality values instead of a single chunk's value. Simulation results are reported on a large variety of real-world network conditions and various throughput prediction algorithms and show the benefit of the proposed method over conventional ABR control mechanisms.

key words: *adaptive bitrate, control mechanism, long-term throughput, quality of experience*

1. Introduction

Streaming videos to a large audience is challenging. Each user has Internet access with different capabilities that can largely fluctuate over time. Therefore, the throughput requirements of the videos need to be dynamically adjusted so as to match the available bandwidth. This task is achieved by using an adaptive bitrate (ABR) video streaming technique that splits videos into short video segments called "chunks". Each chunk is encoded at various quality levels, resulting in different throughput requirements. Therefore, while playing the videos, the video player can select the chunks that will provide the highest quality of experience (QoE) to the users considering the available bandwidth. This is done by providing high-quality pictures, with few quality changes, and importantly, without any stalling events. The chunk selection process is handled by the ABR control mechanism of the video player [1].

Much effort has been put into studying the ABR control mechanism [2]. However, a limitation is the timeframe in which chunks are selected. Indeed, previous works have focused on selecting chunks on a per-chunk basis to match the observed network conditions and buffer filling rate. This paper will show that leveraging knowledge about future throughput even further in time can allow informed decisions to be made that enable stalling to be decreased and high-quality videos to be constantly provided, leading to improved user experience.

Figure 1 depicts this concept by showing an example of a car driving on the road. At the beginning of the example, the car benefits from a high available throughput. However, while driving on the road, the car enters a tunnel, which results in a drop in throughput. In this scenario, the ABR control mechanism observes the drop in throughput and has no choice but to rely on its buffer while waiting for the throughput to increase. As the network condition does not improve, the player is left with no choice but to request low-quality chunks to avoid stalling and may lead to a stalling event. In Fig. 1, such behavior is shown as a dashed line.

Such a passive approach is problematic as it does not leave options for the ABR control mechanism to improve the quality. Therefore, the proposed solution considers that if the player could have anticipated before entering the tunnel that throughput would drop, it would have been able to take different actions. For example, the player would have been able to increase the buffer filling rate at a higher level before entering the tunnel (eventually at the cost of decreasing quality) and then would have been able to maintain sustained high-quality videos while being inside the tunnel. This is represented as a pink continuous line in Fig. 1.

A car entering a tunnel is an extreme example to give an easy-to-understand scenario. However, in this paper, it will be shown that such a principle allows increasing user experience across various network conditions as can be experienced in real-world services.

The rest of this paper is organized as follows. Section 2 describes related work, Sect. 3 introduces the proposed method, Sect. 4 analyzes the performance results, Sect. 5 discusses the results, and Sect. 6 concludes this paper.

2. State of the Art

In the process of ABR streaming, video players aim to maximize the user experience by providing the highest quality possible considering the available throughput while also ensuring that frequent quality changes and stalling do not occur [1]. This process is handled by the ABR control mechanism of the video player and has received a lot of attention. In the following, key algorithms are introduced. These can be categorized as either based on handcrafted rules or learned based approaches.

Manuscript received December 1, 2023.
Manuscript revised April 4, 2024.
Manuscript publicized May 31, 2024.

[†]NTT Network Technology Laboratories, NTT Corporation, Musashino-shi, 180-8585 Japan.
a) E-mail: kazuhisa.yamagishi@ntt.com

DOI: 10.23919/transcom.2023EBP3196

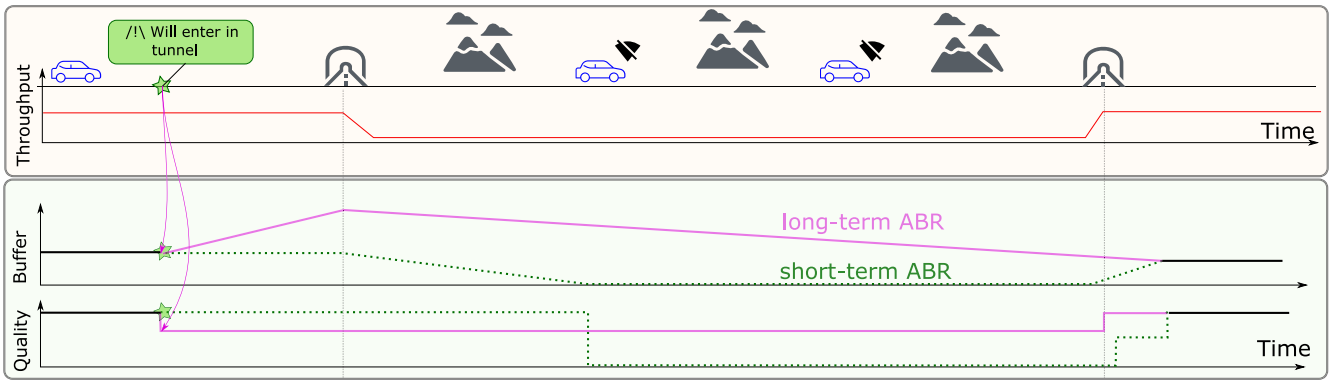


Fig. 1 Using long-term throughput prediction to improve chunk selection mechanism.

2.1 Handcrafted Rule-Based ABR Models

In terms of handcrafted rule-based ABR models, two main categories of models can be found. These can either be only based on buffer filling rate information or can also leverage information on throughput. Regarding buffer-based only approaches, Huang et al. [3] described a method that requests chunks on the basis of a linear function between the buffer filling rate and bitrate required for the considered chunk. By using this principle, as the buffer filling rate becomes low, lower quality chunks are requested to allow the buffer to be filled at a faster rate. In contrast, when the buffer is sufficiently full, higher quality chunks are requested as stalling events are unlikely to occur.

One limit with this approach is that there is no linear relationship between quality and video bitrate. Therefore, in BOLA [4], a more sophisticated approach is used as chunks are selected on the basis of not only the buffer filling rate but also the perceived quality of these chunks. This process is performed by defining threshold values in terms of the buffer filling rate, and if the buffer filling rate reaches the threshold, higher quality is requested.

These methods are of interest and have shown high performance [5], but due to buffer-only dependency, these methods cannot anticipate future events and will only respond to the observed status of the video player. Moreover, having knowledge about throughput can also be expected to improve the chunk selection process. Therefore, previous studies have investigated the use of throughput information when selecting chunks. In the category of handcrafted throughput dependent ABR control mechanisms, Model Predictive Control (MPC) [6]-based methodologies have been heavily investigated. The general principle of this control mechanism is to optimize a quality evaluation function. This function is defined so as to maximize the quality of each downloaded chunk, minimize quality changes between consecutive chunks, and minimize stalling. Future throughput is predicted, and the ABR control mechanism selects chunks on the basis of this quality evaluation function. In MPC [6], throughput is predicted by using the geometric mean and can lead to prediction inaccuracies. To tackle this challenge, RobustMPC was created to account for past throughput prediction errors in the new throughput predictions resulting in a more conservative chunk selection strategy. In Fugu [5], the general optimization criteria used in MPC are preserved while a deep learning-based algorithm is used to improve throughout prediction accuracy. Another alternative is BayesMPC [7], which leverages Bayesian neural networks to predict the distribution of throughput, while chunks are still selected on the basis of the same quality evaluation function as the one defined in MPC [6].

Extensions of ABR control mechanisms to address specific topics such as live streaming have also been proposed [8], [9]. Live streaming presents different challenges as delay should be reduced, requiring the video player to only use a low buffer filling rate to reduce as much as possible the time between a chunk being downloaded and shown to the user.

2.2 Reinforcement Learning-Based ABR Models

The second major category of ABR control mechanisms is the learning-based approaches that leverage reinforcement learning to perform the chunk selection. Pensieve [10] was a precursor for this type of model. Considering that it may be challenging to define proper rules to perform chunk selection, it has been proposed to leverage machine learning techniques to learn a proper set of rules for the ABR control mechanism.

The model uses features such as buffer filling rate information, previous throughput values, and previous downloaded chunk size to predict what chunk needs to be downloaded next. Extensions of this work were proposed by Mao et al. [11] that also account for bandwidth estimates. One limitation is the need to define a loss function while training the model. In these works, models were trained while using a loss function on the basis of the MPC objective function that requires researchers to define weights between coding quality and stalling related features. Considering that setting these weights is not easy, Huang et al. [12] described the use of Generative Adversarial Network (GAN) as it does not rely on the MPC objective function.

Sensor-Augmented Neural ABR Video Streaming on

unmanned aerial vehicles (UAVs), SA-ABR [13], has been proposed that generates ABR video streaming algorithms with the assistance of various kinds of inherent sensor data that are used to pilot UAVs. By incorporating the inherent sensor data with network observations, SA-ABR trains a deep reinforcement learning (DRL) model to extract salient features from the flight state information and automatically learn an ABR algorithm to adapt to the varying UAV channel capacity through the training process.

Finally, it should also be mentioned that the extension of reinforcement learning-based methods has also been proposed in the case of live streaming [14], [15].

2.3 Open Challenges

Based on these previous works, several open challenges remain.

First, chunks in the classical ABR control mechanism are selected on the basis of the MPC's objective function, which does not provide a proper modeling of the user's experience. Indeed, it relies on a weighted linear summation of quality of service (QoS) related metrics such as average bitrate, bitrate variation, stalling count, and stalling duration, which are related to QoE but do not properly model user experience. Modeling of the quality of ABR video streaming has been extensively researched in academia [1], [16]–[18] and standards organizations [19], [20]. Therefore, metrics designed for evaluating the overall quality of ABR video streaming need to be used in the chunk selection process as described in [21], [22].

Another candidate source of improvement is the timeframe of the chunk selection. Most previous works have focused on studying what immediate chunk needs to be requested and have not addressed the long-term aspect of the streaming session. Only reacting to degradation events in terms of network performance results in fewer opportunities for quality management. Indeed, with this short-sighted view, the video player can only decrease quality and rely on the buffer filling rate. Knowing the loss of throughput in advance will allow the player to plan a quality adaptation strategy to maintain high quality when network conditions become challenging. Note that although MPC predicts future throughput as described above, the increase in computational loads has been pointed out [6]. Therefore, meeting client requirements in terms of computational load involves the challenge of reducing computational loads and increasing the quality.

To address this point, only [22], [23] have described how knowledge of future throughput can be applied to improve the user experience. Instead of selecting which chunk should be downloaded on a per-chunk basis, chunk selection is optimized so as to maximize the quality of a path composed of successive chunks. However, these previous works lack two main points. 1) In the case of Bo et al., [23], the quality evaluation function was based on the MPC model that linearly adds the stalling duration, coding quality, and quality adaptation and does not relate well to the user's viewing experience. Furthermore, quality is measured only via bitrate and does not account for video resolution, which results in a very weak quality estimation model. 2) In the case of Lebreton and Yamagishi [22], a function that models the perceived quality of videos with coding and stalling impairment [17] was used and was shown to be able to improve that user experience. However, the work did not describe an end-to-end framework, and the impact of the accuracy of the throughput prediction was only weakly studied. Indeed, the throughput value predictions were based on ground truth data to which Gaussian noise was added. In a real-world scenario, throughput prediction may not follow this type of distribution, and it is also difficult to identify what would be an appropriate noise value. Therefore, throughput predictions algorithms need to be used. Another limitation of this previous work is the computational complexity of identifying the optimal path as computational complexity increases exponentially with path length, so more efficient computational algorithms are needed.

2.4 Contributions

This work addresses the question of designing a low-complexity long-term chunk selection mechanism that performs chunk selection on the basis of future throughput estimates. The method leverages a quality evaluation model and users' experience of the model while watching videos [17]. The quality is not optimized analytically as the parameter-based audiovisual quality estimation model is complex and includes non-linearities, random forest models making it hard to solve in real time on low power devices. Therefore, our contribution was to describe an alternative optimization procedure that consists of evaluating a carefully chosen path among the list of all possible combinations. Unlike previous approaches that considered this type of quality estimation model in the chunk selection process [21], the proposed method proactively responds to low-quality network conditions by identifying a quality adaptation path instead of a per-chunk selection mechanism. Furthermore, different from previous work such as [22], two main novel contributions are made:

- A computationally efficient quality-adapted path identification mechanism is proposed, with computational complexity that increases linearly with the path length. Concretely, to reduce computational loads, the search area for the path is restricted and the path with the stalling event is not used for the quality calculation. The mechanism with low complexity is compared with the state-of-the-art mechanisms in Sect. 4.
- This work uses different throughput prediction algorithms and discusses the impact of throughput prediction accuracy on the benefit of using long-term chunk selection mechanisms.

Note that an existing parameter-based audiovisual quality estimation model, which takes into account resolution, bitrate, framerate, and stalling features to estimate the over-

all user experience, is leveraged to calculate quality in this work. However, developing a quality estimation model is out of our paper's scope.

3. Proposed Method

In this section, the target video streaming and the general procedure of the long-term ABR control mechanism are described.

3.1 Type of Video Streaming Services

There are two main types of video streaming services: video on demand (VoD) and live streaming. In addition, there are two types of live streaming: high- and low-latency live streaming. The streaming server stores all the encoded chunks in VoD before video is provided. In the live streaming, video is encoded and chunks are stored on the streaming server while video is provided. Since an encoded chunk is immediately stored on the streaming server and is requested by clients in low-latency live streaming, the latency is short. On the other hand, even during in sports matches such as baseball and soccer, high-latency live streaming (e.g., about several minutes) is often used. In this case, many chunks are stored on the streaming server although not all chunks of whole content are available due to the live streaming.

From these investigations, the proposed method can be applied to VoD and high-latency-live streaming because the method is based on the long-term throughput prediction.

3.2 Long-Term Scheduling Principle

Figure 1 depicts an example of a user watching a video in a car. At a given time, a user may enjoy a high-throughput network access allowing him to watch high-quality videos. However, when the car enters a tunnel, the available throughput decreases significantly. In this context, the traditional "short-sighted" video player has no choice but to rely on its buffer while also requesting lower quality chunks to avoid stalling.

On the other hand, if the player could have anticipated that throughput would drop to this extent, it would have been able to identify a better solution. For example, it could have used the high throughput available before entering the tunnel to download slightly lower quality videos, allowing it to increase the buffer filling rate and enabling high-quality videos to be played while the car is in the tunnel.

To apply this principle, the video player needs to predict future throughput and compare the quality of different quality adaptation patterns for the period studied. Then, it can select the sequence of chunks that provides the highest quality to the users. In this paper, this sequence of chunks is referred to as a quality adaptation path. The process of finding the right quality adaptation path is shown in Fig. 2. At a certain time t, the video player needs to decide which chunk to request. The video player uses long-term throughput prediction to predict whether the player can successively

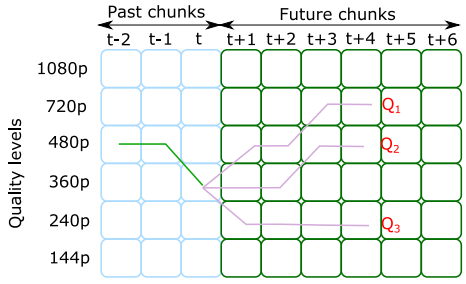


Fig. 2 Search for quality adaptation path. Various quality adaptation paths are compared by using a parameter-based quality estimation model [17], and the path providing the highest overall quality is selected.

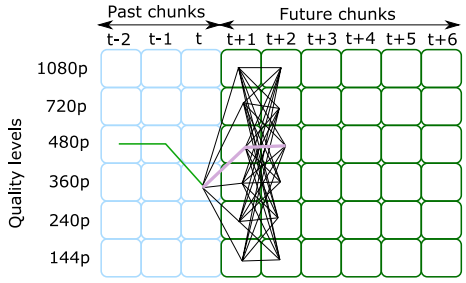


Fig. 3 Evaluation of overall quality for every path. The path providing the highest quality is highlighted in purple. In this example, other paths may deliver higher picture quality but lead to stalling. Therefore, a sequence of lower quality levels would be preferred.

download chunks along a quality adaptation path without involving stalling. As shown in Fig. 2, the player will consider different quality adaptation paths and evaluate the quality of this path by using a parameter-based audiovisual quality estimation model that takes into account resolution, bitrate, framerate, and stalling features to estimate the overall user experience if that path is chosen [17]. On the basis of these criteria, the paths leading to stalling will not be considered, and a path providing constant high-quality video will be selected. In Fig. 2, three paths are compared in terms of quality, but to identify the optimal quality adaptation, numerous paths with various quality adaptation patterns need to be compared to identify which one provides the best user experience. Figure 3 depicts the case where all possible quality adaptations are evaluated. In this figure, the path providing the highest overall quality is highlighted in bold. This figure also shows that a large number of paths can be considered when searching for the one that provides the highest overall quality. In a full-search approach, the number of paths to evaluate is L^D, where L is the number of quality levels and D is the length of the path: the number of future chunks. A less computing intensive search was introduced in [22] by only considering quality adaptations that are bounded within a specific range. For example, if the range $A = 1$, the video player may only increase or decrease the quality of one level or maintain the same quality as the previous step. As shown in Fig. 4, doing so allows the number of paths to be compared to be greatly decreased, resulting in $(2 \times A + 1)^D$ paths to be compared. However, a less computing intensive method is

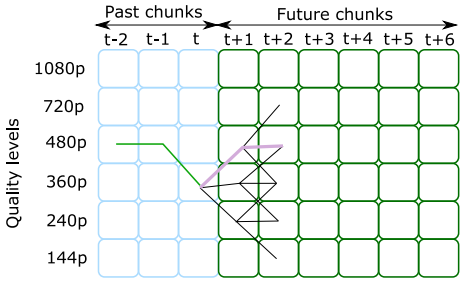


Fig. 4 Evaluation of quality of experience for various paths. The search space is bounded so as to allow only A=1 quality change levels.

still needed. Therefore, one of the main contributions of this paper is to introduce such a low-complexity long-term ABR control mechanism, which will be introduced in Sect. 3.4.

3.3 Long-Term Throughput Prediction Algorithm

First, note that to perform the long-term ABR control mechanism, future throughput values need to be estimated. Future throughput can be predicted on the basis of long-term time series forecasting such as described in [24]–[26]. Furthermore, it should also be considered that in the context of network operators, key features such as radio-related key performance indicators, GPS data, and mobility estimation can also be used for finding out about the network infrastructure as well as user mobility data to improve accuracy. The video player can then leverage information provided by the network leading to higher throughput prediction accuracy as described in Cradio [27].

In this paper, the impact of throughput prediction accuracy on the long-term chunk selection mechanism will be shown by comparing six different baselines with different degrees of accuracy:

- Ground truth: The first baseline will be based on ground truth data. Considering that this work is based on simulations using real-world network traces, accurate future throughput values can be obtained. This will serve as the baseline of an ideal throughput prediction mechanism.
- Last throughput value: This throughput prediction algorithm will consider that throughput is constant over a short period of time and then remains constant to the last observed throughput measurement value. If r_t is the throughput at time t, the estimate of the throughput at $t+1$, $\hat{r}_{t+1}$, is given by $\hat{r}_{t+1} = r_t$.
- Geometric mean: A classical approach for short-term throughput prediction is the geometric means as shown in the MPC algorithm [7]. This throughput predictor performs the geometric mean of all previous measured throughput values. The estimate of the throughput at $t+1$, is given by $\hat{r}_{t+1} = \sqrt[t]{\prod_{i=1}^{t} r_i}$.
- Geometric mean (bounded): Similar to the previous throughput predictor, this predictor only considers the measured throughput while downloading the previous six chunks. The estimate of the throughput at $t+1$ is given by $\hat{r}_{t+1} = \sqrt[6]{\prod_{i=t-5}^{t} r_i}$.
- Kalman filter: The throughput is estimated by using a Kalman filter as described in [23]. The model considers that throughput is constant over a short period plus subject to random noise (w_t), e.g., $r_{t+1} = r_t + w_t$. Similarly, the model also considers that throughput measurements ($v(t)$) are also subject to random measurement noise ($n(t)$), e.g., $v(t) = r(t) + n(t)$. Thus, every time the player downloads a new chunk, the estimate of throughput is updated iteratively on the basis of a Kalman gain that depends on variance of measured and predicted throughput as well as throughput prediction errors. The prediction of the filter is defined by $\hat{r}_{t+1} = \hat{r}_t + K_t \times (r_t - \hat{r}_t)$, where K_t is the Kalman gain at time t defined by $K_t = \frac{P(k-1)+W}{P(k-1)+W+Q}$, where W and Q are respectively the variance of the process noise and measurement noise. Moreover, $P(t)$ is the system error variance, which can be computed recursively at every step by $P(t) = (1 - K_t)(P(t-1) + W)$. On the basis of these equations, throughput prediction is updated iteratively while accounting for prediction and measurement accuracy.
- Random Forest: The algorithm uses a random forest model that takes throughput measurement from downloading the previous six chunks and predicts the throughput when downloading the next chunk. The estimate of the throughput is then given by $\hat{r}_{t+1} = f_{RF}(r_{t-5}, r_{t-4}, r_{t-3}, r_{t-2}, r_{t-1}, r_t)$, with f_{RF} the random forest model. The random forest was pretrained on the basis of ground truth data. Across the 667 network traces, 723 distinct tuples of 7 consecutive throughput values were randomly selected: $(r_{t-5}, r_{t-4}, r_{t-3}, r_{t-2}, r_{t-1}, r_t, r_{t+1})$. Among these 723 traces, 361 tuples were randomly selected and used for training while the remaining 362 tuples were used for validation. A random forest with 100 trees was trained and showed a Pearson correlation coefficient (PCC) of 0.963, and a root mean square error (RMSE) of 63.150 kbps on training data and a PCC of 0.954 and RMSE of 67.492 kbps on validation data.

The different baseline algorithm describes the process for predicting the throughput for downloading the next chunk. Then, the extension to a longer timeframe is obtained by applying these algorithms recursively. Figure 5 depicts the prediction accuracy of the different throughput predictors for different prediction time windows on the same validation data that was used to validate the random forest model. In this figure, as the throughput prediction window increases, prediction error increases. At around 35 seconds, prediction accuracy tends to increase slightly, which must be due to underlying periodicity in the throughput measurements. In this figure, the random forest model generally shows higher prediction accuracy than the other predictors that performed similarly to each other.

The performance of long-term chunk selection mech-

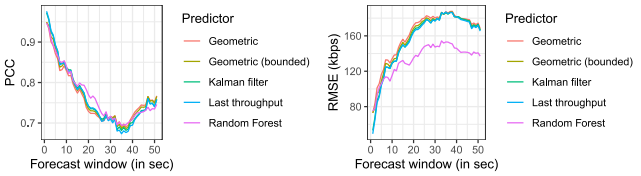


Fig. 5 Prediction accuracy of the throughput estimation algorithms at different timeframes. PCC and RMSE are respectively the Pearson Correlation Coefficient and the Root Mean Square Error.

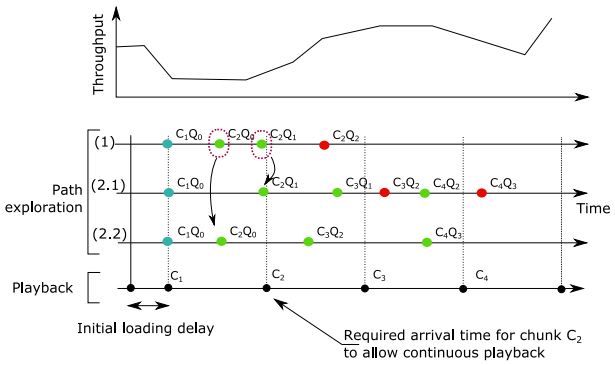


Fig. 6 Dependency between selected chunks while optimizing the quality of a path. Green and red dots correspond respectively to chunks that can be downloaded or not at that time.

anisms combined with these different prediction algorithms will be compared with those of state-of-the-art ABR control mechanisms in Sect. 4.

3.4 Path Quality Optimization

As described in Sect. 3.2, a low-complexity long-term ABR control mechanism is needed. Thus, this section describes how the mechanism identifies the path that is ignored before the quality calculation.

When searching for the path that provides the optimal quality adaptation pattern, a proper search strategy needs to be considered. If the video is available in L quality levels and a path composed of D chunks, a total of L^D possible paths can be compared. The exponential increase in the number of paths depending on the path leads to practical application issues, so optimized search strategies are required.

Figure 6 describes the general idea of the proposed method. In this figure, the dependency between chunks is depicted. Furthermore, it can be seen that throughput varies as a function of time. Therefore, depending on when the download of a chunk starts, the time required to download the chunk may vary. In Fig. 6, the "playback" line corresponds to the constraint of the video player in order to play back video without interruption. The playback line contains points marked as C_1, C_2, C_3, and C_4, which are the times at which chunks C_1, C_2, C_3, and C_4 need to be available so the player can play video continuously. The path-exploration line illustrates the dependencies between consecutive chunks. In path exploration (1), chunk C_1 with the quality level Q_0 is downloaded. To be able to play video

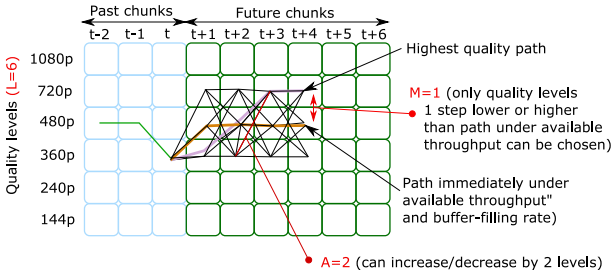


Fig. 7 Search around path directly under throughput. The schema also summarize the roles of parameters L, A, and M.

continuously, chunk C_2 needs to be downloaded before playback of C_2 starts. On the basis of the available throughput, chunk 2 (C_2) can be downloaded in quality level 0 (C_2Q_0) or quality level 1 (C_2Q_1). The chunk with quality level 2 (C_2Q_2) arrives too late, so it cannot be selected (and is marked by a red dot). Quality values higher than Q_2 will only require more bits, so all paths using quality level 2 or higher for the second chunk can be ignored.

For the first step, only two possibilities exist: (C_2Q_0) and (C_2Q_1). First, the proposed algorithm selects the highest quality chunk that can arrive on time: (C_2Q_1). This corresponds to the path exploration (2.1) in Fig. 6. If this chunk is selected, the next highest quality chunk that can be downloaded in time for the third chunk is (C_3Q_1), then it is (C_4Q_2) for the fourth chunk.

However, requesting the highest quality chunk that can be downloaded may not necessarily provide the highest quality to the users. Indeed, in the example of the path exploration (2.2), it is shown that considering the lower throughput available at the time of downloading chunk C_2, requesting a lower quality for the chunk C_2 allows the download to complete earlier. It then enables higher quality chunks to be downloaded for chunks C_3 and C_4 (respectively (C_3Q_2), and (C_4Q_3), resulting in a better overall experience for the users.

Therefore, the proposed search strategy consists of two steps: 1) The identification of the sequence of the highest quality chunk that can be downloaded on the basis of the available throughput and time constraint for uninterrupted video playback. 2) The search for alternative paths around the firstly identified path by increasing or decreasing the quality of the chunks. Note that if there is stalling in all paths, the method will select the path with the lowest coding quality so as to limit the stalling duration.

Figure 7 summarizes the different parameters (L, A, M, D) and shows an example of the candidate paths that are considered using the proposed method. First, the path "immediately under throughput and buffer filling rate" shows the sequence of the highest quality chunks that can be downloaded in time for a smooth playback. Then, other candidate quality adaptation paths are considered by decreasing or increasing chunk quality to around the originally identified path.

The range of the quality deviation from the original path is constrained by a parameter M. A value of $M = 1$ indicates that chunk quality can only be increased or decreased by one

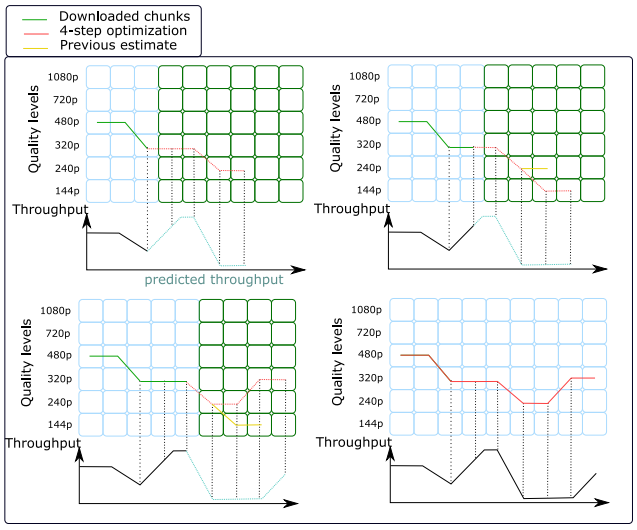


Fig. 8 Iterative process for long-term chunk selection algorithm.

level around the original path. Doing so allows the search space to be decreased while searching for the optimal path. When $L = 6$ is the number of available quality levels, a value of $M = L$ results in searching all possible quality adaptations.

Finally, another refinement is to constrain the quality adaptation between consecutive chunks so as to only allow quality changes from up to a given level ($A = 2$ in Fig. 7).

On the basis of this search strategy, the upper bound of the number of paths that can be considered while searching for the best path increases linearly with the length of the path rather than increasing exponentially as before. Thanks to this improvement, long paths can be considered by the long-term ABR control scheme with a low computational cost.

After paths are identified, the quality per path is calculated on the basis of a parameter-based audiovisual quality estimation model. Finally, the optimal path is selected in terms of quality.

3.5 Chunk Request

Figure 8 depicts the chunk download process. Once the path providing the highest overall quality is identified, the first chunk on this path is requested. Then, by downloading the next chunk, new data on the network performance can be measured. The ABR control mechanism then uses this new measurement to update the throughput prediction, extend the time horizon, update the quality adaptation path, and select the following chunk.

4. Results

In this section, the evaluation procedure as well as simulation results are provided.

4.1 Evaluation Procedure

Although an extreme case (i.e., car entering a tunnel) was

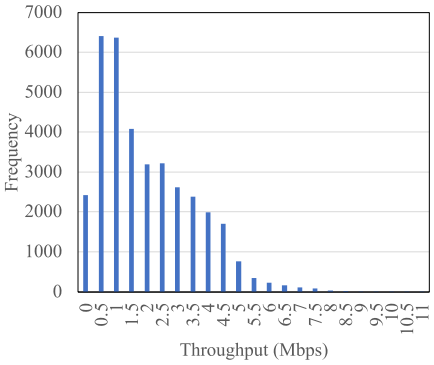


Fig. 9 Throughput histogram.

explained in Sect. 1, the proposed method needs to be used for the various real-world network conditions. To evaluate the performance of the proposed method, three datasets (i.e., 3G/HSDPA [28], FCC [29], and Oboe [30]) are used in this simulation that cover various conditions, including users riding public transportation or being stationary at home, as shown in Fig. 9. Note that the bar is difficult to find when the throughput is higher than 8 Mbps because the amount of data is small and that the bar is drawn on the basis of a bin of every 0.5 Mbps. More concretely, there are throughputs that are basically stable, ramp up and/or down gradually, and ramp up and/or down drastically in the traces. Therefore, these datasets are suitable for evaluating the performance of the proposed method.

Simulations are performed by replicating the traces using a virtual player [7] that allow simulations to be performed at a faster than real-time speed. This allowed numerous conditions to be tested in a short amount of time. Types of service are VoD and high-latency live streaming, as described in Sect. 3.1. Chunk duration is configured to be 2 seconds long, and a bitrate ladder of 6 quality levels ranging from HD to 144p is used. Bitrate values are 4300, 2850, 1850, 1200, 750, and 300 kbps. The framerate was constant to 30 frames per seconds, and the video codec was H.264. Although this configuration was used in our training, note that other configurations of codecs, bitrate, resolutions, and framerate could be supported as well.

On the basis of the real-world throughput measurement datasets, 667 network traces can be tested to evaluate the ABR algorithm performance. Performance evaluation is done by simulating the player behavior while faced with various real-world network conditions. Then, information on selected chunks, chunk quality, initial loading, and stalling is measured and compared with various baselines. In addition, a quality evaluation model designed for ABR video streaming, ITU-T Rec. P.1203 mode 0 [19], [31], [32], was used to predict the overall user's experience by accounting for coding quality and stalling events in a joint manner. This will be the basis for the comparisons between the different models.

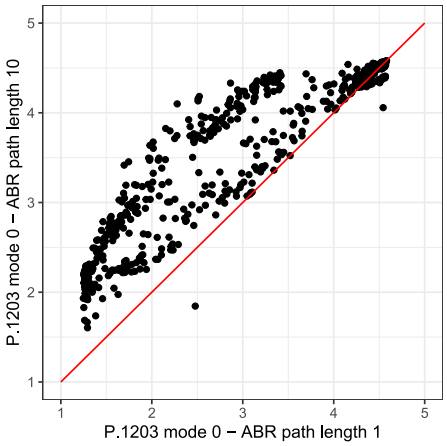


Fig. 10 Comparison of session quality between short- and long-term ABR control mechanism for each of the 667 real-world network traces. Quality is measured using ITU-T P.1203 mode 0 [19], [32].

4.2 Performance Evaluation in Ideal Conditions

First, to demonstrate the benefit of the proposed method, the ideal case where throughput can be estimated accurately is considered. For the search parameters, a value of L=6 (as there are 6 quality levels), A=1, M=6, and various values of D (path length) are used. Other parameter values will be discussed in Sect. 4.5. When this scenario is considered, the differences between short- and long-term ABR control schemes are demonstrated in Fig. 10. In this figure, each point corresponds to the overall session quality of one of the 667 considered network traces while using a short- or long-term ABR control mechanism measured using the ITU- T P.1203 mode 0 model [19]. The ITU-T P.1203 mode 0 model [19] is a parameter-based quality estimation model that takes into account features such as bitrate, resolution, framerate, and stalling features as well as their respective temporal fluctuation to predict quality scores. Quality scores are given in the mean opinion score (MOS) scale that was obtained from intensive subjective testing performed in the ITU-T standardization work. The MOS values are defined on a 5-point scale and correspond to average user ratings that were collected using the 5-point scale single stimulus Absolute Category Rating (ACR) methodology. The five points of this scale correspond respectively to the labels: Bad, Poor, Fair, Good, and Excellent. Figure 10 shows that the quality is improved by using the long-term ABR control scheme.

Next, to deeply analyze the obtained results in Fig. 10, Fig. 11 is discussed next. Figure 11 shows the relationship between path length and percentage of sessions that show a quality improvement of 0.5, 0.75, and 1 unit of MOS, where a long-term chunk selection mechanism is used for the same network condition. It can be observed that as path length increases, the percentage of sessions that show quality improvement increases as well. With a path length of 12, 50% of all sessions show a quality improvement of 0.5 units of

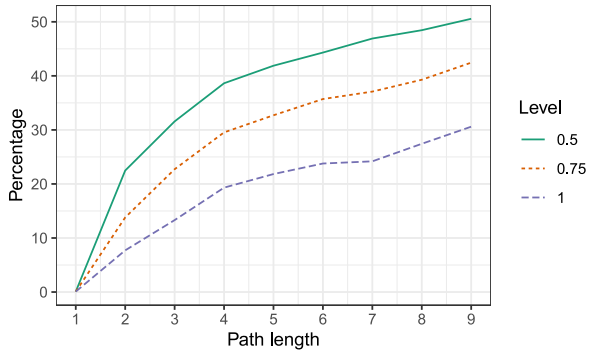


Fig. 11 Percentage of sessions that shows a quality improvement of 0.5, 0.75, and 1 unit of MOS as a function of path length.

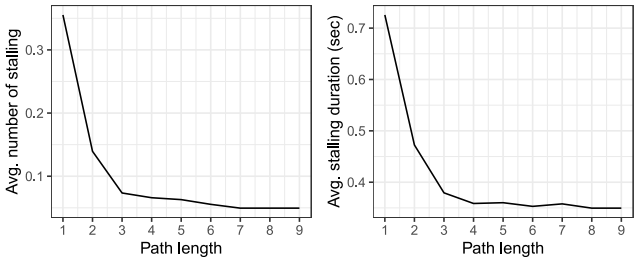


Fig. 12 Average number of stalling events and total stalling duration as a function of path length.

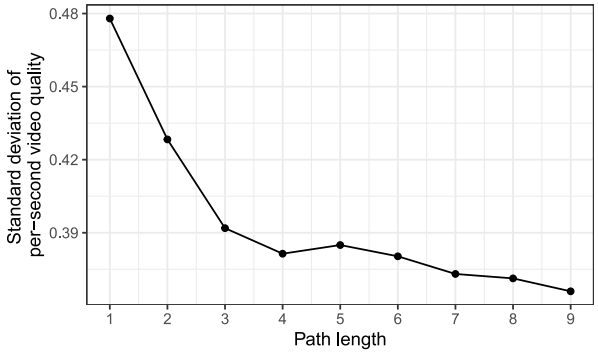


Fig. 13 Standard deviation of per-second video quality scores estimated using ITU-T P.1203 mode 0 as a function of path length.

MOS and 30% of sessions show a quality improvement of 1 unit of MOS. Figure 12 depicts the number of stalling events and stalling duration as a function of path length and shows a similar tendency: as path length increases, the number of stalling events and average stalling duration decrease, and overall user experience improves. Finally, Fig. 13 depicts the standard deviation of per-second video quality estimate from ITU-T P.1203 mode 0 over all 667 streaming sessions as a function of the path length. From this figure, as path length increases, video quality is less subject to quality changes. Concluding this analysis, the results show that optimizing a quality path increases quality, decreases stalling, and decreases the magnitude of quality changes compared with selecting chunks on a one-by-one basis. Figures 11–13 show that the optimal path length is 9 in terms of quality improve-

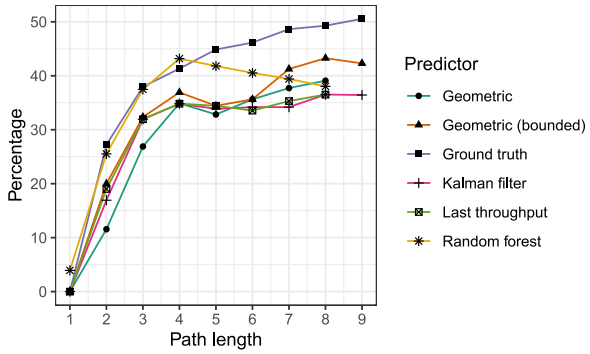


Fig. 14 Percentage of sessions that show a quality improvement of 0.5 units of MOS as a function of path length across different throughput prediction algorithms.

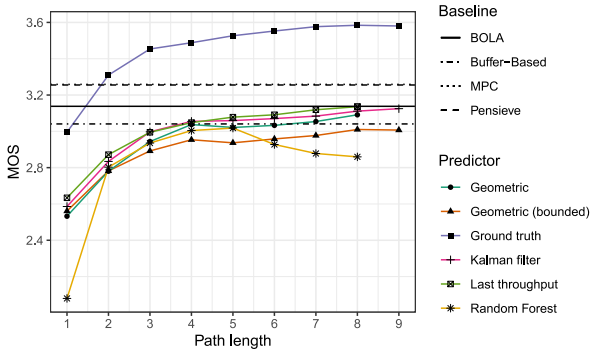


Fig. 15 Evaluation of session quality using ITU-T P.1203 mode 0 across different path length, throughput predictor, and against state-of-the-art methods. State-of-the-art methods do not depend on path length and so are represented as horizontal lines.

ment, stalling events, and quality changes. Therefore, the target duration of bitrate optimization should be less than 18 seconds when the chunk duration is 2 seconds.

From these results, to determine how many path lengths are used, service operators need to take into account the trade-off between quality improvement and latency in VoD and high-latency live streaming. Moreover, if the method does not use many path lengths, the latency can be short but will not be extremely so.

4.3 Comparison between Different Throughput Predictor Models

In the previous section, the benefit of using the long-term control mechanism over short-term ones was shown when throughput is evaluated accurately. However, in real-world scenarios, the estimation of throughput is error-prone. To evaluate the performance of the proposed long-term control mechanism when throughput prediction error occurs, the five different throughput predictors described in Sect. 3.3 are considered and compared with an accurate throughput predictor (based on ground truth data). As for parameter search values, these are left to L=6, A=1, M=6 and various D values (path length).

Figure 14 depicts the percentage of sessions that show a quality improvement of 0.5 units of MOS as a function of the path length. In this figure, increasing the path length up to 4 chunks leads to increased session quality across all predictors. In the case of the random forest model, improvement decreased after a path length of 4. The geometric mean-based predictor shows further slower performance improvement while the Kalman filter-based and last throughput-based predictor are not improved by using a longer path. These results show that the idea of using long-term chunk selection holds even when error-prone throughput predictors are used.

4.4 Comparison with State-of-the-Art Algorithms

To further validate the proposed method, the performance of the long-term chunk selection mechanism is compared with those of several state-of-the-art ABR control mechanisms: Buffer-Based [3], MPC [6], BOLA [4], and Pensive [10]. Figure 15 depicts the average session quality measured using ITU-T Rec. P.1203 mode 0 [19] for the 667 sessions as a function of the path length, throughput predictors, and baselines. In this figure, an error-free throughput predictor in combination with a large path length results in a significant quality improvement compared with the state-of-the-art methods and reached a MOS of 3.58. Among state-of-the-art methods, Pensive and MPC provided the highest quality resulting in overlapping lines in Fig. 15. Average MOS scores were respectively 3.25 for Pensieve and 3.26 for MPC. They were followed by BOLA (MOS=3.14) and Buffer-Based (MOS=3.04). As for the proposed long-term chunk selection mechanism used in combination with error-prone predictors, results show that the last throughput-based and Kalman filter-based algorithms provided quality close to BOLA (MOS=3.13 at a path length of 9). However, a throughput predictor algorithm with higher accuracy would be required to improve quality even further as shown by the ground truth-based predictor.

However, note that the loss of overall session quality as measured by ITU-T Rec. P.1203 mode 0 [19] by the long-term ABR control mechanism was also compensated for by two important aspects: quality was found to be more stable, and shorter stalling was observed.

To illustrate this, the fact that ITU-T Rec. P.1203 mode 0 is able to predict per-second quality scores is used to analyze the temporal evolution of quality. Figure 16 shows the standard deviation of per-second video quality scores (O.22 in ITU-T recommendation P.1203 [19]), as a function of path length and throughput predictors. In this figure, with the exception of random forest and bounded geometric throughput predictors, the long-term ABR control schemes provided a smaller per-second quality score standard deviation and hence more stable quality than the state-of-the-art algorithms. The throughput prediction based on ground truth data provided the most stable quality. Among the state-of-the-art methods, MPC provided the most stable quality, followed by Pensieve, BOLA, and Buffer-Based. As for the error-prone throughput predictors, it was found that more sta-

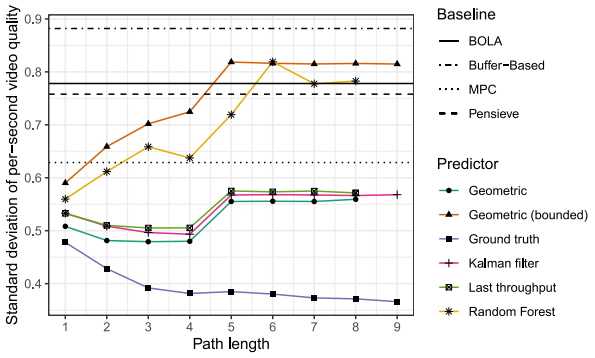


Fig. 16 Average standard deviation of per-second video quality scores estimated using ITU-T Rec. P.1203 mode 0 (noted as O.22 in [19]) across all sessions as a function of path length and throughput prediction algorithms.

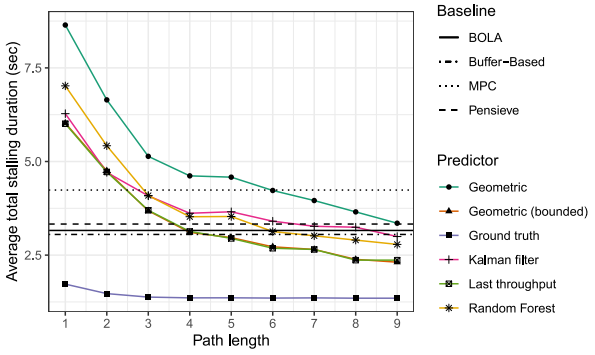


Fig. 17 Average stalling duration across the 667 sessions as a function of path length and throughput prediction algorithms.

ble quality could be achieved by increasing the path length up to four chunks. After this path length, the increasing prediction error resulted in more frequent quality changes (while still being more stable than state-of-the-art algorithms).

With respect to stalling, Fig. 17 depicts the average stalling duration encountered across the 667 sessions. In this figure, as the proposed long-term ABR algorithm avoids paths that include stalling, using a longer path length leads to a shorter total stalling duration. With respect to state-of-the-art algorithms, BOLA and Buffer-Based provided the least amount of stalling, closely followed by Pensieve, while MPC provided a significantly longer total stalling duration. As for the long-term ABR control mechanism, it was shown that using a longer path length allowed for a shorter total stalling duration. The ground truth-based throughput estimator quickly converges to the best solution it can achieve and delivered significantly shorter total stalling duration. As for the error-prone throughput predictors, using a longer path consistently leads to shorter total stalling duration.

Concluding this analysis, note that there exists different components to the overall video QoE: coding quality and stalling. Using long-term ABR control mechanism allowed for a consistent decrease in the total stalling duration (Fig. 17) and frequency of quality change for most throughput predictors (Fig. 16), but resulted in a cost in terms of overall quality as lower quality chunks were selected (Fig. 15). Similar observations were also made on state-of-the-art methods as BOLA and Buffer-based methods provided shorter stalling than MPC and Pensieve, but Pensieve and MPC provided higher coding quality. It was shown that having an accurate throughput predictor with long-term ABR control enables state-of-the-art algorithms to be outperformed by a large margin, but such an accurate throughput predictor is challenging to obtain, and the naive throughput prediction algorithms used in this paper, while being competitive when used with a long path, have room for further improvement. Among state-of-the-art algorithms, Pensive performed best as it allowed almost as short total stalling duration as Buffer-Based and BOLA and also provided high-coding quality. Across the different throughput predictors that were considered, the Kalman filter-based predictor performed consistently across stalling and coding quality and enabled shorter total stalling duration than reference algorithms while providing quality comparable to BOLA.

4.5 Parameters and Search Analysis

As described in Sect. 3, in addition to the path length, the proposed long-term ABR control mechanism depends on several parameters: the range of possible quality adaptation between consecutive chunks (A), and the distance to the base quality path immediately under the available throughput (M). In this section, the influence of these parameters on the performance of the long-term ABR control mechanism will be discussed.

Figure 18 depicts the performance of the proposed optimized search method based on network throughput information (hereafter, network search, NS) illustrated in Fig. 7. Additionally, another baseline that considers all possible paths while restricting the range of quality adaptation to A change in quality levels is also depicted (hereafter, restricted search, RS), which was described in [22] and is illustrated in Fig. 4. To quantify the differences between the different search methods and their respective possible configuration, the Bjøntegaard-Delta [33], [34] is computed and reported in Table I. The Bjøntegaard-Delta is a metric designed to measure the average difference between rate-distortion curves to evaluate differences between codec. In this work, since the quality improvement is a function of the path length, it is proposed to use the Bjøntegaard-Delta to report on the average improvement provided by one algorithm over another one across various path lengths.

4.5.1 Range of Quality Adaptations

Before comparing the search strategies, the impact of the range of quality adaptation allowed between consecutive Chunk (A) onto long-term ABR performance is investigated. In Fig. 18 and Table 1, the performance results of RS are depicted for different values of A (note that RS does not make use of parameter M, so M = NA in Fig. 18). From this result, using lower A-values resulted in higher performance. This can be explained by having lower A-values force the player

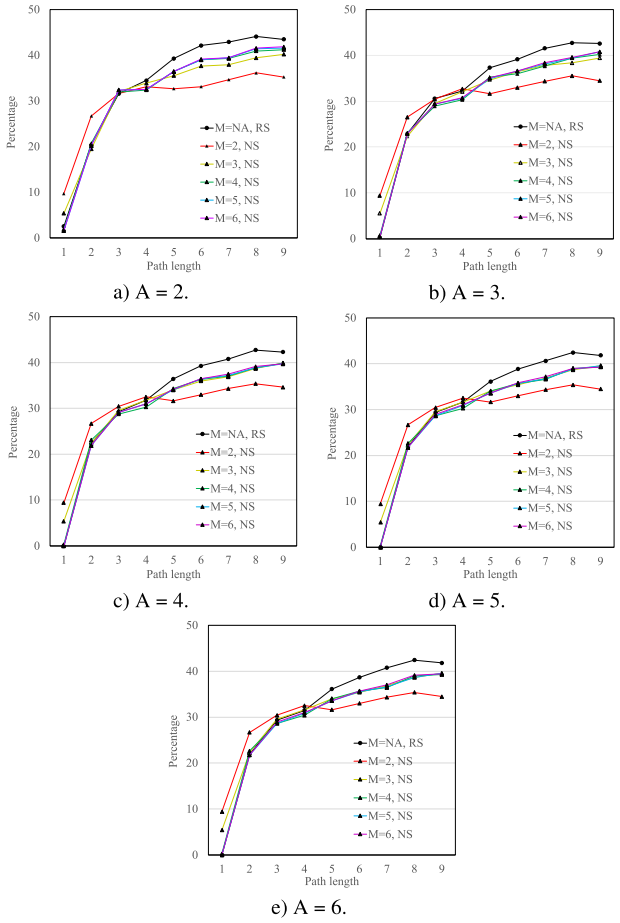


Fig. 18 Comparison of search strategies performance between the restricted search (RS) and the proposed network throughput-based search (NS). Performance results show the percentage of sessions received a quality improvement of at least 0.5 units of MOS. M and A are parameters of the search algorithms introduced in Sect. 3.

Table 1 Bjøntegaard-Delta (BD) between search strategies and their respective configurations.

Ref.	Compared	BD
RS:A2	RS:A3	-0.0111
RS:A2	RS:A4	-0.0166
RS:A2	RS:A5	-0.0192
RS:A2	RS:A6	-0.0193
RS:A2	NS:A2:M1	-0.0342
RS:A2	NS:A3:M1	-0.0398
RS:A2	NS:A4:M1	-0.0396
RS:A2	NS:A5:M1	-0.0398
RS:A2	NS:A6:M1	-0.0398
RS:A2	NS:A2:M2	-0.0241
RS:A2	NS:A3:M2	-0.0299
RS:A2	NS:A4:M2	-0.0321
RS:A2	NS:A5:M2	-0.0331
RS:A2	NS:A6:M2	-0.0331
RS:A2	NS:A2:M3	-0.0189
RS:A2	NS:A3:M3	-0.0321
RS:A2	NS:A4:M3	-0.0350
RS:A2	NS:A5:M3	-0.0375
RS:A2	NS:A6:M3	-0.0375

Ref.	Compared	BD
RS:A2	NS:A2:M4	-0.0180
RS:A2	NS:A3:M4	-0.0303
RS:A2	NS:A4:M4	-0.0356
RS:A2	NS:A5:M4	-0.0376
RS:A2	NS:A6:M4	-0.0376
RS:A2	NS:A2:M5	-0.0180
RS:A2	NS:A3:M5	-0.0303
RS:A2	NS:A4:M5	-0.0356
RS:A2	NS:A5:M5	-0.0387
RS:A2	NS:A6:M5	-0.0393
RS:A2	NS:A2:M6	-0.0175
RS:A2	NS:A3:M6	-0.0294
RS:A2	NS:A4:M6	-0.0347
RS:A2	NS:A5:M6	-0.0373
RS:A2	NS:A6:M6	-0.0374

to perform smooth quality adaptation, resulting in two main benefits. First, smoother quality changes are preferred by the users. Second, offering the player with the ability to perform large quality changes results in cases where request for

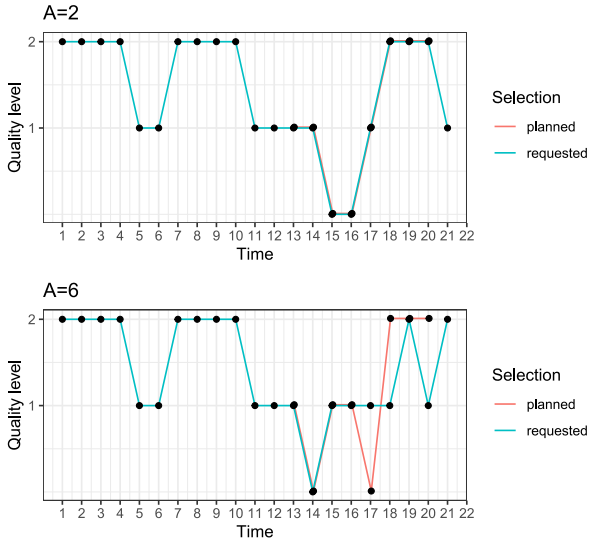


Fig. 19 Effect of the range of quality adaptation on selected path. The figure shows the differences between originally planned path (in red) and actual player actions (in blue). As the path is recomputed, differences between planned and actual actions are observed.

intermediate quality may be delayed so as to directly request even higher quality chunks. However, as illustrated in Fig. 8, the overall chunk selection process is an iterative approach, and as chunks are received, the path becomes continuously recomputed. As the path becomes recomputed, the quality change to the highest quality levels may be reconsidered, leading to an overall lower quality. This phenomenon can occur in conditions where throughput varies largely as found in the real-world network traces used in this work.

Such a scenario can be seen in Fig. 19. In this figure, the blue curve corresponds to the quality levels of the chunks that are requested as a function of time. At the time of requesting the 13th chunk, the long-term ABR model predicts a sequence of chunk quality level that should be selected. Two results from two searches where $A = 2$ or $A = 6$ are depicted as a red path. In the case of $A = 6$, the ABR control mechanism anticipates that for the 18th chunk, it will be able to download in quality level 2. However, as throughput is re-estimated while downloading chunks 13 to 16, new information is taken into account, and the initially scheduled path is not optimal anymore. Therefore, as marked in blue, the final player behavior differs from the path planned at time 13. In the case of $A = 2$, much smoother quality adaptations were planned and resulted in higher chances for the originally planned path at time 13 to be fulfilled. This results in smoother and more constant quality leading to higher overall session quality.

Therefore, this result shows that enforcing smoother quality adaptations, and then lower values of A, leads to higher quality. A recommended value is to use $A = 2$ as it leads to both higher performance and lower computational complexity.

4.5.2 Comparison between Algorithms

With respect to search strategies, it can be seen that RS enables higher performance than the proposed cost-effective search method based on network throughput information (NS). However, it does not do so by a large amount.

To quantify differences, the Bjøntegaard-Delta values are listed in Table 1 and indicate the average difference between methods across various path lengths. From these numbers, the optimal configuration of the network throughput-based search (NS) only differs from RS by 1% while having significantly lower computational cost. Interestingly, the proposed NS also allows higher performance to be reached on lower path length values compared with the RS approach. This result is obtained because NS uses network throughput as a base path, which allows the ABR control mechanism to perform larger quality changes than the RS approach in the short term. However, selecting chunks close to the available throughput results leads to un-optimal solutions when long paths are considered. This can be seen in Fig. 17 when "NS" is used with a value of $M = 2$. In this case, performances are higher for shorter path lengths, while when the path is longer than 5, increasing the value of M leads to performance improvement. This is the result of letting the player deviate from the path directly under the available throughput. This result can be numerically confirmed in Table 1: performance increases as M becomes larger. This result is due to the fact that larger M allows the player to use high throughput values when available to cache higher quality chunks before throughput decreases. Results also show that using a value of $M = 3$ already provides most of the quality improvement over lower M values and should be considered as a trade-off with computational complexity.

5. Discussion

This paper presents an ABR control mechanism that leverages long-term throughput prediction to select chunks. While the proposed method allows the experience of users to be improved, several points should be discussed.

First, with regards to the chunk selection strategy, this paper described how an efficient method that leverages throughput information and a quality evaluation model designed for ABR video streaming [17] were used to identify a subset of quality adaptation paths that should be considered as candidates for providing higher user experience. While the proposed method allows for a low computing cost to identify a path that maximizes the quality measured using a no-reference parameter-based quality estimation model [17], [35], the general idea of a long-term ABR control mechanism does not restrict the proposed ABR control scheme. Considering the performance of reinforcement learning-based ABR control mechanisms such as Pensieve [10], extending state-of-the-art reinforcement learning-based ABR control mechanisms could also be considered to predict a path of quality adaptations instead of only the immediate following chunk. This was not done in this work, but should be considered as a candidate research topic to build upon this work.

Second, with regards to throughput prediction, Sect. 4 shows that even with a simple throughput predictor, the use of a longer path length leads to better user experience by providing shorter total stalling duration and fewer quality changes. The overall quality as evaluated by ITU-T P.1203 mode 0 was found to be comparable to that of the BOLA algorithm. Note that the quality improvement of proposed method for five throughput predictors is about 15% lower than that of ground truth because of the throughput predication error (see Figs. 11 and 18). However, the quality is improved in the meddium and low quality videos, as shown in Fig. 10. In addition, the proposed method with long-term ABR control can suppress quality changes and reduce the stalling duration, as shown in Figs. 16 and 17. Therefore, the proposed method can provide benefits to users of video streaming. If throughput could be predicted accurately, state-of-the-art algorithms could be significantly outperformed both in terms of overall quality (see Fig. 15) and total stalling duration (see Fig. 17). In the literature, ongoing research can be found on the use of radio-related key performance indicators, GPS data, and mobility estimation for taking into knowledge about the network infrastructure and user mobility to provide more accurate throughput predictions [27]. By using such technology, the benefit of long-term ABR control schemes will be even further increased as shown in our simulations based on ground truth data.

Third, Sect. 3 describes how the proposed method reduces computational complexity. Although the full-search approach (i.e., L^D) is ideal to find an optimal path, high computational complexity is needed. Therefore, a less computing intensive search (i.e., $(2\times A+1)^D$) and an ignored path for the quality calculation are introduced (i.e., Sect. 3.4) to reduce the complexity. Since $A = 2$ and $D = 9$ are obtained from the performance evaluation, it results in $5^9 (= 1,953,125)$. However, $6^9 (= 10,077,696)$) is needed in the full-search approach, when 6 quality levels ($L = 6$) are used. If many quality levels ($L > 5$) are used, the proposed method has lower complexity than the full-search approach. Note that th proposed method with long-term ABR control always has higher computational complexity than exisiting models (i.e., Buffer-Based [3], MPC [6], BOLA [4], and Pensive [10]) because it needs to find an optimal path depending on A and D. However, the computational power of recent devices such as smartphones and set-top boxes has increased. Therefore, even when its computational complexity is high, the method will work well. In addition, as described in this section, extending state-of-the-art reinforcement learning-based ABR control mechanisms could also be considered to predict a path of quality adaptations. Therefore, the computational complexities of the proposed method, existing methods, and extended state-of-the-art mechanisms with long-term ABR control could be compared. This was not done in this work, but should be considered as a candidate research topic.

Finally, one could argue that the buffer should only

be filled as much as possible to avoid any sort of stalling. While this is true, the proposed algorithm has the benefit of decreasing load on the network infrastructure in two ways: first, it does not need to download content that the user may not see if he or she terminates the viewing session before the end of the video. Second, it may not be desirable to download numerous chunks at once as at a later stage, throughput may increase and allow higher quality downloading. In such a scenario, a naive player may download a chunk multiple times at different quality levels, while the proposed algorithm will be able to anticipate that a higher quality chunk can be downloaded at a later time and will wait for that time, leading to both increased quality and decreased network usage.

6. Conclusion

In this paper, an adaptive bitrate (ABR) control mechanism was proposed that selects chunks by optimizing a quality path. The effectiveness of the proposed approach was demonstrated in simulations based on real-world throughput measurement. Specifically, a path length of 10 was shown to enable session quality to be increased by 0.5 units of mean opinion score (MOS) on a 5-point scale in 50% of all 667 considered sessions and by 1 unit of MOS in 30% of sessions. This shows the potential for large quality improvement by applying ABR control mechanisms that schedule chunks on the basis of a longer timeframe. Furthermore, the results also show that the number of stalling events was significantly reduced with longer path lengths. Different throughput prediction algorithms were compared, and results showed that a long-term ABR control mechanism could even be achieved with simple throughput mechanisms. An algorithm was also proposed that leverages throughput information to efficiently select the quality path that provides high quality to users. Finally, comparisons with state-of-the-art ABR control mechanisms were performed and showed that the proposed method achieves competitive results even when error-prone throughput predictors are used, while error-free throughput prediction algorithms outperformed state-of-the-art methods by a large margin.

In future work, throughput prediction using ratio information as well as user mobility data shall be studied to improve the throughput prediction accuracy and allow the benefit of the proposed method to be further increased. In addition, although our proposed algorithm selects an optimal path on the basis of an iterative approach, mathematical optimization may be able to be applied to select the optimal path. To verify the performance, this may require large-scale simulations.

References

[1] M. Seufert, S. Egger, M. Slanina, T. Zinner, T. Hoßfeld, and P. Tran-Gia, "A survey on quality of experience of HTTP adaptive streaming," IEEE Commun. Surveys Tuts., vol.17, no.1, pp.469–492, 2015.
[2] A. Bentaleb, B. Taani, A.C. Begen, C. Timmerer, and R. Zimmermann, "A survey on bitrate adaptation schemes for streaming media over HTTP," IEEE Commun. Surveys Tuts., vol.21, no.1, pp.562–585, 2019.
[3] T.Y. Huang, R. Johari, N. McKeown, M. Trunnell, and M. Watson, "A buffer-based approach to rate adaptation: Evidence from a large video streaming service," SIGCOMM, 2014.
[4] K. Spiteri, R. Urgaonkar, and R. Sitaraman, "BOLA: Near-optimal bitrate adaptation for online videos," Proc. 35th Annual IEEE Int. Conf. on Computer Communications, pp.1–9, 2016.
[5] F.Y. Yan, H. Ayers, C. Zhu, S. Fouladi, J. Hong, K. Zhang, P. Levis, and K. Winstein, "Learning in situ: a randomized experiment in video streaming," Symp. on Networked Systems Design and Implementation (NSDI'20), 2020.
[6] X. Yin, A. Jindal, V. Sekar, and B. Sinopoli, "A control-theoretic approach for dynamic adaptive video streaming over HTTP," Proc. 2015 Conference of the ACM SIGCOMM, pp.325–338, 2015.
[7] N. Kan, C. Li, C. Yang, W. Dai, J. Zou, and H. Xiong, "Uncertainty-aware robust adaptive video streaming with bayesian neural network and model predictive control," NOSSDAV, pp.17–24, 2021.
[8] B. Wang, F. Ren, and C. Zhou, "Hybrid control-based abr: Towards low-delay live streaming," 2019 IEEE Int. Conf. on Multimedia and Expo (ICME), pp.754–759, 2019.
[9] A. Erfanian, F. Tashtarian, A. Zabrovskiy, C. Timmerer, and H. Hellwagner, "OSCAR: On optimizing resource utilization in live video streaming," IEEE Trans. Netw. Serv. Manag., vol.18, no.1, pp.552–569, 2021.
[10] H. Mao, R. Netravali, and M. Alizadeh, "Neural adaptive video streaming with pensieve," Proc. Conference of the ACM Special Interest Group on Data Communication, pp.197–210, 2017.
[11] H. Mao, S. Chen, D. Dimmery, S. Singh, D. Blaisdell, Y. Tian, M. Alizadeh, and E. Bakshy, "Real-world video adaptation with reinforcement learning," Workshop in the 36th Int. Conf. on Machine Learning, 2019.
[12] T. Huang, X. Yao, C. Wu, R.X. Zhang, Z. Pang, and L. Sun, "Tiyuntsong: A self-play reinforcement learning approach for ABR video streaming," 2019 IEEE Int. Conf. on Multimedia and Expo (ICME), pp.1678–1683, 2019.
[13] X. Xiao, W. Wang, T. Chen, Y. Cao, T. Jiang, and Q. Zhang, "Sensor-augmented neural adaptive bitrate video streaming on UAVs," IEEE Trans. Multimedia, vol.22, no.6, pp.1567–1576, 2020.
[14] F. Tashtarian, R. Falanji, A. Bentaleb, A. Erfanian, P.S. Mashhadi, C. Timmerer, H. Hellwagner, and R. Zimmermann, "Quality optimization of live streaming services over HTTP with reinforcement learning," 2021 IEEE Global Communications Conference (GLOBECOM), 2021.
[15] L. Cui, D. Su, S. Yang, Z. Wang, and Z. Ming, "TCLiVi: Transmission control in live video streaming based on deep reinforcement learning," IEEE Trans. on Multimedia, vol.23, pp.651–663, 2021.
[16] W. Robitza, M.N. Garcia, and A. Raake, "A modular HTTP adaptive streaming qoe model — Candidate for ITU-T P.1203 ("P.NATS")," 2017 Ninth Int. Conf. on Quality of Multimedia Experience (QoMEX), 2017.
[17] K. Yamagishi and T. Hayashi, "Parametric quality-estimation model for adaptive-bitrate-streaming services," IEEE Trans. Multimedia, vol.19, no.7, pp.1545–1557, 2017.
[18] C.G. Bampis, Z. Li, I. Katsavounidis, T.Y. Huang, C. Ekanadham, and A.C. Bovik, "Towards perceptually optimized adaptive video streaming-a realistic quality of experience database," IEEE Trans. Image Process., vol.30, pp.5182–5197, 2021.
[19] ITU-T Recommendation P.1203, "Parametric bitstream-based quality assessment of progressive download and adaptive audiovisual streaming services over reliable transport," ITU-T, 2017.
[20] ITU-T Recommendation P.1204, "Video quality assessment of streaming services over reliable transport for resolutions up to 4k," ITU-T, 2020.
[21] T. Kimura, T. Kimura, A. Matsumoto, and K. Yamagishi, "Balancing quality of experience and traffic volume in adaptive bitrate streaming," IEEE Access, vol.9, pp.15530–15547, 2021.
[22] P. Lebreton and K. Yamagishi, "Adaptive bitrate control mechanism

based on long-term evaluation," IEEE Int. Conf. on Multimedia and Expo (ICME), 2022.
[23] B. Wang and F. Ren, "Towards forward-looking online bitrate adaptation for DASH," Proc. 25th ACM Int. Conference on Multimedia, MM'17, New York, NY, USA, pp.1122–1129, Association for Computing Machinery, 2017.
[24] X. Dong, W. Fan, and J. Gu, "Predicting lte throughput using traffic time series," ZTE Communications, 2015.
[25] L. Bryan and Z. Stefan, "Time-series forecasting with deep learning: A survey," Philosophical Transactions of the Royal Society a Mathematical, Physical and Engineering Sciences, vol.379, no.2194, 20200209, 2021.
[26] T. Zhou, Z. Ma, Q. Wen, X. Wang, L. Sun, and R. Jin, "FEDformer: Frequency enhanced decomposed transformer for long-term series forecasting," Proc. 39th International Conference on Machine Learning (ICML), 2022.
[27] M. Sasaki, T. Nakahira, T. Moriyama, T. Ogawa, Y. Asai, and Y. Takatori, "Multi-radio proactive control technology (Cradio): A natural communication environment where users do not need to be aware of the wireless network," NTT Technical Review, vol.19, no.8, pp.37–45, 2021.
[28] H. Riiser, P. Vigmostad, C. Griwodz, and P. Halvorsen, "Commute path bandwidth traces from 3G networks: Analysis and applications," ACM MMSys, pp.114–118, 2013.
[29] F.C. Commission., "Raw data — Measuring broadband America," Measuring Broadband America, 2016.
[30] Z. Akhtar, Y.S. Nam, R. Govindan, S. Rao, J. Chen, E. Katz-Bassett, B. Ribeiro, J. Zhan, and H. Zhang, "Oboe: Auto-tuning video ABR algorithms to network conditions," Conf. of the ACM Special Interest Group on Data Communication, pp.44–58, 2018.
[31] A. Raake, M.N. Garcia, W. Robitza, P. List, S. Göring, and B. Feiten, "A bitstream-based, scalable video-quality model for HTTP adaptive streaming: ITU-T P.1203.1," Int. Conf. on Quality of Multimedia Experience (QoMEX), Erfurt, IEEE, May 2017.
[32] W. Robitza, S. Göring, A. Raake, D. Lindegren, G. Heikkilä, J. Gustafsson, P. List, B. Feiten, U. Wüstenhagen, M.N. Garcia, K. Yamagishi, and S. Broom, "HTTP adaptive streaming QoE estimation with ITU-T Rec. P.1203: Open databases and software," 9th ACM Multimedia Systems Conference, Amsterdam, pp.466–471, 2018.
[33] C. Herglotz, M. Kranzler, R. Mons, and A. Kaup, "Beyond Bjøntegaard: Limits of video compression performance comparisons," IEEE Int. Conf. on Image Processing (ICIP), 2022.
[34] G. Bjontegaard, "Calculation of average PSNR differences between RD-curves," VCEG-M33, 2001.
[35] P. Lebreton and K. Yamagishi, "Transferring adaptive bit rate streaming quality models from H.264/HD to H.265/4K-UHD," IEICE Trans. Commun., vol.E102-B, no.12, pp.2226–2242, Dec. 2019.

Pierre Lebreton received his engineering degree in Computer Science from Polytech' Nantes, France in 2009. In 2010, he joined the group Assessment of IP-based Applications at TU-Berlin, where he studied towards his Ph.D. on 3D video QoE. After graduating, he joined in 2015 the group of Audio Visual Technology at the TU-Ilmenau, Germany, in 2015 and the group of Networked Sensing and Control at Zhejiang University, China, in 2016. His work addressed various topics including aesthetic appeal, large scale video quality monitoring, and bike sharing systems. In 2017, he joined NTT Laboratories, where he now focuses on quality and user-engagement prediction for video streaming applications.

Kazuhisa Yamagishi received his B.E. degree in Electrical Engineering from the Tokyo University of Science in 2001 and his M.E. and Ph.D. degrees in Electronics, Information, and Communication Engineering from Waseda University in Japan in 2003 and 2013. Since joining NTT Laboratories in 2003, he has been engaged in the development of objective quality-estimation models for multi-media telecommunications. From 2010 to 2011, he was a visiting researcher at Arizona State University. He received the Young Researchers' Award (IEICE) in Japan in 2007, the Telecommunication Advancement Foundation Award in Japan in 2008, the ITU-AJ Encouragement Award in 2017, and the TTC Award for distinguished service in 2018.