IEICE TRANS. ELECTRON., VOL.E103–C, NO.4 APRIL 2020

INVITED PAPER *Special Section on Solid-State Circuit Design — Architecture, Circuit, Device and Design Methodology*

Essential Roles, Challenges and Development of Embedded MCU Micro-Systems to Innovate Edge Computing for the IoT/AI Age

Takashi KONO[†a)], **Yasuhiko TAITO**[†], ***Nonmembers***, ***and*** **Hideto HIDAKA**[†], ***Member***

SUMMARY Embedded system approaches to edge computing in IoT implementations are proposed and discussed. Rationales of edge computing and essential core capabilities for IoT data supply innovation are identified. Then, innovative roles and development of MCU and embedded flash memory are illustrated by technology and applications, expanding from CPS to big-data and nomadic/autonomous elements of IoT requirements. Conclusively, a technology roadmap construction specific to IoT is proposed.

key words: *edge computing, embedded system, embedded flash memory, IoT, AI*

1. Introduction: Embedded System Challenges for IoT

Requirements for the emerging IoT (Internet of Things) society press embedded systems* with tremendous amount of technology and design challenges, showing inflective changes in the history of embedded systems and roles of MCU (Micro Controller Unit) products. We have entered a drastic upturn rather than conservative extensions in frameworks of "edge computing" in IoT systems. Unprecedented combinations of technology and application inducing innovations are envisioned in coming data-driven society. Accordingly, drastic changes in the embedded computing infrastructure show inflection in technology advancement.

We identify rationales of edge computing, essential IoT capabilities, IoT data supply chain innovations, and justify embedded MCU system approaches to edge computing in Sect. 1.

(*Embedded system in this paper denotes computer system with dedicated functions within a larger mechanical/electrical system.)

1.1 Elements of IoT and Rationales for Edge Computing

The IoT concept in Fig. 1 (a) assumes a lot of computing and communication nodes inter-connected to serve for a society: for sensing, judgments, triggering cloud computing and feedback to work on the real-world through big-data capabilities by IT (Information Technology) and OT (Operational Technology). In this highly interconnected structure, how to organize local computing and communication to better serve for large human systems is still much to be explored.

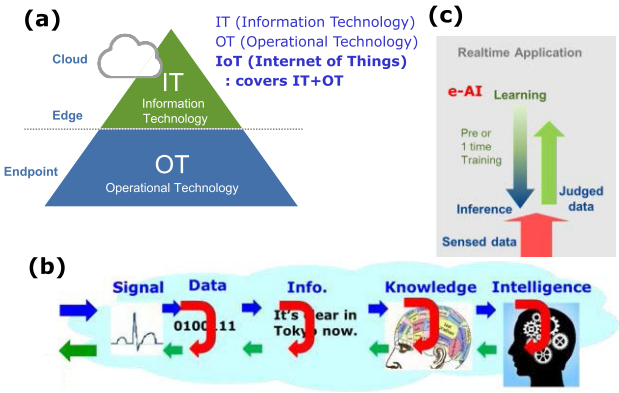


Fig. 1 (a) IoT landscape and (b), (c) important IoT data flows

We see characteristic tendency in data-driven vertical and lateral interconnections in the IoT data supply and flow:

1) Multiple/heterogeneous real-time feedback paths and control for advanced mobility control in Fig. 1 (b) [1].
2) Reduced big-data communication traffic by AI inference at edge computing to cope with discrepancy in available computing power/physical data, and local communication capability in power and cost balance shown in Fig. 1 (c).
3) Upgradable, self-powered systems adaptable to the environment, in nomadic and autonomous ways.

Embedded system integration is re-defined as the key object of edge computing designs in the implementation of IoT concepts. By breaking down into essential elements of IoT concepts in view of IoT data flows above, Fig. 2 (a) depicts three elements of IoT: CPS (Cyber-Physical System), Big Data, and Nomadic/Autonomous. Real implementation of IoT is challenged by these factors. Consequently, embedded systems are evolving from compact, smart, and connected systems to realizing CPS feedback, computation-intensive edge inference, and autonomous operations, all for data supply chain innovations.

Accordingly, MCU products are building up a lot of embedded capabilities unseen in the past, shown in Fig. 2 (b), aiming at development of multiple technology as well as applications. Therefore, embedded systems will encounter with a multi-faceted technical roadmap and development for the IoT/AI age.

Manuscript received October 29, 2019.

[†]The authors are with Renesas Electronics Corp., Tokyo, 135–0061 Japan.
a) E-mail: takashi.kono.fv@renesas.com

DOI: 10.1587/transele.2019CDI0001

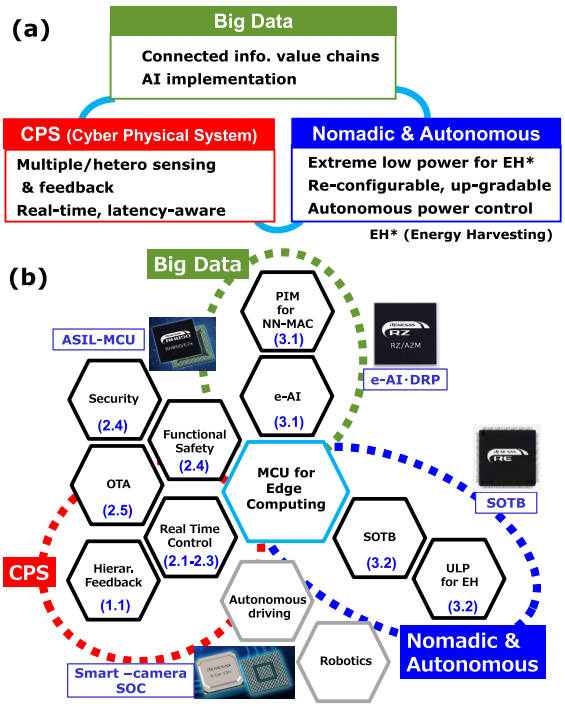


Fig. 2 (a) Three elements of IoT, and (b) technical factors of MCU in edge computing and corresponding sub-section numbers in this paper

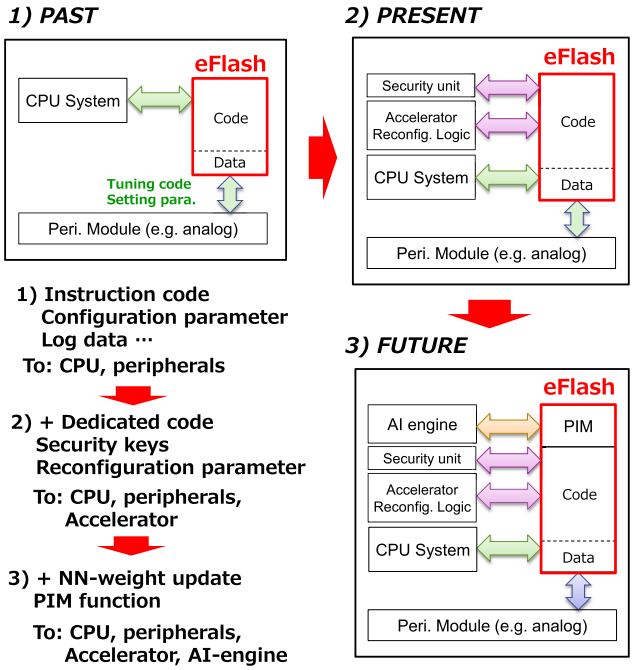


Fig. 3 Development of usages for flash memory storage in MCU

1.2 Roles of MCU and Embedded Flash Memory (eFlash)

Integrated circuits for general-purpose computing uses have split into two major semiconductor products in 1970s: Micro Processor Unit (MPU) and MCU. While MPU is a computation-intensive product for computer systems, MCU is intended for small-scale one-chip integration of the minimum computing functions for digitization of embedded real-time control applications, replacing and/or merging with mixed signal products therein. In general, MCU has integrated element parts and functions in the small system into one-chip system and has adapted to real-time applications most efficiently with small footprint and low power consumption, often under stringent environment conditions. It has penetrated in consumer, industry and automotive applications, realizing embedded systems with optimized performance/power/cost.

On-chip CPU inevitably necessitates program code storage. In a small system with MCU, it has been pursued in a unique way: after trying on on-chip mask-ROM and OTP (One-Time Programmable ROM), coming into embedded flash memory (eFlash) solutions, motivated by value, cost, and performance advancements in embedded systems. In this way, MCU systems have utilized on-chip alterable-code functions for drastically changing the supply chain of MCU products, where eFlash plays an essential role, achieving "flash innovation" in embedded systems around 2000 [2].

Starting from a simple replacement of mask-ROM and OTP by flash memory for alterable code, flash-MCU has found multiple advantages by "programmability," "non-volatility," and "embedded-ness" of eFlash [2]. From code storage expanding to small data, computational coefficients, and security keys, eFlash has been adopted in general purpose MCUs as well as automotive-MCU, security-MCU and DSPs for expanded applications.

Deviating from the flash technology for stand-alone high-density storage, eFlash technology has had a unique evolution path, according to embedded system requirements for low voltage, low-power, high-reliability under higher temperature, and real-time (low latency) operations. Technology as well as circuit and sub-system designs dedicated to eFlash for MCU have met these requirements in the history since 1990s. Extended uses of on-chip storage by eFlash (Fig. 3) have provided program codes then adding small data and computational coefficients to CPU, and then providing them to accelerator and other on-chip IPs. In the future, it is expected to provide large-scale weighting matrix for neural networks in the form of processing-in-memory (PIM).

Figures 4 and 5 describe typical eFlash usages on the chip in expanding applications in the history. In Fig. 4, an EV (Electric Vehicle) motor controller MCU with on-chip inverter-control accelerator off-loads CPU very effectively by employing eFlash to provide program code and computation coefficients to the accelerator [3]. In Fig. 5, a dedicated secure eFlash module stores security keys and secure data on the chip. Together with on-chip cryptography processing, this embedded sub-system works effectively without exposing non-encrypted secure data off the chip. In this way, flash-MCU has expanded applied market by "programmable," "non-volatile," and "embedded" natures of eFlash. Succeeding the flash-MCU innovation, MCU em-

bedded systems now pursue opportunities of innovations in IoT data supply chains according to above-mentioned IoT elements (CPS, Big Data and Nomadic/Autonomous).

In the following Sects. 2 and 3, roles and technical factors of MCU embedded systems in edge computing shown in Fig. 2 (b) are discussed respectively. In Sect. 2, we will discuss how eFlash technology has evolved primarily to advance mobility CPS designs, followed by Sect. 3 to describe future MCU/eFlash application expansion in endpoint artificial intelligence (e-AI) for Big Data, ultra-low power for EH (Energy Harvesting, for Nomadic/Autonomous), and the future prospect of eFlash in view of emerging non-volatile memory. In Sect. 4, we will propose how the roadmap of embedded systems is constructed for coordinated IoT developments.

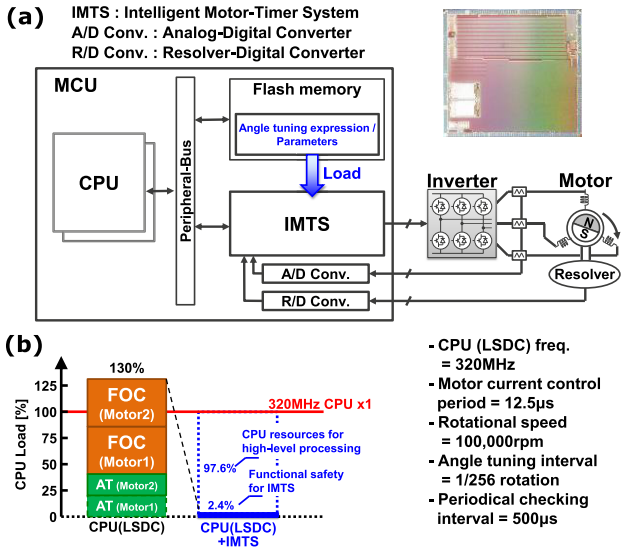


Fig. 4 Embedded flash memory use case in EV motor control [3]: (a) MCU architecture with accelerator (IMTS), and (b) Accelerator offloads CPU very effectively by 97.6%

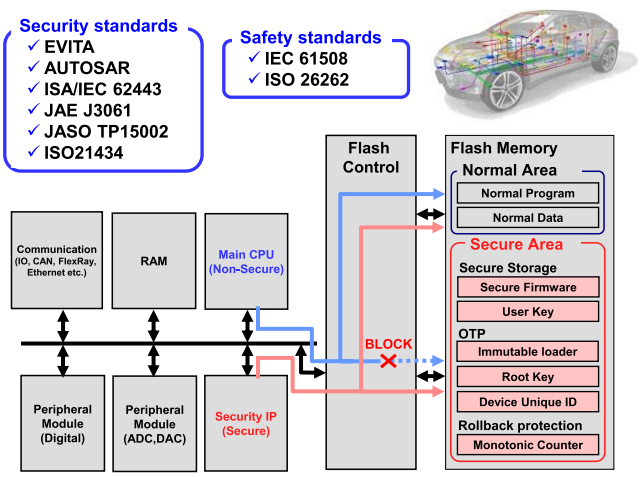


Fig. 5 Embedded flash memory use case in automotive security

2. Embedded Flash Memory Development for CPS: Performance and Power, Security and Safety and OTA

2.1 Process and Design Efforts in Scaling Scenario of eFlash

One of the application areas driving the scaling of embedded non-volatile memories (eNVMs) is automotive control. As shown in the upper half of Fig. 6, traditionally, higher fuel efficiency in combustion engine and more elaborate motor control in HEV/EV have required eNVMs to have larger capacity and faster random read access capability. Furthermore, the advent of new applications such as ADAS, autonomous driving and connected car with over-the-air (OTA) program update demands more powerful MCUs operating at several hundred MHz with large application code, which, accordingly, accelerates the pace of performance and capacity enhancement in eNVM scaling trend [4]. In the future, eNVM-based disruptive design technologies such as PIM or NV-logic will be introduced to meet more stringent requirements in performance and support the technology progress in automotive industry. Meanwhile, automotive applications pose extremely stringent reliability requirements to semiconductor devices. Especially, for eNVMs, stable operations and data retention at extremely high temperature (e.g. Tjmax = 175°C for AEC-Q100 auto grade-0) are mandatory.

Driven by the above-mentioned requirement trends for eNVMs in MCU, eFlash was introduced in 1990s to replace existing mask ROM, EPROM and OTP. Its advantages in scalability and programmability brought "flash innovation" also in automotive applications, and eFlash-base MCU (flash-MCU) obtained strong positions. Many types

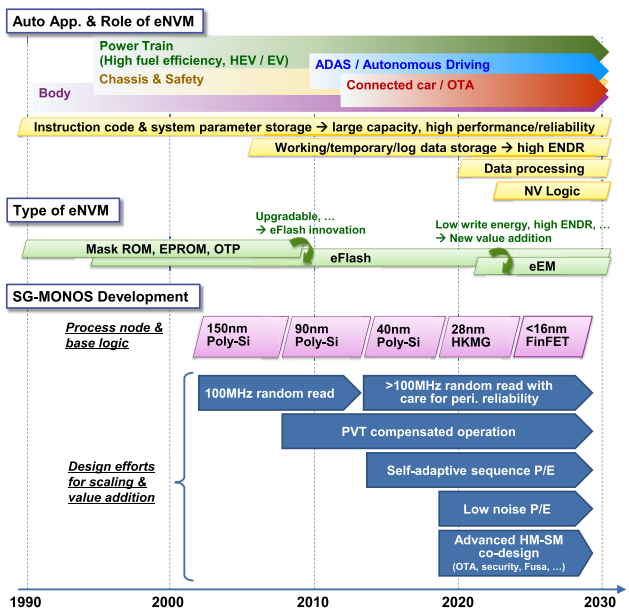


Fig. 6 Scaling scenario by device/circuit (SG-MONOS)

of eFlash memory cells have been proposed [5], but mainly due to the difficulties in cell scalability and affinity with base-line CMOS logic process, most of them reached the limitation of scaling around 90nm. Split-gate (SG)-MONOS still enjoys future prospect thanks to its intrinsic advantages of 1) high reliability based on charge storage in traps separately located in silicon-nitride film [6] and 2) high affinity with CMOS logic process scaling trend (lower gate height or three-dimensional gate shape of logic transistors) owing to thin-film storage by silicon-nitride film. As shown in Fig. 6, SG-MONOS has been successfully scaled down from 150nm to 28nm and it is confirmed at array level that SG-MONOS can be integrated in 16/14nm FinFET CMOS logic process while keeping reliability and program/erase performance [7].

Proper design efforts at an eFlash macro and system level are also essential in conjunction with memory cell scaling [8], as summarized in Fig. 6. They contribute to keeping or enhancing performance and reliability, which, otherwise, would be degraded due to physical scaling of eFlash memory cell and peripheral transistors and metal interconnection. In addition, they play key roles to meet new functional requirements according to technology and standard trends in industries. By taking SG-MONOS as an example, 4 cases of design efforts will be presented in the following sub-sections; 1) random read performance enhancement, 2) adaptive program/erase control, 3) implementation of security and safety functions, and 4) realization of robust OTA software update for automotive.

2.2 Random Read Performance Enhancement

SG-MONOS cell consists of two transistors, an access transistor and a memory transistor, separated by a small separation (~100Å level) formed by a double gate process (Fig. 7 (a)). Thanks to this split-gate structure, SG-MONOS cell is "deplete-free", because access transistors in unselected cells (WL = 0V) cut off the leakage current even when the Vth of erased cells is lower than 0V. Accordingly, MG can be set at 0V in read operations and highly boosted WL level for selected cells is not needed. In addition, no high voltages are applied to WLs and LBLs even during program or erase operations. As a result, all the circuits in the read path can be built with fast low-voltage logic transistors, leading to fast random read access capability. In fact, 100MHz random read access was already achieved at the first generation (150nm) of SG-MONOS.

Along with technology scaling to meet the demands for better performance/power characteristics in CPU-eFlash operations and larger eFlash memory capacity, some design techniques, especially those for WL drivers and sense-amplifiers (SAs), have been introduced with close attention to the reliability issues related to eFlash integration into scaled baseline CMOS logic process. Figure 7 (a) summarizes these techniques and achievements. For example, hierarchical SA scheme with offset cancellation [9] and divided WL scheme with WL repeaters and distributed WL

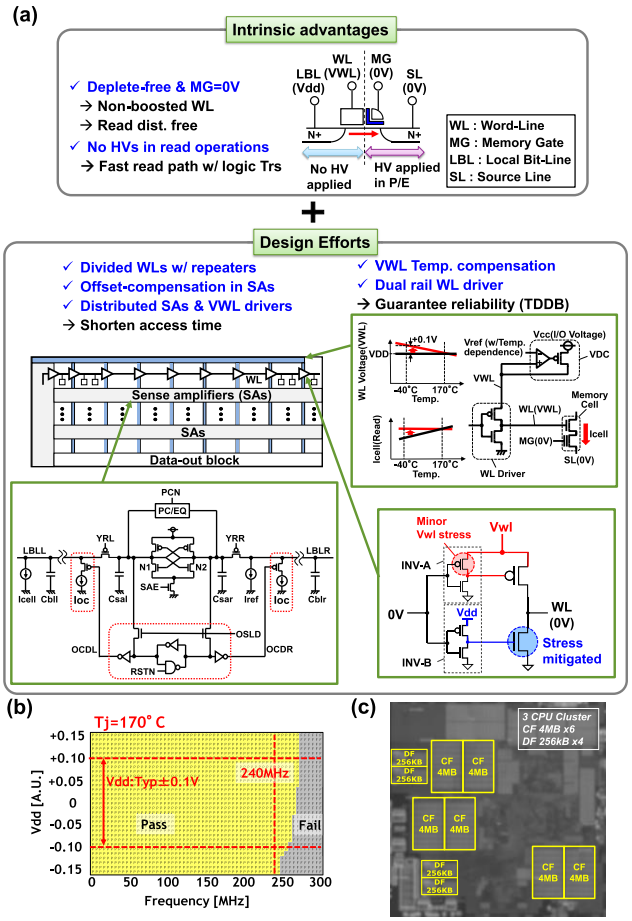


Fig. 7 SG-MONOS read path design [9]–[11]: (a) design efforts in combination with intrinsic advantages of SG-MONOS memory cells, (b) read shmoo plots of 28nm SG-MONOS code flash macro and (c) die photo of 28nm Flash MCU with SG-MONOS flash macros.

voltage (VWL) drivers [10] effectively shorten the random read access time. VWL temperature compensation [11] and dual rail WL driver can mitigate the voltage stress to the gate oxide of WL drivers/repeaters and guarantee time dependent dielectric breakdown (TDDB) lifetime. As shown in Fig. 7 (b) and (c), in 28nm SG-MONOS code macros, 240MHz random read access with less than 1.5ns WL rise time is achieved at junction temperature (Tj) of −40°C ~ 170°C [10].

2.3 Adaptive Program/Erase (P/E) Control

As summarized in Fig. 8 (a), eFlash macro design has faced some challenges for P/E operations along with process scaling and additional requirements from new applications. One of the challenges is the degradation of cell performance and reliability. Intelligent erase scheme (IES) was introduced in 40nm SG-MONOS [9], which focuses on the strong dependence of cell Vth shift on erase current (Iers) and source-line voltage (VSL). Dedicated sequence controller monitors VSL level and dynamically optimizes the erase pulse width in each step according to the results. Erase pulse time is

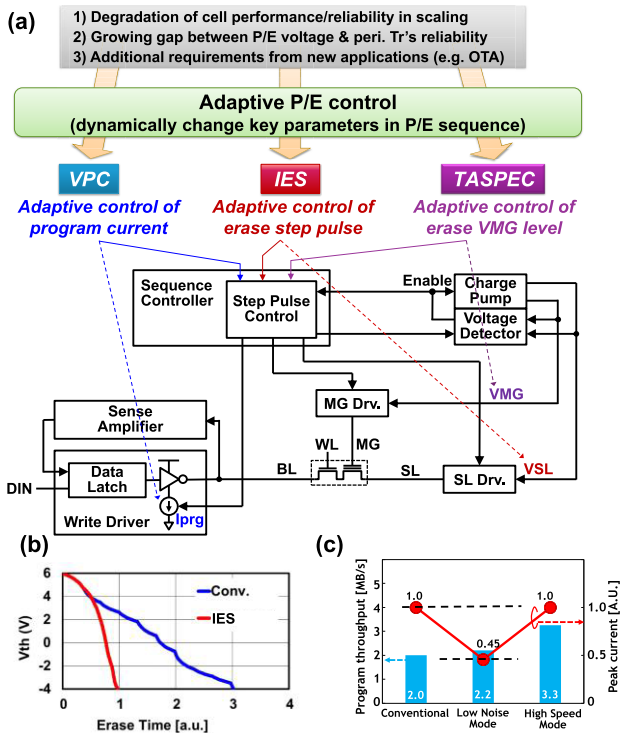


Fig. 8 Design efforts for SG-MONOS P/E operations [9]–[11]: (a) design implementation of adaptive P/E control schemes, (b) adaptive reduction of erase time by IES and (c) peak current reduction by VPC.

reduced by 60% at Tj = 170°C as compared to the conventional scheme (Fig. 8 (b)).

Another challenges in process scaling is the growing gap between cell P/E voltages and peripheral devices' reliability. For example, TDDB lifetime of inter-/intra-metal layer film for high voltage interconnections (MG and SL) gets worse due to narrower metal pitch and non-scaled P/E voltages. Given the different temperature dependency of TDDB (worst at high temperature) and erase speed of memory cells (slowest at low temperature), a temperature-adaptive step pulse erase control (TASPEC) was developed in 28nm SG-MONOS [11]. MG voltage and pulse width are automatically tuned in each step of erase sequence by monitoring equivalent cell Vth shift. As a result, maximum MG voltages at high temperature are relaxed and TDDB lifetime can be improved by a factor of 3 without erase speed degradation at low temperature.

Emerging applications sometimes pose additional challenges on eFlash design. For example, in-field software update by OTA in connected car era requires more careful eFlash design to reduce noise caused by charge pump operations from the viewpoints of EMI (electromagnetic interference) and power integrity. Variable program current (VPC) scheme [10] was proposed to reduce power line noise during program operations. Program current and the number of simultaneously programmed cells are flexibly changed halfway in the sequence and the total current consumption can be suppressed. Measurement results show that the peak current in VPC scheme reduces by 55% without program

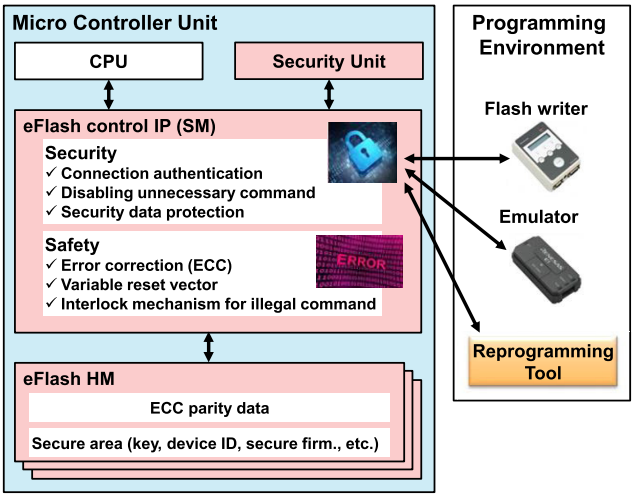


Fig. 9 Security and safety functions for eFlash sub-system

throughput degradation (Fig. 8 (c)).

As seen above, adaptive P/E control techniques effectively solve the difficulties in advanced eFlash development and help to keep or enhance P/E performance within the reliability limitation of memory cells, peripheral devices and metal interconnection.

2.4 Security and Safety Functions for eFlash Sub-System

As the services based on two innovative concepts of "CPS" and "IoT" have been widely prevailing in every corner of the world, security risk (e.g. unauthorized access to and illegal manipulation of important personal/enterprise data) and safety concern (e.g. human and social damage due to occasional but serious malfunctions in autonomous driving or infrastructures) are substantially growing and foster the momentum to prepare the standards for security (e.g. AUTOSAR and ISO21434 for automotive) and safety (e.g. ISO26262 for automotive and IEC61508 for industrial). In this context, proper implementation of security and safety functions in accordance with those standards is one of the key efforts in MCU/SoC products [4], [12], [13] as well as eFlash design for secure data storage and access.

As shown in Fig. 9, eFlash sub-system consists of multiple eFlash hard macros (HMs) and dedicated eFlash soft macro (SM). Especially, SM plays important roles in HM-SM co-design of security and safety functions and brings functional flexibility and diversity to eFlash sub-system.

SECURITY: The main role of eFlash control SM in terms of security is to protect program and parameters stored in eFlash HMs against illegal access or operations in cooperation with security unit which is responsible for the whole security in an MCU. As a "gate keeper" to eFlash HMs, eFlash SM has such specific security functions as connection authentication by ID and detection/invalidation of illegal command. In addition, security unit needs secure firmware, root/user keys, unique device ID, monotonic counter value for rollback protection, etc. Given the impor-

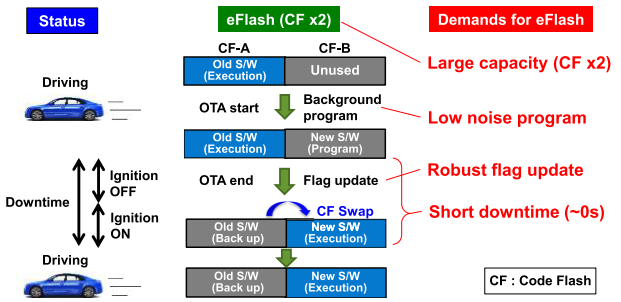


Fig. 10 Background of OTA: challenges and solutions

tance of these information from the viewpoints of chip level security, they are stored in the dedicated areas of eFlash HMs or in the dedicated eFlash HM, which is accessible only by security unit. SG-MONOS based physical unclonable function (PUF) was proposed for implementation of unique device ID [14].

SAFETY: Safety functions in eFlash sub-system is meant to minimize the risks in case of accidental failures (e.g. soft errors or random failures in product lifetime) and improper usage. Error correction with parity bits is one of the most typical examples of safety functions for eFlash. Besides, variable reset vector for secure program update and interlock mechanism against improper commands can be implemented according to the requirements from users and applications.

2.5 Realization of Robust OTA Software (SW) Update for Automotive

Along with the prevalence of connected car technology, OTA SW update is expected to be more widely introduced in cars and offer the benefits of easy function upgrade and time-saving in maintenance to both car owners and dealers. Figure 10 describes Code Flash (CF) swap technique [15] in which new SW is downloaded during driving and old one can be reactivated in case that new SW contains fatal bugs. This accelerates the demand for larger eFlash capacity, because both old and new SW are stored in eFlash at the same time. In addition, low-noise programming (see VPC in Sect. 2.3), short downtime and robust flag update to switch old SW to new one are the key factors for convenient and secure OTA SW update.

Figure 11 (a) shows a block diagram of the MCU and its eFlash sub-system with robust and fast switching (RFS) mechanism [10]. Each CPU cluster is connected to two CFs, CFx-A and CFx-B (x = 0-2), each of which stores either of new or old SW. Conventional CF swap technique decides the valid SW simply by two swap indicators. If either of them is corrupted by abnormal termination during update, MCU may not boot correctly. In contrast, RFS mechanism uses multiple data structures in Data Flash (DF), SWT0/1 and FLG, to select the valid set of option bytes (OPBT-A or B) associated with CFx-A or CFx-B. During the flag update sequence, at least 2 of 3 data sets (SWT0/1 and FLG) always

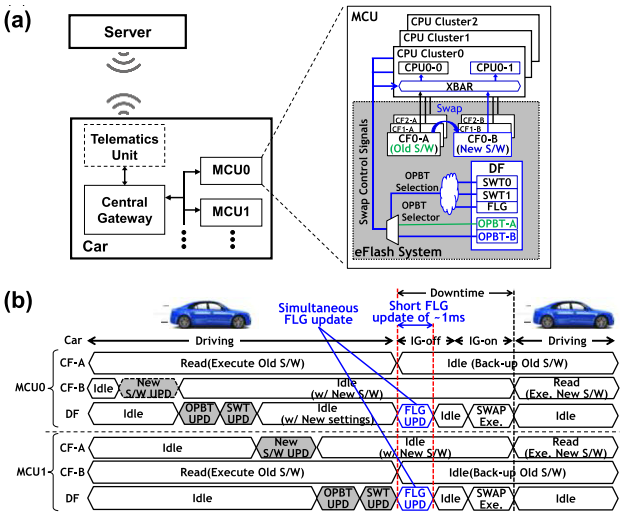


Fig. 11 RFS mechanism for OTA in automotive [10]: (a) system block diagram for OTA and (b) SW update sequence with RFS mechanism.

hold valid values. Therefore, even if unintentional termination occurs during update of OPBT-A or B, SWT0/1 or FLG, the correct swap settings are transferred from OPBT-A or B to CPUs, and MCU can boot correctly.

The timing diagram in Fig. 11 (b) shows the OTA SW update operation with RFS mechanism. Each MCU in a car executes old SW in one CF for driving while new SW and OPBT settings are downloaded from the server to the other CF and DF. After all download is completed and ignition switch turns off (IG-off), the central gateway multicasts FLG update command to all MCUs. FLG update time is dominated by erase. Thanks to fast erase characteristics of SG-MONOS cells by band-to-band tunneling, FLG update takes only ~1ms and can be completed within IG-off period. All MCUs can boot with new SW and OPBT settings even when the switch turns on (IG-on) immediately after IG-off.

3. Challenges for eFlash in Extending Roles to Big Data and Nomadic/Autonomous Elements of IoT

In the upcoming data-centric world, huge amount of data will be sensed, collected, judged and processed autonomously at "endpoint" where natural phenomena and our daily social/financial activities occur in real time. This is a great chance for semiconductor industries to widen their roles. However, at the same time, they are facing unprecedented requirements of performance and energy-efficiency, which demand "disruptive" end-point computing technologies. In this section, we pick up two technologies; PIM for e-AI and extremely low energy solution. eFlash is expected to play key roles here. In addition, the possibility and challenges of embedded emerging memories (eEMs) and eFlash in the future are briefly discussed.

3.1 Processing-in-Memory for E-AI

Recently, AI technology has been making a dramatic

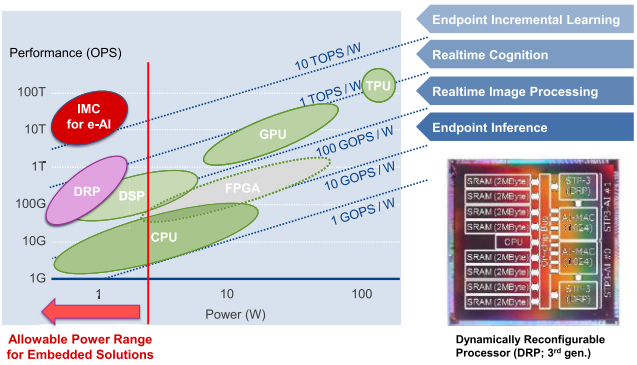
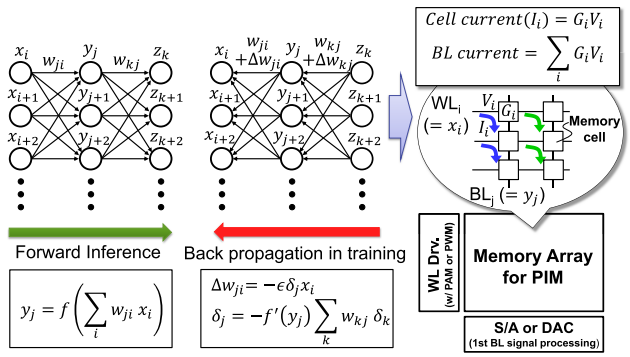


Fig. 12 Performance of processing units and AI-accelerators



	Memory cell type of PIM		
	SRAM	Flash	PCM/ReRAM
Weight/cell	Binary	Multi- or Analog level	Multi- or Analog level
Macro size	Large (x1)	Small (x1/2~1/5)	Small (x1/2~1/5)
Performance (TOPS/W@inference)	10~100	10~100	10~100
Wake-up time	Long (need data load)	Short	Short
Reliability (as NVM)	---	Good	Poor
Process maturity/ availability	Good	Good	Poor

Fig. 13 PIM concept for efficient MAC/VMM operations and comparison table among PIM memory cell candidate.

progress and expanding its application areas. The benefit of e-AI is also proved widely [16]. Figure 12 shows the performance trend of neural network (NN) accelerator candidates and the application requirements. Given the necessity to make on-site judgments and the specific constraints of low power and small form factor, power-efficient and compact implementation of NN models for inference and local optimization is very critical. From this viewpoint, dynamic reconfigurable processor (DRP) [17], a programmable HW core with an array of coarse-grained processing elements and dynamic HW reconfiguration, is the suitable accelerator for e-AI. As another approach, PIM or in-memory computing (IMC) has been getting more attention because of its high power-efficiency in handling NN models.

Figure 13 describes the concept of PIM and its implementation. Vector-by-matrix multiplication (VMM) or multiply-and-accumulation (MAC) is the key operation through each data propagation in NN. This operation is feasible in a memory array based on proper cell Vth or resistance setting (= weight), multiple WL activation (= simultaneous weight multiplication) and summation of cell current on a BL (= summation of all weighed signals). Many ideas of PIMs have been proposed by using SRAM cells [18]–[20] or non-volatile memory (NVM) cells. Especially, NVM based PIMs with such as flash [21], [22], ReRAM [23], [24] or PCM [25], are expected to achieve extremely higher power efficiency thanks to massively parallel execution with multiple weights per cell and no data transfer from outside of the macro to load weigh information. Peripheral circuit design of PIM macros can be quite different from that of normal macros for conventional data storage. In case of multi-level or analog weights, WL driver needs an additional implementation of digital-to-analog converting (DAC) function for WL pulse width or amplitude modulation. To distinguish multiple levels on BLs, sense amplifiers (SAs) with multiple reference levels or analog-to-digital converter (ADC) based BL processing circuits are newly required in place of conventional SAs. These design changes can lead to much larger peripheral areas. Therefore, reasonable balance between performance and cost should be taken into account at the initial stage of PIM macro design.

The PIM performance by memory cell types are summarized in the table in Fig. 13. Cell size and weight/cell are the dominant factors for macro size. For e-AI, wake-up time is also the important factor in device selection for quick response after power-on and short active time in intermittent operation. Given these reasons, NVM based PIMs with wide dynamic range for multiple weights/cell are more suitable in many e-AI applications and have been intensively proposed. Flash based PIMs with multi-level cell capability [21], [22] are also getting attention and some of them utilize commercially available eFlash memory cells with minor revision of node connections to improve cell Vth controllability. In addition to the wide dynamic range suitable for multiple weights/cell or as "analog memory", process maturity and availability in existing nodes is also their advantage over other candidates.

Although NVM based PIMs are attractive, there are some issues to be tackled. The current behavior of these memory cells is not ideal in terms of linearity and variation (e.g. cycle-to-cycle, temperature dependence). In addition, the macro design encounters the trade-off in realizing multi-level weight capability and sacrificing accuracy and macro size.

3.2 Extremely Low Energy Solution with SOTB (Silicon on Thin Buried Oxide) eFlash and Intermittent Operations

Battery-free long-time operation without maintenance of battery recharge or change is one of the most crucial and attractive characteristics for semiconductor devices used at "end-point", the very front line in the real world. To meet

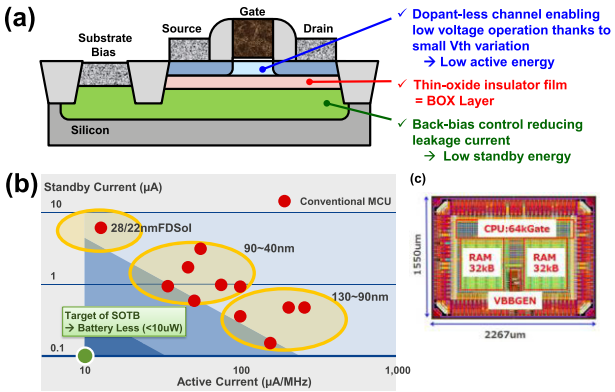


Fig. 14 Advantages and target performance of SOTB: (a) cross-sectional view, (b) comparison of active/standby current with conventional technologies and (c) SOTB MCU test chip [27].

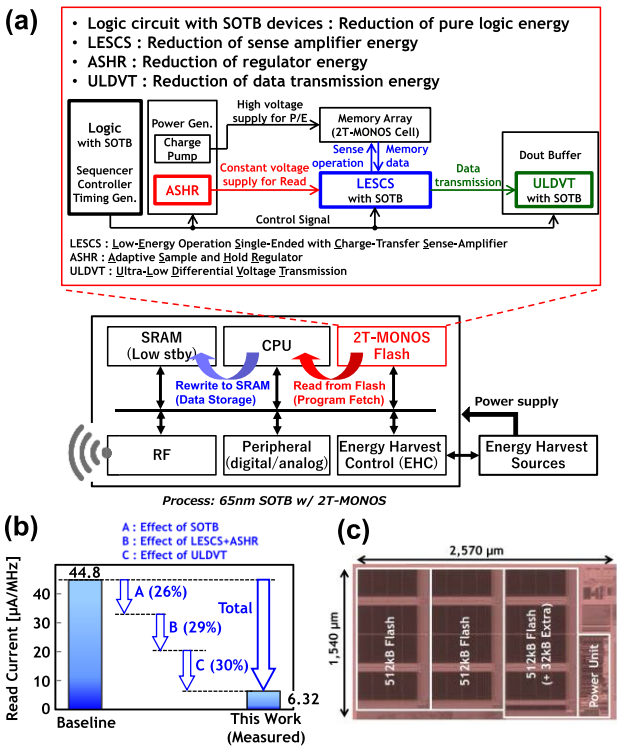


Fig. 15 Design and achievements of 2T-MONOS eFlash macro [29]: (a) configuration of SOTB-based MCU chip and 2T-MONOS flash macro, (b) read current reduction by design efforts in 2T-MONOS flash macro and (c) photo of 2T-MONOS flash macro.

this requirement, proper combination of process and design approach is of extreme importance.

For breaking through the tight trade-off between active (operation) and standby (leakage) current of transistors to reduce total current consumption, SOI (Silicon on Insulator) based process technologies are considered as good candidates. SOTB process [26], [27] is one of the most promising among them and actual shipment of products developed with SOTB process has been already started [28]. As summarized in Fig. 14 (a), SOTB process is characterized by 1) a thin insulating oxide film (BOX, buried oxide) deposited on top of the base silicon substrate and 2) a very thin silicon layer with no impurities deposited on the BOX insulator layer. These features enable to form dopant-less channel transistors with less threshold voltage (Vth) variation, resulting in low voltage operations and very small operation/leakage current. Performance-leakage balance is flexibly optimized at wide range by applying back bias voltage (forward or reverse) to the base silicon substrate underneath BOX layer. In addition, since thin BOX layer is easily removable, conventional bulk transistors for high voltage operations in I/O ports, analog IPs and eFlash can be formed in the same die with SOTB transistors. Figure 14 (b) shows the advantages of 65nm SOTB process in terms of active-standby current trade-off relationship. 65nm SOTB can realize approximately 1/10x smaller active and standby current consumption than conventional low-power processes without sacrificing performance.

Regarding design approach, efficient energy/power generation and adequate reduction of energy consumption should be taken into consideration at the same time. As an alternative of conventional battery, more attentions have been attracted to energy harvesting techniques which generate power from solar/lighting, mechanical vibration or heat, though the amount of energy gained by them is too small. Intermittent operation with periodical status shift between active and standby can substantially reduce energy consumption in some applications, especially those in which much longer standby period is acceptable than active period. One of the preferable examples in the combination of energy harvesting and intermittent operation is real-time sensing application, whose sensing interval (Tsi) is generally in the order of milliseconds (ms). Here, MCU manages data collection, analysis and judgement (data cleansing) and stores the results in eFlash in it (plus, occasionally, sends processed data to cloud via RF/BLE). However, long recharge period by energy harvesting may be needed to compensate energy loss due to read and rewrite operations to eFlash. To prevent the case that the recharge period is longer than Tsi, additional design effort for eFlash is needed.

Figure 15 (a) presents the MCU developed based on the combination of SOTB process, intermittent operation and 2T-MONOS eFlash with intrinsic advantages of low energy operations and less mask adder [29]. SOTB SRAM with very low standby power [30] is utilized as a write buffer to store analyzed data in place of eFlash, removing the energy consumption by write operations to eFlash. Then, the remaining dominant factor in energy consumption is program fetch from eFlash, that is, read operation of eFlash. By introducing energy-efficient S/A, regulator and data transmission design techniques in combination of SOTB intrinsic virtues such as low voltage operation and less Vth variation, total eFlash read current is reduced by 85% and read energy of 6.32μA/MHz (equivalent to read energy of 0.22pJ/bit) is achieved in 512KB macro (Fig. 15 (b) and (c)). This value is sufficiently low to meet the Tsi target even in use of energy

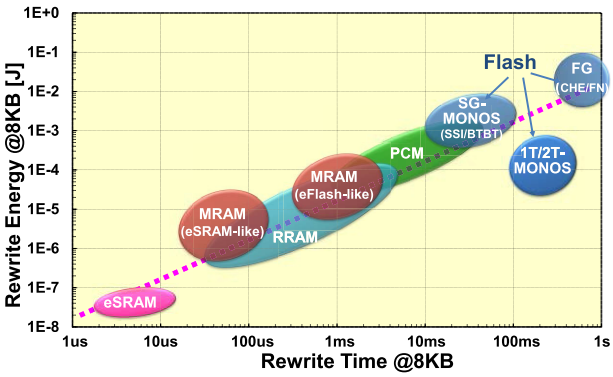


Fig. 16 Rewrite time and energy; eFlash vs eEM candidates

harvesting as a power source.

3.3 Perspective of ENVM Candidates in the Future; What'll Come After eFlash?

Originally, eEMs (e.g. perpendicular STT-MRAM, PCM and ReRAM) have been expected to bring new values in NVM applications thanks to much less rewrite energy and better endurance. Less mask adder to baseline CMOS logic process is also attractive.

Figure 16 presents the relationship of rewrite energy and time in various eNVM types. In case of perpendicular STT-MRAM, cell current for write operation is strongly correlated with data retention capability in principle. This is why eFlash-like MRAM with data retention long enough for MCU applications has larger rewrite energy than embedded SRAM (eSRAM)-like MRAM with much shorter data retention time. Basically, the group of conventional eFlash shows longer rewrite time with larger rewrite energy than eEMs, but MONOS eFlash by FN tunneling program/erase can achieve as low rewrite energy as some of eEMs and high reliability at high temperature for automotive [31]. Lowering the P/E voltages is the key to reduce eFlash's rewrite energy. Some countermeasures for lower P/E voltages have been investigated (e.g. new film material, composition or stack in storage nodes) [32].

The main target of eEMs for near-term applications is to replace the existing embedded memories, eFlash and eSRAM. In case of eFlash replacement, one of the key requirements is data retention capability at high temperature and solder reflow conditions (e.g. 260°C, 3min.). This is quite a big challenge in common to many types of eEMs, because intrinsic temperature dependence of data retention performance is large by large activation energy at the cell, and data retention is inseparably linked with cell current (that is, rewrite power and cell scalability) and endurance. R&Ds to realize high data retention capability in reasonable balance with other trade-off factors have been intensively conducted for eFlash replacement [33]–[36]. Regarding eSRAM replacement, key requirements are 1) higher endurance (e.g. > 1E11~1E12) and 2) faster and balanced read-write speed (e.g. > 100MHz no-wait read and write). Fundamental quality improvement of critical films such as magnetic tunnel junction film in perpendicular STT-MRAM or metal oxide in ReRAM is under way. In common with both applications, small magnetoresistance ratio and read disturb in STT-MRAM, large resistance drift in PCM and random telegraph noise/noise-induced resistance broadening in ReRAM pose critical challenges for sense amplifier and reference circuit design.

After overcoming the above-mentioned issues and proving the maturity for mass production with cost competitiveness, eEMs will grab a big chance especially in advanced end-point consumer applications.

4. Conclusion: Embedded System Roadmap

The basic nature of edge computing is identified in IoT data supply chains to visualize how embedded systems contribute to implement IoT concepts. Based on multi-faceted development of solutions by MCU products, we conclude:
1) Embedded systems are expected to implement IoT elements in CPS, Big Data, and Nomadic/Autonomous, where data supply chain innovations are key driving forces.
2) IoT data supply models provide edge computing rationales in:
- multiple/hetero sensing & feedback control in CPS
- end-to-end connections with balanced local computing and communication with security and by e-AI inference for IoT Big Data schemes.
- re-configurable, upgradable systems augmented by OTA and on/off-line operation by energy harvesting to support distributed autonomous systems.
3) Flash-MCU innovation and on-chip SW-HW co-designs have proliferated by alterable sub-systems, now aiming at adaptive-by-learning by extending on-chip storage usages.
4) Embedded flash memory technology has evolved quite uniquely according to embedded system requirements and will continue to excel in wide temperature range operations because of small activation energy in memory cell physics, compared with emerging non-volatile memories. Multi-level storage and processing in flash memory array for neural-network MAC acceleration is a unique feature provided only by flash memory today, to be exploited for expanding low-power e-AI inference applications.
5) Ultra-low power for EH is pursued by using SOI devices with low-power circuits and sub-system control, to provide energy for accessing and re-writing memory cell data comparable to some of emerging non-volatile memories.

In view of expanding edge computing requirements, the roadmap for embedded system technology is to be established by hierarchy where each new element technology such as security, e-AI, IoT communication etc. is supported by large-scale science based on autonomous approaches and solid SW/HW platforms (Fig. 17), where main points are:
1) Multi-layered technology advancements in each application segment
2) Evolutions in base platforms and system design for local control/computing/communication in balance

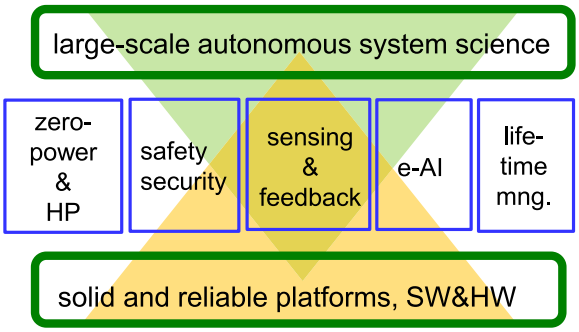


Fig. 17 Proposed organization of embedded micro-system roadmap

3) How edge intelligence will foster autonomy into coming "society of things", a way of realizing IoT/AI society.
By such well-coordinated advancement embedded micro-systems will promote coming IoT implementations in the Society 5.0.

Acknowledgments

The authors are thankful to technical efforts and supports by MCU and embedded flash memory development teams at Renesas Electronics Corp.

References

[1] S. Mochizuki, K. Matsubara, K. Matsumoto, C.L.P. Nguyen, T. Shibayama, K. Iwata, K. Mizumoto, T. Irita, H. Hara, and T. Hattori, "A 197mW 70ms-latency full-HD 12-channel video-processing SoC for car information systems," IEEE Int. Solid-State Circuits Conference (ISSCC) Dig. Tech. Papers, pp.78–79, Feb. 2016.
[2] H. Hidaka, "Chapter 2: Applications and Technology Trend in Embedded Flash Memory," in Embedded Flash Memory for Embedded Systems: Technology, Design for Sub-systems, and Innovations, ed. H. Hidaka, Springer, 2017.
[3] H. Kimura, H. Noda, H. Watanabe, T. Higuchi, R. Kobayashi, M. Utsuno, F. Takami, S. Otani, M. Ito, Y. Shimazaki, N. Yada, and H. Kondo, "A 40nm flash microcontroller with 0.80μs field-oriented-control intelligent motor timer and functional safety system for next-generation EV/HEV," IEEE Int. Solid-State Circuits Conference (ISSCC) Dig. Tech. Papers, pp.58–59, Feb. 2017.
[4] S. Otani, N. Otsuki, Y. Suzuki, N. Okumura, S. Maeda, T. Yanagita, T. Koike, Y. Shimazaki, M. Ito, M. Uemura, T. Hattori, T. Yamauchi, and H. Kondo, "A 28nm 600MHz Automotive Flash Microcontroller with Virtualization-Assisted Processor for Next-Generation Automotive Architecture Complying with ISO26262 ASIL-D," IEEE Int. Solid-State Circuits Conference (ISSCC) Dig. Tech. Papers, pp.54–56, Feb. 2019.
[5] T. Kono, T. Saito, and T. Yamauchi, "Chapter 3: Overview of Embedded Flash Memory Technology," in Embedded Flash Memory for Embedded Systems: Technology, Design for Sub-systems, and Innovations, ed. H. Hidaka, Springer, 2017.
[6] Y. Kawashima, T. Hashimoto, and I. Yamakawa, "Investigation of the data retention mechanism and modeling for the high reliability embedded split-gate MONOS flash memory," International Reliability Physics Symposium (IRPS), pp.MY.6.1–6.5, April 2015.
[7] S. Tsuda, T. Saito, H. Nagase, Y. Kawashima, A. Yoshitomi, S. Okanishi, T. Hayashi, T. Maruyama, M. Inoue, S. Muranaka, S. Kato, T. Hagiwara, H. Saito, T. Yamaguchi, M. Kadoshima, T. Maruyama, T. Mihara, H. Yanagita, K. Sonoda, T. Yamashita, and Y. Yamaguchi, "Reliability and scalability of FinFET split-gate MONOS array with tight Vth distribution for 16/14nm-node embedded flash," IEEE Int. Electron Devices Meeting (IEDM) Dig. Tech. Papers, pp.469–472, Dec. 2017.
[8] T. Kono, "Embedded non-volatile memory system as an enabler of smarter world," International Symposium on VLSI Design, Automation and Test (VLSI-DAT), April 2017.
[9] T. Kono, T. Ito, T. Tsuruda, T. Nishiyama, T. Nagasawa, T. Ogawa, Y. Kawashima, H. Hidaka, and T. Yamauchi, "40nm Embedded Split-gate MONOS (SG-MONOS) Flash Macros for Automotive with 160MHz Random Access for Code and Endurance over 10M Cycles for Data at the Junction Temperature of 170 °C," IEEE J. Solid-State Circuits, vol.49, no.1, pp.154–166, Jan. 2014.
[10] A. Kanda, T. Kurafuji, K. Takeda, T. Ogawa, Y. Taito, K. Yoshihara, M. Nakano, T. Ito, H. Kondo, and T. Kono, "A 24MB Embedded Flash System Based on 28nm SG-MONOS Featuring 240MHz Read Operations and Robust Over-The-Air Software Update for Automotive," 2019 Symposium on VLSI Circuits Dig. Tech. Papers, pp.C210–211, June 2019.
[11] Y. Taito, T. Kono, M. Nakano, T. Saito, T. Ito, K. Noguchi, H. Hidaka, and T. Yamauchi, "A 28 nm Embedded Split-Gate MONOS (SG-MONOS) Flash Macro for Automotive Achieving 6.4 GB/s Read Throughput by 200 MHz No-Wait Read Operation and 2.0 MB/s Write Throughput at Tj of 170°C," IEEE J. Solid-State Circuits, vol.51, no.1, pp.213–221, Jan. 2016.
[12] S. Shibahara, C. Takahashi, K. Fukuoka, Y. Kitaji, T. Irita, H. Hara, Y. Shimazaki, and J. Matsushima, "A 16 nm FinFET Heterogeneous Nona-Core SoC Supporting ISO26262 ASIL B Standard," IEEE J. Solid-State Circuits, vol.52, no.1, pp.77–88, 2017.
[13] C. Takahashi, S. Shibahara, K. Fukuoka, J. Matsushima, Y. Kitaji, Y. Shimazaki, H. Hara, and T. Irita, "A 16nm FinFET heterogeneous nona-core SoC complying with ISO26262 ASIL-B: Achieving 10^-7 random hardware failures per hour reliability," IEEE Int. Solid-State Circuits Conference (ISSCC) Dig. Tech. Papers, pp.80–81, Feb. 2016.
[14] T. Saito, H. Nagase, M. Izuna, T. Shimoi, A. Kanda, T. Ito, and T. Kono, "High-Temperature Stable Physical Unclonable Functions with Error-Free Readout Scheme Based on 28nm SG-MONOS Flash Memory for Security Applications," IEEE International Memory Workshop (IMW), pp.127–130, May 2017.
[15] M. Lobdell, "Robust over-the-air firmware updates using program flash memory swap on kinetis microcontrollers," Freescale Application Note, p.AN4533, 2012.
[16] https://www.renesas.com/us/en/solutions/key-technology/e-ai/business.html
[17] T. Fujii, T. Toi, T. Tanaka, K. Togawa, T. Kitaoka, K. Nishino, N. Nakamura, H. Nakahara, and M. Motomura, "New Generation Dynamically Reconfigurable Processor Technology for Accelerating Embedded AI Applications," 2018 Symposium on VLSI Circuit Dig. Tech. Papers, pp.41–42, 2018.
[18] M. Kang, S.K. Gonugondla, A. Patil, and N.R. Shanbhag, "A Multi-Functional In-Memory Inference Processor Using a Standard 6T SRAM Array," IEEE J. Solid-State Circuits, vol.53, no.2, pp.642–655, Feb. 2018.
[19] J. Zhang, Z. Wang, and N. Verma, "A Machine-learning Classifier Implemented in a Standard 6T SRAM Array," 2016 Symposium on VLSI Circuits Dig. Tech. Papers, pp.252–253, June 2016.
[20] S. Okumura, M. Yabuuchi, K. Hijioka, and K. Nose, "A Ternary Based Bit Scalable, 8.80 TOPS/W CNN accelerator with Many-core Processing-in-memory Architecture with 896K synapses/mm2," 2019 Symposium on VLSI Circuit Dig. Tech. Papers, pp.C248–249, 2019.
[21] F.M. Bayat, X. Guo, M. Klachko, M. Prezioso, K.K. Likharev, and D.B. Strukov, "High-Performance Mixed-Sinal Neurocomputing With Nanoscale Floating-Gate Memory Cell Arrays," IEEE Trans. Neural Netw. Learn. Syst., vol.29, no.10, pp.4782–4790, Oct. 2018.
[22] L. Fick, D. Blaauw, D. Sylvester, S. Skrzyniarz, M. Parikh, and

D. Fick, "Analog In-Memory Subthreshold Deep Neural Network Accelerator," 2017 IEEE Custom Integrated Circuits Conference (CICC), May 2017.

[23] W.-H. Chen, K.-X. Li, W.-Y. Lin, K.-H. Hsu, P.-Y. Li, C.-H. Yang, C.-X. Xue, E.-Y. Yang, Y.-K. Chen, Y.-S. Chang, T.-H. Hsu, Y.-C. King, C.-J. Lin, R.-S. Liu, C.-C. Hsieh, K.-T. Tang, and M.-F. Chang, "A 65nm 1Mb Nonvolatile Computing-in-Memory ReRAM Macro with Sub-16ns Multiply-and-Accumulate for Binary DNN AI Edge Processors," IEEE Int. Solid-State Circuits Conference (ISSCC) Dig. Tech. Papers, pp.494–495, Feb. 2018.

[24] R. Mochida, K. Kouno, Y. Hayata, M. Nakayama, T. Ono, H. Suwa, R. Yasuhara, K. Katayama, T. Mikawa, and Y. Gohou, "A 4M Synapses integrated Analog ReRAM based 66.5 TOPS/W Neural-Network Processor with Cell Current Controlled Writing and Flexible Network Architecture," 2018 Symposium on VLSI Technology Dig. Tech. Papers, pp.175–176, June 2018.

[25] G.W. Burr, R.M. Shelby, C. di Nolfo, J.W. Jang, R.S. Shenoy, P. Narayanan, K. Virwani, E.U. Giacometti, B. Kurdi, and H. Hwang, "Experimental demonstration and tolerancing of a large-scale neural network (165,000 synapses), using phase-change memory as the synaptic weight element," IEEE International Electron Devices Meeting (IEDM), pp.697–700, Dec. 2014.

[26] Y. Yamamoto, H. Makiyama, T. Tsunomura, T. Iwamatsu, H. Oda, N. Sugii, Y. Yamaguchi, T. Mizutani, and T. Hiramoto, "Poly/High-k/SiON Gate Stack and Novel Profile Engineering Dedicated for Ultralow-Voltage Silicon-on-Thin-BOX (SOTB) CMOS Operation," 2012 Symposium on VLSI Technology Digest of Technical Papers, pp.109–110, June 2012.

[27] T. Hasegawa, Y. Yamamoto, H. Makiyama, H. Shinkawata, S. Kamohara, and Y. Yamaguchi, "SOTB (Silicon on Thin Buried Oxide): More than Moore technology for IoT and Automotive," 2017 IEEE International Conference on IC Design and Technology (ICICDT), pp.1–4, May 2017.

[28] https://www.renesas.com/us/en/solutions/key-technology/sotb/products.html

[29] K. Matsubara, T. Nagasawa, Y. Kaneda, H. Mitani, H. Sato, T. Iwase, Y. Aoki, K. Maekawa, H. Yamakoshi, T. Ito, H. Kondo, and T. Kono, "A 65nm Silicon-on-Thin-Box (SOTB) Embedded 2T-MONOS Flash Achieving 0.22 pJ/bit Read Energy with 64 MHz Access for IoT Applications," 2019 Symposium on VLSI Circuit Dig. Tech. Papers, pp.C202–203, June 2019.

[30] M. Yabuuchi, K. Nii, S. Tanaka, Y. Shinozaki, Y. Yamamoto, T. Hasegawa, H. Shinkawata, and S. Kamohara, "A 65 nm 1.0 V 1.84 ns Silicon-on-Thin-Box (SOTB) Embedded SRAM with 13.72 nW/Mbit Standby Power for Smart IoT," 2017 Symposium on VLSI Circuits Digest of Technical Papers, pp.C220–221, June 2017.

[31] H. Mitani, K. Matsubara, H. Yoshida, T. Hashimoto, H. Yamakoshi, S. Abe, T. Kono, Y. Taito, T. Ito, T. Kurafuji, K. Noguchi, H. Hidaka, and T. Yamauchi, "A 90nm Embedded 1T-MONOS Flash Macro for Automotive Applications with 0.07mJ/8kB Rewrite Energy and Endurance Over 100M Cycles Under Tj of 175°C," IEEE Int. Solid-State Circuits Conference (ISSCC) Dig. Tech. Papers, pp.140–141, Feb. 2016.

[32] C. Zhao, C.Z. Zhao, S. Taylor, and P.R. Chalker, "Review on Non-Volatile Memory with High-k Dielectrics: Flash for Generation Beyond 32 nm," Materials, vol.7, no.7, pp.5118–5145, 2014.

[33] K. Lee, R. Chao, K. Yamane, V.B. Naik, H. Yang, J. Kwon, N.L. Chung, S.H. Jang, B. Behin-Aein, J.H. Lim, S. K, B. Liu, E.H. Toh, K.W. Gan, D. Zeng, N. Thiyagarajah, L.C. Goh, T. Ling, J.W. Ting, J. Hwang, L. Zhang, R. Low, R. Krishnan, L. Zhang, S.L. Tan, Y.S. You, C.S. Seet, H. Cong, J. Wong, S.T. Woo, E. Quek, and S.Y. Siah, "22-nm FD-SOI Embedded MRAM Technology for Low-Power Automotive-Grade-1 MCU Applications," IEEE Int. Electron Devices Meeting (IEDM) Dig. Tech. Papers, pp.604–607, Dec. 2018.

[34] O. Golonzka, J.-G. Alzate, U. Arslan, M. Bohr, P. Bai, J. Brockman, B. Buford, C. Connor, N. Das, B. Doyle, T. Ghani, F. Hamzaoglu, P. Heil, P. Hentges, R. Jahan, D. Kencke, B. Lin, M. Lu, M. Mainuddin, M. Meterelliyoz, P. Nguyen, D. Nikonov, K. O'brien, J.O. Donnell, K. Oguz, D. Ouellette, J. Park, J. Pellegren, C. Puls, P. Quintero, T. Rahman, A. Romang, M. Sekhar, A. Selarka, M. Seth, A.J. Smith, A.K. Smith, L. Wei, C. Wiegand, Z. Zhang, and K. Fischer, "MRAM as Embedded Non-Volatile Memory Solution for 22FFL FinFET Technology," IEEE Int. Electron Devices Meeting (IEDM) Dig. Tech. Papers, pp.412–415, Dec. 2018.

[35] Y.J. Song, J.H. Lee, S.H. Han, H.C. Shin, K.H. Lee, K. Suh, D.E. Jeong, G.H. Koh, S.C. Oh, J.H. Park, S.O. Park, B.J. Bae, O.I. Kwon, K.H. Hwang, B.Y. Seo, Y.K. Lee, S.H. Hwang, D.S. Lee, Y. Ji, K.C. Park, G.T. Jeong, H.S. Hong, K.P. Lee, H.K. Kang, and E.S. Jung, "Demonstration of Highly Manufacturable STT-MRAM Embedded in 28nm Logic," IEEE Int. Electron Devices Meeting (IEDM) Dig. Tech. Papers, pp.416–419, Dec. 2018.

[36] M.-C. Shih, C.-Y. Wang, Y.-H. Lee, W. Wang, L. Thomas, H. Liu, J. Zhu, Y.-J. Lee, G. Jan, Y.-J. Wang, T. Zhong, T. Torng, P.-K. Wang, D. Lin, T.-W. Chiang, K.-H. Shen, H. Chuang, and W.J. Gallagher, "Reliability study of perpendicular STT-MRAM as emerging embedded memory qualified for reflow soldering at 260°C," 2016 Symposium on VLSI Technology Dig. Tech. Papers, pp.144–145, June 2016.

Takashi Kono received the B.S. and M.S degrees in Electronic Engineering from University of Tokyo, Tokyo, Japan, in 1992 and 1994, respectively. He performed the research on nano-structure compound semiconductor devices (quantum wire & dot) from theoretical and experimental viewpoints in his graduate study. In 1994, he joined Mitsubishi Electric Corporation. From 1994 to 2002, he was engaged in the research and design of discrete DRAMs from 64M bit to 512M bit including synchronous DRAMs and DDR/DDR2 synchronous DRAMs in ULSI development laboratory, Mitsubishi Electric Corporation. After transferring to Renesas Technology Corporation (RT) in 2003, he worked on the development of low-power pseudo-SRAM and high-density AG-AND Flash memories. Since 2010, he had been responsible for the development of 90nm and 40nm split-gate (SG) MONOS Flash technologies and macros for automotive at Renesas Electronics Corporation. Currently, he belongs to Shared R&D Core Technology Division, Renesas Electronics Corporation, and is in charge of eNVM technology development from design to evaluation, qualification and mass production. In addition, he has been leading the joint projects with partners for developing state-of-the-art MCU platforms. From 2015 to 2018, he had been serving for IEEE International Solid-State Circuits Conference (ISSCC) as a member of memory subcommittee in international technical program committee (ITPC) and chaired memory sessions at ISSCC2015 and 2017. He has 43 United States Patents.

Yasuhiko Taito received the B.S. and M.S. degrees in applied physics from the University of Tokyo, Tokyo, Japan, in 1991 and 1993, respectively. In 1993, he joined Mitsubishi Electric Corporation and engaged in the design of NOR/DINOR flash memories and embedded DRAMs. After transferring to Renesas Technology Corporation in 2003, he worked on the development of embedded flash memories. Currently, he is with Renesas Electronics Corporation, Tokyo, Japan, where he has been engaged in the design and development of the most-advanced embedded flash memories for high-end MCUs and leading 28nm split-gate (SG) MONOS embedded Flash macros for automotive. From 2019, he had been serving for IEEE International Solid-State Circuits Conference (ISSCC) as a member of memory subcommittee in international technical program committee (ITPC). He has 37 United States Patents.

Hideto Hidaka has been engaged in the R&D of high-density and embedded DRAM, flash and non-volatile memory technologies for MCU applications, followed by technical strategy and corporate technical leadership, in Mitsubishi Electric, Renesas Technology and Renesas Electronics, Japan. He led a trend-setting development to create the world's first split-gate SONOS flash memory for MCU products in automotive applications for performance, power and reliability and to make a de-facto standard in MCU products, from 90nm through 28 nm generations, thus leading MCU technology/business trends. He was a Senior Vice-President and Chief Technology Officer after numerous technical and managerial positions, and now is a Fellow of Renesas Electronics Corp. He received the BS, MS, and Ph.D. degrees in electronic engineering from the University of Tokyo, Tokyo, Japan. He was a visiting scientist at the Media Laboratory, Massachusetts Institute of Technology, Cambridge, USA. He is an IEEE Fellow. His professional activity includes ITPC (program committee) Chair of ISSCC (2011–12), IEEE-SSCS Distinguished Lecturer (2015–2016), Associate Editor of IEEE J. Solid-State Circuits (2012–17), IEEE-SSCS AdCom member-elect (2012–2017), IEEE Fellow Committee member (2018–), Chair of IEICE-ICD (Technical Committee on Integrated Circuits and Devices, 2017–19), and technical program committee members for A-SSCC, VLSI-TSA, and ICICDT. He holds 293 United States patents.