INVITED SURVEY PAPER

Survey on Data Center Networking Technologies

Yoshiaki KIRIHA[†a)] *and* **Motoo NISHIHARA**[†], *Members*

SUMMARY In recent years, technologies and markets related to data centers have been rapidly changing and growing. Data centers are playing an important role in ICT infrastructure deployment and promise to become common platforms for almost all social infrastructures. Even though research has focused on networking technologies, various technologies are needed to develop high-performance, cost-efficient, and flexible large-scale data centers. To understand those technologies better, this paper surveys recent research and development efforts and results in accordance with a data center network taxonomy that the authors defined.

key words: *data center network, network virtualization, future Internet, switching & routing, traffic engineering*

1. Introduction

Information and Communication Technology (ICT) is recognized as an important social infrastructure and has been providing large amounts of useful and convenient services that improve people's quality of life. Thanks to the recent popularity of mobile and smart phones as well as the evolution of cloud computing and the Internet, anyone can interact with various ICT services anytime, anywhere. Data centers, which consist of large amounts of servers, storages, and switches, are playing an important role in this rapid growth of ICT related industry and are recognized as common "cloud" platforms for all social infrastructure services.

In accordance with many changes, such as the number of users, the volume of analyzed/processed data, and the complexity of provided service logic, the role and configuration of data centers have changed drastically. In general, data centers were formerly systems dedicated to mainly enterprise customers. They have now become more open and based on commodity technologies, larger scale, greener, and more widely distributed to handle the mixture of huge numbers of cloud service consumers and enterprise customers.

From technical viewpoints, the emerging server consolidation and virtual machine technologies were key to driving the evolution of data centers in terms of both performance and cost. The computing architecture utilized in previous data centers was just a simple client-server model. However, thanks to such advanced technologies, various computing models are emerging and being utilized in current data centers. The examples based on such models are a web service based interaction, distributed file systems like Hadoop Distributed File Systems (HDFS), and MapReduce-based parallel and distributed applications. Data centers are required to provide efficient, scalable, and reliable data access and processing.

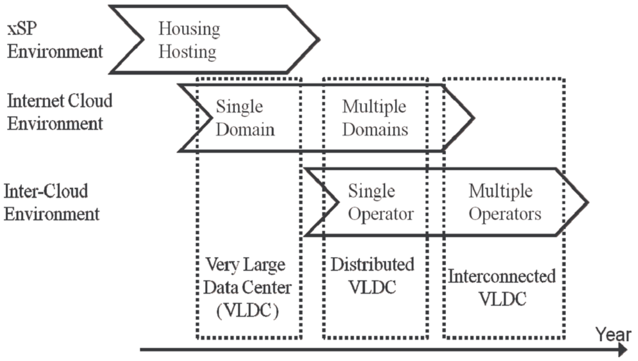

Fig. 1 Transition of data center networking.

Such computing models heavily rely on networking capabilities due to their distributed processing nature. Because of this, data center networks are becoming important and play critical roles in achieving such dynamic and flexible capability in larger-scale, wider-distributed, and more complex networking environments.

In accordance with the changing purpose of data centers that correspond to an xSP environment, Internet Cloud environment, and inter-cloud environment, data center networks change the scale of networks, the degree of distribution, and collaborative complexity. Such transition of data center networks is illustrated in Fig. 1. Requirements of earlier data center networking were on-demand resource assignment for public clouds and VLAN enlargement for private clouds. Next, data center networks were required to deal with wide-area interaction and collaboration among distributed environments. Furthermore, multi-domain/multi-operator interactions are expected to support making data centers safer, more reliable, greener (i.e., lower and efficient power usage) in a collaborative environment.

In this paper, we survey recent research and development activities related to data center networks that include widely spread technologies. The rest of this paper is organized as follows. The requirements for data center networks and a taxonomy of data center network technologies are discussed in Sect. 2. In accordance with the taxonomy, each research trend with selected research paper abstracts is introduced in Sects. 3, 4, and 5. Finally, concluding remarks

Manuscript received July 5, 2012.
Manuscript revised November 2, 2012.
[†]The authors are with NEC, Tokyo, 108-8001 Japan.
a) E-mail: y-kiriha@ay.jp.nec.com
DOI: 10.1587/transcom.E96.B.713

and future research directions are given in Sect. 6.

2. Data Center Networks

Before discussing data center networking technologies in more detail, we categorize and define generations of data center networking. For each data center generation, requirements and technical issues on networking are discussed.

Figure 1 illustrates a transition of data center roles and required types of distribution. A data center was originally started by xSPs (any type of Service Provider) mainly to reduce users' total cost ownership (TCO). Various types of services, such as Server Housing, Server/Application Hosting, were offered and operated. Almost all services were delivered in accordance with client-server computing.

Technical trends, such as the emergence of Virtual Machine (VM), expectation of server consolidation, growth of the Internet, and high-spec commodity servers, brought forth the next generation of data centers. Data centers became larger and larger, containing tens of thousands of servers. We refer to this type of data center as a "1st generation: Very Large Data Center (VLDC)", whose main criteria are high performance and efficient resource utilization.

Data centers are continuing to grow rapidly. For example, cloud service providers have multiple data centers in different locations, and are required to manage hundreds of thousands of servers at maximum. Such distributed data centers are interconnected and collaborative. We refer to this type of datacenter as a "2nd generation: Distributed VLDC", whose main criteria are availability (or service continuity) and managed QoS/SLA. Furthermore, data centers of different cloud providers are required to interact and collaborate in order to improve the quality of end-user experience, flexibility for new customers/services, and adaptability for newly emerging innovation. We refer to this type of datacenter as a "3rd generation: Interconnected VLDC", whose main criteria are dynamic optimization, flexible customization, and integrated management in totally millions of servers environment.

2.1 Common Requirements of Data Center Networking

In accordance with previously defined data center generations, common networking requirements are discussed in this section. These requirements are needed to categorize data center networking technologies in the following sections. Our examination results are depicted in Fig. 2.

In 1st Generation: Very Large Data Center (VLDC), the main networking requirements are high throughput, high reliability, and high resource utilization. Since computer/network resources are expanding and consuming more and more power, the deployment and operation costs have become a serious problem. Because of this, a scale-out architecture that enables performance to be improved only by adding Common Off The Shelf (COTS) products is strongly demanded in this generation. Data center networking needs throughput, reliability, and utilization to be achieved simultaneously under the scale-out architecture.

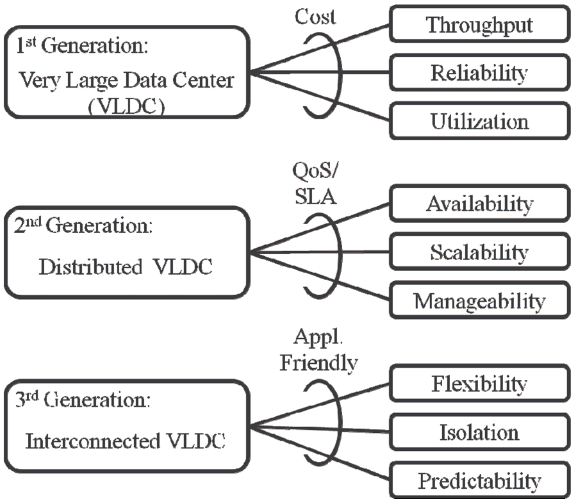

Fig. 2 Common requirements on data center networking.

Although the internal resource utilization with lower cost was crucial in major large data centers of Google, Amazon, Facebook, etc., this situation was changed by the shift to 2nd Generation: Distributed VLDC. The important requirements were to keep and ensure the quality of services and service level to customers. Due to the limitation of enlarging a single data center, distributed large data center architecture became common. In this distributed environment, improving availability, scalability, and manageability for whole huge systems was generally difficult, and such challenging research issues needed to be tackled.

The processing demand continues to grow, and more and more evolutions are expected in order to prepare for coming big data and M2M/IoT environments where multiple players (i.e., consumers, infrastructure operators, service application providers, etc.) are interconnected and collaborative. The ability to deal with unknown and unexpected requirements from emerging and future applications is highly demanded in 3rd Generation: Interconnected VLDC. In this environment, data center networking must be flexible and agile, predictable, and virtually isolated by introducing novel concepts and technologies.

2.2 Taxonomy of Data Center Network Solutions

To understand data center networking technologies, a wide variety of technologies and their mutual relationships should be considered. From the consideration discussed in Sect. 2.1, we recognized that requirements in each data center networking generation are closely related with communication layer hierarchy. On the basis of this observation, we can divide the data center design space in accordance with the communication layer hierarchy as shown in Fig. 3.

The goal of 1st Generation: Very Large Data Center (VLDC) was to achieve high performance and high resource utilization by using commodity hardware. The most effective technologies to achieve these were thought to be topol-

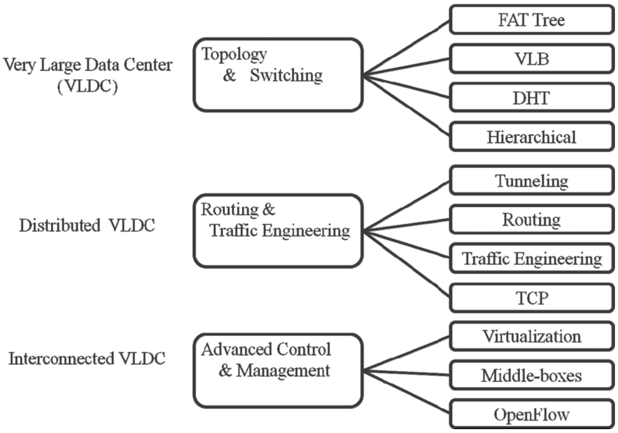

Fig. 3 Taxonomy of data center networking solutions.

ogy and switching. The goal of 2nd Generation: Distributed VLDC was to improve service quality level by utilizing geographically distributed resources as a whole. The most effective technologies to solve this were thought to be routing and traffic engineering. Finally, the goal of 3rd Generation: Interconnected VLDC is to achieve a heterogeneity transparent control and management scheme. The most effective technologies to tackle this are thought to be advanced control and management.

According to the above consideration, derived taxonomy is as follows: Topology and Switching (Layer 2 capabilities), Routing and Traffic Engineering (Layer 3 capabilities), and Advanced Control and Management including optical and wireless communication infrastructures. "Topology and Switching" can be subdivided by the type of tree setup: Fat Tree, Variant Load Balancing (VLB), Distributed Hash Table (DHT), and Hierarchical Architecture. "Routing and Traffic Engineering" can be subdivided: Tunneling, Routing, Traffic engineering, and TCP Congestion Control & Multi-path. "Advanced Control and Management" can be subdivided: Network Virtualization, Middle-boxes (Appliances), and OpenFlow.

In accordance with this taxonomy, each research trend with selected research paper abstracts is introduced in Sects. 3, 4, and 5. We note that although each surveyed paper includes and discusses the multiple technologies listed in Fig. 3, we categorized each paper by selecting one major, novel topic.

3. Topologies and Switching (Layer 2 Control)

The network of a traditional data center is hierarchical and possesses 1+1 redundancy. Since equipment higher in the hierarchy needs to handle more traffic, is more expensive, and requires more effort to control, the system architecture is a scale-up type that costs too much for improvements. It suffers the following impairments: Internal fragmentation (i.e., prevention of applications dynamic growing/shrinking), no performance isolation, and limited server-to-server capacity (i.e., oversubscription).

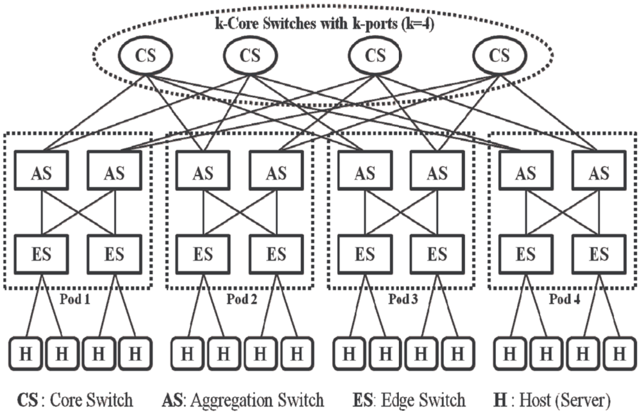

Fig. 4 k-ary fat tree topology (k=4)

It was generally said that 80% of the packets stay inside the data centers for data mining, index computations, and back end to front end, even though 10–30% CPU utilization is considered "good" in the above-described "scale-up" designed data centers.

To deal with growing data processing demands with lower installation and operational cost, maximizing performance, reliability, and utilization are crucial metrics in 1st generation: VLDC. In this section, topologies and switching related technologies are investigated to improve performance, reliability, and utilization.

3.1 Clos/Fat-Tree

The Clos network and its fat-tree topology configuration have been proposed as data center network technologies by Al-Fares et al. [2] and Greenberg et al. [4]. The benefits of a Clos/fat-tree topology are that all paths between ToR pairs have the same length and it is inexpensive to build using commodity switches. In addition, the Clos/fat-tree topology structure is well-suited to randomized load balancing (also described in Sect. 3.2), where the path for a flow is selected randomly.

As illustrated in Fig. 4, the k-ary fat tree enables k pods (or clusters), which are organized by $k/2$ aggregation switches and $k/2$ edge switches, to be interconnected through k core switches with k ports. Each edge switch directly connects to $k/2$ hosts. As a topological characteristic, fat tree is able to interconnect $k^3/4$ hosts as a whole. In the case of $k = 48$ (i.e., the number of switch port is 48), this fat tree network enables the support of 27648 hosts ($k^3/4$: k=48) and to the provision of 576 equal-cost paths (24×24: combination of 24 Edge and Core switches in each) between any given pair of hosts in different pods.

Al-Fares et al. [2] proposed an architecture to leverage large commodity Ethernet switches to support the full aggregate bandwidth of clusters consisting of tens of thousands of elements by introducing a fat tree. Since the k-ary fat tree has $(k/2)^2$ shortest paths between any two hosts on different pods, a special treatment is required on routing in order to utilize effectively the high fan-out available from the fat tree. For extensions of IP forwarding, they proposed the following: two-level routing table, flow classification and scheduling scheme, and Bidirectional Forwarding Detection (BFD)-based fault tolerant capability. In the two-level routing table, first-level prefixes are utilized to identify and forward packets that are sent to hosts in the same pods, and a second-level routing table is used to route packets to hosts in the different pods by utilizing suffixes. A flow classification and scheduling scheme aim to enhance the traffic engineering capability by considering a subsequent packet's flow and duration. They completed a prototype switch using NetFPGA [32] and Click [33] and reported their evaluation results in a four-port fat tree environment (i.e., 20 switches and 16 hosts).

Mysore et al. [3] proposed PortLand: a scalable, fault tolerant layer 2 routing and forwarding protocol for data center environments. PortLand also assumes a fat tree topology such as the traditional data center multi-rooted topology, and adds some new techniques like Pseudo MAC (PMAC) based hierarchical addressing, fabric manager with OpenFlow protocol, distributed location discovery, loop free forwarding, and fault tolerant routing. PortLand edge switches learn a unique pod number and a unique position number within a pod that are assigned by their original location discovery protocol. Because of this, edge switches of the fat tree can be assigned to directly connected hosts with a 48-bit PMAC of the form of pod.position.port.vmid, where vmid is used to multiplex multiple virtual machines on the same physical machine. By using this host location encoded PMAC, PortLand efficiently achieves packet forwarding and routing as well as VM migration. Fabric Manager is responsible for PMAC and actual MAC transformation, ARP handling, and OpenFlow flow table maintenance.

3.2 Variant Load Balancing (VLB)

The Valiant load balancing (VLB) architecture, which was first proposed by L. G. Valiant for processor interconnection networks, is an oblivious routing strategy known to handle arbitrary traffic variations that obey the hose model. VLB has been shown to be capacity efficient for handling traffic variation. VLB distributes traffic across a set of intermediate nodes that correspond to core switches in a fat-tree topology configuration. By leveraging random distribution of traffic into equal cost multi paths, VLB can achieve a hot-spot-free core fabric for data centers.

Greenberg et al. [4] proposed Monsoon, which implements VLB by adding an additional encapsulation header to frames that directs them to a randomly chosen switch. Monsoon utilizes commodity hardware that can scale out to achieve large data centers and introduces source routing as a control plane and hot-spot-free multi-path routing as a data plane. To achieve such capabilities, Monsoon requires a modification of server network stack (e.g., ARP off) and an introduction of original directory services to handle packet encapsulation (i.e., MAC-in-MAC encapsulation defined by IEEE802.ah).

For an extended/advanced version of Monsoon, Greenberg et al. [5] proposed VL2: a practical network architecture that scales to support huge data centers. The network model that VL2 creates is a virtual layer 2 per service and is able to provide uniform high bandwidth and performance isolation while keeping Ethernet semantics. VL2 implements the network model by introducing VLB, equal-cost multi-path (ECMP), separation of ID/Locator, and end host information directory. VL2 uses two different IP address families: one is a location specific IP addresses (LAs) for all switches like ToR switches and Intermediate (Core) switches, and another is an application specific IP addresses (AAs) for servers and VMs. The host information directory and each host-installed VL2 agent are responsible for mapping LAs and AAs. As a result of this, VL2 is able to keep layer 2 semantics. VL2 is based on IP routing and forwarding in order to maintain the switch level topology (i.e. VLB enabled fat-tree) and a single large IP subnet (i.e., AAs' address space). Every switch is required to run link-state routing, ECMP forwarding, IP anycasting, and IP multicasting. Their performance evaluation shows that performance at 40, 80, and 300 servers is excellent.

3.3 Distribute Hash Table (DHT)

DHT research was originally motivated by Peer-to-Peer (P2P) systems, and DHT is a distributed management technology to provide a lookup service similar to a hash table. Key-value pairs stored in DHT, and any participating node can efficiently retrieve the value associated with a given key. In the context of packet forwarding and routing, mapping host information to a switch is a hash function. Examples of such key-value pairs are (MAC address, location) for a directory service and (IP address, MAC address) for an address resolution service.

Kim et al. [6] proposed SEATTLE: an architecture that integrates the best of IP scalability and Ethernet simplicity. SEATTLE utilizes flat addressing, plug and play services, shortest path routing, and hash based resolution of host information. To improve flexibility while keeping performance scalability, SEATTLE supported hierarchical configuration by leveraging a multi-level, one-hop DHT. Since SEATLE uses the global switch-level view provided by a link state routing protocol, it can form a DHT where each node can find any nodes in one-hop by utilizing such global information. Kim et al.'s prototype system was developed by utilizing open-source routing software platforms: Click [33] and XORP [34]. Their experiments show that SEATLE efficiently handles network failure and host mobility, while reducing control overhead by roughly two orders of magnitude compared with Ethernet bridging.

3.4 Hierarchical

Hierarchical and/or recursive structure are generally appropriate in order to avoid the existence of a single point of failure as well as to increase networking capacity. Tree-based data center network architecture cannot do the above actions simultaneously.

Motivated by such issues, Guo et al. [7] proposed DCell: a novel network structure to improve scalability and fault tolerance in a data center network. DCell is a recursively defined structure, a high level DCell is constructed from many low level DCells, and same level DCells are fully connected with each other. The Lowest level Dcell has *n* servers and a mini switch for interconnection. Based on this environment, Dcell Fault Tolerant Routing (DFR) protocol, and one-to-one IP/DCN address mapping scheme were introduced. Their experimental results on a 20-server DCell test-bed showed that DCell provides twice as much throughput as the conventional tree-based structure for MapReduce traffic patterns.

As an extended/advanced version of DCell, Guo et al. [8] proposed BCube: an architecture for modular data centers, and the core of BCube is its server centric network structure. Servers with multiple network ports connect to multiple layers of Commodity Off The Shelf (COTS) mini-switches, similar to DCell hierarchical construction. To utilize multi-paths for load balance and fault tolerance, BCube introduces an original source routing with path selection and path adaptation. The authors reported that the construction and power costs of BCube and fat tree are similar. However, BCube uses fewer wires than fat tree.

4. Routing (Layer 3 Control) and Traffic Engineering

Since demands are growing and growing, data centers were expected to be enlarged by utilizing geographical distribution. In such widely distributed environments, improvements in quality of services (QoS) and service continuity (e.g., disaster recovery, multiple faults recovery, and backup services) are strongly demanded. Because of this, in 2nd generation: Distributed VLDC, data center interconnected WAN related topics have become more important. How to utilize distributed computing and networking resource pools is a crucial point.

In this section, routing and traffic engineering related technologies are investigated to improve scalability, availability, and manageability.

4.1 Tunneling/Encapsulation

From the service continuity and disaster recovery viewpoints, data centers should generally be distributed. Such distributed data centers, however, should be required to interconnect, keeping layer 2 semantics logically. In addition to such enlargement of layer 2 broadcast domains, introducing server virtualization technology (e.g., Virtual Machine) drastically increases the number of MAC addresses in a single layer 2 domain. As a result, it takes longest with MAC learning, then causes of frequent flooding, and last reduction of network stability.

The Virtual Private LAN Service (VPLS), which encapsulate Ethernet frame by MPLS tag, is a solution. However, it has a disadvantage: it requires more interconnection links, because it assumes mesh topology. To cope with the above problem, two prevalent solutions have emerged over the years and are becoming viable. One is an IETF's Transparent Interconnection of Lots of Links (TRILL), and another is an IEEE 802.1aq: Shortest Path Bridging (SPB). They aim to enlarge and extend the capability of Layer 2 Ethernet networks by introducing frame encapsulation, multiple paths calculation and selection, and so on.

SPB supports Ethernet data planes 802.1ah/ad and Ethernet OAM 802.1ag, which are widely deployed industry standards. It introduces widely deployed IS-IS link state protocol with only minor TLV extensions. SPB's new route calculation enables multiple shortest equal cost paths to be produced for both unicast and multicast traffic in Layer 2 VPNs. Thanks to 802.1ah ISID on the data path, SPB can support tens of thousands of services.

TRILL is an IETF specified standard protocol that performs layer 2 bridging using IS-IS link state protocol. Since IS-IS is able to run directly at layer 2, then no IP addresses are needed. A device that implements TRILL is called an RBridge (Routing Bridge). The RBridge has features like transparency, plug & play, virtual LANs, frame priorities, and virtualization support as a bridge. In addition, it also has features like multi-paths, optimal paths, rapid failover, and the safety of a TTL as a router. TRILL data frames are encapsulated in a local link header and a newly defined 8-byte TRILL header.

4.2 Routing

Improving routing efficiency and flexibility leads to distributed and interconnected systems that are robust and constantly available from a service continuity viewpoint.

Levchenko et al. [9] proposed XL: approXimate Link state routing algorithm, which aims at increasing routing efficiency by suppressing updates from parts of the network. XL works by propagating only some of the link updates. At the heart of the algorithm are three rules describing when an update should be propagated. These conditions are 1) when the update is a cost increase, 2) when the link is used in the node's shortest path tree, and 3) when it improves the cost to any destination by more than a $1+\varepsilon$ cost factor where ε is a design parameter of the algorithm. They showed that XL significantly outperforms standard link-state and distance vector algorithms through their simulation studies.

Motiwala et al. proposed [10] path splicing, a new routing primitive that allows network paths to be constructed by combining multiple routing trees (slices) to each destination over a single network topology. Path splicing allows traffic to switch trees at any hop on route to the destination. End systems can change the path on which traffic is forwarded by changing a small number of additional bits in the packet header. Path splicing can be deployed on existing routers with small modifications to existing multi-topology routing functions. Moreover, it is applicable to other routing protocols such as wireless, overlay routing, and so on.

4.3 Traffic Engineering

For consumers to keep their negotiated SLAs, traffic engineering capability plays an important role in distributed data center environment.

Wilson et al. [12] proposed D^3: Deadline aware control protocol, which aims to achieve application throughput maximization, burst tolerance, and high utilization. D^3 uses explicit rate control to apportion bandwidth in accordance with flow deadlines. Their evaluation results show that D^3, even without any deadline information, easily outperforms TCP in terms of short flow latency and burst tolerance.

Laoutaris et al. [13] proposed NetStitcher: a system that employs a network of storage nodes to stitch together unutilized bandwidth across different data centers, whenever and wherever it exists. NetStitcher gathers information about leftover resources, uses a store-and-forward algorithm to schedule multipath and multi-hop data transfers, and adapts to resource fluctuations. Their experimental evaluation showed that NetStitcher outperforms other mechanisms like overlay, BitTorrent, and Random Store-and-Foward. It also showed that NetStitcher can rescue up to five times additional data center bandwidth.

Zohar et al. [14] proposed PACK: Predictive ACKs, an end-to-end Traffic Redundancy Elimination (TRE) system, designed for cloud computing customers. PACK is based on receiver-based cloud friendly TRE that allows the client to use newly received chunks to identify previously received chunk chains, which in turn can be used as reliable predictors of future transmitted chunks. Moreover, PACK can eliminate redundancy based on content arriving at the client from multiple servers without applying a three-way handshake. Their evaluation using a wide collection of content types showed that PACK had clear advantages over conventional (i.e., sender-based) TRE.

4.4 TCP: Congestion Control & Multipath

TCP in the data center does not meet demands of applications due to suffering from bursty packet drops and in-cast [16]. Operators' work on such TCP problems seems to be ad-hoc, inefficient, and often expensive. Furthermore, consequences and tradeoffs are not properly understood, and building up large queues will result in adding significant latency.

The problem, a drastic throughput reduction when multiple senders communicate with a single receiver, is known as (TCP) In-cast. The overflow in small Ethernet switch buffers due to the burst and fast data traffic causes intense packet loss and leads to TCP timeouts and delays in application level protocols. Furthermore, these timeouts and delays reduce the throughput as a result.

Such pathological behavior of TCP results in gross under-utilization of link capacity in certain many-to-one communication patterns in distributed storage, MapReduce, and Web-search workloads. Chen et al. [15] reported analytic results concerning dynamics of TCP In-cast. They used empirical data to reason about the dynamic system of simultaneously communicating TCP entities.

Vasudevan et al. [16] proposed a solution for In-cast that utilizes high-resolution timers in TCP to allow for microsecond granularity timeouts: the TCP retransmission timeout (RTO). In their simulation and real-world experiments, their proposed technique effectively avoided In-cast. By using a combination of microsecond granularity timeouts, randomized retransmission timers, and disabling delayed acknowledgements, the proposed technique allowed high-fan-in data center communication.

Alizadeh et al. [17] proposed DCTCP: Data Center TCP. The problem is, for example, that bandwidth hungry background flows build up queues at the switches and thus affect the performance of latency sensitive foreground traffic. DCTCP is a solution for achieving the following three together: high burst tolerance, low latency, and high throughput. DCTCP leverages Explicit Congestion Notification (ECN) in the network to provide multi-bit feedback to the end hosts.

Raiciu et al. [18] proposed multi-path TCP that can effectively and seamlessly use available bandwidth, giving improved throughput and better fairness on many topologies. MPTCP enables topologies that single path TCP cannot utilize. They showed that their dual homed variant of the fat-tree topology with MPTCP outperforms the fat-tree for a wide range of workloads.

5. Advanced Control and Management

To handle future big data and M2M/IoT environments, the current data centers are expected to evolve further and deal with various kinds of business models and applications. Some trends have already emerged in accordance with the context of Future Internet, for example, network virtualization and OpenFlow. The crucial criteria in this 3rd generation: Interconnected VLDC are resource isolation (i.e., virtualization), flexibility/agility for future application/services, and predictability for autonomous/adaptive management and control. In this section, advanced control and management related technologies are investigated to approach new capabilities described above as well as other topics on optical and wireless networking for data centers.

5.1 Network Virtualization

Network virtualization is useful for interconnecting distributed data centers that are operated by different organizations. To achieve such network virtualization, there are two standardization activities for enlarging layer 2 (broadcast) domains by using layer 3 tunneling: VXLAN and Network Virtualization over GRE (NVGRE). Both technologies aim to provide a large-scale multi-tenant service, a virtual tenant service over inter data centers, and disaster recovery service in a geographically distributed cloud environment.

VXLAN is an IETF's draft proposed by VMware, Cisco, Citrix, et al. In the case of utilizing VLAN, the number of virtual tenant segments is actually limited to 4096 due to the 12-bit tag fields. To enhance this limitation, VXLAN's encapsulation format is newly defined, and 24-bit VXLAN Network Identifier (VNI) is introduced to support around 16 million virtual segments.

NVGRE is also an IETF's draft proposed by Microsoft. Virtual switches are interconnected by GRE encapsulated tunnels to keep the compatibility with current standards. Since this specification cannot identify UDP/TCP ports, it is difficult to achieve a fair load balancing among equal-cost multi-paths (ECMPs).

Concerning research activities, the following points were discussed from an architectural viewpoint.

Ballani et al. [19] proposed Okutopus: a system that implements the abstractions that capture the trade-off between the performance guarantees offered to tenants, their costs, and the provider revenue. The features of Okutopus are virtual over-subscribed clusters, centralized network manager, rate limiting at end hosts, and so forth.

Carapinha et al. [20] discussed the impact of network virtualization as a result of FP7 4 WARD project. They proposed a network virtualization reference model, business roles from different provider viewpoints, and a design of virtual network enabled networks. They provide a broad overview of the main challenges ahead, such as virtual network design parameters, mapping scheme of virtual resources into physical resources, and virtual network re-optimization.

Schaffrath et al. [21] proposed VNET: a control and management architecture of virtual networks. They discussed network virtualization architecture for enabling Future Internet innovation. The architecture is developed from both business and technology perspectives and comprises four main players: a physical infrastructure provider, a virtual network provider, a virtual network operator, and a service provider. VNET control plane architecture provides the control and management functions for the virtual network architecture to the various players described above. They introduce interfaces definitions, an instantiation scenario, and an end-system attachment scenario. By using simple Xen and Click [33] based virtual node, the VNET Control Plane prototype was implemented. Evaluation of virtual network instantiation showed that it can scale linearly with the number of nodes.

Mudigonda et al. [22] proposed NetLord: an architecture that provides a scalable multiple-tenant capability for virtualized datacenters. Most existing network architectures cannot meet all the needs simultaneously: which support multi-tenancy, scale, and ease of operation. NetLoad provides tenants with simple and flexible network abstractions, by fully and efficiently virtualizing the address space at both layer two and layer three. NetLoad agents (NLA) reside in the hypervisor of each physical server and perform two major tasks: one is the VM transparent encapsulation/decapsulation of packets from and to the local VMs, and another is the collection and maintenance of all the information needed for the encapsulation. Their experimentation showed that NetLoard architecture scales to several thousands of virtual tenants and hundreds of thousands of VMs.

5.2 Middle-Boxes/Appliances

Chen et al. [23] proposed DAC: a generic and automatic Data center Address Configuration system. DAC automates the assignment of IDs to network devices. It begins with a network blue print that specifies the logical ID of switches and servers. DAC then automatically learns device IDs (e.g., MAC address) and then finds a mapping of logical to device IDs. This mapping is found by modeling the problem as a graph isomorphism problem. The authors propose some heuristics to speed up the graph isomorphism search.

Dilip et al. [24] proposed PLayer and pswitch: a policy-aware switching layer architecture consisting of interconnected policy-aware switches. In accordance with policies specified by administrators, pswitches explicitly forward different types of traffic through different sequence middle-boxes (e.g., Firewalls and Load Balancers). Through a small experimentation by using a NetFPGA [32]-based prototype pswitch, it was validated to enable correct middle-box traversal to be guaranteed and more flexible networking to be achieved.

Chowdhury et al. [25] proposed Orchestra architecture for improving the data transfer times of common communication patterns such as broadcast and shuffle (i.e., MapReduce distributed processing). An Inter Transfer Controller (ITC) and Transfer Controllers (TCs) control and manage data transfer scheduling and algorithms. In this architecture, the ITC manages TCs that are responsible for choosing appropriate transfer mechanisms depending on data size, number of nodes, and other factors.

Abu-Libdeh et al. [26] proposed CamCube, which is a platform for developing application specific routing protocols in 3D torus topology environments. The API is inspired by the Key-Based Routing (KBR) API, which is used in many structured overlay. Examples of network services discussed are a multi-hop routing service and a TCP/IP service that enables unmodified TCP/IP applications that are run on top of CamCube, VM/File distribution service, an aggregation service, and in-memory object cache service.

5.3 OpenFlow

OpenFlow is a protocol that aims to separate a data plane and a control plane in networking. One of the motivations for developing OpenFlow was to open traditionally closed designs of commercial switches in order to enable fast network innovation.

OpenFlow switches form the data plane, and the control plane is implemented as a distributed system running on commodity servers.

Yu et al. [27] proposed DIFANE: a scalable and efficient solution that keeps all traffic in the data plane by selectively directing packets through intermediate switches that store the necessary rules. To reduce overhead of cache misses and to scale to large networks with richer policies, DIFANE relegates the controller to the simpler task of partitioning these rules over the switches. All data plane functions required in DIFANE can be expressed with three sets of wildcard rules of various granularity with simple actions: Cache rules, authority rules, and partition rules. DIFANE introduces authority switches that can distribute cache rules. Then the controller partitions and distributes the flow rules to OpenFlow switches through the authority switches.

Curtis et al. [28] proposed DevoFlow: a modification of the OpenFlow model that gently breaks the coupling between control and global visibility, in a way that maintains a useful amount of visibility without imposing unnecessary costs. DevoFlow introduces two new mechanisms for devolving/delegating control to a switch, rule cloning, and local actions including multipath support and rapid re-routing. Furthermore, DevoFlow provides three different ways to improve the efficiency of OpenFlow statistics collection: sampling, triggers and reports, and approximate counters. From their experimentation, DevoFlow uses 10-53 times fewer flow table entries at an average switch and uses 10-42 times fewer control messages.

5.4 Other Topics (Optical and Wireless Networking)

Below are two optical and one wireless networking activities for enhancing data center network performance.

Wang et al. [29] proposed c-Through: a prototype system based on a hybrid packet and circuit switched data center network architecture. The packet switch network uses a traditional hierarchy of Ethernet switches arranged in a tree. The circuit switched network connects the ToR switches. c-through solves the traffic de-multiplexing problem by using VLAN based routing. Each server controls its outgoing traffic using a per-rack output traffic scheduler. On the basis of their destination rack, the packets are then assigned to either the electrical or optical VLAN output queue.

Farrington et al. [30] proposed Helios: a hybrid electrical/optical switch architecture that can deliver significant reductions in the number of switching elements, cabling, cost, and power consumption. Helios implements its traffic estimation and traffic de-multiplexing features on switches. This approach makes traffic control transparent to end hosts.

Halperin et al. [31] proposed 60 GHz wireless links that are utilized to relieve hotspots in oversubscribed data center networks. By using directional antennas, many wireless links can run concurrently at multi-Gbps rates on Top of Rack (ToR) switches.

6. Conclusion and Future Directions

As common platforms for all ICT social infrastructure, data centers comprising huge numbers of servers, storages, and switches are playing an important and crucial role. There has been a tremendous amount of research into data centers, and research and development efforts will continue towards further data center evolution.

This paper focused on the networking aspect of data centers and surveyed typical research and development activities. Specifically, after introducing three historical generations of data center networking, this paper discussed requirements and research challenges in each generation. In accordance with such trends and changes of research, a data center networking taxonomy was defined, and we investigated research topics, ideas, and results for every item in the taxonomy.

For the future, to make data centers more sustainable and greener, more research is needed on lifecycle management of evolving distributed systems and applications/ services aware autonomous management, for example. In the coming big data and IoT era, especially for handling physical data, data center networking must efficiently handle stream data processing, real-time networking, and dynamic customization. The real-time and dynamic capability will be crucial for future generations of data centers.

Furthermore, as a recent hot topic, the Software Defined Network is now attracting tremendous attention from both academia and industry. The SDN white paper [35] published by the Open Network Foundation (ONF) describes the following key computing trends driving the need for a new network paradigm: Changing traffic patterns, the consumerization of IT, the rise of cloud services, and big data means more bandwidth. It also said that SDN will be able to deliver the following substantial benefits to both enterprise and carriers: Centralized management and control, improved automation and management, rapid innovation, programmability for all stakeholders, increased network reliability and security, more granular network control, and so on. We believe that SDN will strongly affect the improvement of data center networking.

Lastly, note that we hope this survey is useful and valuable for many researchers and developers who will tackle current and future issues and will contribute to the progress in the field of data center networking technologies.

References

[1] A.R. Curtis, Reducing the Cost of Operating a Datacenter Network, A thesis presented to the University of Waterloo in fulfillment of the thesis requirement for the degree of Doctor of Philosophy in Computer Science, Waterloo, Ontario, Canada, 2012.

[2] M. Al-Fares, et al., "A scalable, commodity data center network architecture," Proc. ACM SIGCOMM 2008, Seattle, Washington, USA, Aug. 2008.

[3] Radhika Niranjan Mysore, et al., "PortLand: A scalable fault-tolerant layer 2 data center network fabric," Proc. ACM SIGCOMM 2009, Barcelona, Spain, Aug. 2009.

[4] A. Greenberg, et al., "Towards a next generation data center architecture: Scalability and commoditization," Proc. PRESTO workshop at ACM SIGCOMM 2008, Seattle, Washington, USA, Aug. 2008.

[5] A. Greenberg, et al., "VL2: A scalable and flexible data center network," Proc. ACM SIGCOMM 2009, Barcelona, Spain, Aug. 2009.

[6] C. Kim, et al., "Floodless in SEATTLE: A scalable ethernet architecture for large enterprises," Proc. ACM SIGCOMM 2008, Seattle, Washington, USA, Aug. 2008.

[7] C. Guo, et al., "DCell: A scalable and fault-tolerant network struc-

ture for data centers," Proc. ACM SIGCOMM 2008, Seattle, Washington, USA, Aug. 2008.

[8] C. Guo, et al., "Bcube: A high performance, server-centric network architecture for modular data centers," Proc. ACM SIGCOMM 2009, Barcelona, Spain, Aug. 2009.

[9] K. Levchenko, et al., "XL: An efficient network routing algorithm," Proc. ACM SIGCOMM 2008, Seattle, Washington, USA, Aug. 2008.

[10] M. Motiwala, et al., "Path splicing," Proc. ACM SIGCOMM 2008, Seattle, Washington, USA, Aug. 2008.

[11] T. Benson, et al., "Understanding data center traffic characteristics," Proc. WREN Workshop at ACM SIGCOMM 2009, Barcelona, Spain, Aug. 2009.

[12] C. Wilson, et al., "Better never than late: Meeting deadlines in datacenter networks," Proc. ACM SIGCOMM 2011, Toronto, Ontario, Canada, Aug. 2011.

[13] N. Laoutaris, et al., "Inter-datacenter bulk transfers with NetStitcher," Proc. ACM SIGCOMM 2011, Toronto, Ontario, Canada, Aug. 2011.

[14] E. Zohar, et al., "The power of prediction: Cloud bandwidth and cost reduction," Proc. ACM SIGCOMM 2011, Toronto, Ontario, Canada, Aug. 2011.

[15] Y. Chen, et al., "Understanding TCP incast throughput collapse in dataceter networks," Proc. WREN Workshop at ACM SIGCOMM 2009, Barcelona, Spain, Aug. 2009.

[16] V. Vasudevan, et al., "A (In)cast of thousands: Scaling datacenter TCP to kiloservers and gigabits," CMU-PDL-09-101, Parallel Data Laboratory, Carnegie Mellon University, Feb. 2009.

[17] M. Alizadeh, et al., "Data center TCP (DCTCP)," Proc. ACM SIGCOMM 2010, New Delhi, India, Aug.-Sept. 2010.

[18] C. Raiciu, et al., "Improving datacenter performance and robustness with multipath TCP," Proc. ACM SIGCOMM 2011, Toronto, Ontario, Canada, Aug. 2011.

[19] H. Ballani, et al., "Towards predictable datacenter networks," Proc. ACM SIGCOMM 2011, Toronto, Ontario, Canada, Aug. 2011.

[20] J. Carapinha, et al., "Network virtualization — A view from the bottom," Proc. VISA Workshop at ACM SIGCOMM 2009, Barcelona, Spain, Aug. 2009.

[21] G. Schaffrath, et al., "Network virtualization architecture: Proposal and initial prototype," Proc. VISA Workshop at ACM SIGCOMM 2009, Barcelona, Spain, Aug. 2009.

[22] J. Mudigonda, et al., "NetLord: A scalable multi-tenant network architecture for virtualized datacenters," Proc. ACM SIGCOMM 2011, Toronto, Ontario, Canada, Aug. 2011.

[23] K. Chen, et al., "Generic and automatic address configuration for data center networks," Proc. ACM SIGCOMM 2010, New Delhi, India, Aug.-Sept. 2010.

[24] D.A. Joseph, et al., "A policy-aware switching layer for data centers," Proc. ACM SIGCOMM 2008, Seattle, Washington, USA, Aug. 2008.

[25] M. Chowdhury, et al., "Managing data transfers in computer clusters with orchestra," Proc. ACM SIGCOMM 2011, Toronto, Ontario, Canada, Aug. 2011.

[26] H. Abu-Libdeh, et al., "Symbiotic routing in future data centers," Proc. ACM SIGCOMM 2010, New Delhi, India, Aug.-Sept. 2010.

[27] M. Yu, et al., "Scalable flow-based networking with DIFANE," Proc. ACM SIGCOMM 2010, New Delhi, India, Aug.-Sept. 2010.

[28] A.R. Curtis, et al., "DevoFlow: Scaling flow management for high-performance networks," Proc. ACM SIGCOMM 2011, Toronto, Ontario, Canada, Aug. 2011.

[29] G. Wang, et al., "c-Through: Part-time optics in data cetners," Proc. ACM SIGCOMM 2010, New Delhi, India, Aug.-Sept. 2010.

[30] N. Farrington, et al., "Helios: A hybrid electrical/optical switch architecture for modular data centers," Proc. ACM SIGCOMM 2010, New Delhi, India, Aug.-Sept. 2010.

[31] D. Halperin, et al., "Augmenting data center networks with multi-gigabit wireless links," Proc. ACM SIGCOMM 2011, Toronto, Ontario, Canada, Aug. 2011.

[32] J.W. Lockwood, et al., "NetFPGA — An open platform for gigabit-rate network switching and routing," Proc. IEEE International Conference on Microelectronic Systems Education, 2007.

[33] E. Kohler, et al., "The click modular router," ACM Trans. Comput. Syst., vol.18, no.3, pp.263–297, 2000.

[34] M. Handley, et al., "XORP: An open platform for network research," Proc. 1st Workshop on Hot Topics in Networks, Princeton, USA, Oct. 2002.

[35] Open Networking Foundation, "Software-Defined Networking: The new norm for networks," ONF White Paper, April 2012.

Yoshiaki Kiriha is a senior manager, in the Cloud System Research Laboratories, NEC. He received an M.S. degree in Electrical Engineering from Waseda University in 1987. He then joined NEC, where he worked in the R&D division for over 20 years. He has been involved in many projects in NEC and has transferred core technologies to the product division. His research interests include distributed database systems, real-time systems, as well as Future Internet service and management. He has contributed continuously as a TPC for almost of all IM/NOMS/DSOM conferences from 2000, and has served as a chair of TC on Information Communication Management, IEICE 2010–2011.

Motoo Nishihara is a general manager in the Cloud System Research Laboratories, NEC. He received a B.S. degree from University of Tokyo and an M.S. degree in Electrical Computer Engineering from Carnegie Mellon University in 1985 and 1996, respectively. He has been engaged in R&D in the product division for over 20 years and involved in many projects, including ATM node systems, IP router, dedicated wide area Ethernet systems, network front-end systems, security appliances, and inter-cloud systems. He received the 18th Advanced Technology Award "The Prize of Fuji Sankei Business i." in 2004.