Enabling Technologies for Future Data Center Networking: A Primer

Min Chen and Hai Jin, Huazhong University of Science and Technology
Yonggang Wen, Nanyang Technological University
Victor C. M. Leung, The University of British Columbia

Abstract

The increasing adoption of cloud services is demanding the deployment of more data centers. Data centers typically house a huge amount of storage and computing resources, in turn dictating better networking technologies to connect the large number of computing and storage nodes. Data center networking (DCN) is an emerging field to study networking challenges in data centers. In this article, we present a survey on enabling DCN technologies for future cloud infrastructures through which the huge amount of resources in data centers can be efficiently managed. Specifically, we start with a detailed investigation of the architecture, technologies, and design principles for future DCN. Following that, we highlight some of the design challenges and open issues that should be addressed for future DCN to improve its energy efficiency and increase its throughput while lowering its cost.

Currently, an increasing number of science and engineering applications involve big data and intensive computing, which present increasingly high demands for network bandwidth, response speed, and data storage. Due to complicated management and low operation efficiency, it is difficult for existing computing and service modes to meet these demands.

As a novel computing and service mode, cloud computing has already become prevalent, and is attracting extensive attention from both academia and industry. Companies find such a mode appealing as it requires smaller investments to deploy new businesses, and extensively reduces operation and maintenance cost with a lower risk to support new services. The core idea of cloud computing is the unification and dispatch of networked resources via a resource pool to provide virtual processing, storage, bandwidth, and so on. In order to achieve this goal, it is critical to evolve the existing network architecture to a cloud network platform that has high performance with large bandwidth capacity and good scalability while maintaining a high level of quality of service (QoS). The cloud network platform where applications obtain services is referred to as a data center network (DCN).[1]

A traditional DCN interconnects servers by electronic switching with a limited number of switching ports, and usually employs a multitier interconnection architecture to extend the number of ports to provide full connectivity without blocking, which suffers from poor scalability. However, with the convergence of cloud computing with social media and mobile communications, the types of data traffic are becoming more diverse while the number of clients is increasing exponentially. Thus, the traditional DCN meets two major downfalls in terms of scalability and flexibility:

- From the interconnection architecture point of view, it is hard to extend the network capacity of the traditional DCN physically to scale with the fluctuating traffic volumes and satisfy the increasing traffic demand.
- From the multi-client QoS support point of view, the traditional design is insensitive to various QoS requirements for a large number of clients. Also, it is challenging to virtualize a private network for each individual client to meet specific QoS requirements while minimizing resource redundancy.

The modern data center (DC) can contain as many as 100,000 servers, and the required peak communication bandwidth can reach up to 100 Tb/s [1]. Meanwhile, the supported number of users can be very large. As predicted in [2], the number of mobile cloud computing subscribers worldwide is expected to grow rapidly over the next five years, rising from 42.8 million subscribers in 2008 (approximately 1.1 percent of all mobile subscribers) to just over 998 million in 2014 (nearly 19 percent). With the requirements to provide huge bandwidth capacity and support a very large number of clients, how to design a future DCN is a hot topic, and the following challenging issues should be addressed.

Scalable bandwidth capacity with low cost: In order to satisfy the requirement of non-blocking bisection bandwidth among servers, huge bandwidth capacity should be provided by an efficient interconnection architecture, while the cost and complexity should be decreased as much as possible.

Energy efficiency: DCs consume a huge amount of power and account for about 2 percent of the greenhouse gas emissions that are exacerbating global warming. Typically, the annual energy use of a DC (2 MW) is equal to the amount of

This work is supported in part by the 1000 Young Talents Plan, and China National Natural Science Foundation under grant No. 61133006. We would like to thank Professor Limei Peng for her very helpful suggestions and comments during the delicate stages of concluding this article.

[1] *In this article, depending on the context, DCN represents data center network and data center networking interchangeably.*

0890-8044/13/25.00 © 2013 IEEE

energy consumed by around 5000 U.S. cars in the same period. The power consumption of DCs mainly comes from several aspects, including switching/transmitting data traffic, storage/computation within numerous servers, cooling systems, and power distribution loss. Energy-aware optimization policies are critical for green DCN.

User-oriented QoS provisioning: A large-scale DC carries various kinds of requests with different importance or priority levels from many individual users. The QoS provisioning should be differentiated among different users. Even for the same user, the QoS requirements can change dynamically over time.

Survivability/reliability: In case of system failures, uninterrupted communications should be guaranteed to offer almost uninterrupted services. Thus, it is very crucial to design finely tuned redundancy to achieve the desired reliability and stability with the lowest resource waste.

In this article, the technologies for building a future DCN are mainly classified into three categories: DCN architecture, inter-DCN communications, and large-scale clients supporting technologies. The organization of the rest of this article addresses the above three categories of technologies. We present various DCN interconnection architectures. The emerging communication techniques to connect multiple DCNs are then described. The design issues to support QoS requirements for large-scale clients are given. We provide a detailed description of a novel testbed, Cloud3DView, for modular DC, as well as outline some future research issues and trends. Finally, we give our concluding remarks.

Architecture for Data Center Networking

A large DCN may comprise hundreds of thousands or even more servers. These servers are typically connected through a two-level hierarchical architecture (i.e., a fat-tree topology). In the first level, the servers in the same rack are connected to the top of the rack (ToR) switch. In the second level, ToR switches are interconnected through higher-layer switches. The key to meeting the requirements of huge bandwidth capacity and high-speed communications for DCN is to design an efficient interconnecting architecture.

In this section, we first classify the networking architecture inside a DC into four categories — electronic switching, wireless, all-optical switching, and hybrid electronic/optical switching — which are detailed below.

Electronic Switching Technologies

Although researchers have conducted extensive investigations on various structures for DCNs recently, most of the designs are based on electronic switching [3, 4]. In electronic-switching-based DCN, the number of switching ports supported by an electronic switch is limited. In order to provide a sufficient number of ports to satisfy the requirement of non-blocking communications among a huge number of servers, the server-oriented multi-tier interconnection architecture is usually employed.

Due to the hierarchical structure, oversubscription and unbalanced traffic are the intrinsic problems in electronic-switching-based DCN. The limitation of such an architecture is that the number of required network devices is very large, and the corresponding construction cost is expensive; also, the network energy consumption is high. Therefore, the key to tackling the challenge is to provide balanced communication bandwidth between any arbitrary pair of servers. Thus, any server in the DCN is able to communicate with any other server at full network interface card (NIC) bandwidth. In order to ensure that the aggregation/core layer of the DCN is not oversubscribed, more links and switches are added to facilitate multipath routing [3, 4]. However, the increment of the performance is traded off for the increased cost of a larger amount of hardware and greater networking complexity.

Wireless Data Center

Recently, wireless capacity of 60 GHz spectrum has been utilized to tackle the hotspot problem caused by oversubscription and unbalanced traffic [5]. The 60 GHz transceivers are deployed to connect ToR for providing supplemental routing paths in addition to traditional wired links in DCNs. However, due to the intrinsic line-of-sight limitation of the 60 GHz wireless links, the realization of wireless DC is quite challenging. To alleviate this problem, a novel 3D wireless DC is proposed to solve the link blockage and radio interference problems [6]. In 3D wireless DC, wireless signals bounce off DC ceilings to establish non-blockage wireless connections. In [7], a wireless flyway system is designed to set up the most beneficial flyways and routes over them both directly and indirectly to reduce congestion on hot links.

The scheduling problem in wireless DCN is first found and formulated as an important foundation for further work on this area [8]. Then an ideal theoretical model is developed by considering both the wireless interference and the adaptive transmission rate [9]. A novel solution to combining throughput and job completion time is proposed to efficiently improve the global performance of wireless transmissions. As pointed out in [9], channel allocation is a critical research issue for wireless DCNs.

All-Optical Switching

Due to super-high switching/transmission capacity, the optical fiber transmission system is considered as one of the most appropriate transmission technologies for DCN. Moreover, the super-high switching capacity and flexible multiplexing capability of all-optical switching technology provide the possibility of flattening the DCN architecture, even for a large-scale DC.

The all-optical switching techniques can be mainly divided into two categories: optical circuit switching (OCS) and optical packet switching (OPS).

- *OCS* is a relatively mature technology with market readiness, and can be used for the core switching layer to increase the switching capacity of DCNs while significantly alleviating the traffic burden. However, OCS is designed to support the deployment of static routing with pre-established lightpaths, but statically planned lightpaths cannot handle bursty DC traffic patterns, which leads to congestion on overloaded links. Since OCS is a coarse-grained switching technology at the level of a wavelength channel, it exhibits low flexibility and inefficiency on switching bursty and fine-grained DCN traffic.
- *OPS* has the advantage of fine-grained and adaptive switching capability but is subject to some serious problems in the aspect of technological maturity due to the lack of high-speed OPS fabrics and all-optical buffers.

Hybrid Technologies

Compared to an optical-switching-based network, an electronic network exhibits better expandability but poorer energy efficiency. Hybrid electronic/optical-switching-based DCN tries to combine the advantages of both electronic-switching-based and optical-switching-based architectures [10, 11]. An electronic-switching-based network is used to transmit a small amount of delay-sensitive data, while an optical-switching-based network is used to transmit a large amount of traffic. Table 1 compares the features of various representative DCN

Name	Networking architecture	Switching granularity	Scalability	Energy consumption
Fat-Tree	Electronic	Fine	Low	High
BCube	Electronic	Fine	Low	High
DCell	Electronic	Fine	Low	High
VL2	Electronic	Fine	Low	High
Schemes in [5]	Wireless	Fine	Low	High
Schemes in [6]	Wireless	Fine	Low	High
HyPaC	Hybrid	Medium	Medium	Medium
Helios	Hybrid	Medium	Medium	Medium
DOS [12]	Optical	Coarse	High	Low
Scheme in [13]	Optical	Coarse	High	Low

Table 1. *A comparison of DCN architectures.*

architectures in terms of different interconnecting technologies. As seen in Fig. 1, suitable DCN architecture design trade-offs should attempt to satisfy application-specific bandwidth and scalability requirements while keeping deployment cost as low as possible.

Inter-DCN Communications

Nowadays, due to the wide deployment of rich media applications via social networks and content delivery networks, the number of DCNs around the world is increasing rapidly, and a DC seldom works alone. Thus, there is a demand to connect multiple DCNs placed in various strategic locations, and the communications among them is defined as *inter-DCN communications* in this section. When inter-DCN communications and intra-DCN communications are jointly considered, a two-level structure emerges. Inside a DC, any architecture presented earlier can be selected for intra-DCN communications. In this section, we first survey alternative architectures for inter-DCN communications and then the joint design between the two levels.

Optical-Switching-Based Backbone for Inter-DCN Communications

Since optical fiber can provide large bandwidth, it can be utilized to interconnect multiple DCNs to solve the problems of traffic congestion and unbalancing. Similar to intra-DCN communications, it is hard to deploy OPS for inter-DCN communications due to the lack of high-speed optical packet switching fabrics and all-optical buffers. Instead, OCS is the better choice because of the maturity of the technology. Although OCS is characterized by slow reconfiguration on the order of a millisecond, the traffic flow in the backbone network is relatively stable, and thus the cost of lightpath configuration can be amortized over backbone traffic streams that have sufficiently long durations. In this section, two OCS technologies are considered for this purpose, wavelength-division multiplexing (WDM) and coherent optical-orthogonal frequency-division multiplexing (CO-OFDM) technology [13].

Wavelength-Division Multiplexing — Traditionally, a single-carrier laser optical source is used as the WDM light source. In order to meet the requirements for mass data transmissions, the number of laser sources needs to be increased, resulting in a sharp rise in cost and energy consumption. Thus, the demand for the increase of carriers in a WDM light source is critical. In recent years, multicarrier source generation technology [14] and point-to-point WDM transmission systems have attracted much attention. Thousand-channel dense WDM (DWDM) has also been demonstrated successfully. Based on a centralized multicarrier source, an optical broadcast and select network architecture is suggested in [14].

CO-OFDM Technology — As one of the OCS technologies, CO-OFDM shows great potential in reducing the construction cost for future DCNs. One great advantage of CO-OFDM is its all-optical traffic grooming capability compared with the legacy OCS networks where optical-to-electronic-to-optical (O/E/O) conversion is required. This is a critical feature desired in the interconnection of DCNs where traffic is heterogeneous with extremely large diversity. Thus, it can improve bandwidth capacity with high efficiency and allocation flexibility. However, it suffers from the intrinsic network agility problem that is common to all of the existing OCS-based technologies.

Software-Defined Networking Switch

GatorCloud [15] was proposed to leverage the flexibility of software-defined networking (SDN) techniques to dramatically boost DCNs to over 100 Gb/s bandwidth with cutting-edge SDN switches by OpenFlow techniques. SDN switches separate the data path (packet forwarding fabrics) and the control path (high-level routing decisions) to allow advanced networking and novel protocol designs, which decouples decision-making logic and enables switches remotely programmable via SDN protocols. In such a way, SDN makes switches economic commodities. With SDN, networks can be abstracted and sliced for better control and optimization for many demanding data-intensive or computation-intensive applications over DCNs.

Hence, SDN makes it possible to conceive a fundamentally different parallel and distributed computing engine that is deeply embedded in the DCNs, realizing application-aware service provisioning. Also, SDN-enabled networks in DCNs can be adaptively provisioned to boost data throughput for specific applications for reserved time periods. Currently, the main research issues are SDN-based flow scheduling and workload balancing, which remain unsolved.

Joint Intra-DCN and Inter-DCN Design

For large-scale DCNs where many servers and multiple DCNs are interconnected, traditional solutions exhibit the undesirable features of various kinds of load-unbalancing problems. In this section, the whole architecture is jointly designed in three levels according to the characteristics of intra-DCN's internal servers and inter-DCN's traffic flows. At the rack level, hybrid electronic/optical switching is proposed to interconnect similar servers; at the intra-DCN level, optical switching is suggested to connect ToR switches; at the inter-DCN level, multiple DCNs are connected by main backbone optical networks.

Rack Level — In the same rack, the servers are more apt to communicate with nearby servers only. Compared to communications with higher layers, the communications between servers in the same rack have a higher tendency to burst. To

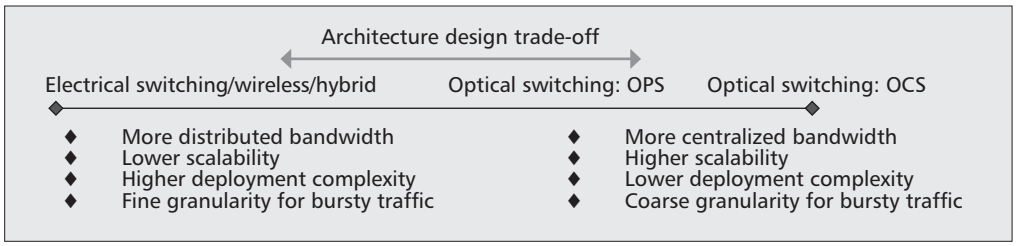

Figure 1. *Trade-off map for DCN architecture design.*

handle dynamic and bursty traffic flows between similar servers, packet switching technology is most suitable. Since it is challenging to deploy OPS networks, electronic switching or hybrid electronic/optical switching are recommended for rack-level communications.

Intra-DCN Level — As mentioned earlier, a typical DCN contains a certain number of racks, and a typical rack accommodates several dozen servers. The hierarchical tree structure's top bandwidth is much smaller than the available bandwidth of the servers' NIC, resulting in bottlenecks and the hotspot problem of oversubscription to aggregation/core switches. Optical switching technology can provide large communication bandwidth, which can effectively solve the bandwidth shortage and facilitate the improvement of the DCN's design and implementation. At the intra-DCN level, the traffic flows between racks have no distinguishable dynamic change. Although the deployment of an optical switch incurs a higher cost initially, the device also has a longer life span, which compensates for the higher deployment cost.

Multi-DCN Level — At the multi-DCN level, the amount of traffic flows on the downstream and upstream links are usually asymmetric. The upstream data flow from a terminal server to a DCN consists mostly of simple request messages, while the returned content flow (e.g., video streaming) from the DCN to the terminal server may be very large. Such asymmetry causes a serious imbalance in capacity consumption between upstream and downstream links, possibly resulting in wastage of bandwidth resources.

While WDM with a multicarrier optical source is a suitable technology for the main backbone of DCNs, it has several issues that need to be addressed. For example, optical carrier distribution algorithms should be designed to meet the demands of dynamic flow change. The effective use of optical carriers produced by a multicarrier source can maximize spectrum efficiency. Then a highly efficient and reliable optical carrier distribution method can be used to improve carrier quality in terms of optical SNR in order to satisfy the demand of data transmissions.

Large-Scale Client-Supporting Technologies

A large-scale DC may process various kinds of requests with different levels of importance or priority from lots of individual users. In order to support large-scale clients with different QoS requirements, this section mainly investigates two categories of technologies, virtualization and routing. Two kinds of virtualization technologies (i.e., server consolidation and virtual machine migration) are introduced. Then routing issues are investigated by considering intra-DCN traffic characteristics.

Virtualization for Large-Scale Applications

At present, DCs are power hungry: their energy cost accounts for 75 percent of the total operating cost [16]. Servers consume the largest fraction of overall power (up to 50 percent), while approximately the other half of the power is consumed by cooling systems. This is why many research efforts have focused on increasing both server utilization and power consumption efficiency. This includes virtual machine (VM) consolidation (migration of VMs to the minimum amount of servers and switching off the rest of the servers), dynamic voltage and frequency scaling of servers' central processing units (CPUs), air flow management (to optimize cooling), and the adaptation of air conditioning systems to current operating conditions.

Server Consolidation — In server consolidation, multiple instances of different applications are moved to the same server, with the aim of freeing up some servers for shutdown. Server consolidation can save the energy consumed by a DC. However, such a use case will challenge the existing DCN solutions, because a large amount of application data might need to be shifted among different servers, possibly located on different racks.

VM Migration — Another technology suitable for serving a large number of clients is to migrate VMs in response to user demand. Similar to the server consolidation use case, this case will also result in additional data communications among physical servers in different racks. As such, new technologies and strategies should be developed in order to mitigate potential detrimental effects.

One possibility is to leverage the existing named-data networking (NDN) technology for distributed file system (DFS) management and VM migration. First, DFS is a crucial component for DC and cloud computing. In current architectures, metadata management systems are becoming a bottleneck, limiting their scalability. In our Cloud3DView project, two novel ideas based on NDN and information-centric networking (ICN) have been proposed for DFS metadata management systems. We have shown that our new solution can reduce the overhead from $O(\log n)$ to $O(1)$. Initial numerical results verified that the throughput and response delay can be reduced significantly. Second, VM migration is the most popular approach for load balancing in DCs nowadays. However, the traditional approach often suffers from overwhelming overhead. In this research, we leverage the emerging NDN framework to decouple object locality from its physical infrastructure. In this case, an object can easily be routed without constantly updating the metadata system. We expect that overhead cost can be reduced significantly.

Routing for Dynamic Flow Demands

Today's DCs are shared among multiple clients running a wide range of applications. These applications require a network with flexibility to meet various flow demands. Given a flow matrix, an ideal routing algorithm should be able to find the optimal routing distribution. However, as the flow in the DCN changes dynamically and randomly, the routing algorithm cannot possibly know the actual flow matrix.

Currently, most DCN routing algorithms are only designed

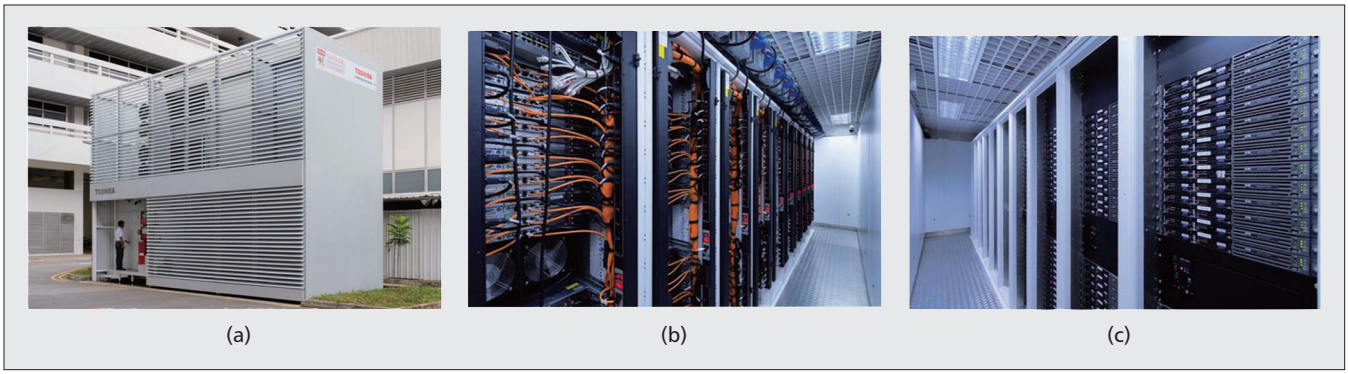

Figure 2. *The infrastructure of Cloud3DView: a) data center testbed; b) hot isle; c) cold isle.*

for specific DCN architectures (e.g., electronic-switching-based or optical-switching-based), and either demand oblivious routing (DOR) [3] or demand specific routing (DSR) [11] is used. As presented earlier, hybrid electronic-/optical-switching-based DCN can exhibit higher integrated performance. How to design a highly efficient algorithm for a hybrid electronic-/optical-switching-based DCN needs to be further explored. In order to address this issue, this section analyzes DOR and DSR, and proposes a hybrid scheme.

Demand Oblivious Routing — DOR is a conservative approach to address flow randomness. Since it assumes that the routing algorithm does not know the flow characteristics, it cannot make highly efficient routing selections.

Demand-Specific Routing — By comparison, DSR mechanisms need a given flow matrix for an optimal routing decision. DSR has higher performance than DOR when the flow demand is known. However, due to the dynamic changes in actual DCN flows causing the flow matrix to be outdated, DSR may suffer from low efficiency when the flows change dynamically.

Hybrid Scheme — This section presents a hybrid algorithm of DOR and DSR to maximize network performance for a large DCN with uncertain dynamic flows. The hybrid scheme is motivated by combining the advantages of both routing mechanisms. Although the DC flows change dynamically, it can be divided into slowly changing flows (e.g., big but stable data flows) and rapidly changing flows (e.g., small bursts of data). Thus, if slowly changing data flows could be estimated accurately, DSR would gain better performance. On the other hand, DOR could be used for rapidly changing flows. The design of the hybrid scheme needs to address the following issues.

Flow matrix composition and representation: Understanding flow characteristics is key to the development of a highly efficient routing algorithm. It includes several topics, such as whether the DC flows can be decomposed into steady flows and burst flows, how to decompose the flows, and which mathematical model could be used to represent both steady flows and burst flows.

Determining the proportion of DSR and DOR: An algorithm is needed to decide which parts of the flows and how much of the flows should be transferred to the DSR or DOR paths. The proportion is related to network capacity and link congestion probability.

Algorithm design to meet the congestion probability condition: Assuming that congestion probability on any link should be lower than a certain threshold value, the design of a hybrid DOR and DSR algorithm should maximize the network flows accommodated while meeting the congestion probability when a probability density function of congestion probability for the flow matrix is given.

CLOUD3DVIEW: A Testbed for a Modular Data Center

In this section, as an example, we briefly introduce a modular DC testbed operated at Nanyang Technological University. The purpose of this testbed is to develop dynamic power management technologies, some of which are based on the aforementioned DCN technologies, to reduce the electricity consumption of DC operations.

Testbed Infrastructure

The testbed consists of three subsystems, including information and communication technology (ICT), cooling, and power distribution. In this subsection, we illustrate its overall structure and ICT capacity.

Overall Structure — As Fig. 2 shows, the modular DC consists of two modules, including an air conditioning module and a server module. The former module is stacked over the latter to save the precious resource of space in Singapore.

The air conditioning module makes use of outside air for cooling purposes whenever feasible, such as when the outside air temperature is lower than the temperature inside the DC. The intake fresh air flows from the cold area to the hot area through the servers and takes away the heat generated by the servers. It saves energy by preventing the DC from running colder for longer hours.

In the modular DC of Cloud3DView, the server module is organized into 10 racks, each of which contains up to 27 servers and two layer 2 switches. We adopt a fat tree topology for rack-level internetworking, and all the ToR switches are connected to the aggregate router, which is connected to the Internet. Cloud3DView, a software system developed at NTU, monitors and manages all the equipment.

ICT Capacity — The modular DC contains 270 servers, which consist of a 394.2 Tbyte disk, 1080 Gbyte memory, and 540 CPUs. All of these servers are connected to the ToR switches via Gigabit Ethernet. The operation system for the servers is Centos 6.3 for stability.

System Management Suite

One goal of this project is to create a novel gamification DC management system. Cloud3DView aims to eliminate the traditional command-line-based DC management style and provide a gamification solution to control DC with high efficiency. To achieve this goal, Cloud3DView contains five subsystems.

Data Collection Subsystem — This subsystem focuses on collecting data from DC equipment (e.g., servers, switches, power distribution units, and air conditioner) and applica-

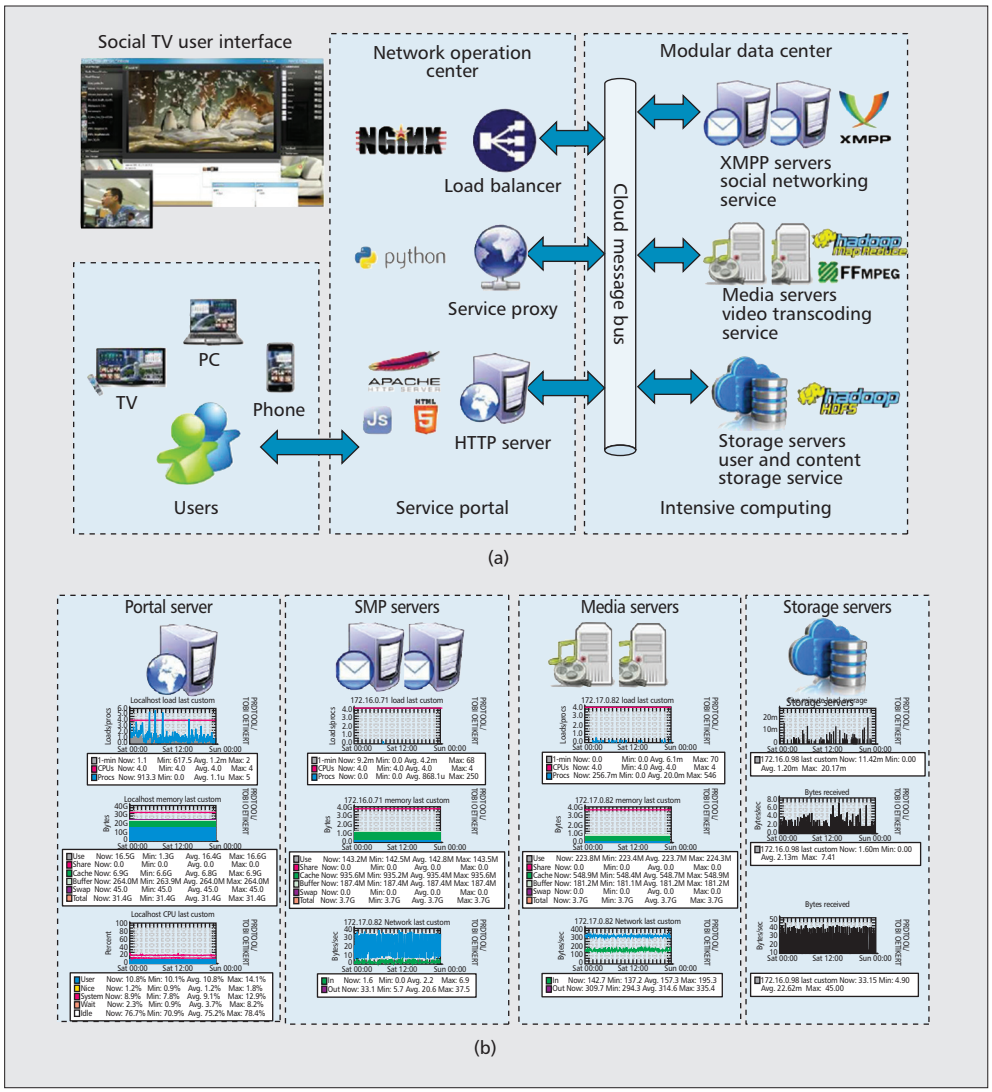

Figure 3. *Multiscreen cloud social TV application deployment: a) social TV deployment architecture; b) data center performance for social TV.*

tions (e.g., Hadoop, Apache, and Mysql). The collected data will be used for monitoring, analysis, visualization, and management.

Data Storage Subsystem —This subsystem focuses on storing data efficiently in both a traditional relational database management system (RDBMS) and NoSQL. It also provides effective application program interfaces (APIs) for data visualization and data mining.

Data Mining Subsystem —This subsystem focuses on analyzing collected data and providing solutions to various design problems (e.g., green DC).

System Monitoring Subsystem — This subsystem provides visualization of the collected data in 3D and a game-like experience. Administrators can view the real-time status of DC equipment and applications. If anything goes wrong, an alert message can be sent to the administrators.

System Management Subsystem — This subsystem mainly focuses on managing physical servers, VMs, applications, and network devices. Traditionally, in order to deploy an application in a DC, administrators and clients may take a lot of time to do background tasks (e.g., creating VMs and installing software). Cloud3DView can eliminate these background tasks and deploy applications to fulfill clients' requirements automatically.

Two Trial Applications

The testbed is primarily used to support two applications, multi-screen cloud social TV [20], and big data analytics.

Multiscreen Cloud Social TV — Our multiscreen Social TV technology has been touted by global media (1600+ news articles from 29+ countries) as an innovative technology to transform traditional "laid-back" TV viewing behavior into a proactive "lean-forward" social networking experience, marrying TV to the social networking lifestyle of today. This platform, when fully developed and commercialized, would transform the value of TV and potentially save it from a downfall similar to that of newspapers. In our system, examples of salient and sticky features include, but are not limited to, a virtual living room experience that allows remote viewers to watch TV programs together with text, audio, and video communication modalities, a video teleportation experience that allows viewers to seamlessly migrate programs across different screens (e.g., TV, smartphone, and tablet) with minimum learning. Moreover, to meet the requirements of various

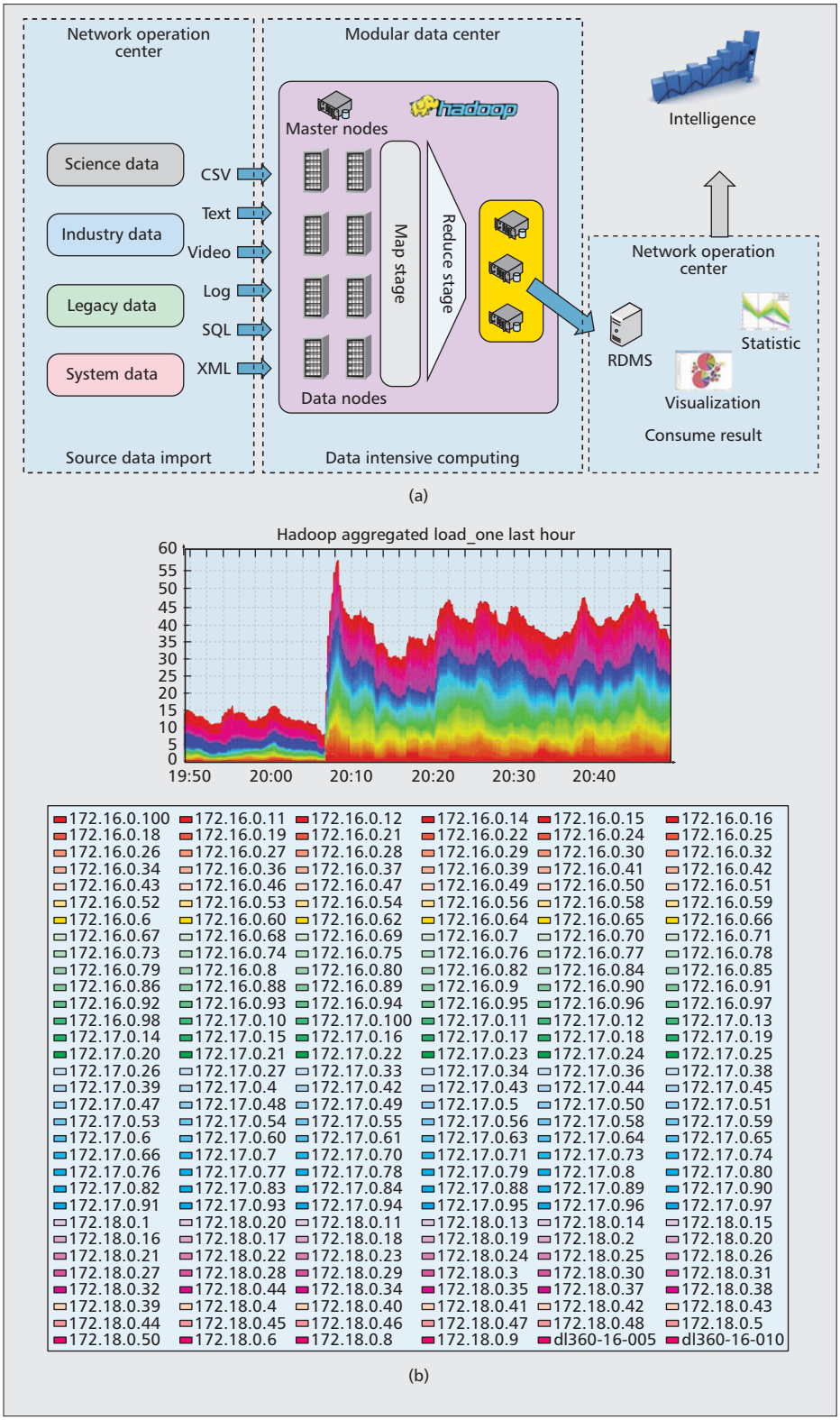

Figure 4. *Big data analytics application deployment: a) big data analytic reference architecture; b) big data performance.*

customers, the platform will provide a set of APIs for other developers to design, implement, and deploy novel value-added content services for specific customer needs (e.g., home care for the elderly, TV ad workflow redesign, real-time TV shopping, collaborative e-learning, autism diagnosis, and assistive treatment, to name a few).

In Fig. 3, we present the deployment architecture for multiscreen cloud social TV into the modular DC and its performance in system measurement. In the deployment architecture, as shown in Fig. 3a, different servers are configured for specific functions, including portal, XMPP messaging, media players, storage, and so on. Using the Cloud3DView management system, we have been able to monitor the performance statistics of each individual server. As shown in Fig. 3b, for each column, the three figures, from top to bottom, show the performance for CPU, storage, and networking. The visualization of these performance metrics is instrumental for the operators to dynamically control the workload for better QoS or quality of experience (QoE).

Big-Data Analytics — In the testbed, we have also implemented a reference architecture for big data analytics, based on the very popular Hadoop framework. Our reference architecture is illustrated in Fig. 4a, in which the key component is a cross-rack Hadoop deployment. The reference architecture includes four major steps in the big data analytics process, including:

- Source data input
- Data-intensive computing
- Result consumption
- Intelligence

In Fig. 4b, we present an alternative view of the system status, by combining the computation loads from all the servers into the same diagram. In this diagram, each server is labeled with a specific color, and the total computation load is the sum of all the individual servers (or colors). With this aggregate load, it is easy to keep track of the total load of the computation task. This aggregate load metric is a good indicator when more resources are acquired from the DC.

Conclusions

In this article, we have surveyed the major aspects for building a future DCN in terms of architecture, multiple-DCN interconnection, user-oriented QoS provisioning, and other supporting technologies. In Fig. 5, we summarize the core design issues with respect to several DCN system components. With regard to large-scale applications and clients, more open issues for building a future DCN, such as efficient interconnection architecture, resource assignment, reliability, energy-efficiency, and traffic characteristics, need to be further investigated in future work.

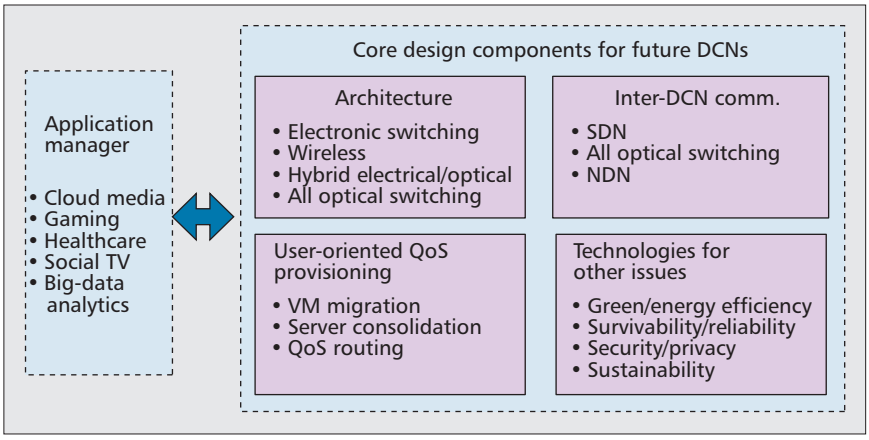

Figure 5. *Core functional components for future DCN design.*

References

[1] http://www.datacenterknowledge.com/.
[2] ABI, "Mobile Cloud Computing Subscribers to Total Nearly one Billion by 2014," ABI, http://www.abiresearch.com/press/1484/, Tech. Rep., 2009.
[3] A. Greenberg *et al.*, "VL2: A Scalable and Flexible Data Center Network," *SIGCOMM 2009*, Barcelona, Spain.
[4] C. Guo *et al.*, "BCube: A High Performance, Server-Centric Network Architecture for Modular Data Centers," *SIGCOMM 2009*, Barcelona, Spain.
[5] D. Halperin *et al.*, "Augmenting Data Center Networks with Multi-Gigabit Wireless Links," *ACM SIGCOMM 2011*, Toronto, Canada.
[6] X. Zhou *et al.*, "Mirror Mirror on the Ceiling: Flexible Wireless Links for Data Centers," *ACM SIGCOMM 2012*, Helsinki, Finland.
[7] J. Kandula and P. Bahl, "Flyways to De-Congest Data Center Networks," *HotNets '09*, 2009.
[8] Y. Cui, H. Wang, and X. Cheng, "Channel Allocation in Wireless Data Center Networks," *IEEE INFOCOM 2011*, 2011.
[9] Y. Cui *et al.*, "Dynamic Scheduling for Wireless Data Center Networks," *IEEE Trans. Distrib. Systems*, http://doi.ieeecomputersociety.org/10.1109/TPDS.2013.5, Jan. 2013.
[10] N. Farrington *et al.*, "Helios: A Hybrid Electronic/Optical Switch Architecture for Modular Data Centers," *SIGCOMM 2010*, New Delhi, India.
[11] G. Wang *et al.*, "c-Through: Part-Time Optics in Data Centers," *SIGCOMM 2010*, New Delhi, India.
[12] X. Ye *et al.*, "DOSA Scalable Optical Switch for Datacenters," *ANCS 2010*, La Jolla, CA, 2010.
[13] L. Peng *et al.*, "A Novel Approach to Optical Switching for Intra-Datacenter Networking," *IEEE/OSA J. Lightwave Tech.*, vol. 30, no. 2, Jan. 2012, pp. 252–66.
[14] Y. Cai *et al.*, "Design and Evaluation of an Optical Broadcast-and-Select Network Architecture with a Centralized Multi-Carrier Light Source," *IEEE/OSA J. Lightwave Tech.*, vol. 27, no. 21, Nov. 2009, pp. 4897–4906.
[15] "GatorCloud," http://www.s3lab.ece.ufl.edu/projects/
[16] S. L. Sams, "Discovering Hidden Costs in Your Data Centre: A CFO Perspective," IBM white paper.
[17] D. Li *et al.*, "Exploring Efficient and Scalable Multicast Routing in Future Data Center Networks," *Proc. IEEE INFOCOM*, Shanghai, China, 2011.
[18] R. Dai, L. Li and S. Wang, "Adaptive Load-Balancing in WDM Mesh Networks with Performance Guarantees," *Photonic Network Communications*.
[19] Y.G. Wen, G. Y. Shi and G.Q. Wang, "Designing an Inter-Cloud Messaging Protocol for Content Distribution as A Service (CoDaas) over Future Internet," *6th Int'l. Conf. Future Internet Technologies 2011*, Seoul, Korea, June 2011.
[20] Y. C. Jin *et al.*, "On Monetary Cost Minimization for Content Placement in Cloud Centric Media Network," accepted to the *2013 IEEE Int'l. Conf. Multimedia and Expo*, July 15–19, 2013, San Jose, CA.

Biographies

MIN CHEN [M'08, SM'09] (minchen@ieee.org) is a professor at the School of Computer Science and Technology of Huazhong University of Science and Technology (HUST). He was an assistant professor at the School of Computer Science and Engineering of Seoul National University (SNU) from September 2009 to February 2012. He worked as a postdoctoral fellow in the Department of Electrical and Computer Engineering of the University of British Columbia (UBC) for three years. Before joining UBC, he was a postdoctoral fellow at SNU for one and half years. He has more than 180 paper publications. He received the Best Paper Award from IEEE ICC 2012 and was the Best Paper Runner-Up from QShine 2008. He has been a Guest Editor for *IEEE Network*, *IEEE Wireless Communications*, and other publications. He was Symposium Co-Chair for IEEE ICC 2012 and 2013. He was General Co-Chair for IEEE CIT 2012. He is a TPC member for IEEE INFOCOM 2014. He was a Keynote Speaker for CyberC 2012 and Mobiquitous 2012.

YONGGANG WEN [S'99,M'08] (ygwen@ntu.edu.sg) is an assistant professor with the School of Computer Engineering of Nanyang Technological University, Singapore He received his Ph.D. degree in electrical engineering and computer science (with a minor in Western Literature) from the Massachusetts Institute of Technology (MIT), Cambridge, in 2008; his M.Phil. degree (with honor) in information engineering from the Chinese University of Hong Kong in 2001, and his B.Eng. degree (with honor) in electronic engineering and information science from the University of Science and Technology of China (USTC), Hefei, Anhui, in 1999. Previously, he worked at Cisco as a senior software engineer and a system architect for content networking products. He also worked as a research intern at Bell Laboratories, Sycamore Networks, and Mitsubishi Electric Research Laboratory (MERL). He has published more than 60 papers in top and prestigious conferences. His system research has resulted in two patents and has been featured in international media (e.g., *Straits Times*, *Business Times*, *Lianhe Zaobao*, Channel News Asia, ZDNet, CNet, ACM Tech News, United Press International, Times of India, Yahoo News). His research interests include cloud computing, mobile computing, multimedia networking, cyber security, and green ICT.

HAI JIN [M'99, SM'06] (hjin@hust.edu.cn) is a Cheung Kung Scholars Chair Professor of computer science and engineering at HUST. He is now dean of the School of Computer Science and Technology at HUST. He received his Ph.D. degree in computer engineering from HUST in 1994. In 1996, he was awarded a German Academic Exchange Service fellowship to visit the Technical University of Chemnitz in Germany. He worked at the University of Hong Kong between 1998 and 2000, and as a visiting scholar at the University of Southern California between 1999 and 2000. He was awarded an Excellent Youth Award from the National Science Foundation of China in 2001. He is the chief scientist of ChinaGrid, the largest grid computing project in China, and the chief scientist of the National 973 Basic Research Program Project of Virtualization Technology of Computing Systems. He is a member of the ACM. His research interests include computer architecture, virtualization technology, cluster computing and grid computing, peer-to-peer computing, network storage, and network security.

VICTOR C. M. LEUNG [S'75, M'89, SM'97, F'03] (vleung@ece.ubc.ca) is a professor of electrical and computer engineering and holder of the TELUS Mobility Research Chair at the University of British Columbia, Vancouver, Canada. He has contributed some 650 technical papers, 25 book chapters, and five books in the areas of wireless networks and mobile systems. He was a Distinguished Lecturer of the IEEE Communications Society. Several papers he co-authored have won best paper awards. He has been serving on the Editorial Boards of *IEEE Transactions on Computers*, *IEEE Wireless Communications Letters*, and several other journals, and has contributed to the organizing and technical program committees of numerous conferences. He was a winner of the 2012 UBC Killam Research Prize and the IEEE Vancouver Section Centennial Award. He is a Fellow of the Canadian Academy of Engineering and the Engineering Institute of Canada.