Disaggregated Data Centers: Challenges and Tradeoffs

Yuxin Cheng[1], Rui Lin[1], Marilet De Andrade[2], Lena Wosinska[3], and Jiajia Chen[1]
School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Sweden
Ericsson Research, Sweden
Department of Electrical Engineering, Chalmers University of Technology, Sweden

Abstract—**Resource utilization of modern data centers is significantly limited by the mismatch between the diversity of the resources required by running applications and the fixed amount of hardwired resources (e.g., number of central processing unit CPU cores, size of memory) in the server blades. In this regard, the concept of function disaggregation is introduced, where the integrated server blades containing all types of resources are replaced by the resource blades including only one specific function. Therefore, disaggregated data centers can offer high flexibility for resource allocation and hence their resource utilization can be largely improved. In addition, introducing function disaggregation simplifies the system upgrade, allowing for a quick adoption of new generation components in data centers. However, the communication between different resources faces severe problems in terms of latency and transmission bandwidth required. In particular, the CPU-memory interconnects in fully disaggregated data centers require ultra-low latency and ultra-high transmission bandwidth in order to prevent performance degradation for running applications. Optical fiber communication is a promising technique to offer high capacity and low latency, but it is still very challenging for the state-of-the-art optical transmission technologies to meet the requirements of the fully disaggregated data centers. In this paper, different levels of function disaggregation are investigated. For the fully disaggregated data centers, two architectural options are presented, where optical interconnects are necessary for CPU-memory communications. We review the state-of-the-art optical transmission technologies and carry out performance assessment when employing them to support function disaggregation in data centers. The results reveal that function disaggregation does improve the efficiency of resource usage in the data centers, although the bandwidth provided by the state-of-the-art optical transmission technologies is not always sufficient for the fully disaggregated data centers. It calls for research in optical transmission to fully utilize the advantages of function disaggregation in data centers.**

I. INTRODUCTION

Cloud computing is one of the major services provided by modern data centers (DCs), where users are able to freely choose resources and operating systems (OSs) for running their applications without considering the underlying hardware setup. According to Cisco [1], the total amount of installed cloud computing workload instances (e.g., virtual machine VM, container) in global DCs were already about 150 million in 2016, and is expected to continue growing to 500 million in 2021, representing a 19% compound annual growth rate (CAGR). DC operators have to increase the total capacity of DC resources including computing, storage, networking in order to serve this large amount of workload. For example, the total amount of storage of the global DC was about 1000 Exabytes (1 EB = 10^{18} Bytes) in 2016 and is expected to be 2500 EB in 2021 [1]. On the other hand, the resource utilization of modern DCs is relatively low. It is reported that the utilization of central processing unit (CPU) and memory in Googles' DCs are only about 35% and 55%, respectively [2], implying that a large amount of the installed resources cannot be fully utilized. If keeping business as usual, DC operators have to either install more hardware or replace the existing equipment with more advanced one in order to handle the workload growth, leading to extremely high cost and power consumption.

Low resource utilization in modern DCs can be related to the mismatch between the diversity of the running applications' resource usage and the fixed amount of resources integrated in the physical blade servers (known as integrated server) in DCs. In modern DCs, thousands of integrated servers are located in different racks and connected to top-of-rack switches (Fig. 1 left) by the network interface cards (NICs), and they are communicating with each other through the Ethernet/IP traffic. Each integrated server has a fixed amount of resources (e.g., a HP ProLiant BL660c Gen8 blade server with 8 cores CPU, 16 GB memory, 600 GB hard drive, 1 Gb/s Ethernet NIC). Such a static hardware configuration leads to 'resource stranding' [3], i.e., a server that has used up one type of resource cannot carry out more workload even though there is still a big amount of leftover of other types of resources. For example, a computing-intensive task like video processing may consume all the CPU resources in a server, and memory in the same sever cannot be assigned to the other tasks. Moreover, integrating all resources within a server chassis makes it unpractical and not economical for the DC operator to change and/or upgrade only one or a few types of resources. Instead, DC operator has to discard old servers and buy new ones. It may cause high cost for maintenance and upgrade, and also postpone to adopt new-generation hardware [3].

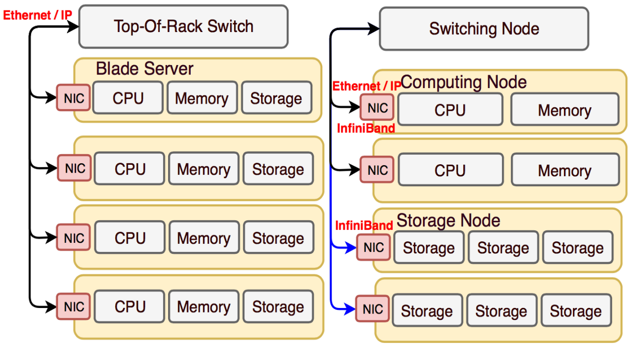

Figure 1. Modern data center racks: integrated server (left) and partially disaggregated architecture (right)

Resource disaggregation is one possible solution to solve the resource stranding issue in DCs. 'Disaggregation' means that different types of resources are decoupled from each other and hence can be allocated individually when a new application or service is deployed, which is in contrast to the modern integrated server. Depending on the level of resources to be disaggregated, we further categorized disaggregated data center into 'partially disaggregated' and 'fully disaggregated'. In the recent decade, partial resource disaggregation has been widely used in modern DCs. Specifically, storage resources are decoupled from the integrated server, and are interconnected to the rest of computing resources (including CPU and memory) through external switch fabrics (Fig. 1 right). In this case, a special NIC (e.g., InfiniBand) might be required in order to support the computing-storage communication. Such a partially disaggregated architecture allows DC operators to separately upgrade the storage (e.g., from hard disk drive HDD to solid-state drive SSD) without affecting the computing nodes. Moreover, comparing with the integrated server, the partially disaggregated architecture brings more flexibility on the data management, such as data backup and migration. There are already commercially available partially disaggregated solutions, such as Ericsson's Hyperscale Data Center 8000.

While partially disaggregated solutions have already been in use for several years, it should be noted that in this type of architecture, the CPU and memory resources are still coupled as the computing node, causing the issue of limited resource utilization of CPU/memory. Recently, the concept of fully disaggregated architecture has been proposed, where there are no more physical 'boxes' integrating different types of resources. Instead, the same type of resources forms a unit, i.e., a resource blade, rack, or even cluster. Such units are interconnected to allow communication between the different types of resources. The fully disaggregated architecture enables DC operators to replace/upgrade any type of resource when necessary, and have a great potential to improve resource utilization [4][5]. However, most existing works on disaggregated DC have ignored the transmission capacity limitation for the communication between different types of resources, such as data read and write between CPUs and memories. In the modern computer architecture, the latency requirements for communications among different types of resources vary from milliseconds (e.g., CPU-storage) to nanoseconds (e.g., CPU-memory) [6]. Meanwhile, the peak bandwidth can range from only a few Gb/s to several hundreds of Gb/s. Failing to meet these requirements could cause significant performance degradation of running applications [6]. Although there are already mature technologies (e.g., InfiniBand) to support the resource communication with moderate bandwidth requirement (i.e., CPU-storage), it is hard to find commercially available technologies that are able to support ultra-high peak data rate and ultra-low latency required by CPU-memory communications. Optical communications are promising to offer high bandwidth and low latency. However, they cannot be assumed to provide infinite capacity. To the best of our knowledge, it is still very challenging for the state-of-the-art optical transmission technologies to achieve 400+ Gb/s while keeping the latency in the magnitude of nanoseconds.

In this regard, we present different architectural options for disaggregated data centers and review the state-of-the-art optical transmission technologies for short-reach scenarios. Key performance, including blocking probability, resource utilization, and revenue, is evaluated when employing the state-of-the-art optical transmission technologies in disaggregated DCs, with a special focus on investigating if the capacity provided by these optical transmission techniques can properly support the interconnect between the different types of resources. Our results reveal a negative impact of the limited capacity provided by the state-of-the-art optical transmission technologies on the performance of fully disaggregated DCs, which calls for the future breakthrough research.

II. Disaggregated Data Centers

Rack-scale function disaggregation in DCs, where each type of resource is held on the resource blade and interconnected within a rack, is the most common one to be considered in the literature [4-6]. Therefore, in this section we concentrate on rack-scale disaggregated DCs and present two candidates for the fully disaggregated DC architectures. Due to the evolution of the underlying physical hardware, there are new challenges such as resource allocation and management, which will be discussed afterwards.

Architectures

Figure 2 shows two architectures for rack-scale fully disaggregated data centers. In each architecture, every type of resources, such as CPU, memory, storage, network interface card (NIC), and accelerator such as graphics processing unit GPU (not shown in the figure) are fully decoupled from each other. Instead of server blades that contain all types of resources, the resource blades that only includes one specific type of resources are interconnected through the optical interfaces in a rack. Depending on the type of interconnects used in the rack, these two architectures can be categorized as having either an all-optical or a hybrid interconnect (see Figure 2).

In the case of an all-optical interconnect, one optical interface is required in every resource

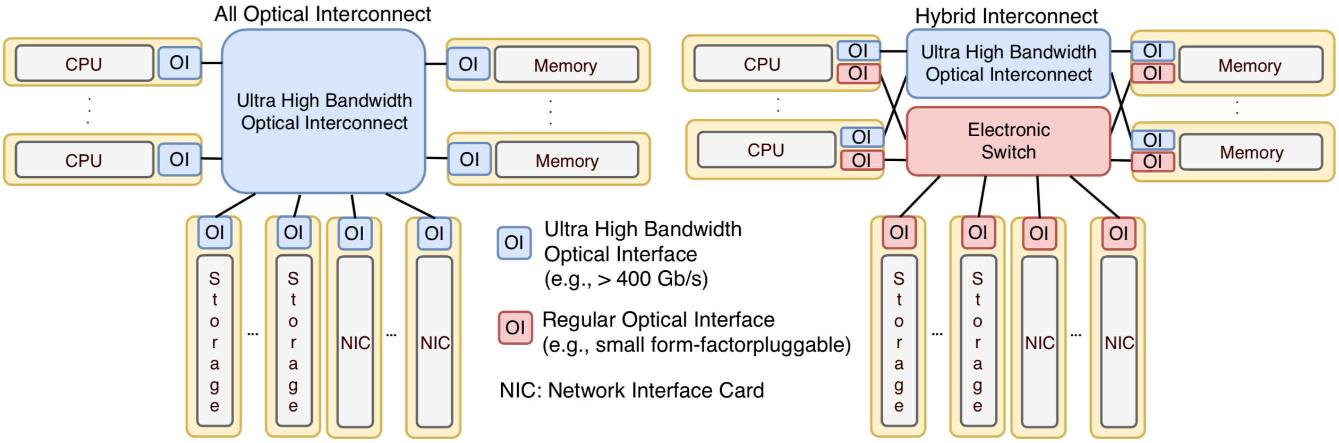
All Optical Interconnect
CPU
OI
CPU
OI
Ultra High Bandwidth Optical Interconnect
OI
Memory
OI
Memory
OI
Storage
OI
Storage
OI
NIC
OI
NIC
...
...
OI
Ultra High Bandwidth Optical Interface (e.g., > 400 Gb/s)
OI
Regular Optical Interface (e.g., small form-factorpluggable)
NIC: Network Interface Card
Hybrid Interconnect
CPU
OI
OI
CPU
OI
OI
Ultra High Bandwidth Optical Interconnect
Electronic Switch
OI
OI
Memory
OI
OI
Memory
OI
Storage
OI
Storage
OI
NIC
OI
NIC

Figure 2. Fully disaggregated rack with all-optical interconnect (left) and hybrid interconnect (right)

blade, and all the blades are connected to an optical interconnect by the optical link. All types of resource communications (including CPU-memory, memory-storage, memory-NIC, etc.), which were used to be on the motherboard buses of the integrated server, are now carried out on the external optical paths established between resource blades. This means that the optical interfaces on resource blades must satisfy the critical requirements in terms of latency and bandwidth of communications between the resources to avoid performance degradation in running applications. In Figure 2, all the optical interfaces for the all-optical interconnect as well as the interconnect itself are shown in blue, representing their capabilities to support all type of resource communications, especially the most bandwidth-hungry ones, i.e., CPU-memory, where new generation memory with high performance usually requires a peak bandwidth higher than 400 Gb/s.

In contrast to the first architecture, the second one includes two types of interconnects in a rack. One is ultra-high bandwidth optical interconnect dedicated to CPU-memory communications and the other is an electronic switch for the resource communications that typically do not have the performance requirements as stringent as CPU-memory communications. For the CPU and memory blades, two types of optical interfaces are needed, namely ultra high bandwidth optical interface (>400 Gb/s, shown in blue in Fig. 2) serving the optical interconnect, and regular optical interface (i.e., small form-factor pluggable SFP, shown in red in Fig. 2) connecting to the electronic switch. On the other hand, the storage and NIC blades can be equipped with the regular optical interfaces only, and the resource communication related to these two types are handled exclusively by the electronic switch. It should be noted that regular optical interfaces are also required at the port of electronic switch (not shown in Fig.2) for the optical-electronic signal conversion.

The main difference between these two architectures is the additional regular optical interfaces and the electronic switch in the second architecture. The cabling in the first architecture is less complex than in the second one, since there can be only one fiber for every resource blade. However, due to the fact that every communication from/to a resource blade is handled by the single optical interface, the communication coordination is more complex. For example, extra efforts should be taken on the memory blade so that the ultra high bandwidth CPU-memory communication does not use up the optical interface bandwidth all the time, and leave the memory-storage and memory-NIC communications unserved and starved. In the hybrid architecture, the coordination of the resource communications is simpler, thanks to the dedicated connections for the lower bandwidth resource communications. Moreover, there are already standard and commercial products (e.g., InfiniBand remote direct memory access (RDMA) from Mellanox) that can be applied in this architecture, since they are able to meet the requirements of latency and bandwidth of storage- and NIC-related communications.

Resource Management

The VMs are widely used in modern DCs. The VMs allow the DC operators and users to utilize any operating systems that are suitable for their applications without considering the details of hardware setup. The hypervisor is used to monitor and manage the VMs running on the integrated servers. In addition, the hypervisor also allocates the requested resources in the integrated server to the new incoming VM request.

In disaggregated DC, the hardware changes in the rack should be transparent to the VMs. Otherwise, there would be a tremendous work on modifying the existing applications and it is unpractical to ask DC users to change their running applications due to the upgrade of DC's hardware. In disaggregated DC, it is hypervisor's work to hide all the hardware changes and provide the consistent resource abstraction to the VMs utilized by the DC users. Figure 3 illustrates an example of hypervisor for disaggregated DC. Instead of running utilizing the individual integrated servers, the hypervisor is now running on the top of the resource rack. It has the access to all the resource blades, and also monitors the resource usage of each blade. When a new VM request comes, the hypervisor allocates the resources from the most suitable resource blades

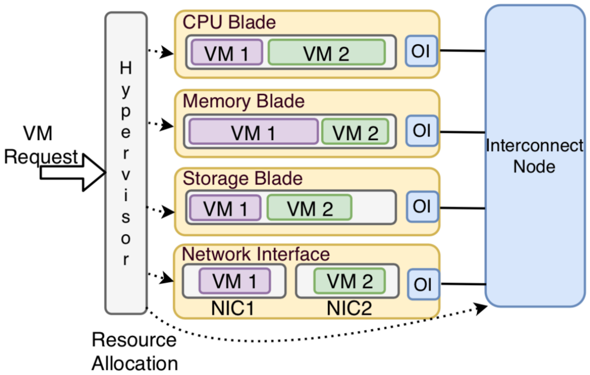

Figure 3. Resource Allocation in fully disaggregated architecture.

based on the current resource utilization of all the blades in the rack. It should be noted that the optical interface bandwidth of the resource blade is limited. Moreover, the switching time of the most advanced optical interconnect switch is in the scale of micro-seconds, which is much longer than the allowed latency of CPU-memory (nanoseconds). The lack of bandwidth and the longer latency may bring serious performance degradation in the running applications. Consequently, when the hypervisor deploys the VM, not only the sufficient amount of resource should be available on the resource blade, but also there should be enough available bandwidth on the optical interfaces, taking into account the already running other VMs. Regarding the configuration of the optical interconnect, the state-of-the-art optical switching technology cannot support switching time in nanoseconds, so the circuit switching is more applicable than packet switching in disaggregated data center. The hypervisor should configure the optical interconnect nodes so that dedicated channels are established from port to port for the resource communication.

III. Communication Between Resources

In this section, the network requirements for communication between resources in disaggregated data centers are presented. In this regard, we review the state-of-the-art optical transmission technologies and discuss their applicability in disaggregated DCs.

Network Requirements of Communication Between Resources

In the rack-scale disaggregated data center, communication between resources is performed either by an optical interconnect or by an electronic switch. It is important for the transmission links and the optical interconnects to meet the requirements in terms of latency and bandwidth. Table 1 lists the latency and bandwidth requirements of three major types of resource interconnects in a modern integrated server [3][6]. It can be seen that for storage and NIC related communications, the latency requirement is in the scale of microseconds (or even longer), and the bandwidth requirement is less or equal to 10 Gb/s. To support these two types of communication between resources, various commercially available approaches can be used, such as low latency Ethernet switch from Cisco (100Gb/s, <1µs), InfiniBand switch from Mellanox (100 Gb/s, <1µs), and PCIe switch from H3 Platform (~60 Gb/s, <1µs) [6]. All these products can be adopted as the 'Electronic Switch' in the hyrid architecture shown in Fig. 2.

Table 1 Network Requirements of Common Resource Communication [3][6]

Type of Communication Between Resources	Latency	Bandwidth Gb/s
CPU-Memory	<100 ns	200-400+
CPU-Storage	10-1000 µs	1-5
CPU-10G NIC	1-10 µs	10

On the other hand, the requirements of CPU-memory communications are very strict. Considering the required bandwidth, for example, the overall CPU-memory bandwidth is highly dependent on the performance of CPU and memory, and it is calculated by multiplication of the word size of CPU, the memory clock speed, and the number of memory controllers in CPU. For a double data rate 4rd generation (DDR4) memory with a clock speed of 2133 MHz, given a common 64-bit CPU with 3 memory controllers, the peak data rate required by CPU-memory communications is 400 Gb/s. It is extremely challenging for the aforementioned commercial products to support such ultra-high bandwidth.

Optical Transmission for Communication Between Resources

To meet the critical requirements of communication between resources, especially the communications between CPU and memory, optical transmission technology is considered the only possible solution due to its feasibility to offer ultra-high bandwidth and low latency. Optical transmission can be categorized as two major types: 1) intensity modulation and direct detection (IM/DD) system, and 2) coherent system. While coherent system has been widely applied in the long-haul transmission, its high cost and system complexity makes it difficult to be affordable for short-reach applications. Meanwhile, complex digital signal processing required at transponders causes a large delay, which may be a problem for Datacom. On the other hand, IM/DD has the advantages of simple system setup and has been considered promising to provide high bandwidth for DCs. Therefore, we focus on IM/DD transmission systems hereafter.

Table II lists up-to-date works for short-reach optical communication beyond 200Gb/s, where different modulation formats, multiplexing approaches, types of transceivers, signal processing techniques, and forward error corrections (FECs) are employed, indicating possible enabling techniques for the communication between resources in the disaggregated DCs. In order to achieve a low cost and low energy consumption per bit, high per-lane data rate is preferable. Beyond 100 Gb/s per-lane transmission has been achieved by using simplest modulation formats, e.g., non-return to zero on-off-keying (NRZ-OOK) and partially responding signaling electrical duo-binary (EDB) [10]. 4-level pulse amplitude (PAM4) [12] and discrete

Table 2 State-of-the-art optical short-reach transmission

Modulation	Wavelength band (nm)	Data rate per fiber	Multiplexing	Reach	Optical link	Transceiver	Pre-FEC BER	Reference
DMT	1550	4 x 87 Gb/s	WDM	20km	SMF	SiP	3.8e-2	[7]
NRZ	850	6x40 Gb/s	SDM	7m	MMF	VCSEL	1e-12	[9]
NRZ/EDB	1550	7x100 Gb/s	SDM	10km	MCF	EAM	5e-5	[10]
NRZ	1310	8x4x25 Gb/s	SDM/WDM	1.1km	MCF	VCSEL	1e-12	[11]
PAM4	1550	7X149 Gb/s	SDM	1 km	MCF	VCSEL	3.8e-3	[12]

DMT: Discrete multitone modulation; WDM: wavelength division multiplexing; SMF: single mode fiber; SiP: silicon photonics
NRZ: Non-return-to-zero; SDM: spatial division multiplexing; MMF: multi mode fiber; VCSEL: vertical-cavity surface emitting laser
EDB: electrical duo-binary; MCF: multi core fiber; EAM: electro-absorption modulator; PAM4: 4-level pulse amplitude modulation

multi-tone (DMT) [7] are another two options for modulation, which can alleviate the baud rate and achieve high bandwidth efficiency.

By exploiting multiplexing techniques, such as wavelength division multiplexing (WDM) [7], spatial division multiplexing (SDM) based on multicore/multimode fiber [12] and their combination [11], the per-fiber capacity of the interconnect can be further boosted. The WDM system implies high cost of the transceiver while the SDM approach may be expensive due to the utilization of advanced fiber technologies. Single mode fiber (SMF) link enables the relatively long distance communication with fine transceivers while low cost transceivers can be used together with multi-mode fiber (MMF), nevertheless signal bandwidth and transmission distance are limited [9].

In the disaggregated DC, the transceivers should be small and simple to be implemented or integrated on the resource unit in a cost-efficient manner. For the transceiver side, vertical-cavity surface emitting lasers (VCSELs) [9, 11, 12] and silicon photonic (SiP) integrated circuit techniques [7] are two main candidates to address the challenges in terms of cost and footprint. VCSELs are being widely used in short-reach optical communications, usually integrated with SDM systems. Properly designed VCSELs are able to operate without additional monitoring over a wide range of temperatures with minimal change in performance, which is very suitable for DC since the temperatures might vary a lot depending on the different work load. The character of surface emission also enables dense 2-D fabrication of VCSELs and vertical integration with other elements, and therefore the packaging is simplified, which makes the whole transmission module small and easy enough to be integrated and implemented on the resource unit. It also allows for wafer level testing and screening, which lowers the cost of manufacturing, reducing the total cost of the DC infrastructure. SiP together with WDM enables high data rate by utilizing the efficiency of high-volume silicon manufacturing and good reliability. 100 Gb/s per single channel IM/DD link was already demonstrated [7]. The evolution from 100G Ethernet to 400G Ethernet [8] in DC networks makes the advantage of SiP more obvious. There are already 400G solutions based on SiP from industries, including but not limited to Intel, Luxtera and Acacia, indicating its potential supporting the transmissions in disaggregated DC.

Moreover, minimizing the latency is essential for the practical deployment of the optical interconnect for disaggregated DC. Comparing to long-haul links, the propagation delay is obviously lower in DCs. Besides extra processing time introduced by DSP modules, typical FEC latency becomes a major contributor to overall system latency. Standard hard decision FEC (HD-FEC) (at level of tens of nanoseconds, e.g., 51 ns of 802.3bj KR FEC [13]) or innovative low-latency codes are more suitable, which in turn imposes stringent requirements in terms of pre-FEC bit error rates and receiver sensitivity.

400 Gb/s and 800 Gb/s may become the next standardized data rates [8], which will allow higher per lane speed, e.g., 400 Gb/s per lane might be available. Though the state-of-the-art optical transmissions listed in Table II are able to achieve a data rate up to 800 Gb/s per fiber, whether these data rates are sufficient for CPU-memory communications required in the fully disaggregated DC remains an open question.

IV. Performance Evaluation

In this section, we evaluate the performance in terms of resource utilization and VM request blocking probability for different DC architectures using a customized Python-based simulator [15]. CPU-memory resource communications in the fully disaggregated architecture are carried out by the optical interconnect, while the other types of communication between resources can be supported by the electronic switches. Three scenarios are considered: 1) integrated server, in total 32 blades within a rack, each with 16 cores, 64 GB memory, 1024 GB storage, 2) partially disaggregated, 32 computing nodes, each with 16 cores and 64 GB memory and 16 storage nodes, each with 2048 GB, and 3) fully disaggregated, 16 CPU blades, each with 32 cores, 16 memory blades, each with 128 GB and 16 storage blades, each with 2048 GB. The total amount of resources in three scenarios are the same. In the fully disaggregated scenario, we consider 400 Gb/s and 800 Gb/s optical interfaces for CPU-memory communications, representing two data rates that will be probably standardized soon [8]. In addition, when a VM is deployed in the fully disaggregated scenario, two types of CPU-memory peak capacity requirements are considered, i.e., 200 Gb/s and 400 Gb/s, which are equivalent to the bandwidth of the common memory (DDR3-1600 MHz) and the high performance memory (DDR4-3200 MHz) with double memory controllers. In the fully disaggregated scenario, we apply the first-fit algorithm for the VM request deployment. Note that in this scenario the requests might be blocked due to either lack of the resource on the blades or the required bandwidth on the optical interfaces. For benchmark, we further relax the constraints on the maximum bandwidth of the optical interfaces, assuming no bandwidth limit in the fully disaggregated DCs, so that the blocking is only caused by the shortage of resources on the blades. On the other hand, in the integrated server scenario and partially disaggregated scenario, there is no need for the external CPU-memory communications and the bandwidth of the optical interfaces is not a constraint any more. Besides average resource utilization of three types of resources and VM request blocking probability, we also show the revenue difference between running VMs in the considered disaggregated scenarios (i.e., partially disaggregated and fully disaggregated) and the integrated server scenario, which is calculated according to the Google's VM pricing system [14], reflecting revenue either gain or loss from the operators' perspective. The VM required resource and request arrival pattern we refer to [15].

Figures 4(a) and 4(b) show the VM request blocking probability and average resource utilization, respectively. As it can be seen, there is an obvious impact of optical interface bandwidth on the fully disaggregated scenario. In the scenario of DDR3 fully disaggregated, the performance with 400 Gb/s is even worse than that in the scenarios of the integrated server and partially disaggregated (i.e., higher blocking probability and lower resource utilization). Given an optical interface of 800 Gb/s or higher bandwidth, the performance of DDR3 fully disaggregated can be much better than that of the integrated server and partially disaggregated. On the other hand, the DDR4 fully disaggregated with 400 Gb/s optical interfaces shows an extremely high blocking probability and low resource utilization. In this scenario, most of the resources on the blades cannot be utilized due to the shortage of bandwidth on optical interfaces. With 800 Gb/s optical interfaces, the DDR4 fully disaggregated scenario is only able to achieve a similar performance as that of the integrated server case, slightly lower than that of the partially disaggregated case. If there is no bandwidth limit on the optical interfaces, i.e., there is always sufficient bandwidth for resource communications, the fully disaggregated scenario can outperform all the others regardless the types of memory

Figure 4(c) shows the revenue difference, i.e., how much revenue the DC operator can gain or lose after upgrading the integrated server based architecture to different disaggregated architectures. Without sufficient bandwidth, the fully disaggregated scenario is only able to achieve a similar revenue as the integrated server (DDR3 with 400 Gb/s, DDR4 with 800 Gb/s), or possibly much worse (DDR4 with 400 Gb/s), meaning that the fully disaggregated is totally not desirable. On the other hand, with sufficient bandwidth on the optical interfaces, the full function disaggregation triples the revenue difference compared to the partially disaggregated scenario.

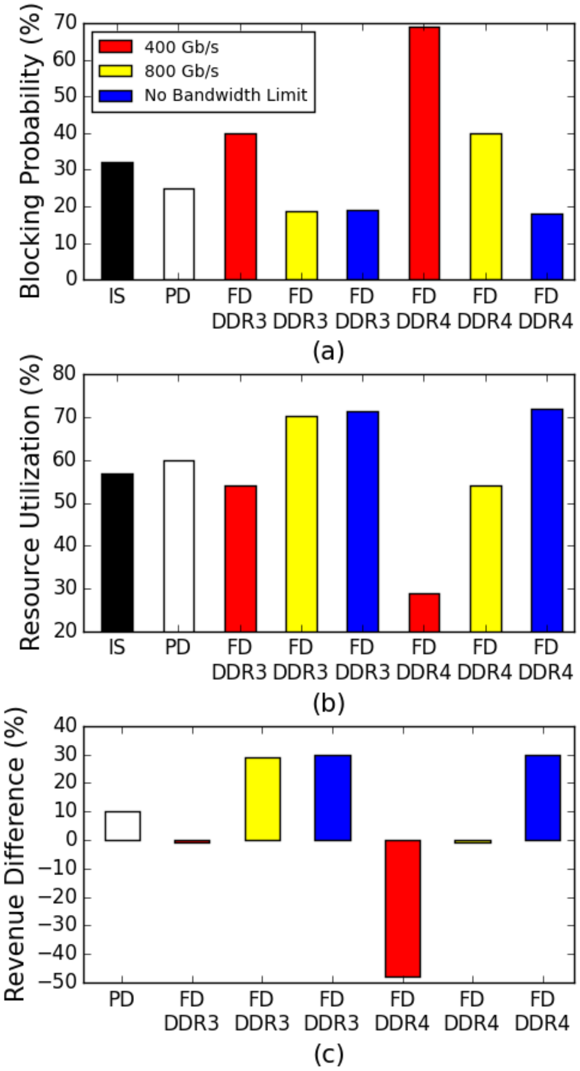

Figure 4. (a) VM request blocking probability (b) resource utilization of all scenarios (c) revenue difference between traditional DC and disaggregated DC. IS: Integrated server; PD: Partially disaggregated; FD: Fully disaggregated.

V. Conclusions and Future Work

In this paper, we introduce and evaluate the concept of disaggregated DCs, which is expected to achieve much better resource utilization compared to the data centers based on the integrated servers. However, even with ultra-high speed optical transmission, the capacity of communication between resources cannot be assumed unlimited. We have found that with advanced CPU and memory (e.g., DDR4), the benefits of fully disaggregated DCs may be reduced or even not existing, due to the fact that the bandwidth provided by the state-of-the-art optical fiber communication technologies is not sufficient. This definitely calls for research on cost-efficient short-reach optical transmission with higher bandwidth (e.g., over 1 Tb/s). In addition, the impacts on latency and energy consumption introduced by function disaggregation are not included and need to be further examined.

Furthermore, the proper resource allocation algorithms for the applications deployment in the disaggregated DCs should also be investigated. In this paper, the simple first-fit algorithm is used to allocate resources for the VM requests, whereas advanced strategies (e.g., machine learning for VM request prediction and resource usage optimization) may have a great potential to further improve the performance of the disaggregated DC.

Acknowledgment

This work was supported by the Swedish Foundation for Strategic Research, Swedish Research Council, Göran Gustafssons Stiftelse, and SENDATE-EXTEND funded by Vinnova (Celtic-Plus project ID C2015/3-3).

References

[1]Cisco Global Cloud Index: Forecast and Methodology, 2016–2021, White Paper, 2018, accessed on Sep. 2018.

[2]C. Reiss et al., "Google cluster-usage traces: format+ schema," Google Inc., White Paper, 2011, accessed on Sep. 2018.

[3]A. Roozbeh et al., "Software-Defined "Hardware" Infrastructures: A Survey on Enabling Technologies and Open Research Directions," IEEE Communications Surveys & Tutorials, VOL. 20, NO. 3, 2018.

[4]G. Zervas et al., "Optically Disaggregated Data Centers with Minimal Remote Memory Latency: Technologies, Architectures, and Resource Allocation," J. OPT. COMMUN. Vol. 10, no. 2, p. 270 (2018).

[5]A. Pages et al., "Optimal VDC Service Provisioning in Optically Interconnected Disaggregated Data Centers," COMMUN. Letters, Vol. 20 (2016).

[6]P. Gao et al., "Network requirements for resource disaggregation," OSDI' 16, Savannah (2016).

[7]P. Dong et al., "Four-channel 100-Gb/s per channel discrete multi-tone modulation using silicon photonic integrated circuits," 2015 Optical Fiber Communications Conference and Exhibition (OFC), Los Angeles, CA, 2015.

[8] Ethernet Alliance, 'The 2018 Ethernet Roadmap', https://ethernetalliance.org/the-2018-ethernet-roadmap/, access on Sep. 2018.

[9]P. Westbergh et al., "VCSEL arrays for multicore fiber interconnects with an aggregate capacity of 240 Gbit/s", IEEE Photon. Techn. Lett, 2014

[10]R. Lin et al., "Real-time 100 Gbps/lambda/core NRZ and EDB IM/DD transmission over multicore fiber for intra-datacenter communication networks," Optics Express, vol. 26, no. 8, s. 10519-10526, 2018.

[11]T. Hayashi et al., "125-µm-Cladding Eight-Core Multi-Core Fiber Realizing Ultra-High-Density Cable Suitable for O-Band Short-Reach Optical Interconnects," in Journal of Lightwave Technology, vol. 34, no. 1, pp. 85-92, 1 Jan.1, 2016.

[12]O. Ozolins et al., "7×149 Gbit/s PAM4 Transmission over 1 km Multicore Fiber for Short-Reach Optical Interconnects," in Conference on Lasers and Electro-Optics, OSA Technical Digest (online) 2018.

[13]FEC Codes for 400Gbps 802.3bs, IEEE 802 Group, accessed on Sep. 2018.

[14] Google Cloud, https://cloud.google.com/pricing/list accessed on Sep. 2018.

[15]Y. Cheng et al., "Resource Disaggregation versus Integrated Servers in Data Center: Impact of Internal Transmission Capacity Limitation", in Proc. of 44th European Conference on Optical Communication (ECOC), September 2018.

Biographies

Yuxin Cheng (yuxinc@kth.se) is currently a PhD student in KTH, Sweden. His research interests include optical networking and data center networks.

Rui Lin is a postdoc researcher in KTH. Her research interests include techniques for datacenter interconnect and cybersecurity.

Marilet De Andrade received her Ph.D. degree from Universitat Politecnica de Catalunya (UPC), Spain, in 2010. From 1998, Marilet has worked for several institutions such as Movistar Telefonica Venezuela (former Telcel Bellsouth) as engineer, UPC as lecturer, Politecnico di Milano and the Royal Institute of Technology (KTH) as researcher. Her research work covers a variety of topics in passive optical network evolution, resource management of broadband access networks, mobile and fixed networks convergence, software defined networks, and network function virtualization in datacenter networks. Recently, she joined Ericsson, in Kista, Sweden, as a researcher, contributing in aspects related to the 5G core standardization. Her contribution in this paper has been performed while working at KTH.

Lena Wosinska received her PhD degree in Photonics and Docent degree in Optical Networks from KTH Royal Institute of Technology, Sweden where she has been a Full Professor of Telecommunication by October 2018. Currently she is a Research Professor in Chalmers University of Technology, Sweden. She is founder and leader of the Optical Networks Lab (ONLab). She has been working in several EU projects and coordinating a number of national and international research projects. She has been involved in activities including serving in the panels evaluating research project proposals for many funding agencies, guest editorship of IEEE, OSA, Elsevier and Springer journals, serving as General Chair and Co-Chair of several IEEE, OSA and SPIE conferences and workshops. She has been an Associate Editor of OSA Journal of Optical Networking and IEEE/OSA Journal of Optical Communications and Networking. Currently she is serving on the Editorial Board of Springer Photonic Networks Communication Journal and of Wiley Transactions on Emerging Telecommunications Technologies.

Jiajia Chen (jiajiac@kth.se) is currently an associate professor at KTH Royal Institute of Technology, Sweden.