Analysis of Service Blocking Reduction Strategies in Capacity-limited Disaggregated Datacenters

Albert Pagès[1, *], Fernando Agraz[1], Salvatore Spadaro[1]
1: Optical Communications Group (GCO), Universitat Politècnica de Catalunya (UPC), Barcelona, Spain
**e-mail address: albertpages@tsc.upc.edu*

Abstract: Disaggregated DCs offer multiple benefits. However, transmission capacity limitations at blade level can severely degrade their performance. We analyze several strategies to enhance their service acceptance. © 2020 The Authors

1. Introduction

Datacenter (DC) infrastructures are a key element for the provisioning of network services (NSs), especially in softwarized/virtualized environments, in which several of the network functions are deployed in the form of virtual resources [1] (e.g. Virtual Machines (VMs) in DCs). Then, connectivity across the deployed functions is achieved thanks to DC network (DCN) fabrics. In regards of the allocation of computational resources (CPU cores, memory, storage), traditional integrated DC architectures, in which resources are hosted at integrated servers, have been shown to face limitations. Due to the poor modularity of servers, computational resources may not be fully exploited, leading to server underutilization and, ultimately, VM allocation blocking. The adoption of disaggregated DC (DDC) architectures [2] may overcome such limitation. In DDCs, computational resources are distributed across blades, which then are mounted at rack level. Such architecture enhances the modularity of DCs, since computational resources are no longer tied to a specific server. Then, connectivity requirements across blades are achieved thanks to high performance optical networks.

However, DDCs are not exempt of limitations, being the main one the transmission capacity at the blades [3]. Besides computational resources, to allocate a VM it is necessary to also provision connectivity across them. In integrated architectures, this is achieved through the motherboard while on DDCs this is achieved through the intra-DCN. This imposes a high utilization of the network resources, since huge capacities are required between provisioned blades. This may result in the exhaustion of the network resources, leading to higher service blocking. Nevertheless, recent works have shown that this effect can be compensated by either increasing the number of blades per rack or the number of optical channels per blade [4].

However, such solutions suffer in terms of scalability and they may become too expensive, since substantial extra hardware may be required. In light of this, it may be possible to alternatively approach the problem by relaying on control layer-based strategies for the provisioning and re-allocation of connectivity between blades; however, at the expenses of a higher burden for the control layer.

2. Datacenter architecture and service blocking reduction strategies

Following the architectures presented in [2], [5], Figure 1 depicts the assumed DDC. Multiple specialized blades (CPU cores, memory, storage) are hosted at trays and placed within a rack. To achieve the inter-blade connectivity, each of the trays is equipped with an Edge of the Tray (EoT) switch, which interconnects the blades. Then, the several EoTs are connected to a Top of the Rack (ToR) switch, which provides the inter-trays connectivity. Finally, ToRs are connected to a switch, which then is connected to other switches to offer connectivity across clusters.

As said before, substantial capacities are required between blades (several hundreds of Gb/s for CPU-to-memory and up to tens of Gb/s for CPU-to-storage) [6]. For this, optical technologies are assumed for the intra-DCN. In addition, a parallel electronic packet switching (EPS) network is also considered. This stems from the disparity of capacity requirements. To materialize a VM, it is necessary to connect the assigned CPU blade with the assigned memory and storage blades, respectively. Hence, CPU-to-memory connectivity will be achieved by the optical DCN, satisfying its high capacity requirements, while the lower requirements of CPU-to-storage connectivity will be achieved thanks to the EPS fabric. As such, each blade is equipped with both optical and electrical interfaces connected to an optical or electronic EoT. Then, the EoTs connect to the corresponding network infrastructure.

It is in such architecture in which we assume the placement of NSs. A NS consists on multiple VMs to be deployed at the blades, with some connectivity requirements between VMs. Such connectivity requires that the CPU blades of the multiple VMs are interconnected, following the service virtual graph (Figure 1 depicts an example of NS allocation). These connectivity requirements will be either allocated through the optical or the electrical network depending on the inter-VM capacity requirements. In this regard, we assume that connectivity requirements that are

below the electronic interface capacity at the blades will be served by EPS whereas the other cases will be served through the optical network. Despite that, because a significant portion of both network fabrics is consumed to materialize a single VM, it may be impossible to allocate the connectivity requirements between VMs. As a result, NSs may be blocked due to the lack of network resources, either for intra- or inter-VM networking requirements.

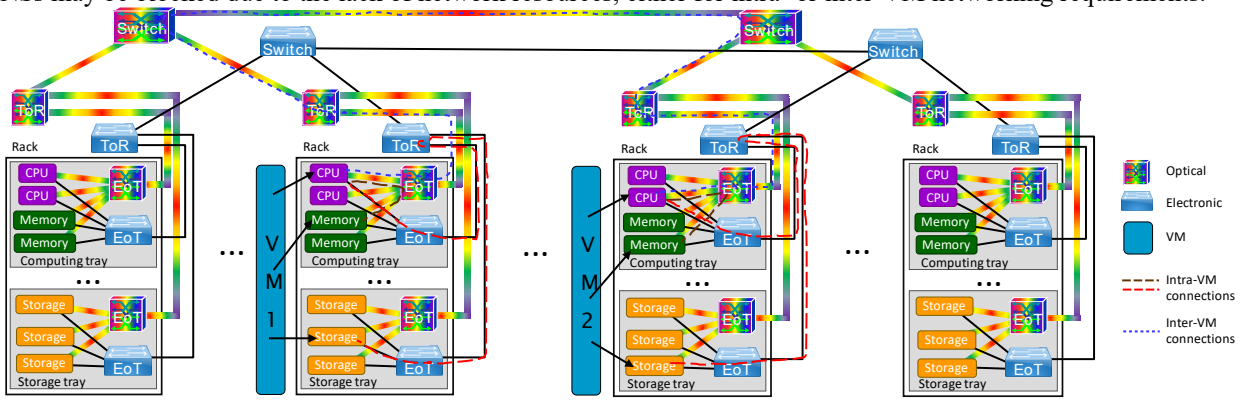

Fig. 1. Assumed hybrid disaggregated datacenter architecture.

In light of this, it is essential to devise strategies to overcome transmission capacity limitations at blade level. In [4], authors demonstrated that these limitations may be overcome thanks to proper dimensioning of the hardware resources. Analyzing their relationship with NS blocking, it becomes evident that the lack of network resources is the prime factor towards NS blocking. As such, one strategy involves the allocation of extra network resources at the blades to better digest the capacity requirements for blade-to-blade communications. This translates to having both extra optical and electronic interfaces at the blades. While such strategy allows drastically reducing the NS blocking, it also incurs in significant extra capital expenditures (CAPEX). As an alternative, it may be viable to deploy extra blades at the trays, since each extra blade will be equipped with their corresponding network interfaces. Indeed, this effectively increases the overall network capacity of the intra-DCN fabric at blade level, nevertheless with again additional CAPEX impact. In order to compensate for that, authors showed that less performant blades in terms of computational resources may be deployed, although at higher numbers, to compensate for the higher CAPEX.

Moving away from these hardware-based reduction strategies, control layer capabilities may be exploited to lower NS blocking. Indeed, it is necessary to have some control/management layer to provision both resources blades and network connections. By further exploiting the control layer capabilities', it may be possible to monitor the network utilization at blade levels. When reaching critical utilization levels, data-path re-allocation techniques may be triggered to re-allocate connections between blades. In this regard, we define a blade saturation metric as the percentage of electronic bandwidth used and the standard deviation of the utilization of optical channels. This is motivated from the fact that network blocking at the blades comes from insufficient electronic capacity or the unbalanced utilization of wavelength channels due to the dynamic regime of NSs arrival, leading to optical capacity per channel fragmentation. Once the blade saturation reaches a defined threshold, data-path reallocation techniques are triggered. On one hand, to alleviate electronic capacity saturation, EPS connections may be offloaded to optical channels. In combination to that, optical connections may be re-allocated to other channels in order to consolidate the available capacity, keeping as much fully free optical channels as possible. This can help on reducing NS blocking, however, with an increased operational expenditure (OPEX) cost due to the extra operations that need to be supported. Next section evaluates how the different strategies allow for reducing NS blocking.

3. Test scenarios, results and discussion

The employed test scenario consists on a DDC of three clusters, with each cluster consisting in 8 racks. Then, in all racks, it is assumed that 2 computing trays (CPU and memory blades) and 2 storage trays are present. The ToRs (both electronic and optical) are connected to a central switch in a tree fashion, one per cluster, with the switches being connected in a ring. For the baseline scenario, it is assumed that the trays are equipped with 6 blades of the corresponding type with 24 CPU cores, 64 GB of memory and 2 TB of storage. In addition, we assume that each blade is equipped with 16 optical channels at 300 Gb/s and one electrical interface at 10 Gb/s.

We assume the dynamic arrival of NSs allocation requests following a Poisson distribution with exponentially distributed mean holding and inter-arrival times (HT, IAT). Each NS consist in 2 interconnected VMs randomly requesting one of the resources profiles in [7]. Capacity requirements between VMs are uniformly distributed between 1-50 Gb/s. We also assume that the required capacities for connecting the blades within a VM are of 100 Gb/s and 1 Gb/s for the CPU-to-memory and CPU-to-storage communications, respectively.

We analyze the experienced NS blocking probability (BP) for increasing loads. To this, we compare the baseline scenario with three different strategies. For the first one, labeled "network-based (NB)", extra network resources per blade are equipped, specifically each blade will be equipped with 20 wavelength channels and 2 electrical interfaces. For the second strategy, labeled "blade-based (BB)", each tray is equipped with an increased number of blades, that is, 9 blades per tray, with characteristics being 20 cores, 54 GB of memory and 1700 GB of storage. The last strategy, labeled "control-based (CB)", keeps the same physical resources as in the baseline scenario and employs the re-allocation strategy previously described. In this regard, we explore two values for the blade saturation threshold (T), namely 0.2 and 0.6. With this, Figure 2 (left) depicts the obtained results, considering 10^5 NS arrivals per data point.

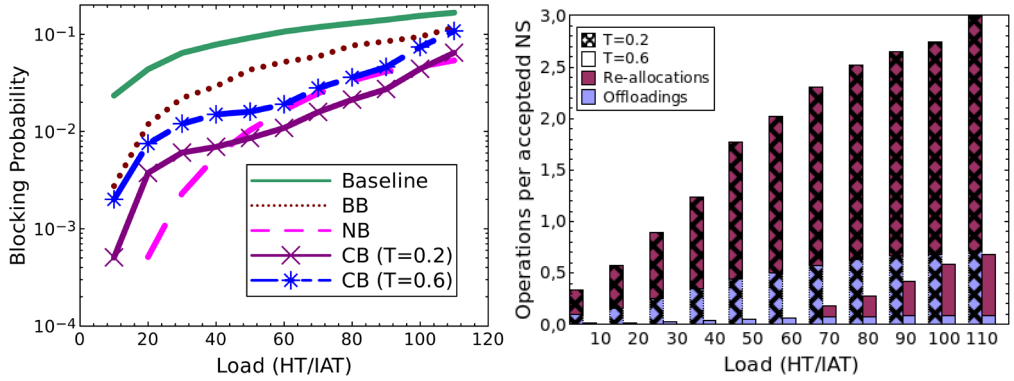

Fig. 2. BP as a function of the load (left); Offloading and re-allocation operations per accepted NS (right).

It can be appreciated that the different strategies reduce significantly the BP. Both hardware-based strategies achieve up to around one and two orders of magnitude reductions for the BB and NB cases, respectively. This is due to the extra hardware that is equipped, which allows to bypass the network resources bottleneck. However, substantial increases on hardware resources are needed. For instance, for the NB strategy, 4 more wavelength channels and one more electrical interface per blade is needed, impacting the CAPEX costs in large DDCs. We can see that the CB strategy is also able to achieve significant reductions, similar to the ones achievable through hardware-base strategies, and even achieve slightly lower BPs for some loads. This is due to the fact that it allows for a better network utilization through re-arrangement of the connections, without requiring any extra hardware.

However, such capabilities come to a price, which translates to more complex operations at the control layer. In order to analyze this, in Figure 2 (right) we depict the average number of NSs that some of its connections have been offloaded or re-allocated per accepted NS. It can be appreciated that low levels of the threshold translate to a significant number of operations, since the mechanism is triggered early; however, with better BP figures. On the other hand, with more conservative thresholds (T=0.6), the number of operations required is substantially lower and is still able to maintain good performance in terms of BP, without significant added CAPEX costs.

4. Conclusions

In this paper, we analyzed several strategies to reduce service blocking in disaggregated datacenter infrastructures. We showed that control-based strategies are a promising candidate to overcome such limitations. Thanks to connection re-arrangements between blades, it is possible to alleviate bandwidth starvation and increase service acceptance to comparable levels as if deploying extra hardware. In addition, the extra required operations to achieve this may be kept at minimum by properly tuning the reallocation strategy, achieving good trade-offs between complexity and blocking reductions. *The presented work has been supported by the Spanish Government through project ALLIANCE-B (TEC2017-90034-C2-2-R) with FEDER contribution.*

5. References

[1] ETSI, "Network Functions Virtualisation – White Paper on NFV priorities for 5G", White Paper, Feb. 2017.
[2] A. Peters, et al., "In Compute/Memory Dynamic Packet/Circuit Switch Placement for Optically Disaggregated Data Centers', OSA J. Opt. Commun. Netw., 10 (7), pp. B164-B178, July 2018.
[3] Y. Cheng, et al., "Resource Disaggregation versus Integrated Servers in Data Centers: Impact of Internal Transmission Capacity Limitation", in Proc. ECOC 2018, Sept. 2018.
[4] A. Pagès, et al., "On the Impact of IT Resources Disaggregation in Optically Interconnected Data Centres", in Proc. ECOC 2019, Sept. 2019.
[5] X. Guo, et al, "Performance Assessment of a Novel Rack-scale Disaggregated Data Center with Fast Optical Switch", in Proc. OFC 2019, March 2019.
[6] P. X. Gao, et al., "Network Requirements for Resource Disaggregation", in Proc. OSDI 2016, Nov. 2016.
[7] "OpenStack – Manage flavors", [online]: https://docs.openstack.org/horizon/latest/admin/manage-flavors.html, accessed 1st Oct. 2019.