Optical Interconnection of CDN Caches with Tb/s Sliceable Bandwidth-Variable Transceivers featuring Dynamic Restoration

G. Otero, D. Larrabeiti,
J. A. Hernández, P. Reviriego
Universidad Carlos III de Madrid, Spain
{goterop, dlarra, jahgutie, revirieg}@it.uc3m.es

J. P. Fernández-Palacios, V. López
Telefonica Global CTO, Madrid, Spain
Email: {jpfpg, vlopez}@tid.es

M. Svaluto Moreolo, J. M. Fabrega
L. Nadal and R. Martinez
CTTC, Castelldefells, Spain
Email:{michela.svaluto, jmfabrega,
laia.nadal, ricardo.martinez}@cttc.es

***Abstract*—A scalable cost-effective solution for high service availability featuring low latency is required for CDN caching. Recently, a field of study is that MAN data centers, within the latency budget, act as a backup for other local data centers of lesser reliability. Given the low latency target of caching, the optical layer is the preferred option to interconnect caches. However, carrying the backup traffic from one data center (DC) to another with a permanent optical circuit based on Fixed Transceivers (FT) features low utilization and no statistical multiplexing gain on the path, which makes the backup network resources costly. In this paper we compare several approaches to implement this scenario with dynamic circuits, considering both inter-cache and backup traffic with FTs featuring both permanent and switched optical circuits, and with the Tb/s sliceable bandwidth-variable transceivers (S-BVT) developed in the EU project PASSION. As we show, S-BVTs can be key devices to improve backup-network scalability in terms of IT resources and transceivers, thanks to their capability to adapt to the actual traffic demand and to obtain multiplexing gains at the optical layer.**

Keywords*—*Sliceable Bandwidth-Variable Transceiver, EU project* PASSION*, Edge Computing, Data Center Protection, Metropolitan Area Network,* CDN *backup

I. Introduction

Telecommunications operators (aka telcos) are concerned about the cost and scalability of the upcoming multiple edge computing capabilities such as CDN (Content Delivery Networks) caching, MEC (Mobile Edge Computing) or NFV (Network Function Virtualisation) schemes running on edge cloud architectures such as CORD (Central Office Re-Architected as a Data Center, (https://opencord.org/)). However, providing carrier-grade data center services means upgrading the numerous edge facilities of a telecom operator with costly redundant computing, storage and communication equipment, as well as dual power supply and air conditioning. Given the large amount of network edges, the only scalable solution for high service availability seems to be making remote data centers backup other data centers of usually lesser reliability. This is the case of hierarchical CDN caching, where we focus this study, although the use case is generalisable to any other edge computing service that needs low-latency communication with another server (e.g. augmented reality). A CDN cache can save a lot of traffic in the core but it needs a certain permanent connectivity for edge cache update from a cache at a higher hierarchical level, which may also take over the role of the edge cache in the event of data center outage. It should be noted, that backing up a whole data center with another assumes that the communication equipment necessary to switch the traffic over to another data center has its own protection mechanisms and remains up and running while the local data center is down. This is a major challenge to be addressed by MAN network designers and service engineers due to the amount of capacity to be provided at a given edge of the network during the edge data cloud outage time.

One way to deal with both inter-cache traffic and backup traffic is the IP layer. However, using IP routers equipped with fixed transceivers (FT) to move the whole data traffic of a CO from one point to another may not be the most effective approach, given the low utilization of the backup capacity and the additional queuing latency. Given the ultra-low latency targets of caching and other edge computing services, the optical layer is the preferred technical option to interconnect caches. On the other hand, carrying the backup traffic from one data center (DC) to another with a permanent optical circuit based on Fixed Transceivers (FT) features ultra-low latency but low utilization and no statistical multiplexing gain on the path, which makes the backup network costly. In this paper we compare several approaches to implement this scenario, considering both inter-cache and backup traffic with FTs featuring both permanent and dynamic connections, and high-capacity sliceable bandwidth-variable transceivers (S-BVT). As we shall show, S-BVTs can be key devices to improve backup-network scalability in terms of IT computing resources and transceivers, thanks to their capability to adapt to the actual traffic demand and to obtain multiplexing gains at the optical layer. In the next section we introduce the architecture and functionality of the target S-BVT design used in Section III to perform the comparison.

II. Sliceable Bandwidth-Variable Transceivers (S-BVTs)

Sliceable Bandwidth-Variable Transceivers have been proposed recently to support multiple flow transmission. They are able to serve a high number of users and to leverage the void fibre bandwidth in an efficient way (Flexi-grid). Sliceable variable allocation enables us not only to move the traffic from a failing data center to another backup node but also to quickly populate the caches at the startup phase or in case of a failure in local storage. We can do this without having to provision

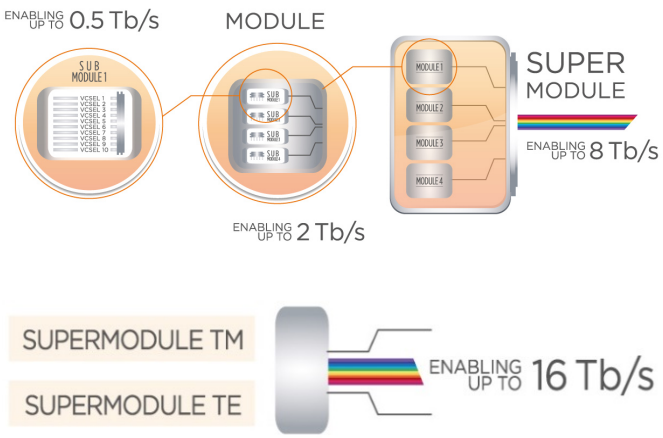

Fig. 1: S-BVT modular design at the transmitter according to PASSION project approach: sub-module, fundamental module and maximum capacity aggregation for single and dual polarization

a permanent capacity to interconnect local and centralized caches. Once that the caches reach their steady state, the provided capacity can be released and further allocated to more convenient needs.

Vertical-cavity surface-emitting laser (VCSEL) technology and dense photonic integration appear as good candidates to build cost-effective S-BVTs [7] with many applications in MAN scenarios and, in particular, for edge computing node interconnection. The S-BVT being designed in the framework of EU project PASSION [8] is equipped with direct modulated large bandwidth (up to 20 GHz) VCSELs operating at long wavelengths, with an integrated modular design, able to support terabit capacities [1]. This S-BVT has been conceived to address the challenges of agile and high capacity future optical metro networks, and distributed computing is one of the main target applications for such network segment. The S-BVT fundamental module at the transmitter integrates 40 VCSELs on a single chip. Each VCSEL is directly modulated to achieve up to 50 Gb/s , as shown in Figure 1, and operates at a different wavelength (within the C-band). Thus, a sub-module of 10 VCSELs can provide up to 500 Gb/s and 4 such sub-modules (forming the fundamental chip) provide an aggregated flow of up to 2 Tb/s covering the C-band. Higher capacity S-BVTs can be implemented adding fundamental modules. Assuming that the minimum channel spacing is 25 GHz, the maximum S-BVT capacity at a single polarization fully exploiting the C-band is 8 Tb/s and it will be obtained including a total of 4 fundamental modules, with 2 Tb/s each, as shown in Fig. 1. The capacity of this S-BVT can be doubled to 16 Tb/s considering the polarization dimension and further enhanced by including the spatial dimension. Coherent reception, more robust to transmission impairments compared to direct detection, can be envisioned at the receiver side for extended reach connections at the expense of cost-effectiveness. The main feature of S-BVT compared to other types of transceivers [1] is its slicing capability which enables the S-BVT to host simultaneous WDM connections from other S-BVTs with a granularity of 50 Gb/s. This means that, driven by an appropriate control plane, a DC equipped with an S-BVT can be dynamically configured to send/receive backup traffic from other DCs simultaneously with a single transceiver [2].

III. SBVT-BASED NETWORK ARCHITECTURE AND COMPARISON WITH FIXED-TRANSCEIVER SOLUTIONS

In this section, we compare the benefits and drawbacks of three different architectures for handling DC downtimes/failures and show how S-BVTs can bring significant improvements by using dynamic circuits that are set up in the event of a DC failure. The network is structured in two levels such that local DCs can serve most of the traffic (assumed to be 1 Tb/s as a target case) using local caching. This target traffic corresponds to 70,000 active subscribers of individual IPTV contents watching a 15Mb/s 4K-video (Netflix recommended reservation rate) attached to the edge node, accessing content available at their local cache. On the other hand, caches are connected to a higher hierarchical level that serves the contents not available locally (assumed to be 100 Gb/s in our target example).

The architectures are shown in Fig.2. The first one is a pair-wise backup system (see Fig.2, **Scenario A**). The traffic of each metropolitan area is served by a local cache running at the CO's DC on a first attempt. In case of a failure, demands can be satisfied using the IT available resources (i.e., VMs, storage, etc) in the paired DC. These can be satisfied using the IT resources available in the paired DC over a 1 Tbps dynamic circuit. If the latter is not possible due to a failure in the backup DC, the traffic is lost, as not enough capacity is provisioned toward the core for the complete traffic demand, that is, no 1Tbps connectivity is foreseen from edge to core and only 100 Gbps of would be supported. This option provides low latency but comes at a high cost in terms of IT resources as each local DC needs to provision resources to handle the failure of its pair.

In the second approach, **Scenario B**, each local DC has its backup on a central DC, using an independent dynamic

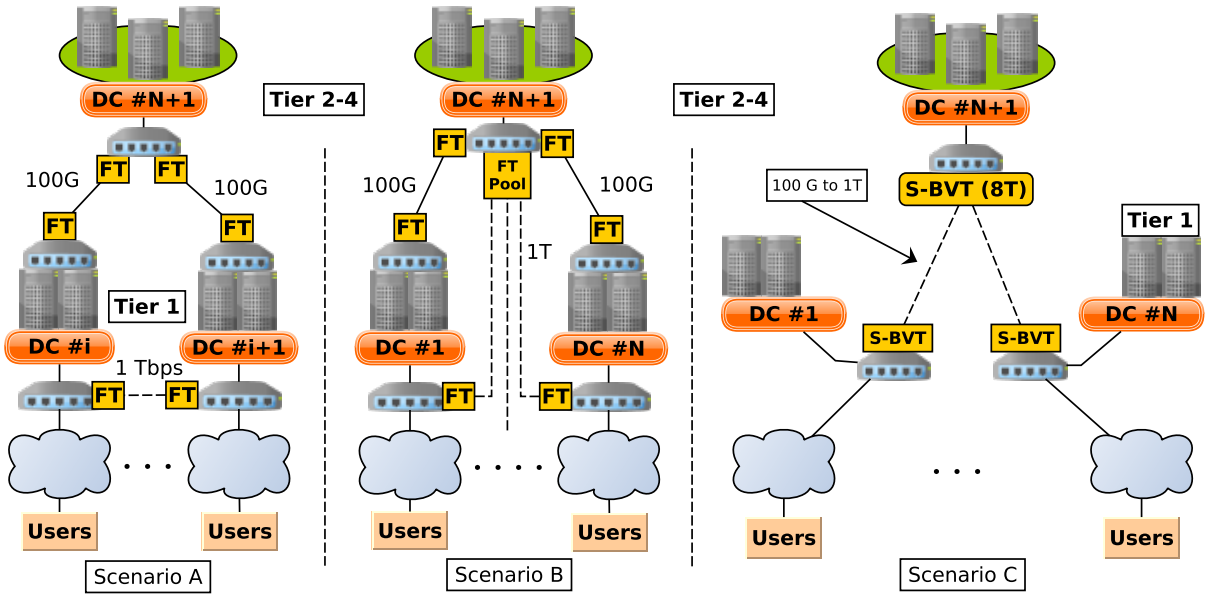

Fig. 2: Architecture diagrams for dynamic restoration (A) Pair-wise backup, (B) Hierarchical backup with FT,(C) Hierarchical backup with S-BVT (C). Dashed lines represent dynamic connections and the solid lines refer to the permanent ones.

connection that is set up upon a DC failure. This however implies a larger latency. From the cost point of view, this scenario benefits from statistical multiplexing and can significantly reduce the IT resources needed to support the backup as discussed in the following. However, a major drawback is that a large number of optical circuits and transceivers are needed when fixed transceivers are used. The third option in Fig. 2, **Scenario C**, implements also a backup to a central node but using S-BVTs. This enables significant reduction of both optical circuits and transceivers, as well as flexibility to assign additional bandwidth to DCs when needed.

In the rest of the section we analyze the probability of failure for the three options and compare the IT and optical resources needed to show the advantages of the S-BVTs centralized architecture. Let us next use the Tier classification of DCs [3] fostered by the Uptime Institute, the data center classification standard most adopted by IT industry. For the sake of cost, we shall assume that edge DCs are the simplest data center infrastructure considered by this standard: Tier 1, and the most sophisticated DC technology is in place in the core, i.e. Tier 4. Let $P_{\text{failure}_{\text{T1}}}$ be the failure probability of a Tier 1 DC estimated as the unavailability of a Tier 1 DC (99.671%). Then, the probability of not being able to serve the traffic of any city of the pair in Scenario A is $P_{\text{No service A}} = P^2_{\text{failure}_{\text{T1}}}$. Considering a number of N cities and $\frac{N}{2}$ pairs of DCs, we would need to double the resources at each DC $\#i$, $\forall i \in [1, N]$, in order to have enough resources to cope with the traffic of two cities, just in case one of the DCs in the pair fails.

On the other hand, Scenarios B and C (see Fig.2, middle and right) represent a centralized backup architecture. Here, when a local DC fails, the traffic demand is directed to a higher tier DC. This applies to every single local DC in the lower tier. Let $P_{\text{failure T1}}$ and $P_{\text{failure T4}}$ be the failure probabilities of a Tier 1 and Tier 4 data center, respectively. In this context, the probability of being unable to find available resources for user's requests of any city in Scenario B can be expressed as $P_{\text{No service B,C}} = P_{\text{failure T1}} \cdot P_{\text{T4 Not available}}$, where $P_{\text{T4 not available}}$ can be computed as

$$P_{\text{T4 not available}} = P_{\text{failure T4}} + P_{\text{Not enough resources T4}} - \\ - P_{\text{failure T4}} \cdot P_{\text{Not enough resources T4}}. \quad (1)$$

In order to compute $P_{\text{Not enough resources T4}}$, we need to have in mind that N Tier 1 DCs can potentially fail. Also, we assume that a given number of IT resources (M) are available to cope with the incoming traffic, each one equivalent to the resources of one Tier 1 DC. Substituting the appropriate expressions and rearranging them we get:

$$P_{\text{No service B,C}} = P_{\text{failure T1}} \Big[P_{\text{failure T4}} + \\ + \sum_{i=M+1}^{N} \binom{N}{i} P^i_{\text{failure T1}} (1 - P_{\text{failure T1}})^{N-i} (1 - P_{\text{failure T4}}) \Big] \quad (2)$$

From here, we may compute how many resources (M) we need in the Tier 4 DC in order to achieve the same overall service availability as that of Scenario A by solving $P_{\text{No service B,C}} = P_{\text{No service A}}$. Let us consider that the Tier 1 DC's availability is 99.67% and that of a Tier 4 DC is 99.99% [3]. Also, assume that we want to dimension both scenarios to support $N = 40$ metropolitan areas. Turning aforementioned availability times into no service probabilities by using $P_{\text{No service}} = 1 - (\frac{\text{Availability}}{100})$, and solving $P_{\text{No service B,C}} = P_{\text{No service A}}$ we get that the number of resources that we need at the Tier 4 DC driven by the S-BVT is $M = 3$. Therefore, in Scenario A, we would need a total number of $2 \cdot N = 80$ IT resources while, in Scenarios B and C, we would need $N + M = 43$ IT resources (i.e., DC's resources). This represents a total saving of $\simeq 46\%$

		Scenario A	Scenario B	Scenario C
IT Resources (Availab.$_{T1} = 99.67\%$; Availab.$_{T4} = 99.99\%$)		$2 \cdot N = 80$	$N + M = 43$	$N + M = 43$
IT Resources (Availab.$_{T1} = 99.67\%$; Availab.$_{T4} = 99.75\%$)		$2 \cdot N = 80$	$N + M = 43$	$N + M = 43$
Number of Fixed Transceivers	100G	$2 \cdot N = 80$	$2 \cdot N = 80$	-
	1T	$2 \cdot \frac{N}{2} = 40$	$N + M = 43$	
Number of Bandwidth Variable Transceivers	S-BVT (2T)	-	-	$N = 40$
	S-BVT (8T)			1
Wavelength occupancy		*Fixed*	*Fixed*	$\propto$ *Load*
One-Way propagation delay of backup path		43.5 μs	125 μs	125 μs

TABLE I: Comparison of resources for optical connectivity+restoration architectures for $N = 40$ edge CDN caches

in the required resources. The same reasoning applies to the number of transceivers in the central DC. Table I summarizes the comparison between the two architectures, regarding the amount of required IT resources and the needed fixed and variable bandwidth optical transceivers.

The number of required IT resources in Scenarios B and C to meet the same availability as in Scenario A is dramatically reduced by means of statistical multiplexing. We include an extra case where Scenarios B,C are supported by a lower Tier DC. Observe that having a Tier 4 DC in the upper level does not reduce the number of IT resources we need to provision compared to staying with a cheaper Tier 2 DC. Also, we show the required hardware to implement each architecture. It is worth highlighting that the amount of needed transceivers does not scale well with the number of edge CDN nodes for Scenarios A and B. Conversely, we achieve savings in the number of needed transceivers for Scenario C by using N S-BVTs at the local DCs and one S-BVT at the core. Furthermore, for the example considered, the S-BVT at hte central site with 140 channels configured at 50 Gb/s each, can provide 7 Tb/s and the S-BVTs at the edges can provide 1 Tb/s by enabling half of the VCSELs. The amount of required FTs w.r.t. S-BVTs gives an idea of how much more costly an S-BVT can be with respect to a FT for Scenario C being a more cost-effective choice than B. From the wavelength occupancy point of view, the S-BVT is the best choice as it can fit the real load of the network with finer granularity (50 Gb/s). Finally, the last row of Table I shows the expected latency for each architecture having in mind the average distances between levels of aggregation of a real world metropolitan network deployment reported in [8] and assuming a delay of $5\mu s/Km$. Despite the fact that Scenarios B and C suffer from a higher delay, 125 μs is not a heavy burden for most applications.

IV. Conclusions

The only scalable solution for high service availability of low latency CDN caching is making other MAN data centers within the target latency budget backup other data centers of usually lesser reliability. In this paper, we compared several approaches to implement this scenario using dynamic optical protection circuits, considering both inter-cache and backup traffic both with FTs and with the PASSION Tb/s sliceable bandwidth-variable transceivers (S-BVT) to build flexi-grid channels. As we showed, S-BVTs can be key devices to improve backup-network scalability in terms of IT resources and transceivers, thanks to their capability to adapt to the actual traffic demand and to obtain multiplexing gains at the optical layer.

The advantage of sliceable variable allocation of bandwidth through the network is clear not only when there is a need to move the traffic from a failing data center to another backup node. It is also an essential capability to quickly populate the caches when they are first started or after a general storage failure, without having to devote a large amount of permanent capacity to interconnect edge and central caches. Once the caches are updated, the capacity of the network and the transceivers are released and are available for other purposes.

Acknowledgment

UC3M authors would like to acknowledge the support of the Spanish project TEXEO (grant no. TEC2016-80339-R) and TAPIR-CM (grant no. P2018/TCS-4496). Telefonica and CTTC would like to acknowledge the support of EU project PASSION (grant no. 780326).

References

[1] M. Svaluto Moreolo et al., "VCSEL-based sliceable bandwidth/bitrate variable transceivers," SPIE-PWO, San Francisco, CA (USA), Feb. 2019.

[2] R. Martinez et al., "Proof-of-Concept validation of SDN-controlled VCSEL-based S-BVTs in flexi-grid optical metro networks," in Proc. OFC 2019, San Diego, CA (USA), March 2019.

[3] R. Arno et al., "Reliability of Data Centers by Tier Classification," in IEEE Trans. Industry Applications, 48(2), 2012

[4] O. Gonzalez de Dios et al., "First Demonstration of Multi-vendor and Multi-domain EON with S-BVT and Control Interoperability over Pan-European Testbed," ECOC 2015.

[5] M. Svaluto Moreolo et al., "Modular SDN-enabled S-BVT Adopting Widely Tunable MEMS VCSEL for Flexible-Elastic Optical Metro Networks," OFC 2018.

[6] C. Xie et al., "Single VCSEL 100-Gb/s short reach system using discrete multi-tone modulation and direct detection" OFC 2015.

[7] M. Svaluto Moreolo, et al., Exploring the Potential of VCSEL Technology for Agile and High Capacity Optical Metro Networks, in Proc. ONDM 2018, 14-17 May 2018, Dublin (Ireland).

[8] EU PASSION PROJECT: Photonic technologies for progrAmmable transmission and switching modular systems based on Scalable Spectrum/space aggregation for future agIle high capacity metrO Networks (EU H2020 KET Project). URL: http://www.passion-project.eu, last access: Feb. 2019.

[9] G. O. Pérez, J. A. Hernández and D. Larrabeiti, "Fronthaul network modeling and dimensioning meeting ultra-low latency requirements for 5G," in IEEE/OSA Journal of Optical Communications and Networking, vol. 10, no. 6, pp. 573-581, June 2018. doi: 10.1364/JOCN.10.000573