

The Role of Metro Transport Node Architectures in Optimized Edge Data-Center Dimensioning

António Eira¹, João Pedro^{1,2}

¹ Infinera Unipessoal Lda, Rua da Garagem 1, 2790-078 Carnaxide, Portugal

² Instituto de Telecomunicações, Instituto Superior Técnico, Avenida Rovisco Pais 1, 1049-001 Lisboa, Portugal
{AEira, JPedro}@Infinera.com

Abstract—The combination of traditional optical transport network paradigms with the advent and necessities of edge computing opens up new challenges for optimizing both the optical layer and the data-center (DC) infrastructure. In the specific case of ring/horseshoe metro access topologies, the flexibility to divert services to a specific DC with low latency is highly dependent on the node architecture that is employed. Consequently, the optimal trade-off between consolidating services in larger DCs and the number of transponders required for transporting traffic will shift depending on the possibilities enabled by the network itself. Furthermore, traditional north/south traffic to/from the core is still expected to be significant in these topologies, providing opportunities to leverage existing optical bandwidth for improved cost effectiveness. This process requires a careful dimensioning in order to optimally consider all impacting factors and select which services to assign where. We model this problem with a mixed integer quadratically constrained problem (MIQCP) that optimizes DC resources for a given availability level according to services' mean and peak loads. This optimization accounts for the interdependencies with the required optical transponder costs and the constraints imposed by service latency, as well as the specific restrictions imposed by optical nodes based on fixed/reconfigurable add/drops, or filterless solutions. The resulting analysis not only provides optimal solutions to specific network/DC planning instances, but also identifies general deployment guidelines for metro access networks in terms of desired node architectures for each network scenario.

Keywords—optical transport networks, network optimization, edge computing

I. INTRODUCTION

The support of new service profiles in the 5G ecosystem is changing how access and transport networks are architected. Rather than built merely with bulk capacity in mind, transport networks must also evolve to provide flexibility in handling diverse service requirements. A key illustration of this shift is the growing importance of jointly designing the optical transport and data-center (DC) infrastructure, actively shaping traffic patterns to achieve the optimal trade-offs between bandwidth and processing/storage resources [1]. In the metro and access space, edge DCs are a vital solution to, on one hand, meet the latency requirements of more sensitive applications, and on the other hand alleviate the massive transport bandwidth needed to process services entirely in core DCs [2].

Metro transport networks, particularly those closest to the access segment, are uniquely positioned to reach an efficient compromise for co-location of DC sites in terms of proximity to traffic sources and scale of the DCs. Consolidation of DC sites is especially appealing due to the fixed costs that can be saved [3], but also because it enables higher statistical multiplexing gains that can be exploited to reduce overprovisioned resources [4]. Unlike DC interconnection networks in the core segment, metro/access topologies do not typically possess highly meshed topologies that enable the straightforward any-to-any connectivity needed to optimize DC dimensioning. Existing topologies based on chains or rings, and legacy optical transport nodes based on fixed optical add/drop multiplexers (FOADMs) were generally built to support capacity between tributary access points to a central core-facing node. The metro network/node architecture thus directly impacts how efficiently edge DCs can be dimensioned, and how optical bandwidth and IT resources can be traded-off in joint network/DC design [5].

In this paper, we evaluate the joint dimensioning of the metro transport network and the DC infrastructure, assuming metro networks based on either FOADM, filterless or reconfigurable OADM (ROADM) nodes. The optimization relies on a mixed integer quadratically constrained problem (MIQCP) model that solves planning instances supporting a mix of legacy traffic and edge DC bound services. It optimally balances optical and IT costs, consolidating services in the same DCs based on their load and variability, which reduces the overhead IT resources needed to ensure a given availability at the potential expense of additional optical bandwidth. The application of this model to various network scenarios and architectures also identifies key design strategies and trade-offs for the deployment of optical metro transport networks with co-located edge DCs.

II. NETWORK ARCHITECTURES AND PROBLEM STATEMENT

A. Metro Transport Node Architectures

The basic architecture assumed in this work comprises a wavelength-division-multiplexed (WDM) network with a horseshoe/ring topology, where multiple tributary nodes aggregate access traffic from their coverage area, and one hub node interfaces with core network layers (Fig. 1a). The architecture of the tributary nodes defines the type and amount of channels that can be established in the network. In the case of legacy FOADM nodes, each tributary communicates with the hub node on a pre-defined set of frequencies determined by the

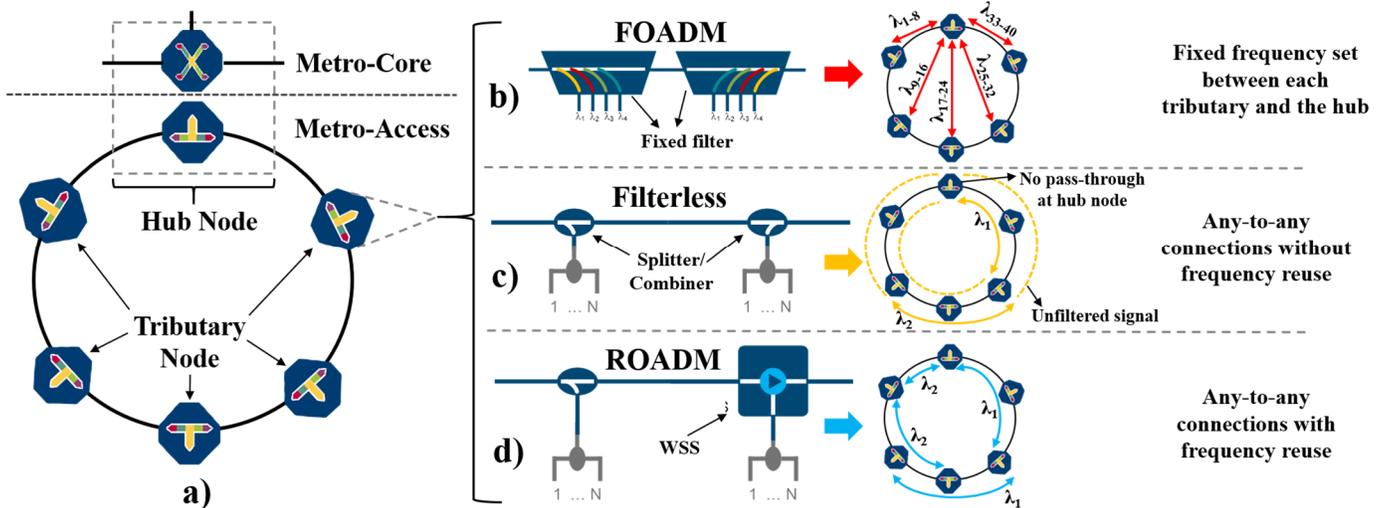


Fig. 1 – Network scenario and node architectures: (a) Network model; (b) FOADM architecture; (c) Filterless architecture; (d) ROADM architecture.

fixed filter structures deployed at each node. As Fig. 1b illustrates, transparent lightpath connections between tributary nodes are not possible. Therefore, flows between tributaries must be routed via the hub on more than one lightpath, incurring additional latency and potentially extra transponders. Alternatively, the nodes may be based on filterless structures as exemplified in Fig. 1c. Here, the express layer is built solely with splitters/combiners, meaning that direct connections between any node-pair are possible, but the same frequency can be used only once in the entire network [5]. Note that the hub node must block express channels to avoid signal loops (i.e., a ring topology becomes a logical horseshoe in the filterless case). The frequency reuse restriction can be removed by resorting to ROADM structures based on broadcast-and-select nodes [5], where channels are selectively dropped/expressed on any frequency (Fig. 1d). This architecture is costliest due to the need of active wavelength selective switch (WSS) components but provides the highest capacity and adaptability to different traffic patterns. The different architectures also have an impact on the performance and reach of the optical channels, due mainly to varying degrees of filtering effects and component insertion losses. The modeling of these effects and respective channel reach data is reported in [6].

B. Traffic and Service Model

The traffic generated in the tributary nodes is split into two profiles, as outlined in Fig. 2. The first profile refers to legacy or background traffic that is forwarded from each tributary node towards the hub (and vice-versa). This traffic encapsulates

traditional fixed/wireless residential and business traffic that is forwarded outside the metro-access segment (e.g., user-to-user traffic, content delivery on remote DCs, etc.). The second profile refers to services to be processed at an edge DC within the metro access segment. These may refer to various 5G vertical use-cases (e.g. smart factories, crowdsourced video, sensor networks, etc.), which have varying requirements with respect to transport bandwidth, IT resource consumption (processing, memory and storage) and maximum latency.

In the scope of this work, the legacy traffic is defined by its load (in Gb/s) and source tributary node, and the optical metro access network must ensure each of these demands is routed to the hub node. The edge services, in turn, are characterized by a source node, a bandwidth requirement for transport to an edge DC, a maximum latency between the source node and the destination DC (accounting for propagation latency and optical/electrical conversions at each used lightpath), and the IT resource requirements to process the service at a DC.

Typically, joint optical/DC optimization attempts to minimize the amount of DCs in order to reduce all the setup costs associated to a deployment [1,3]. In the metro access space, all nodes should have some co-located mini DC capabilities to address the most latency-sensitive applications. However, even with DCs at potentially every node there is still an incentive to consolidate flows from a pure DC dimensioning perspective. In this work, we assume services must be provisioned with a guaranteed availability p . Hence, each DC must possess sufficient resources to meet the aggregate

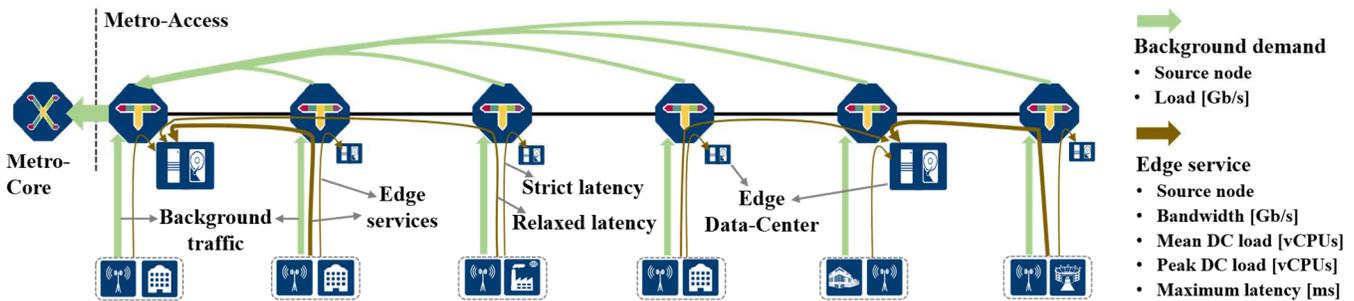


Fig. 2 – Traffic types and parameters.

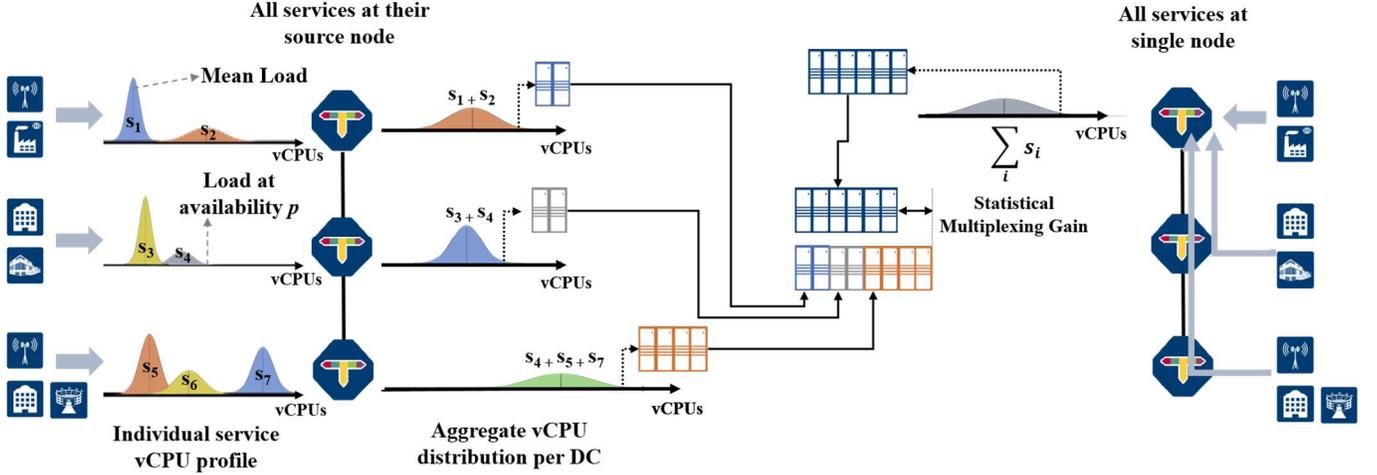


Fig. 3 – Service processing load characterization and DC dimensioning options.

distribution of the services assigned to it with probability higher or equal to p . Figure 3 showcases the edge service characterization and its role in the DC resource dimensioning. Without loss of generality, we model services through their virtual CPU (vCPU) requirements for DC dimensioning purposes in terms of needed server racks (in general memory and storage units need also to be considered). The vCPU requirements per service are modeled according to a normal distribution with mean μ and standard deviation σ . Under this simplifying assumption, the aggregate flow distribution at a DC has a mean $\sum_i \mu_i$, and a variance $\sum_i \sigma_i^2$, where i are all the individual edge services assigned to a given DC. As Fig. 3 exemplifies, consolidating services at the same DC sites yields statistical multiplexing benefits by pooling DC resources together to cover the variability of multiple services simultaneously. Figure 3 also illustrates two basic dimensioning approaches setting the extreme options for DC consolidation: if all services are handled at their respective source nodes, no extra optical bandwidth is needed, but additional DC resources are scattered at each node. Conversely, pooling all services in a single DC minimizes the global vCPU requirements, but needs optical connectivity from all nodes to the DC node. The optimal cost solution must balance these conflicting requirements, further accounting for the existing optical bandwidth serving legacy traffic and the constraints introduced by edge service latency.

III. OPTIMIZATION MODEL

As the previous Section described, the problem to be solved is quite complex due to the sheer number of interconnected parameters. The selection of which services to process where (given latency restrictions and transponder costs) is typically not trivial even for a small network with a diverse set of edge service profiles. Once services do need to be distributed over several DCs, the statistical multiplexing gains depend (non-linearly) on the specific set of services at each DC (e.g. there's a higher gain from mixing higher variance flows). Finally, the lightpaths used to transport legacy traffic may provide opportunities for piggybacking edge services at no extra transponder cost.

The cost-optimized planning is performed resorting to a MIQCP model, since it requires quadratic constraints on top of a traditional mixed-integer linear programming model to

correctly account for the variance of the sum of normally distributed services at a given DC. The objective of the model is to minimize the joint cost of transponders and vCPU resources, while serving all background traffic and assigning each service to a DC node. The model uses the following variables and parameters:

Parameters and Sets

LP_cost	Cost of a lightpath (transponder pair).
$vCPU_cost$	Cost of a vCPU unit.
$Source(s)$	Source node of service s .
Cap_l	Capacity of lightpath using path l [Gb/s].
BW_s / BW_d	Bandwidth requirement of service s / background demand d .
Lat_l	Propagation and optical/electrical conversion latency of lightpath with path l [ms].
$MaxLat_s$	Maximum allowable latency for service s .
$Src(l)/Dst(l)$	Source/Destination node of lightpath with path l .
$vCPU_s$	Mean vCPU requirements of service s .
Var_s	Variance vCPU requirements of service s .
$NChs$	Maximum number of channels per link.
$NFreq_l$	Number of reserved frequencies for lightpaths with path l .

Variables

$x_{s,n} \in [0,1]$	Binary variable equal to 1 if service s is assigned to a DC at node n .
$y_l \in \mathbb{N}^0$	Number of provisioned lightpaths on path l .
$v_n \in \mathbb{N}^0$	Number of vCPUs required at the DC in node n to support the mean load of the services assigned to it.
$z_n \in \mathbb{N}^0$	Number of vCPUs required at the DC in node n to support the variance of the aggregate services with probability p .
$w_{s,l} \in [0,1]$	Binary variable equal to 1 if service s uses a lightpath with path l .

$t_{d,l} \in [0,1]$ Binary variable equal to 1 if background demand d uses a lightpath with path l .

The model formulation is given as follows:

$$\min \sum_l y_l LP_cost + \sum_n (v_n + z_n) vCPU_cost \quad (1)$$

Subject to:

$$\sum_n x_{s,n} = 1, \quad \forall s \quad (2)$$

$$\sum_{l:Dst(l)=n} w_{s,l} = \begin{cases} 1 - x_{s,n}, & n = Source(s) \\ \sum_{l:Src(l)=n} w_{s,l} - x_{s,n}, & n \neq Source(s), \forall s, n \end{cases} \quad (3)$$

$$y_l Cap_l \geq \sum_s w_{s,l} BW_s + \sum_d t_{d,l} BW_d, \quad \forall l \quad (4)$$

$$\sum_l w_{s,l} Lat_l \leq MaxLat_s, \quad \forall s \quad (5)$$

$$\sum_{l:Src(l)=n} t_{d,l} - \sum_{l:Dst(l)=n} t_{d,l} = \begin{cases} 1, & n = Source(d) \\ -1, & n = Destination(d), \forall d, n \\ 0, & otherwise \end{cases} \quad (6)$$

$$v_n \geq \sum_s x_{s,n} vCPU_s, \quad \forall n \quad (7)$$

$$\frac{z_n^2}{erf\text{inv}(p * 2 - 1) * \sqrt{2}} \geq \sum_s x_{s,n} Var_s, \quad \forall n \quad (8)$$

The objective function (1) minimizes the sum of transponder and vCPU costs. Constraint (2) assigns each service to exactly one DC. Constraint (3) implements flow conservation for edge services, stating a service must exit a node unless it is assigned to a DC there. Constraint (4) provisions sufficient lightpaths of type l to cover the background and edge service bandwidth assigned to them. Constraint (5) ensures that a service cannot be assigned to a set of lightpaths that cumulatively exceed its latency threshold. Constraint (6) implements traditional flow conservation for the background traffic demands. Constraint (7) calculates the minimum amount of vCPUs at DC n to cover the aggregate mean of the services assigned to it. Finally, (8) is a quadratic constraint that calculates the amount of additional vCPUs at node n needed, such that there is at least a probability p that the combined service load assigned to the DC (each with variance Var_s) is supported.

Constraint (8) is a special case of a quadratic constraint that can be represented as a second-order cone program (SOCP) in the form $a = \sqrt{\sum_i b_i^2 c_i}$ [7]. The standard deviation of the aggregate flows assigned to node n is given by:

$$\sigma_n \geq \sqrt{\sum_s x_{s,n} Var_s} \quad (9)$$

This expression can be converted to the canonical SOCP form observing that $x_{s,n} = x_{s,n}^2$ (due to its binary nature), and Var_s is a known fixed coefficient. Solving for σ_n with probability p in the normal distribution error function, we obtain

the coefficient of the z_n variable in constraint (8) that models the excess resources needed to accommodate a given set of services with Var_s . This transformation is useful because it enables commercial solvers to employ SOCP specific methods that are usually more efficient than solving the general quadratic program or linearizing the constraints [7].

The model described thus far is agnostic to the specific optical node architecture used (i.e. it has no limits on channel capacity). Because we want to evaluate different architectures in parallel, the model must be augmented with architecture specific constraints for each node type. Thus, in the case of ROADM nodes, we need to impose that the number of lightpaths on any single link e cannot exceed the maximum link capacity:

$$\sum_{l \ni e} y_l \leq NChs, \quad \forall e \quad (10)$$

Note that spectrum continuity constraints are not included in the model for scalability purposes. Instead, the result of the ROADM-based scenario is post-processed with a first-fit assignment heuristic. In case of infeasibilities, the MIQCP model is executed again lowering $NChs$ by 1 channel in all the links where no feasible frequency was found. In the filterless case, it is sufficient to set an upper bound on the total number of channels in the network:

$$\sum_l y_l \leq NChs \quad (11)$$

It is not necessary to explicitly assign spectrum to channels, as a valid coloring solution is guaranteed to exist in this case. Finally, in the FOADM case we assume that the available frequencies are evenly distributed across all tributary nodes. Hence, the required constraint is:

$$y_l \leq NFreq_l, \quad \forall l \quad (12)$$

where $NFreq_l$ may vary slightly per node when the tributary node count is not a divisor of $NChs$.

In order to improve the average solving time of the model, we seed the solver with trivial initial heuristic solutions obtained by: (1) processing all flows at their respective source nodes and using direct lightpaths for the background traffic, and (2) assigning all services (not bound by latency) to each of the network nodes, provisioning the necessary lightpaths in each case (this yields an heuristic solution per network node).

IV. SIMULATION PARAMETERS AND RESULTS

The planning model was applied to a wide set of network scenarios, seeking to extract the most meaningful relationships between the different parameters, such as topology data, traffic loads, node architecture and edge service profiles. We evaluate metro/access horseshoe/ring topologies with 5, 10 and 15 nodes, assuming Gaussian distributed link lengths with mean of 25 km and standard deviation of 5 km. For each topology and node architecture combination, we test several base traffic loads, between 2 Tb/s and 10 Tb/s, where 70% of the load is background traffic randomly distributed amongst tributary nodes, and 30% are edge services generated with a uniform distribution amongst all nodes. There are on average 5 edge services generated per node, with their mean bandwidth requirement scaled to the 30% of offered traffic load divided by

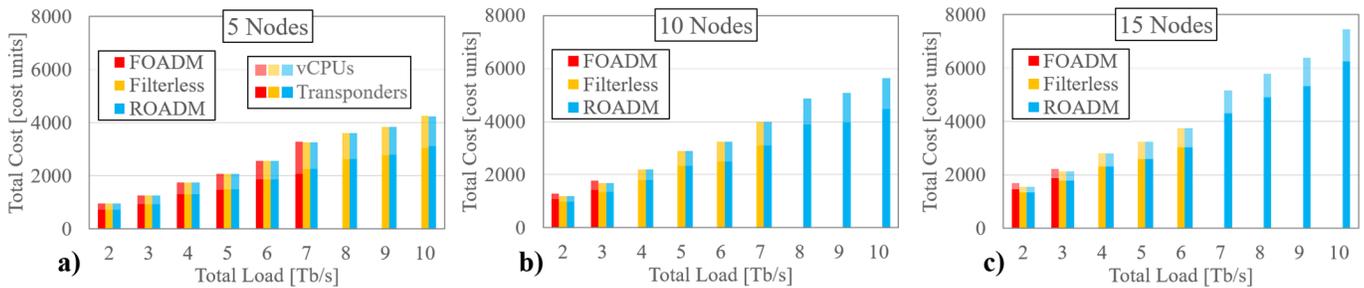


Fig. 4 – Total network cost with $LP_{cost} = 40 vCPU_{cost}$ per architecture and traffic load for: (a) 5-node network; (b) 10-node network; (c) 15-node network.

all flows. The mean vCPU requirements per flow are derived from the required bandwidth according to the service characterization data in [9], assuming one of the services is randomly selected. The standard deviation of each flow’s vCPU requirements is randomly generated from a uniform distribution between 5% and 50% of the flow’s mean vCPU load. The latency of each service follows a normal distribution with 1 ms mean and 0.2 ms standard deviation. The latency associated to each lightpath (electrical/optical/electrical conversion) is 0.1 ms. A DC is expected to have an availability level $p=0.999$ for the set of services it supports. There are 40 channels per link, which can employ either 100 or 200 Gb/s rates based on the calculated performance value. The ratio between the optimization weights LP_{cost} and $vCPU_{cost}$ is assessed with values of 40 and 4, based on cost estimate ranges from [10]. Each scenario was evaluated over 20 runs of the model with random traffic and link lengths, with average results being reported for a scenario only when at least half of those runs returned feasible solutions (i.e. there was sufficient optical bandwidth for the background traffic). The model was solved using CPLEX v12.9.

Figure 4 reports the total network cost for each architecture with increasing traffic loads in 5, 10 and 15 node networks. The costs are subdivided (different shadings) into the transponder and vCPU components. This scenario assumes a transponder costs 40 times as much as deploying one vCPU. The comparison shows that the raw capacity of the FOADM system decreases significantly as the number of nodes increases, due to the available frequencies being more thinly spread across multiple tributary nodes. The filterless architecture improves this result somewhat, mimicking the ROADM scenario for 5 nodes, but still lagging behind in achievable capacity for 10/15 nodes. With a cost ratio that assumes higher transponder costs, we find that, for feasible capacities, the comparative costs between the architectures are quite similar. The notable exceptions are FOADM architectures with 10 or 15 nodes, which exhibit on

average 6% extra cost over filterless or ROADM scenarios. In its achievable capacity range, the filterless case is very similar to the ROADM one in terms of cost, often achieving the exact same solution. The intuition is that, with higher transponder costs, minimizing the number of lightpaths is the dominant optimization focus, leading the solution towards serving background traffic with the least possible transponders, and then using whatever spare lightpath capacity is left to piggyback edge services towards the hub node. The FOADM scenario is more expensive in some cases because in filterless/ROADM architectures the model can reduce the lightpath count via selective intermediate grooming, which also opens more direct optical connectivity to nodes other than the hub that can further consolidate edge services that cannot be cost effectively routed to the hub node.

Figure 5 shows the same analysis as in Fig. 4, but this time assuming a cost ratio of 4 between transponders and vCPUs. Here, there is a much more intricate trade-off between optical and DC costs. The FOADM case can be up to 20% costlier than the alternatives. Essentially, the new cost structure implies that it is often attractive to provision lightpaths specifically to optimize the consolidation of flows, while opening up multiple larger DCs at different edges of the network to geographically cover the stricter latency applications. Hence, Fig. 5 generally shows the transponder part of costs actually being lower in FOADM scenarios than the other architectures, at the expense of much greater vCPU overprovisioning. Once again, the filterless scenario is equivalent to a ROADM network up to the point where the maximum channel count is exhausted. From thereon, the ROADM case is able to keep adding more lightpaths to reduce the vCPU overhead even further, achieving up to 16% savings over a filterless scenario for a loaded network.

Figure 6 helps illustrate how the different architectures influence the provisioning policy for edge services, showing the fraction of services that are not processed at their source node DC for a 10-node network. Almost half of the edge services are

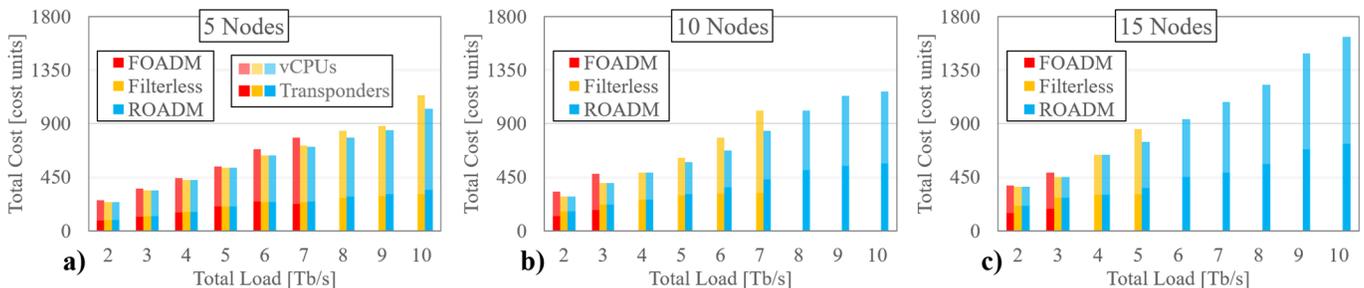


Fig. 5 – Total network cost with $LP_{cost} = 4 vCPU_{cost}$ per architecture and traffic load for: (a) 5-node network; (b) 10-node network; (c) 15-node network.

confined to their source node in the FOADM architecture due to the inability of the optical layer to provide low-latency connectivity to nearby DC nodes. The filterless architecture is on par with the ROADM scenario up to its capacity exhaustion point. Once over 40 lightpaths are (optimally) required to transport background and edge services, the filterless scenario gradually reduces the offloaded edge service share. When changing the cost ratios, the optimum service offload share jumps from around 50% to 85% by decreasing the relative weight of the transponders.

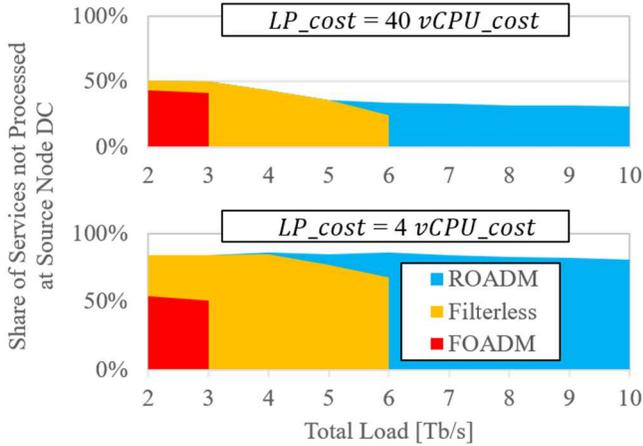


Fig. 6 – Share of services not locally processed for a 10-node network.

Another variable that may influence the comparative network design is the latency profile of the edge services. In order to evaluate this effect, we reduced the average service latency from 1ms to 0.5ms and repeated the cost optimization process. Table I shows the relative cost difference with 0.5ms average latency versus 1ms for the optical (transponder) and vCPU components, along with the total cost difference. The results shown are averaged over all the feasible loads for each scenario assuming a transponder/vCPU cost ratio of 4. There are basically two network effects contributing to the results in Table I. On one hand, with fewer nodes it is still possible to consolidate some flows, albeit much harder to do so substantially at a single DC location. Hence, the FOADM architecture suffers the most, as the hub is its only viable location for aggregating edge services. On the other hand, as the rings/horseshoes grow bigger a larger fraction of flows are latency constrained to their source node only (path lengths increase). Thus, the ability of the filterless/ROADM architectures to trade transponders for more consolidated DCs reduces, as evidenced by a general decrease in optical costs and a substantial increase in vCPU requirements.

V. CONCLUSION

This paper evaluated the relationship between the node architecture in metro-access optical transport networks and the ability to provide optimized cost deployments balancing edge DC and transponder costs. The MIQCP model provided optimal solutions for FOADM, filterless and ROADM-based networks demonstrating how introducing more flexibility in the optical nodes can lead to an improvement in DC dimensioning without unduly increasing transponder costs. The results suggest that for

higher relative transponder costs, filterless and ROADM-based designs are matched up to the frequency exhaustion point of the filterless architecture, while FOADM networks lead to larger DCs due to lack of optical connectivity. When transponder and DC costs are more balanced, there is higher demand for provisioning lightpaths specifically for transporting edge services, which further stresses the filterless network capacity, leading to substantial cost differences versus a ROADM network for higher loads and larger networks. The combination of these results with accurate modeling of the optical nodes' costs, will be useful to determine the ideal architecture for each network and traffic combination.

TABLE I. ADDITIONAL NETWORK COST PER ARCHITECTURE WITH REDUCED SERVICE LATENCY THRESHOLD

Nodes	FOADM			Filterless			ROADM		
	Optical	vCPU	Total	Optical	vCPU	Total	Optical	vCPU	Total
5	-10%	53%	31%	-3%	32%	20%	-3%	35%	21%
10	-11%	30%	15%	-17%	52%	16%	-18%	60%	19%
15	-2%	25%	15%	-9%	49%	12%	-9%	31%	10%

ACKNOWLEDGEMENT

This work was supported by the H2020 Metro-Haul project under grant agreement 761727

REFERENCES

- [1] Xia et al, "Network function placement for NFV chaining in packet/optical datacenters," *Journal of Lightwave Technology*, vol.33, no.8, pp.1565-1570, Apr. 2015.
- [2] Yu et al, "A survey on the edge computing for the Internet of Things," *IEEE Access*, vol.6, pp. 6900-6919, Nov. 2017.
- [3] Garrich et al., "Open-source network optimization software in the open SDN/NFV transport ecosystem," *Journal of Lightwave Technology*, vol.37, no.1, pp.75-88, Jan. 2019.
- [4] Chen et al, "Effective VM sizing in virtualized data centers," in *Proc. International Symposium and Workshops on Integrated Network Management*, May 2011.
- [5] Eira et al, "On the capacity and scalability of metro transport architectures for ubiquitous service delivery," in *Proc. International Conference on Transparent Optical Networks (ICTON)*, paper Mo.D3.5, July 2018.
- [6] Emmerich et al, "Capacity limits of C+L metro transport networks exploiting dual-band node architectures," accepted in *Optical Fiber Communications Conference (OFC)*, paper M2G.5, Mar. 2020
- [7] Lobo et al, "Applications of second-order cone programming," *Linear Algebra and its Applications*, vol.284, no.1-3, pp.193-228, Nov. 1998.
- [8] Luo et al, "An introduction to convex optimization for communications and signal processing," *IEEE Journal on Selected Areas in Communications*, vol.24, no.8, pp.1426-1438, Aug 2006.
- [9] Askari et al, "Virtual-network-function-placement for dynamic service chaining in metro-area networks," in *Proc. International Conference on Optical Network Design and Modeling (ONDM)*, May 2018.
- [10] Metro-Haul Deliverable D2.3: Network architecture definition, design methods and performance evaluation, Apr 2019.