Network Slicing and Workload Placement in Megacities

Polyzois Soumplis[1,2], Panagiotis Kokkinos[1,3], Dimitrios Lagos[1,2], Aristotelis Kretsis[1,2], Vasileios Sourlas[1], Emmanouel (Manos) Varvarigos[1,2]

[1]Institute of Communication and Computer Systems (ICCS-NTUA), Athens, Greece.

[2]School of Electrical and Computer Engineering, National Technical University of Athens, Greece

[3]Department of Digital Systems, University of Peloponnese

e-mail: soumplis@mail.ntua.gr

ABSTRACT

The number and size of megacities and their respective networking and computing infrastructures constantly increases, while a large part of the network traffic originates and terminates locally. In such a dense and heterogeneous environment, network slices are formed, utilizing various optical networking infrastructures that carry data from end users to distributed, intra-city edge computing units, in which applications' workload is appropriately offloaded. In our work, we examine mechanisms for the joint resource allocation and the applications' workload placement in the access and metro network segments.

Keywords: network slicing, optical networks, edge resources, application workload placement

1. INTRODUCTION

The United Nations projects that 60% of the global population will be urbanized by 2030, while the number of megacities - cities with population greater than 10 million - grew from 28 in 2014 to 33 in 2018, with the largest one, Tokyo, Japan, reaching 37.5 population. Also, global IP network traffic is constantly increasing [1], driven by the plethora and popularity of cloud and 5G-based services. To ensure high quality services, content distribution networks and service providers, tend to push more and more content, data and services in the edge, increasing the regional traffic and the provisioned local capacity even with a faster pace than the core-capacity.

The importance of intra-city networking and processing operations that involves the allocation of networking resources in the access and metro segments and the proper placement of application's workload becomes critical. Hence, to minimize the experienced latency and the volume of traffic moved to the core network, processing resources in edge computing [2], are placed near the end-users. These edge resources consist of generic mini- and full-sized datacentres build inside the city or in the outskirts and specialized hardware accelerator units (e.g., FPGA, GPUs) [3]. A process that can mostly benefit from edge computing and hardware accelerators is the training and update of machine learning models, utilized from AI-related applications [4].

The efficient networking interconnection of the end-devices and the distributed computing resources is necessary. Today's metropolitan networks are based on IP over Wavelength Division Multiplexing (WDM) optical technology, implementing various topologies: point to point, rings, stars, meshes, and trees of interconnected rings or meshes. For the access part, Next Generation Passive Optical Networks (NG-PON) [5], also known as WDM PONs, are the most prominent access technology.

In this context, network slicing [6] is a network methodology to embed logical networks on the same physical network infrastructure. Each network slice is an isolated end-to-end network that interconnects computing resources tailored to fulfil diverse and stringent application requirements. A number of network slicing mechanisms have been proposed for serving offline and online demands in an optimal or near-optimal manner, utilizing infrastructures of various technologies (e.g., optical and wireless). In [7] the authors propose offline algorithms in order to create network slices, considering the physical layer constraints of the optical network. In [8] the authors formulate a network slicing problem consisting of two parts: the static virtual network mapping and the dynamic virtual network reconfiguration for intelligently adapting resources provided to each slice.

A number of works also consider the joint allocation of computing and networking resources in the form of network slices. The authors in [9] suggest schemes to coordinate the reservation of computing and networking resources among different network flows so as to minimize queuing latency. [10] proposes a network slicing mechanism considering a 2-tier architecture that consists of a central office and a transport network in the upper tier and a multi-access edge and radio access network in the lower tier. In [11] the authors propose heuristic algorithms for network slicing in a joint 5G radio access and WDM metro-aggregation network. In [12] a hierarchical edge cloud fronthaul slicing framework is proposed that considers jointly bandwidth and computing resources, so as to satisfy the latency constraints of the demands.

In this work, we present an Integer Linear Programming (ILP) -based algorithm and a heuristic to efficiently address the problem of jointly (i) allocating access and metro optical network resources and (ii) placing applications' workload to edge and core processing nodes. These mechanisms serve network slice requests' latency and capacity constraints and can be executed both offline and online. The rest of our work is organized as follows. In Section 2, we describe the network model under consideration. In Section 3, we present the problem formulation and the respective offline Integer Linear Programming (ILP)-based algorithm and the offline/online heuristic mechanism, while simulation results are discussed in Section 4. Finally, our work is concluded in Section 5.

2. CONVERGED MEGACITY INFRASTRUCTURE

A significant part of the generated megacity traffic flows originates and terminates in the same city, demanding low latency and high capacity services. This requires the convergence of the access and metro network segments and of the co-located computation resources. In our work, we consider such a megacity infrastructure, organized into layers, leveraging ultra-high capacity, programmable and flexible optical technologies (Figure 1). Computation resources co-exist with access and metro nodes and are distinguished to edge nodes for lower layer computation resources and to core nodes for higher layer ones.

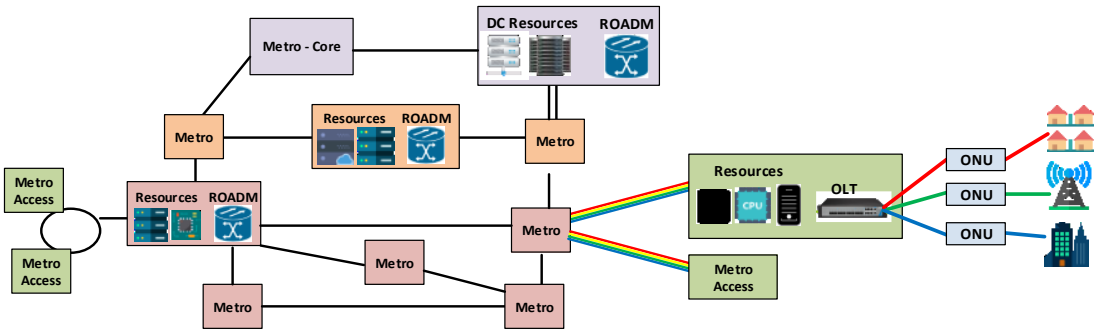

Figure 1: A multilayer megacity network and computing infrastructure.

WDM-PON is the most prominent technology in the access network segment. PON's topology structure enables operators to deliver in a flexible manner, high bandwidth connections to multiple endpoints over long distances and low latency. WDM-PONs consist of an optical line terminal (OLT), located at the operators' central office, and multiple optical network units (ONU) close to the end-users. The wavelengths are assigned to the ONUs exclusively, over the physical point-to-multipoint (OLT-ONUss) fiber infrastructure, based on the demanded capacity, offering also traffic isolation.

Metropolitan networks that interconnect the access networks with the core are becoming the focus of attention. In order to efficiently serve the high access demands, they are organized in hierarchical layers. The exact number of these layers may depend on several parameters such as the coverage area, the number and distribution of users, the traffic characteristics, the capabilities of the networking devices and other. Thus, the nodes of the lower layer, the metro access nodes, are typically co-located with the OLT nodes inside the operators' central offices and are organized in ring and mesh topologies. In the intermediate metro layers, the supported transmission rates increase as more powerful transceivers are employed to transfer the aggregated traffic. The higher layers of the metro infrastructure consist of fewer nodes, namely the metro core nodes, and are organized in mesh topologies.

Selected nodes of the metro hierarchy are also equipped with storage and computational resources that range from general purpose equipment (e.g., servers, mini-racks) to special purpose hardware accelerators (e.g., FPGA, GPU). So, a hierarchical computing infrastructure is created and interconnected with high-speed optical connections through the metropolitan network. This enables the offloading of computational operations close to the data origin, introducing intelligence at the edge and low latency computations for critical applications.

3. PROBLEM FORMULATION

The network slicing and application workload placement problem can be stated as follows. Let $\mathrm{G} = (V, E)$ be a directed graph that represents the network topology. The set V denotes the access and metro network nodes, while the set E represents the physical connections (links) between two nodes. Each physical link $e \in E$, is characterized by its length l_e and a latency factor τ_e due to signal propagation.

Nodes are equipped with processing capacity of type $t \in T$ (generic or hardware accelerator), $c_{v,t}$, and a number of optical transceivers M_v, whose feasible transmission configurations are described by the transmission rate b over distance h. We assume that the metro network is organized into layers, with different processing and networking capabilities and thus different processing $\xi_{v,t}$ cost for the nodes of different layers.

Each slice demand $r \in R$ from a source node s to a destination node d, is characterized by its requested network capacity ζ_r, processing capacity $\varepsilon_{r,t}$ (measured in FLOPS) of type t and maximum latency τ_r. The slices' demanded network capacity is routed through the network and processing power is allocated at the traversed nodes, with the objective to minimize the processing cost of the utilized by the slice resources or the propagation latency of the selected network path.

To speed up the calculations, we make use of a pre-processing phase in which the required number of wavelengths are assigned to the ONUs at the different WDM PON access networks. Next, the feasible connections that make use of a transceiver and its different transmission configurations are calculated for each node $v \in V$. These connections are the optical paths that can be established in the network and will be used to form the end-to-end connections. The traversed links/nodes of each such connection a are described in the sets $\delta^+(v)$, $\delta^-(v)$, for the different flow directions, each one having a length d_a and introducing latency τ_a .

Table 1. The ILP formulation for the megacity slicing and applications' workload assignment.

Variables

$\boldsymbol{x^r_{a,\lambda,b}} \in \{0,1\}$, equal to 1 if slice request r utilizes optical path a over wavelength λ with transmission rate b

$\mathbf{y^r_{v,t}} \in \{0,1\}$, equal to 1 if processing of slice r is performed at node v

$\mathbf{z_{e,\lambda}} \in \{0,1\}$ equal to 1 if transmission over wavelength λ is performed at link $e \in E$

$0 \leq \boldsymbol{\pi^r_{v,t}} \leq C_{v,t}$, an integer variable that denotes the processing of slice r performed at node v

$0 \leq \boldsymbol{\gamma} \leq 1$ the objective's weighting coefficient.

Objective

- $\min \gamma \sum_{r\in R} \tau_r + (1-\gamma) \cdot \sum_{r\in R} \sum_{v\in V} \sum_{t\in T} \pi^r_{v,t} \cdot \xi_{v,t}$,

Subject to the following constraints

- Routing of slices $\forall r \in R, v \in V, \sum_{a\in\delta^+(v)} \sum_{\lambda\in\Lambda} \sum_{b\in B} x^r_{a,\lambda,b} - \sum_{a\in\delta^-(v)} \sum_{\lambda\in\Lambda} \sum_{b\in B} x^r_{a,\lambda,b} = \begin{cases} -1, if\ v = s \\ 1, if\ v = d \\ 0, else \end{cases}$
- Nodes where processing can be performed: $\forall r \in R, a \in \delta^+(v), v \in \alpha, t \in T\ \ y^r_{v,t} \leq \sum_{\lambda\in\Lambda} \sum_{b\in B} x^r_{a,\lambda,b}$
- Processing capacity per node: $\forall r \in R, v \in V, t \in T, \pi^r_{v,t} \leq \left(y^r_{v,t} - 1\right) \cdot C_{v,t}$
- Satisfy the processing requirements of a slice: $\forall r \in R, \varepsilon_{r,t} \leq \sum_{v\in V} \sum_{t\in T} \pi^r_{v,t}$
- Bandwidth constraints: $\forall r \in R, a \in \delta^+(v), \zeta_r \leq \sum_{\lambda\in\Lambda} \sum_{b\in B} x^r_{a,\lambda,b} \cdot b$
- Latency Constraints: $\forall r \in R, \tau_r \leq \sum_{a\in\delta^+(v)} \sum_{\lambda\in\Lambda} \sum_{b\in B} x^r_{a,\lambda,b} \cdot \tau_a$
- Transmission reach constraint: $\forall r \in R, v \in V, a \in \delta^+(v)\ \sum_{\lambda\in\Lambda} \sum_{b\in B} x^r_{a,\lambda,b} \cdot h_b \leq d_a$
- Available transceivers per node: $\forall v \in V, \sum_{r\in R} \sum_{a\in\delta^-(v)} \sum_{\lambda\in\Lambda} \sum_{b\in B} x^r_{a,\lambda,b} \leq M_v$
- Distinct wavelength assignment constraint: $\forall e \in a : a \in \delta^+(v), z_{e,\lambda} \geq \sum_{r\in R} \sum_{b\in B} x^r_{a,\lambda,b}, \forall e \in E, \lambda \in \Lambda, z_{e,\lambda} \leq 1$

We also developed a heuristic algorithm that can efficiently serve a large number of slice requests for a large size network topology significantly faster than the ILP-based optimal algorithm. The heuristic serves the slice requests sequentially. It begins with assigning wavelengths to the different WDM-PONs. Then, based on the networking demands of the slice requests, metro demands are formed, which are sorted in descending order according to their latency requirements. The sorted metro demands are served by searching for a path with available processing capacity in the metro region between the source and destination access metro node. To make faster the algorithm's execution, pre-calculated shortest paths between the metro nodes are used. The candidate combinations of networking and processing allocations among the different paths are searched, selecting the one that minimizes the objective function and fulfils the latency requirements. Due to space limitations, our heuristic algorithm will be presented in detail in future work.

4. SIMULATION RESULTS

We evaluated the performance of the proposed network slicing and application workload placement mechanisms, implementing them in MATLAB, using the IBM ILOG CPLEX optimizer. Simulations were performed over a megacity network topology that consists of 7 WDM PONs, 7 metro access nodes and 4 metro core nodes organized hierarchically into two layers (layer 1 and 2 respectively), with the physical link distances drawn uniformly on the interval [10-200] km. The processing capacity of the resources co-located with the layer 2 metro nodes was taken to be 50% higher compared to the capacity of the edge resources in layer 1 metro nodes. Also, the cost of the computing nodes in layer 1 was 30% higher than in layer 2.

The OLT of each WDM PON supports transmission rates up to 50 Gbps in the uplink and downlink and handles up to 30 wavelengths. The metro nodes are equipped with elastic optical transceivers that transmit at 12.5, 25 and 50 Gbps up to 200, 100 and 50 km respectively and 40 wavelengths are available at each link. Traffic scenarios are based on offline slice requests whose number increases from 80 to 200 and their capacity requirements are drawn from a uniform distribution on the interval [0-50] Gbps for network and [1-10] GFlops for processing capacity. Each traffic request has a latency constraint, which in practice dependd on the application it serves. The proposed mechanisms were evaluated based on the average cost of the utilized computing resources for the workload processing and on the experienced propagation latency of the selected network path per slice request.

Initially, we examined the optimality of the heuristic in a small-scale topology with a number of offline network slice requests that varied from 50 to 100. Under these settings, we were able to track the optimal solution using the ILP-based algorithm in a 5-hour time window and to record the optimality gap with the heuristic to 9%.

More extensive simulation experiments were conducted on the larger network topology and traffic scenarios described above, using the heuristic algorithm that provides good resource allocation solutions in reasonable time. We examined under various traffic scenarios the use of edge and core computing nodes, for the execution of application workload specified by the requesting slices' (Figure 2.a), considering two different objective functions: (i) minimize latency (ML) and (ii) minimize processing cost (MC). When the objective is the minimization of the total cost, the core resources are preferred compared to the edge ones. In this case, only the slices with most strict latency requirements offload their workload in the edge nodes, while slices with relaxed latency requirements select processing nodes located in the metro core, with lower cost. One the other hand, the use of the lower cost

core processing nodes, means that longer networking paths and more optical network resources are utilized. In particular, we measured that in the 200 slices requests scenario, 7% more optical transponders are used. When the objective set is the minimization of the average latency then edge resources are fully utilized by the respective slices, for all the examined traffic scenarios, while core processing resources are used only due to exhaustion of the edge nodes' capacity.

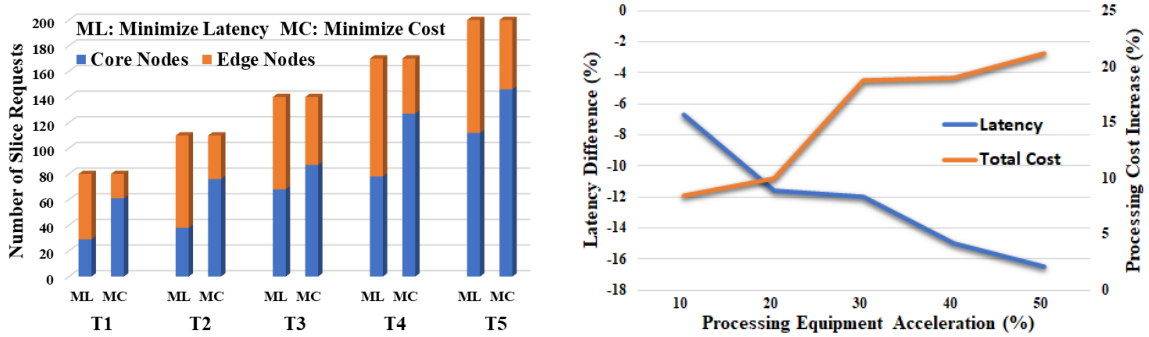

Figure 2. a) The number of slice requests served by edge and core computing nodes for different traffic scenarios (T1=80, T2=110, T3=140, T4=170, T5=200 slice requests) and objective functions: minimize latency (ML), minimize cost (MC), b) The average latency decrease and the average cost increase of the resource allocation decisions when faster processing equipment is used in the edge nodes.

Figure 2.b, illustrates the average latency and the average cost of the resource allocation decisions when faster processing equipment (e.g., hardware accelerators) is used in the edge nodes. We assumed that the special purpose equipment can offer better processing performance compared to the generic resources up to 50%. Thus, we examined for the T3 (140 slice requests) traffic scenario and the latency minimization objective, the trade-off between latency and cost, with the illustrated values being compared (as %) to the case where generic computing resources are used. Faster processing equipment in the edge resources, results in a decrease in the average latency up to 16%, while the average processing cost is increased up to 20%.

5. CONCLUSION

The wide range of 5G and cloud services in one hand and the vast size of megacities on the other, set new challenges for the deployed networking and computing infrastructures, making their convergence necessary. In our work, we present a multilayer metropolitan network model that interconnects access networks, while computing resources are placed in lower (edge) and higher (core) layers. We also propose an ILP-based mechanism and a heuristic for network slicing in the access and metro segments and for the proper placement of applications' workload. As our simulations indicate, the developed mechanisms serve efficiently the slice requests based on the objective set, while also take advantage of the improved processing capabilities of the edge resources, for reducing latency up to 16% at the expense of increased cost.

ACKNOWLEDGEMENTS

This research has been co-financed by the European Regional Development Fund of the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH – CREATE – INNOVATE (ARMONIA, project code: T1EDK-05061)

REFERENCES

[1] Cisco Visual Networking Index: Forecast and Trends, 2017–2022, 2018.
[2] T. Taleb, et al., "On Multi-Access Edge Computing: A Survey of the Emerging 5G Network Edge Cloud Architecture and Orchestration," *IEEE Communications Surveys & Tutorials*, Vol. 19, No. 3, 2017.
[3] S. Biookaghazadeh, et al.,"Are FPGAs Suitable for Edge Computing?", *USENIX Workshop on Hot Topics in Edge Computing*, 2018.
[4] X. Wang, et al.,"In-edge ai: Intelligentizing mobile edge computing, caching and communication by federated learning", *IEEE Network*, Vol. 33, No. 5, 2019.
[5] D. Nesset, "PON roadmap", *IEEE/OSA Journal of Optical Communications and Networking*, Vol. 9, 2017.
[6] S. Vassilaras et al., "The algorithmic aspects of network slicing", *IEEE Communications Magazine*, Vol. 55, 2017.
[7] A. Buttaboni, et al.,"Virtual PON assignment for fixed-mobile convergent access-aggregation networks", *International Conference on Optical Network Design and Modeling (ONDM)*, pp. 198-203. 2015.
[8] M. R. Raza, et al.,"Dynamic slicing approach for multi-tenant 5G transport networks", *Journal of Optical Communications and Networking*, Vol. 10, 2018.
[9] W. Wang, et al. "Coordinating Multi-access Edge Computing with Mobile Fronthaul for Optimizing 5G End-to-End Latency", *IEEE OFC*, 2018.
[10] H. Chien, et al. "End-to-End Slicing With Optimized Communication and Computing Resource Allocation in Multi-Tenant 5G Systems," IEEE Transactions on Vehicular Technology, vol. 69, 2020.
[11] H. Yu, et al. "Isolation-Aware 5G RAN Slice Mapping Over WDM Metro-Aggregation Networks," *Journal of Lightwave Technology*, Vol. 38, 2020.
[12] C. Song et al., "Hierarchical edge cloud enabling network slicing for 5G optical fronthaul," *IEEE/OSA Journal of Optical Communications and Networking*, vol. 11, no. 4, pp. B60-B70, 2019.