QoS-Aware Inter-Domain Connectivity: Control Plane Design and Operational Considerations

Stanislav Lange*, Jane Frances Pajo§, Thomas Zinner*, Håkon Lønsethagen§, Min Xie§

*NTNU - Norwegian University of Science and Technology, Department of Information Security and Communication Technology
{stanislav.lange,thomas.zinner}@ntnu.no

§Telenor Research, Fornebu, Norway
{jane-frances.pajo,hakon.lonsethagen,min.xie}@telenor.com

***Abstract*—Scenarios related to 5G and beyond give rise to a high degree of heterogeneity in terms of applications, services, and user expectations as well as more demanding QoS requirements with an end-to-end scope that can cover multiple operator domains. In this work, we design an inter-domain component that addresses the control plane challenges of establishing such end-to-end connectivity in a scalable, efficient, and automated way. Furthermore, we provide insights into operational aspects by investigating to which extent different traffic aggregation mechanisms can be used to benefit from economies of scale while meeting QoS constraints.**

***Index Terms*—B5G, QoS, Inter-Domain, Traffic Aggregation.**

I. Introduction

5G and beyond 5G (B5G) networks are facing an increasing number of different vertical applications, network services, as well as user devices and corresponding performance expectations. This heterogeneity is paired with even stricter Quality of Service (QoS) constraints in terms of criteria like bandwidth, delay, or jitter, with disruptive use cases such as automotive, augmented / virtual reality (AR / VR), and Industry 4.0 coming into play. Furthermore, these constraints need to be met in an end-to-end (E2E) fashion, potentially across the boundaries of multiple operator domains, in order to provide the desired Quality of Experience (QoE) to the end-users.

These requirements introduce challenges in two key directions. On the one hand, novel control plane entities need to be designed and should be able to establish and negotiate such E2E connections. On the other hand, the corresponding processes need to be performed in a scalable, resource efficient, and manageable manner.

In this work, we address both aspects. On the design side, we present the considerations behind the TeraFlow[1] Inter-Domain Component (IDC). This component is part of the TeraFlow Network OS whose targeted use cases specifically include inter-domain scenarios from the beginning as opposed to being an add-on to existing frameworks. Nonetheless, existing building blocks such as L3VPNs, slicing, and related 5G mechanisms such as 5G QoS Identifier (5QI) classes are leveraged to facilitate development and allow fine-grained traffic differentiation. The definition of appropriate interfaces towards internal and external components ensures programmability and automation which are essential for dealing with temporal dynamics.

On the operational side, it is clear that an exclusive and on-demand allocation of resources on an E2E and per-flow level is not feasible from a scalability, (control plane) complexity, and resource efficiency perspective. Hence, to address the second aspect, we need to identify appropriate traffic aggregates and abstractions regarding traffic types and patterns so as to benefit from multiplexing gains, while also addressing scalability and reducing complexity. This also includes finding the appropriate granularity at which to aggregate traffic and perform control plane actions.

In summary, the contributions of this work are twofold. First, we present the design of the TeraFlow IDC and how it fits into the TeraFlow and overall networking ecosystem. Second, we analyze the impact of different traffic aggregation mechanisms on delay and maximum tolerable link load when mixing traffic profiles that are representative of typical applications such as VoIP, video streaming, or file download. To this end, we perform queueing simulations using aggregation mechanisms that correspond to *best effort* traffic handling, *slicing* with hard isolation, and *prioritization*.

The remainder of this work is structured as follows. We provide an overview of related work on inter-domain connectivity and traffic aggregation in Section II. The inter-domain component and its role in the TeraFlow ecosystem are outlined in Section III. After presenting the chosen abstractions and simulation methodology in Section IV, we discuss evaluation results in Section V. Finally, we conclude the paper in Section VI.

II. Related Work

Since this work covers both the control plane design of the IDC as well as its operational aspects, we also approach related literature from these angles.

The rise of next-generation verticals with highly heterogeneous requirements has undoubtedly driven recent advance-

[1]https://www.teraflow-h2020.eu/

978-1-6654-0601-7/22/$31.00

ments in network slicing, where each slice can correspond to a QoS profile that has been carefully aligned with the vertical's key performance indicators (KPIs) [1]. In line with this, the potentials of network slicing for QoE-aware resource allocation have been evaluated in [2] in terms of slice dimensioning and fine-tuning of per-flow bitrates with respect to both application requirements and resource efficiency.

Closely related to inter-domain QoS provisioning [3], [4], multi-domain slicing has been a hot topic in recent years as vertical applications and services start to span multiple technological and / or administrative domains. From this perspective, a field trial has been presented in [5] to demonstrate the viability of multi-domain service provisioning based on Software-defined Networking (SDN) over a multi-layer and multi-vendor network environment. The authors in [6] proposed a multi-domain slicing architecture and operational procedures for slice lifecycle management, also indicating open challenges such as service profiling, resource sharing, and isolation. Similarly, the work in [7] addressed the problem of cross-domain slice orchestration where domains may adopt distinct orchestration solutions, suggesting a peer-to-peer federation of administrative domains for management scalability, privacy, auditability, etc. On this note, the decentralization of the 5G slice resource allocation has been considered in [8]. Moreover, the TeraFlow H2020 project seeks to advance this state-of-the-art through a cloud-native SDN controller, acting as a network OS that is able to bridge various stakeholders - ranging from Telco operators to Edge and hyperscale Cloud providers - and deploy multi-domain services in a secured and autonomic way [9].

Considering the analysis of traffic aggregates, several works in the domain of queueing theory have dealt with mixing or superimposing traffic from different sources [10], [11]. However, several strong assumptions and / or limitations narrow the scope of scenarios for which closed-form or stable numerical solutions can be obtained. Such assumptions include modeling arrival processes as renewal processes as well as no or limited consideration of inter-dependencies between traffic streams and their autocorrelation. Hence, we use a queueing simulation in this work to extract qualitative relationships which will guide subsequent detailed simulations and testbed-based experiments.

Finally, the authors of [12] also pointed out that existing traffic aggregation methods are mostly based on quantitative QoS requirements and static QoS classes, and hence, proposed to also consider qualitative requirements and dynamically group flows using Artificial Intelligence (AI).

III. Inter-Domain Component (IDC)

Telco infrastructures can be segmented into different *technological*, i.e., (Radio) Access, Transport, Edge, Core, as well as *administrative*, i.e., per-operator and / or per-geography, domains. In this work, the former will be referred to as "technology segments", and the latter as "domains". While the TeraFlow OS will cover both inter-segment as well as inter-domain functionality, we focus on the latter in this work.

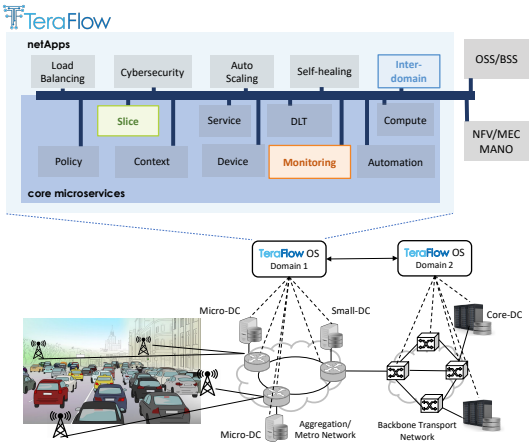

Fig. 1. Overview of the TeraFlow architecture with inter-domain and closely related slicing and monitoring components highlighted. Graphic extended based on [9].

From an individual domain's perspective, an operator needs to accommodate heterogeneous requests from a wide variety of clients while ensuring that Service Level Agreements (SLAs) and QoS requirements are met, subject to the resources' availability and capabilities, as well as the corresponding costs and control plane complexities. Appropriate trade-offs between these dimensions should be established to ensure efficient and effective operations. While this is challenging enough in a single domain scenario with services spanning multiple technology segments, next-generation verticals are now further pressing for multi-domain deployments and a "converged" infrastructure perception.

This section briefly describes the approach that is taken towards inter-domain connectivity in the TeraFlow project, and the operational challenges that it seeks to address.

A. TeraFlow Network OS and IDC Design

In a nutshell, TeraFlow aims at realizing a novel, cloud-native network OS that seeks to advance traffic flow management in (multi-domain) B5G networks through innovative features enabled by its *core microservices* and *netApps*. The TeraFlow OS has been designed to interact with other network management elements such as the Network Functions Virtualization (NFV) and Multi-access Edge Computing (MEC) Orchestrators, as well as the Operations / Business Support Systems (OSS / BSS), in order to support multi-tenancy and coordinate (geo-distributed) service deployments in a fully automated manner. Moreover, interactions between peer TeraFlow OS instances (each managing different operators' domains) will be enabled to support inter-domain connectivity services. This work focuses on the latter, specifically on the IDC design and SLA-driven traffic aggregation. Figure 1 shows how the IDC fits into the TeraFlow architecture.

The IDC is designed with key functionalities closely linked to service lifecycle management – from the preparation and activation of a connectivity service, to its runtime modification.

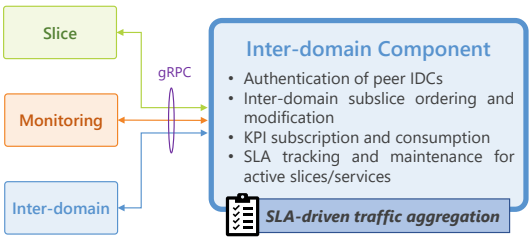

Fig. 2. IDC design considerations and interfaces with the TeraFlow core microservices. With its E2E view, the IDC has the potential of optimizing traffic aggregates so as to meet QoS constraints and leverage multiplexing gains where possible.

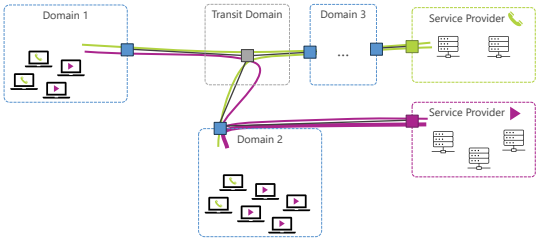

Fig. 3. Inter-domain traffic aggregation. Heterogeneous traffic can be aggregated in different ways at each station, resulting in trade-offs w.r.t. performance, costs, and scalability.

With per-domain subslices as basic building blocks, an inter-domain connectivity service is then realized as a transport network (TN) slice [13] by coordinating the interconnections among the involved subslices. Such TN slices go beyond traditional L2/L3 VPNs in so far as they also include a specification of connectivity requirements such as a guaranteed minimum bandwidth or maximum latency. Furthermore, monitoring of service KPIs across domains and, in case of violations, triggering mitigation actions, ensures that the E2E QoS requirements are met.

As illustrated in Figure 2, the IDC also interfaces with TeraFlow's *Slice* and *Monitoring* microservices, facilitating the necessary workflows towards the aforementioned ambitions. Being the recipient of customer-initiated service requests, the Slice component needs to first assess whether each request is intra- or inter-domain, and in the latter case, forward the inter-domain subslice request to the IDC, which is then communicated to next-hop peer IDC(s) involved in the E2E TN slice. In turn, each IDC also communicates to its domain's Slice component any subslice (modification) requests received from peer IDCs, in order to establish a service or enforce the necessary changes within the domain. As regards QoS assurance, the IDC subscribes to relevant KPIs from its domain's Monitoring component in order to track the SLAs of active services/slices and then trigger mitigation actions in case of violations.

Following a microservice approach, Google's *gRPC* [14] and *protobuf* [15] have been used to enable language-agnostic, efficient, and programmable communications among TeraFlow OS components, while maintaining the OS's modularity. This allows components to evolve independently as long as interfaces have been agreed upon.

B. Operational Challenges

In a multi-domain scenario, meeting E2E QoS requirements remains the key goal. However, there are still a number of operational challenges that also need to be addressed, such as the scalability and control plane complexity, among others. As a first step, this work looks into the SLA-driven traffic aggregation at domain-crossings, which is one of the requirements defined for the IDC.

Figure 3 illustrates a simple example of multi-domain service deployments, where traffic flows are (re-)aggregated at each domain-crossing that they traverse. In particular, the corresponding E2E path can contain domains of other operators who actively participate in the inter-domain negotiations as well as transit domains that are outside the IDC's reach, but still might affect traffic characteristics along the way. It is important to note that the domain operator can have different options in handling traffic flows, each one representing an intra-domain trade-off in terms of cost, complexity, scalability, and the expected SLA compliance. Furthermore, depending on the chosen option, each domain-crossing can involve changes to traffic characteristics such as delay or jitter.

With these in mind, we investigate how traffic with different characteristics and SLA profiles behaves when aggregated in different ways. Particularly, we seek to quantify the benefits of different aggregation techniques - e.g., potentials for increasing sustainable link load - when aggregating heterogeneous traffic profiles.

IV. Methodology

We employ a queueing simulation to carry out investigations about the aggregation of heterogeneous traffic profiles. To this end, we use a number of abstractions regarding the aggregation mechanisms as well as traffic characteristics that are discussed in this section. Furthermore, we present the simulation scenarios and evaluation metrics that are covered in this work.

A. Aggregation Mechanisms

We consider a total of three traffic aggregation mechanisms that are visualized in Figure 4. Given two arrival streams of packets from two traffic profiles, the aggregation mechanisms determine the allocation of the available service capacity as well as packets' trajectory through the system. In the following, we discuss how these mechanisms work and their operational implications.

The first and most simple approach is referred to as Best Effort (BE) and consists of putting all packet arrivals into a shared queue that is serviced at maximum rate in a FIFO manner. Since differentiated treatment per traffic class is not necessary when using this approach, it does not incur control plane overhead. Furthermore, it allows leveraging economies of scale since the service unit is active whenever a packet of any kind is in the system. However, the only way to improve the delay performance of the aggregate traffic stream consists

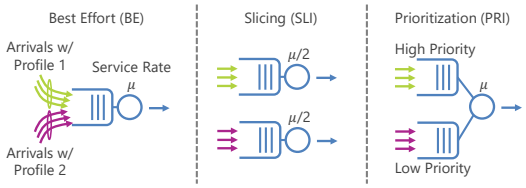

Fig. 4. The chosen abstractions regarding traffic aggregation mechanisms differ in terms of their potential for multiplexing gains and degree of isolation while being representative of available configuration options.

of lowering the utilization, either by lowering the number of admitted clients or overprovisioning.

In 5G and B5G networks, slicing plays a crucial role for multiplexing heterogeneous traffic. Hence, the second abstraction mimics traffic handling in the presence of slicing (SLI) with hard isolation. In this context, the available capacity is split between two systems with fully isolated queues and service units, each responsible for one of the two arrival streams. Such an approach would require corresponding hardware capabilities such as virtual routers and ports to be implemented on a shared physical infrastructure. Additionally, the granularity of the resource split, the number of virtual subsystems, and the degree of isolation can vary between devices. As a consequence, the slicing approach involves control plane overhead and requires fine-tuning of - potentially device-specific - parameters. Although strict isolation with resource reservation does not offer multiplexing gains, elevated security and robustness requirements can be met which are crucial in contexts such as Public Protection and Disaster Relief (PPDR).

The final abstraction is based on prioritization (PRI) with shared capacity, and represents a middle ground between the first two options. While it still requires some hardware capabilities and configuration effort, differentiated treatment is possible without losing economies of scale. However, it is worth noting that performance guarantees are less strict so that low-priority traffic might suffer unless more complex shaping is employed, e.g, by means of hierarchical QoS (hQos) using a Hierarchical Token Bucket (HTB). In our simulations, the first traffic class in a combination receives prioritized treatment whereas the second is treated with a low priority.

B. Traffic Profiles

In order to allow covering a wide range of applications, we start out by characterizing three typical applications with respect to their traffic properties. In the process, we define archetypal traffic profiles whose parameters such as rate or packet size can be fine-tuned to fit and more closely represent other applications or variations of existing ones. For instance, parameters of a video streaming archetype could be adjusted to represent video streaming with content of different bitrates while a VoIP archetype could serve as starting point for real-time conversational video services. Since we are primarily interested in general relationships between traffic profiles and aggregation techniques, we omit in-depth application- and protocol-level behavior such as dynamic bit rate adaptation or congestion control.

TABLE I
TRAFFIC PROFILES UNDER STUDY.

Traffic Profile	Packet Size	Arrival Pattern	Rate
VoIP	Small (75 B)	CBR (20 ms IAT)	30 kbps
File download (FDL)	Big (1,500 B)	CBR (6 ms IAT)	2 Mbps
Video stream (VID)	Big (1,500 B)	Bursty on / off	2 Mbps

A total of three traffic profiles are under consideration in this work. These include VoIP, file download (FDL), and video streaming (VID). They are listed alongside their main traffic characteristics in Table I. Both VoIP and FDL exhibit constant bit rate (CBR) arrival patterns with packet interarrival times (IATs) of 20 / 6 ms, using small / big packets, and resulting rates of 30 kbps / 2 Mbps, respectively. In contrast, the VID profile captures the typical behavior of video streaming where a video file is split into segments that are watched and downloaded as the session progresses. Assuming a segment duration of 2 sec and an average rate of 2 Mbps, this results in a bursty on / off pattern with a 4 Mbit burst every 2 s with no activity in between.

We deliberately do not impose delay or other QoS requirements a priori, but leave them variable for the evaluation, so that it is possible to check which constraints can be met with each aggregation mechanism.

C. Simulation Setup and Evaluation Metrics

Using our three traffic profiles, we run simulations with each possible two-profile combination in conjunction with each of the three aggregation mechanisms. We control the load that is offered to the system by varying the number of clients n we simulate for each profile. The traffic mix is dimensioned so that the total rate per profile is identical. Hence, FDL and VID are mixed with a 1:1 ratio whereas the number of VoIP clients is multiplied by 66 to compensate for the rate difference. We use link capacities of 100 Mbps, 1 Gbps, and 10 Gbps, and by varying n appropriately, the offered load covers a range from 40 to 100 % of the capacity in increments of 4 %. Furthermore, we ensure independence between individual clients by randomly offsetting their starting time, i.e., the time of the first packet.

To obtain statistically significant results, each scenario is simulated ten times. The duration of individual simulation runs is chosen in a way that we record the statistics of at least two million packets per traffic profile per run during the steady phase of the simulation. We ignore events from the first 20 s of each run to avoid transient behavior during initialization and compute the mean alongside 95 % confidence intervals of relevant statistics across the ten repetitions. All simulation parameters are summarized in Table II.

While it is possible to extract a multitude of per-packet, per-client, and per-profile performance metrics, our main interest in this work is on packets' *sojourn time* per traffic profile,

TABLE II
SIMULATION PARAMETERS.

Parameter	Value(s)
Link capacity	100 Mbps, 1 Gbps, 10 Gbps
Aggregation mechanism	$\{BE, SLI, PRI\}$
Traffic combination	$\{VoIP, FDL, VID\}^2$
Offered load	$\{40, 44, \ldots, 100\%\}$
Number of simulated packets	At least 2e6 per profile per run
Repetitions per scenario	10

i.e., the time difference between the moment a packet arrives at the system and the moment it departs after having been processed by the service unit. This allows us to derive insights regarding the trade-offs between different traffic aggregation mechanisms under different load conditions and traffic mixes.

V. EVALUATION

In this section, we discuss numerical results of our queueing simulation. After analyzing the sojourn times in the different simulation scenarios, we turn our focus to operational aspects by studying the effects of traffic aggregation strategies, traffic composition, and delay requirements on the maximum tolerable load level.

A. Sojourn Time

Figure 5 displays the results for scenarios in which VoIP and VID traffic profiles are mixed. While the sub-plots correspond to different aggregation strategies and traffic profiles, the x-axis denotes the offered load, and the logarithmically scaled y-axis shows the mean sojourn time including 95 % confidence intervals. Differently colored curves represent different link capacities.

A general observation can be made across all plots, namely that the link capacity is inversely proportional to the resulting mean sojourn time. This is in line with the fact that a higher link capacity leads to lower packet service times and therefore also to faster recovery times from burst arrivals.

In the case of BE, we can observe a sojourn time increase of up to three orders of magnitude for both traffic types when going from the lowest to the highest load levels. In contrast, PRI keeps the sojourn time increase of VoIP packets linear since they are prioritized and processed at full capacity while making up only half of the arriving load. With slicing-based aggregation, a sharper sojourn time increase can be observed towards the highest load levels. This can be explained by the fact that each slice operates at half the original link capacity. However, since VoIP clients exhibit a CBR traffic pattern, the increase is not as steep as for VID whose clients have bursty arrivals. Furthermore, we can observe that while the delay performance of VID traffic does not change as much as that of VoIP, the resulting sojourn times decrease from BE to PRI to SLI. Although VID traffic is handled with low priority in the PRI case, the packets are still processed at full capacity whereas in the case of SLI, only half the link capacity is available, which especially affects bursty arrivals.

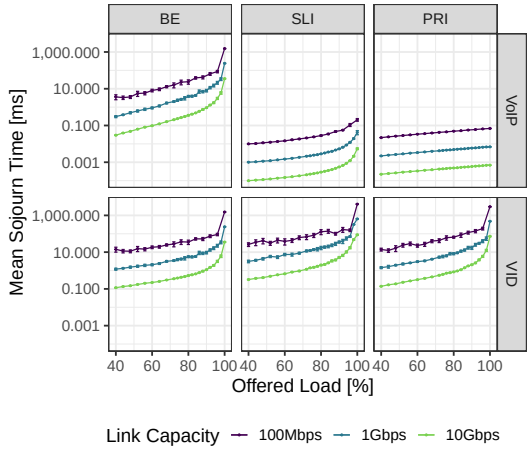

Fig. 5. Impact of offered load on mean per-application sojourn time when mixing VoIP and VID under different aggregation mechanisms and link capacities. While best effort (BE) results in sojourn time values of all traffic covering several orders of magnitude, slicing (SLI) and prioritization (PRI) allow trading off the delay performance of one traffic class against the other.

B. Tolerable Load under Delay Constraints

Since we are particularly interested in the effects of traffic aggregation strategies, traffic composition, and delay requirements on operational aspects such as the maximum tolerable load level, we perform the following preprocessing steps on the simulation outputs in order to obtain the visualizations in this subsection: given delay constraints in the range $\{0, 1, \ldots, 100\}$ ms for each application, we find the scenarios in which these delay constraints are met, and from those extract the scenario with the highest number of clients, i.e., the highest offered load.

As an example, we present results for scenarios with a link capacity of 1 Gbps and the traffic mix of VoIP+VID in Figure 6. While the sub-plots in each graphic represent different aggregation strategies, the logarithmically scaled x- and y-axes denote delay constraints for the first and second traffic profile in the mix, respectively. Using the outlined preprocessing procedure, the color of each tile corresponds to the maximum load that can be handled while meeting the respective delay constraints.

In the case of BE-based aggregation, the maximum tolerable load level of 98 % can only be reached if both traffic classes exhibit a delay tolerance of at least 35 ms each. Additionally, the maximum tolerable load gradually declines when stricter delay constraints are introduced on any of the two traffic types. This is in line with the traditional strategy of overprovisioning resources to achieve targeted QoS constraints.

In contrast, when using priority-based aggregation (PRI), the 98 % load level can be handled with significantly stricter delay constraints for the prioritized class, allowing 1 ms delays. However, this comes at the price of higher delays for the second traffic type in the mix and therefore requires it to be able to sustain delays up to 56 ms.

With slicing-based isolation (SLI), the lack of multiplexing gains manifests itself in a narrower range of delay constraint

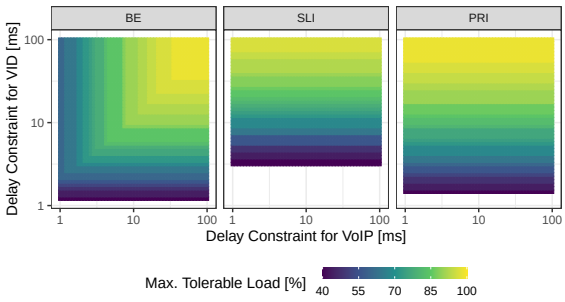

Fig. 6. Impact of delay constraints and aggregation mechanisms on maximum tolerable link load when superimposing VoIP and video streaming (VID) traffic profiles on a 1 Gbps link. While best effort (BE) can only achieve high utilization when both traffic types are delay-tolerant, slicing-(SLI) and prioritization-based (PRI) aggregation manage to meet tighter delay bounds for one of the traffic types even at high link loads.

combinations that allow for its maximum load level of 96 %. However, SLI provides isolation-related benefits with respect to security and robustness against interference between classes. It is also worth noting that splitting the available processing resources between slices results in longer service times, so that the lowest delay constraints can not be met in case of the bursty VID traffic. This leads to the gap at the bottom of the corresponding sub-plot.

When considering the pairwise difference of cells between different aggregation scenarios, gains w.r.t. the maximum tolerable load for each combination of delay constraints can be derived. For instance, if constraints of 5 and 60 ms were chosen for VoIP and VID, respectively, BE, SLI, and PRI would have maximum tolerable load levels of 84, 94, and 98 %. Hence, using PRI over BE for this particular constellation would allow for a 14 % higher link utilization while meeting the delay performance requirements of both applications. Since such considerations are performed at each aggregation step along an E2E path, they might be even stricter. We also note that while the absolute numbers in terms of tolerable load and delay constraints vary depending on the application mix and traffic parameters, the qualitative relationships between aggregation mechanisms are stable between scenarios.

In summary, our simulations allow quantifying the impact of and trade-offs between various aggregation strategies on different heterogeneous traffic mixes and identify feasible regions for efficient operation. Future investigations will focus on how changing traffic profiles and profile parameters affect the gradients in the presented heatmaps, and whether there is a generalizable relationship between them.

VI. Conclusion

Providing inter-domain E2E connectivity with QoS guarantees plays an increasingly important role for Telco operators who face heterogeneous demands from a multitude of users, services, and vertical applications. In this work, we have presented the design of the TeraFlow IDC which is an enabler for such QoS-aware inter-domain connectivity. Additionally, we have discussed how the IDC addresses key control plane requirements regarding automation, scalability, and efficiency.

Furthermore, we have explored operational aspects by investigating the impact of three traffic aggregation strategies on the delay performance and maximum tolerable load levels when merging different traffic profiles. The insights allow us to choose an appropriate aggregation mechanism that is tailored to the specific network and traffic conditions of interest. Furthermore, it allows identifying cases where we can leverage economies of scale while meeting diverse customer demands.

As future work, we plan to confirm the results from the simple queueing simulation in a more detailed, packet-level simulation that covers more complex application and protocol behavior as well as in a physical testbed using components of the TeraFlow OS. Finally, considering temporal dynamics within traffic aggregates can add room for further operational improvements.

Acknowledgments

This work has been performed in the framework of EC H2020 TeraFlow (101015857).

References

[1] P. Alemany, A. Román, R. Vilalta *et al.*, "A KPI-Enabled NFV MANO Architecture for Network Slicing with QoS," *IEEE Communications Magazine*, vol. 59, no. 7, pp. 44–50, 2021.

[2] M. Bosk, M. Gajić, S. Schwarzmann *et al.*, "Using 5G QoS Mechanisms to Achieve QoE-Aware Resource Allocation," in *International Conference on Network and Service Management (CNSM)*. IEEE, 2021.

[3] M. Chamania and A. Jukan, "A Survey of Inter-Domain Peering and Provisioning Solutions for the Next Generation Optical Networks," *IEEE Communications Surveys & Tutorials*, vol. 11, no. 1, pp. 33–51, 2009.

[4] P. Jacob and B. Davie, "Technical Challenges in the Delivery of Interprovider QoS," *IEEE Communications Magazine*, vol. 43, no. 6, pp. 112–118, 2005.

[5] S. Barguil, V. Lopez, C. Manta-Caro *et al.*, "Field Trial of Programmable L3 VPN Service Deployment Using SDN-Based Multi-domain Service Provisioning over IP/Optical Networks," *IEEE Network*, 2021.

[6] T. Taleb, I. Afolabi, K. Samdanis *et al.*, "On Multi-Domain Network Slicing Orchestration Architecture and Federated Resource Control," *IEEE Network*, vol. 33, no. 5, pp. 242–252, 2019.

[7] J. Ordonez-Lucena, C. Tranoris, J. Rodrigues *et al.*, "Cross-domain Slice Orchestration for Advanced Vertical Trials in a Multi-Vendor 5G Facility," in *2020 European Conference on Networks and Communications (EuCNC)*, 2020, pp. 40–45.

[8] F. Fossati, S. Moretti, S. Rovedakis *et al.*, "Decentralization of 5G Slice Resource Allocation," in *2020 IEEE/IFIP Network Operations and Management Symposium (NOMS)*, 2020, pp. 1–9.

[9] R. Vilalta, R. Muñoz, R. Casellas *et al.*, "TeraFlow: Secured Autonomic Traffic Management for a Tera of SDN Flows," in *Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit)*. IEEE, 2021, pp. 377–382.

[10] J. F. Shortle, J. M. Thompson, D. Gross *et al.*, *Fundamentals of Queueing Theory*. John Wiley & Sons, 2018, vol. 399.

[11] S. Kim, "The heavy-traffic bottleneck phenomenon under splitting and superposition," *European Journal of Operational Research*, vol. 157, no. 3, pp. 736–745, 2004.

[12] P. Tang, Y. Dong, Y. Chen *et al.*, "QoE-Aware Traffic Aggregation Using Preference Logic for Edge Intelligence," *IEEE Transactions on Wireless Communications*, vol. 20, no. 9, pp. 6093–6106, 2021.

[13] R. Rokui, S. Homma, K. Makhijani *et al.*, "IETF Definition of Transport Slice," Internet Engineering Task Force, Internet-Draft draft-nsdt-teas-transport-slice-definition-04, Sep. 2020, work in Progress. [Online]. Available: https://datatracker.ietf.org/doc/html/draft-nsdt-teas-transport-slice-definition-04

[14] "Google Remote Procedure Call," https://grpc.io/.

[15] "Protocol Buffers," https://developers.google.com/protocol-buffers.