Asynchronous Time-Sensitive Networking for 5G Backhauling

Jonathan Prados-Garzon and Tarik Taleb

***Abstract*—Fifth Generation (5G) phase 2 rollouts are around the corner to make mobile ultra-reliable and low-latency services a reality. However, to realize that scenario, besides the new 5G built-in Ultra-Reliable Low-Latency Communication (URLLC) capabilities, it is required to provide a substrate network with deterministic Quality-of-Service support for interconnecting the different 5G network functions and services. Time-Sensitive Networking (TSN) appears as an appealing network technology to meet the 5G connectivity needs in many scenarios involving critical services and their coexistence with Mobile Broadband traffic. In this article, we delve into the adoption of asynchronous TSN for 5G backhauling and some of the relevant related aspects. We start motivating TSN and introducing its mainstays. Then, we provide a comprehensive overview of the architecture and operation of the IEEE 802.1Qcr Asynchronous Traffic Shaper (ATS), the building block of asynchronous TSN. Next, a management framework based on ETSI Zero-touch network and Service Management (ZSM) and Abstraction and Control of Traffic Engineered Networks (ACTN) reference models is presented for enabling the TSN transport network slicing and its interworking with Fifth Generation (5G) for backhauling. After, we cover the flow allocation problem in asynchronous TSNs and the importance of Machine Learning techniques for assisting it. Last, we present a simulation-based proof-of-concept (PoC) to assess the capacity of ATS-based forwarding planes for accommodating 5G data flows.**

Backhaul Networks, Transport Networks, Time-Sensitive Networking, Asynchronous Traffic Shaper, Urgency-Based Shaper, Deterministic Quality-of-Service, Machine-Learning, Flow Allocation, 5G

I. Introduction

5G technology is here to accelerate the digitalization of economies and society. Over the last decade, the combined efforts from academy and industry have materialized in matured 5G standards (e.g., Third Generation Parnetship Project (3GPP) Releases 15 and 16), and the 5G commercialization is now a reality. The initial deployments of 5G can be considered as a natural evolution of Fourth Generation (4G). They are based on 3GPP Release 15 and intended for enhanced Mobile Broadband (eMBB) communications, i.e., user-centric services with unprecedented capacity demand and enhanced connectivity and mobility requisites. However, the most challenging mobile network upgrading is yet to come with phase 2 5G rollouts (based on Release 16). Services with delay and reliability constraints never seen before will be brought to mobile networks, thus opening new market opportunities to Mobile Network Operators (MNOs).

Jonathan Prados-Garzon and Tarik Taleb are with the Department of Communications and Networking, School of Electrical Engineering, Aalto University, Finland (e-mail: jonathan.prados-garzon@aalto.fi; tarik.taleb@aalto.fi). Tarik Taleb is also with the Faculty of Information Technology and Electrical Engineering, Oulu University, and with the Department of Computer and Information Security, Sejong University, Seoul 05006, South Korea.

Albeit 5G Release 16 standard includes capabilities for supporting Ultra-Reliable Low-Latency Communications (URLLCs) at both the Radio Access Network (RAN) and the packet core domains, the Transport Network (TN) also plays an important role to guarantee end-to-end (E2E) SLAs (Service Level Agreements) and KPIs (Key Performance Indicators) especially for those applications that require strict latency and bandwidth guarantee [1], [2]. The TN is the network domain providing connectivity among the 5G network functions and out of the 3GPP scope. There are two major requirements for the TN technology in order to enable 5G to support URLLCs affordably:

i) The TN shall provide low-latency and ultra-reliability to convey 5G traffic in a deterministic way.
ii) The same TN infrastructure must be able to accommodate all the 5G heterogeneous services.

TSN meets all the requisites referred to above and is, therefore, an attractive solution for connectivity in 5G. TSN is a set of standards specified by IEEE 802 aiming to provide deterministic services via IEEE 802 networks, i.e., assured streams transport with bounded latency, jitter, and frame loss. Also, TSN is acknowledged as a converged layer 2 (L2) network technology in all respects [3], harmonizing operation principles from both packet and circuit switching and thus fitting the necessities for conveying any traffic.

In this article, we delve into the application of asynchronous TSN networks for 5G backhauling, i.e., the use of TSN enabled networks not requiring the synchronization of their constituent forwarding plane devices (TSN bridges) for interconnecting the 5G system RAN and Core Network (CN). Although asynchronous scheduling increases the latency compared to synchronous one, it improves the network scalability and the link utilization as it does not require a network-wide coordinated time to schedule the traffic transmission of each stream over reserved time slots. Asynchronous TSN is ideal for conveying sporadic traffic with real-time constraints, such as that generated by the management planes in fog computing-enabled IoT services [4], and allowing its coexistence with best-effort streams.

The building block of asynchronous TSN is the IEEE 802.1Qcr Asynchronous Traffic Shaper (ATS), which is based on the Urgency-Based Shaper (UBS) proposed by Specht and Samii [5]. ATS enhances traditional asynchronous scheduling, in which a set of First Come, First Served (FCFS) queues, each associated with a traffic class and a priority level, are arbitrated by a strict priority transmission selection scheme. Specifically, ATS adds traffic regulation to conventional asyn-

chronous schedulers cost-effectively. In this way, per-hop traffic regulation is enabled in the network, thus avoiding the burst size or burstiness of the streams grows when they traverse the network, and the worst-case delay becomes arbitrarily large [6].

To the best of the authors' knowledge, there is no work addressing either the aspects related to the adoption of asynchronous TSN for 5G backhauling or the assessment of the capacity of this technology to accommodate 5G streams considering a realistic setup. In this work, we cover this gap. Specifically, the contribution of this article is threefold. First, we provide a comprehensive overview of the asynchronous TSN operation and its integration with 5G for backhauling. Regarding the integration, we sketch a blueprint of an orchestration framework for ATS-based TNs, relying on ETSI ZSM and IETF ACTN reference models. Second, we shed light on the flow allocation problem in ATS-based 5G Backhaul Networks (BNs), underlying the differences with flow allocation in synchronous networks. Also, we propose a deep Reinforcement Learning (RL)-based solution for handling the configuration complexity of these networks. Last, we evaluate the asynchronous TSN networks capacity to accommodate 5G data flows while assuring their performance constraints via a simulation-based proof-of-concept (PoC). Results show that the flow rejection probability exhibits a logarithmic dependency on bottleneck links utilization for our setup.

This article is organized as follows. We commence with an overview on TSN and its synergies with 5G, and clarify the IEEE 802.1Qcr Asynchronous Traffic Shaper (ATS) and Urgency-Based Shaper (UBS) operations. Next, we identify the key requisites for interoperating TSN BNs with 5G system. Following that, we address the flow allocation in 5G TSN networks, highlighting the possible implementation options and comparing them. After, we discuss the use of Machine Learning (ML) for handling the configuration complexity and leveraging the flexibility of TSN networks. Then we present the results of our performance evaluation. Finally, we draw the main conclusions.

II. Background

A. TSN and Its Synergies with 5G

TSN standards come to satisfy the needs of many industries, such as professional audio/video, electrical utilities, building automation systems, wireless industrial applications, and cellular transport networks, among others, for deterministic network services [3]. That is the ability to establish a multi-hop path over the network for a given flow with deterministic Quality of Service (QoS) guarantees in terms of latency, jitter, frame loss, and reliability.

One of the main TSN components is to define flow control mechanisms, which are in charge of handling the frames within the TSN bridges, with bounded low latency. The performance of these flow control mechanisms is mathematically analyzable, i.e., we can derive analytical expressions for the end-to-end (E2E) maximum delay, jitter, and frame loss experienced by any flow given its characteristics (e.g., committed data rate and burst size), the current state of the network (e.g., ongoing flows), and its allocation setup (e.g., priority level or Traffic Class (TC) for the ATS). Then, the flow allocation algorithms reverse somehow these expressions to find an allocation setup for every incoming flow while ensuring its QoS requirements. Other TSN pillars are

1) time synchronization for enabling a precise and common time reference among the bridges of a synchronous TSN network,
2) capabilities for ultra-reliability such as frame replication and elimination, and
3) resource management functionalities like the Stream Reservation Protocol and YANG models.

We refer the interested reader to [3] for a detailed and didactic explanation of all these concepts.

This work focuses on the use of asynchronous TSN for the deterministic transport of 5G data flows between the RAN and the CN. However, it is noteworthy that two well-known use cases encompass the combined functioning of TSN and 5G. The first use case is TSN and 5G integration for industrial automation as specified in 3GPP TS 23501 version 16.2.0 Release 16. For this use case, a whole 5G system behaves as a virtual TSN bridge through including translation capabilities at the edge. On the other hand, TSN IEEE 802.1CM standard defines a profile for conveying fronthaul streams between the radio heads and the baseband units for both the Common Public Radio Interface (CPRI) and enhanced CPRI (eCPRI).

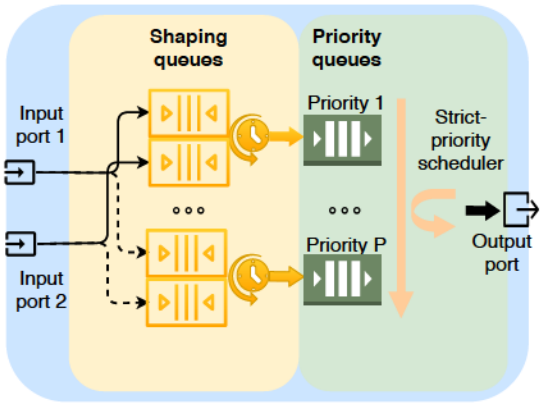
B. Asynchronous Traffic Shaper

The ATS defines an asynchronous method for handling the frames at the egress ports of the TSN bridges [3], [7]. The ATS specified in TSN standards [7] is based on the UBS originally proposed by Specht and Samii in [5]. Figure 1 sketches the internal architecture operation for both the UBS and the ATS.

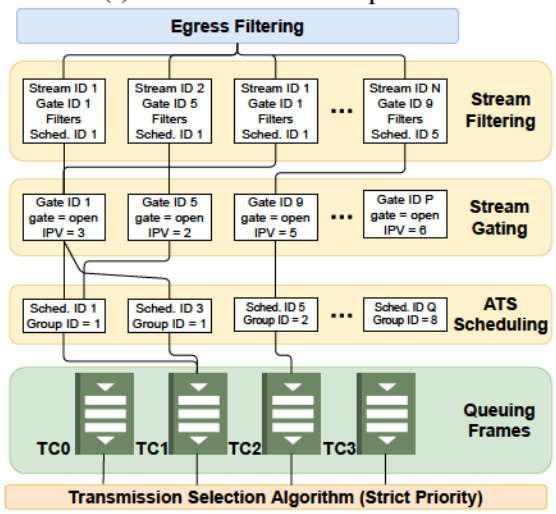
Figure 1a includes the UBS queuing model. The figure shows only one egress port for simplicity, but it shall be noted that there is an UBS instance per bridge egress port. The UBS consists of two stages of queuing: i) a set of shaped queues for interleaved shaping, and ii) a set of priority queues. All these queues follow a FCFS discipline.

The interleaved shaping is the novel and critical concept behind the ATS for achieving per-flow and per-hop traffic regulation cost-effectively. It enables the use of a single queue (shaped queue) for realizing the traffic regulation of a set flows, each with its own constraints. To that end, only the eligibility of the head-of-line (HOL) frame is checked, i.e., to examine whether the HOL frame is eligible for transmission according to the regulation constraints of its flow. If so, the frame is released for transmission to the following queuing level. Interestingly, the interleaved shaping does not increase the worst-case latency of the UBS [5], [6]. UBS supports per-flow leaky bucket shaping constraints to enforce the respective committed data rate and burst size for each flow.

The flows-to-shaped buffers assignment is subject to a set of rules, referred to as queue allocation rules (QARs). In a nutshell, each shaped buffer can be associated with only one ingress port (QAR1 rule), one priority level in the previous hop (QAR2 rule), and one internal priority level (QAR3 rule).



(a) UBS architecture and operation.



(b) ATS architecture and operation [7].

Fig. 1: UBS and ATS architectures and operation.

QAR2 and *QAR3* rules are required to provide deterministic QoS, whereas *QAR1* isolates the flows from different nodes, avoiding the propagation of non-conformant traffic overloads. These rules determine the required number of shaped queues to implement P priority levels. By way of illustration, an ATS with P internal priority levels and receiving traffic from N input ports requires at least $P \cdot N$ shaped queues.

The second stage in the UBS queuing hierarchy includes one FCFS queue per priority level in the scheduler. Each queue merges the output of all shaped queues assigned to the same priority level. The transmission selection algorithm at this stage is strict priorities, i.e., a given Traffic Class (TC) has preference over those with lower priority levels in accessing the physical medium. In other words, frames ready for transmission at a given TC have to wait as long as the higher priority TCs buffers have data.

Figure 1b shows the ATS architecture and its operation. The ATS can be regarded as the implementation of the UBS in TSN standards. The ATS forwarding process consists of the following steps:

1) *Stream filtering*. It refers to the ATS per stream classification and filtering. Each received frame is mapped to one filter using its priority and part of its connection identifier. Each filter is associated with a gate and ATS scheduler and might block a stream if any of its frames exceed the maximum service data unit (SDU).
2) *Stream Gating*. For ATS, this stage optionally assigns an internal priority value (IPV) that overrides the actual priority of a stream in step 4 of the forwarding process. Its purpose is to ease the adjustments in the ATS per-hop delay bound to meet the E2E flow requirements.
3) *ATS scheduling*. This stage realizes the interleaved shaping concept previously explained. The ATS schedulers associated with the same input port and TC in the previous hop (QAR1 and QAR2 rules) might be grouped.
4) *Queuing and Transmission*. As in the UBS's second queuing level, the frames are queued to the corresponding TC (priority level) buffer to be transmitted according to a strict priority scheduler. Observe that IPV assignment are enabled in Fig. 1b. Then, the flows to TCs mapping is done according to the IPV previously assigned. At every egress port, there is a manageable table to map flow priorities to TCs.

III. Flow Allocation in ATS-Based Networks

Given an optimization goal like the maximization of the flow acceptance ratio, the flow allocation involves the optimal selection of one or several paths for every TC and optimally finding the configuration for every ATS included in paths. An ATS flow allocation configuration primarily defines the flow-to-shaped buffer and flows-to-priority level assignments. These decisions are subject to the QoS constraints fulfillment of the incoming flows and all the ongoing flows. For critical flows, typical E2E performance requisites are the following:

- Frame delay budget: the upper bound for the time the network takes to transport a packet between the source and the destination.
- Maximum jitter delay: the permitted delay variation in the frame delivery from the source to the destination.
- Frame loss ratio: the fraction of the frames that are lost when they traverse the network.
- Reliability: the probability of network success to carry out the communication and fulfill the flow's required service level during its entire lifetime.

Besides the QoS requirements, the problem is also subject to technology constraints, such as the available links capacities and ATSs buffers size.

The flow allocation problem in asynchronous TSN networks is in its early stages. To the best of our knowledge, this combinatorial optimization problem has been only tackled by Specht and Samii in [8]. Specifically, they consider the maximization of delay surplus as the optimization objective and explore the following two approaches to solve it: i) a pure Satisfiability Modulo Theories (SMT) solver and ii) a heuristic approach dubbed Topology Rank Solver (TRS) to cope with the computational complexity of the SMT solver.

The flow allocation in synchronous TSN networks has attracted more attention from academia. For instance, in [9] Steiner et al. address the flow allocation in a IEEE 802.1Qbv Time-Aware Shaper (TAS)-based network. TAS schedules the

transfer of the traffic from the different flows over synchronized time slots (time-division multiplexing [10]). For this purpose, it uses a gate control list whose entries are binary vectors indicating which gates are open for transmission during the respective time window (see stream gating in Fig. 1b). Then the flow allocation problem in a TAS-based network consists of finding a feasible configuration for all the involved TASs. The configurable parameters for each TAS are the number of time windows and their duration, and the gate control entries.

Synchronous TSN is suitable to transport performance-sensitive traffic with periodic patterns like that one generated by closed-loop control systems in Industry 4.0 [11]. Conversely, asynchronous TSN networks perform well in scenarios where deterministic aperiodic (or sporadic) and best-effort traffics are predominant. However, the exact number of flows to be allocated, and their features are often unknown in these scenarios. Thus, the flow allocation in asynchronous TSN networks is a stochastic optimization problem in nature. We can harness historical data and predictive analytics to foresee traffic volume between all possible pairs of sources and destinations (traffic matrix) and its stochastic features (e.g., distribution of the sustainable rate per-5G QoS Identifier (5QI)) to use them as optimization process inputs.

We can distinguish two approaches for performing the flow allocation in TSN networks, namely, offline and online methods. Online methods compute the flow's allocation configuration right after it arrives at the network. Hence, they might run an optimization algorithm to find the allocation for every incoming flow. Conversely, offline methods compute a long-term configuration for the whole network for each type of traffic. Specifically, the flows are clustered into classes, e.g., according to their 5QI or 5QI and slice in 5G TNs, then, the allocation configuration is computed for each class. Offline methods require less state information (same configuration for all the flows of a traffic type), and the access control mechanism becomes a lightweight process that only needs to check whether there are enough resources (links capacities and buffer space) for the incoming flow. Conversely, online methods offer higher flexibility (flows with the same 5QIs might have different configurations) and greater agility to adapt to the changing network conditions.

IV. ZSM-based Transport Network Architecture

Figure 2 sketches a blueprint of a possible management and orchestration framework for TNs based on ETSI ZSM [12] and IETF ACTN [13] reference models. This architecture enables the customer to create and operate Virtual Networks (VNs) (TN slicing) while hiding the complexity of the underlying physical infrastructure. Also, it provides cross-domain coordination, which is crucial to ensure the cohesion and satisfiability of the configurations applied to the distinct domains. For instance, the E2E delay budgets imposed by the services need to be distributed among the different network domains (e.g., RAN, TN, and CN). We consider a fully centralized (SDN-like) TSN network. Since we target deterministic single-digit delays (i.e., less than 9 ms), it becomes trivial that the geographical area is of small size. Hence, considering

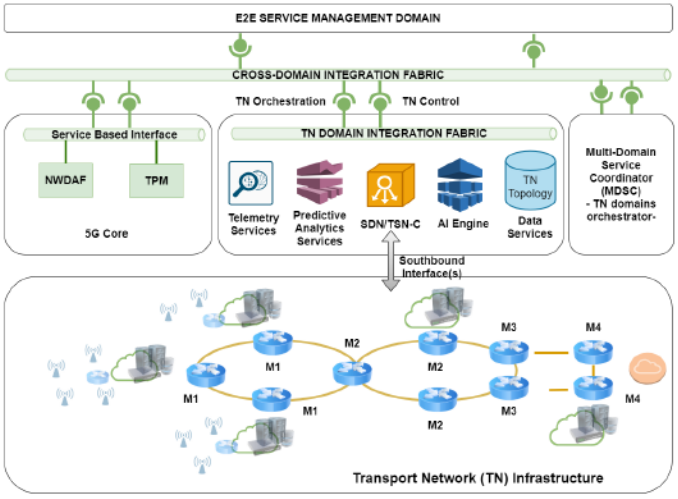


Fig. 2: Transport network management and orchestration architecture.

a centralized SDN-like architecture specific to that area is technically sound and shall not incur any scalability issues. Please refer to [3] for a detailed description of its components and operation.

The ACTN defines a three-tier reference model to realize the concept of VN as a service in TNs [13]. ACTN model comprises the following three components (see Fig. 2):

- *Traffic Provisioning Manager (TPM)*: VN service consumer in charge of issuing requests/commands to create and manage VNs (TN slices). This functionality might be part of the 5G core control plane. TPM also estimates the foreseen traffic volume between all pairs of sources and destinations in the VNs. To that end, it might harness historical data and predictive data analytics. Last, it maps 5G flows onto TSN traffic classes according to 3GPP slice and 5QI defined for each flow so that the TN can grant the service level agreed. The TSN traffic class assigned to each 5G flow is encoded in a packet header's immutable field.
- *Multi-Domain Service Coordinator (MDSC)*: It is responsible for orchestrating the different TN domains. Although Fig. 2 shows only one TN domain for simplicity, the TN might comprises multiple domains defined, for example, according to the underlying vendor technology or administrative zones. It translates the TPM requests/commands into a set of parameters specific to the underlying TN infrastructure so that the TSN controller can use them as input to configure the network. Last, it abstracts the underlay TN resources to hide the network configuration complexity from the TPM.
- *Software-Defined Networking (SDN)/TSN Controller (SDN/TSN-C)*: it acts as provisioning network controller. It controls the network devices and monitors and collects telemetry data about the network. It is also responsible for finding feasible configurations of the network and allocating enough resources to the VNs to fulfill the performance constraints of the 5G services. Specifically, the TSN-C has to solve the flow allocation problem

stated in the previous section to configure all ATSs within a given VN properly. For that purpose, TSN-C might leverage predictive data analytics and Artificial Intelligence (AI) techniques.

V. Machine Learning for TSN Networks Operation Optimization

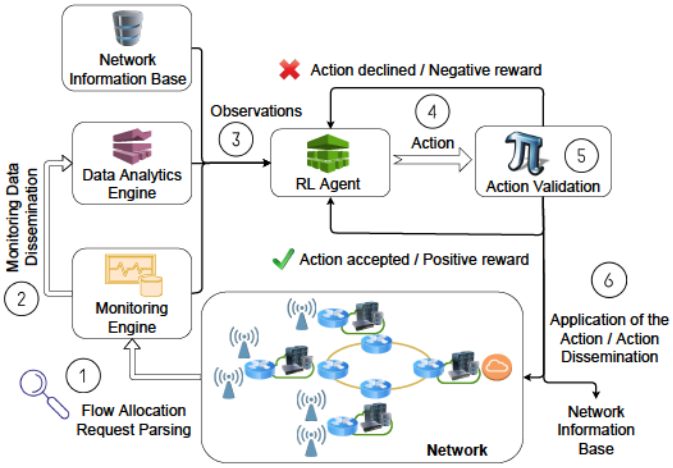


Fig. 3: Reinforcement Learning (RL) for flow allocation and time-sensitive networks optimization [14].

Due to the ever-increasing complexity of the wireless access networks, ML is envisioned as a cornerstone for achieving their full automation through assisting the different decision engines at the management planes [2], [15]. The composition of E2E services across the diverse infrastructure components easily yields overly complex or even intractable decision problems. Particularly, in the context of ATS-based BNs, the primary applications of ML are:

- Predictive data analytics on the upcoming workload to adapt the network configuration and drive the flow allocation process according to the working conditions.
- Assistance in the flow allocation and configuration processes. TSN networks offer a large space of possible configurations, and their optimization via analytical solvers exhibits high computational complexity. On the other hand, modeling the associated optimization programs might require a high domain knowledge for some optimization objectives.

The main concern for applying ML in this context is that we cannot ensure the validity of the decisions taken by the current ML methods. A valid decision corresponds to a flow allocation configuration that fulfills all the problem constraints. To that end, the feasibility check of the actions made by ML agents is required using the TSN analytical performance models. This approach is a practical solution as the per-hop performance bounds offered by ATS schedulers are known [5], [6]. Further, the feasibility checking process using these performance models is lightweight even for large networks.

Here, we consider deep Reinforcement Learning (RL) to solve the flow allocation problem in asynchronous TSN networks as its features are well suited for that problem. In contrast to alternative ML techniques, deep RL supports online learning efficiently, which is advisable for the model adaptability to the changing network conditions. In the same way, the RL exploration capability also allows adapting the agent's decision policy. On the other side, deep RL can handle large state-action spaces as required in medium and large scale TSN networks. Last, RL might act alone to output the solution directly from the input without any restriction on the optimization objective.

Figure 3 shows an online RL-based solution for the flow allocation policy-making in ATS-based networks proposed in our previous work [14]. First, every incoming flow allocation request is parsed to determine the flow type and characteristics (step 1). Then the flow characteristics, along with the traffic predictive data analytics and the network information and status, are taken by the agent as observations. Next, the agent outputs an action, which is validated through verifying the following conditions analytically:

- The E2E delay/jitter experienced by the incoming flow has to be lower than its E2E delay/jitter budget.
- If the flow were allocated, the ongoing flows must keep experiencing an E2E delay lower than their E2E delay budgets.
- The aggregated burstiness allocated to any shaping buffer has to be lower than its size.
- The aggregated rate allocated to any link must be lower than its capacity.
- The flow to shaped queue allocation rules (QAR) have to be met.

If the action is validated, the agent will be positively rewarded, and the action applied. Otherwise, if it would impact anyhow the deterministic performance requirements, it is simply not applied. In this way, the analytical models' information is transferred to the agent, and, most importantly, the flow allocation process becomes fully reliable. Also, it is remarkable that some performance metrics like the worst-case delay can be only efficiently estimated through analytical models.

Although deep RL fits the necessities of the flow allocation problem, at least theoretically, it has several practical issues (e.g., low interpretability, complicated configuration, inefficient training process, and lack of robustness). Then, it is still interesting to explore alternative ML techniques to find feasible configurations. For instance, robust decision trees or support vector machines based classifiers could be used to assist a master algorithm in lower-level decisions, such as the mapping of 5QI onto TSN TCs. Finally, it shall be noted that the ML method that yields the best performance for the flow allocation problem is still an open question. Thus, further research is required to address it.

VI. Results

This section includes simulation results for evaluating the capacity offered by an ATS-based BN to accommodate data flows. To that end, we carried out a simulation-based PoC. The simulator consists of the following five main modules:

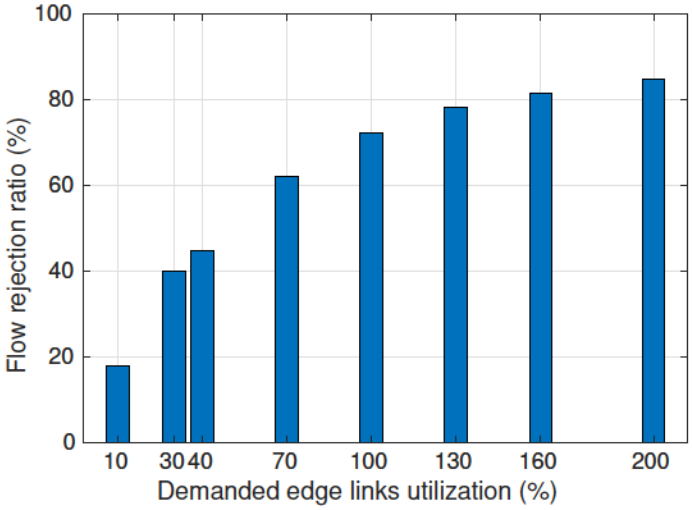
- *Network generator*. It is in charge of the network topology generation. In particular, we consider a backhaul

network topology like that shown in Fig. 2, where each M1 node provides access to six gNodeBs (5G base stations) through 1 Gbps point-to-point links. The link capacities for the access ring (M1 devices), aggregation ring (M2 devices), and core ring (M3 and M4 devices) were set to 10 Gbps, 40 Gbps, and 100 Gbps, respectively. All link capacities of the access ring (M1 devices) and aggregation ring (M2 devices) were set to 10 Gbps, 40 Gbps, and 100 Gbps, respectively. All the links were duplicated for reliability. We consider both M4 nodes to provide access to the 5G system core and the data network. All ATSs include eight shaped buffers without buffer space limitation and eight priority levels.

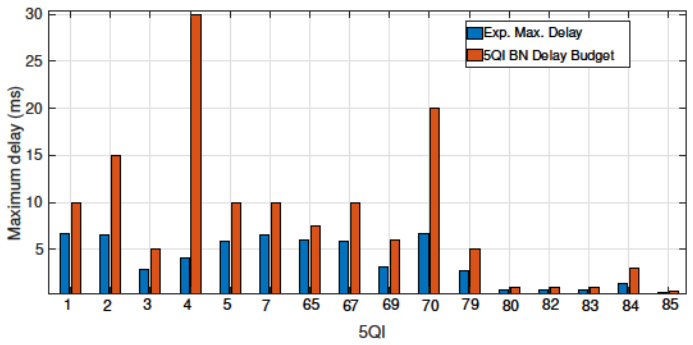
- *Flow arrival process generator*. It simulates the flow arrivals according to a renewal process. The flows inter-arrival times distributions implemented were: i) exponential, ii) Erlang-2, and iii) hyperexponential. The Erlang-2 and hyperexponential distributions can approximate any probability function with a coefficient of variation lower and greater than one, respectively. We observed a similar network performance regardless of the arrival process type. The 5QI, source (M4 devices), and destination (gNodeB) for each flow are selected randomly following discrete uniform distributions. The demand of resources and QoS constraints for every 5QI are contained in Table I.
- *Network optimizer*. This module aims at optimizing the configuration of the BN for a given goal. In particular, for this PoC, we consider an offline solution (see Section III) mixing analytical and heuristic methods for the flow rejection ratio minimization. The solution heuristically assigns predefined paths for each combination of 5QI, source, and destination, relying on predictive data analytics to minimize the workload imbalances throughout the network. For the delay critical Guaranteed Bit Rate (GBR) flows (e.g., 5QIs 82, 83, 84, and 85), frame duplication through disjoint paths was set to ensure their reliability. The BN delay budget for each 5QI is distributed among the links of the respective path inversely proportional to the link capacity. The 5QI BN delay budget was set to 10% of the E2E one defined in 3GPP standards. Last, a convex program for minimizing the flow rejection probability subject to the constraints listed in Section Section III is run to find the optimal configuration of every ATS in the network.
- *Flow admission control*. It is executed at runtime for every incoming flow for checking whether there are enough resources to accept the flow.
- *Flow allocator/releasing*. It updates the status of the network-related variables (e.g., available resources) for every flow entering or leaving the network.

Figure 4a depicts the flow rejection ratio as a function of the demanded link utilization at the access links interconnecting M1 devices and gNodeBs. Every point shown in Fig. 4a was obtained via simulating the arrivals and departures of 1.8M of flows. As observed, the flow rejection ratio depends logarithmically on the demanded edge link for our setup. Specifically, the logarithmic function $24 \cdot ln(U_l) - 39$ accurately fits the measured points in Fig. 4a, where $ln()$ and U_l denote the natural logarithm function and the access link demanded utilization, respectively. We observed that our algorithm offers high rejection probability (penalizes) flows with high data rate demands, e.g., those with 5QIs 2, 7, and 5 (see Table I), as it seeks for maximizing the number of accepted flows.

Figure 4b shows both the BN delay budget per 5QI (labeled as "5QI BN Delay Budget"), which is 10% of the E2E delay budget defined in 3GPP standards (see fifth column in Table I), and the worst-case delay per 5QI obtained through simulation (labeled as "Exp. Max. Delay"). As observed, the delay constraint is met for every 5QI. The maximum delay experienced by each 5QI primarily depends on its priority level in the TSN network, which is assigned by our algorithm. This fact explains the variability observed in the obtained maximum delay for the different 5QIs.



(a) Flow rejection ratio.



(b) Worst-case delay.

Fig. 4: ATS-based BN performance.

VII. Concluding Remarks

The imminent upgrading of Fifth Generation (5G) mobile networks for supporting ultra-reliable and delay-sensitive services will require breakthrough architectural changes, operation optimizations, and technologies to every mobile network segment, including the transport network. In this regard, Time-Sensitive Networking (TSN) is envisaged to play a central

TABLE I: Flow types characteristics

5QI	Prio	Rate (Mbps)	Burstiness (bits)	Dmax (ms)	Rmin (%)	Income	Avg. Dur. (s)	Lmax (bits)	Ex. service
1	20	0.064	6400	100	95	1	130	6400	Conv. voice
2	40	1.5	2250000	150	95	1	130	10832	Conv. video
3	30	0.1	5000	50	95	1	1200	5000	Real Time gaming
4	50	0.083	2500	300	95	1	231	2500	Non-conv. video
5	10	0.01	2040	100	95	1	130	2040	IMS signaling
7	70	2	10832	100	95	2	231	10832	Live video
65	7	0.064	6400	75	95	1.5	600	6400	MCPTT
67	15	0.5	10832	100	95	1.5	300	10832	MCV
69	5	0.01	2040	60	95	1.5	600	2040	MCPTT signalling
70	55	0.01	2040	200	95	1.5	600	2040	Mission Critical Data
79	65	0.0003	2040	50	95	1.5	300	2040	V2X messages
80	68	5	32496	10	95	1.5	600	10832	Augmented Reality
82	19	0.1	2040	10	99.999	2.5	1200	2040	Discrete Automation
83	22	0.2	10832	10	99.999	2.5	1200	10832	Discrete Automation
84	24	0.3	10832	30	99.999	4	1200	10832	Intelligent transport systems
85	21	0.3	2040	5	99.999	3	1200	2040	Electricity distribution HV

Internet Multimedia Subsystem (IMS); Mission Critical user plane Push To Talk voice (MCPTT); Mission Critical Video user plane (MCV); Vehicle-to-Everything (V2X); High Voltage (HV). Some data included in this table (e.g., E2E delay budget, critical flows burstiness, default priorities, and services) were extracted from 3GPP TS 23501 version 16.2.0 Release 16.

role in many scenarios for providing connectivity among the network functions and services of the upcoming mobile networks. In this article, we have motivated and offered fresh insights into the use of asynchronous TSN networks for 5G backhauling. Currently, the ATS is the building block of the asynchronous TSN enabled bridges. The relevance of this bridge egress port scheduler lies in cost-effectively enabling per-hop traffic regulation. Per-hop shaping is a solution against the large worst-case delays exhibited by FIFO buffering. On the other site, asynchronous networks are preferred over synchronous ones because of its lower complexity and better scalability as they do not need a precise and common reference time reference among all the network devices. Besides, synchronous networks offer worse network resource utilization.

As the main contributions of this work, we have provided a comprehensive overview of the key concepts of asynchronous TSN networks. Second, we have discussed the flow allocation problem in ATS-based networks and presented the different approaches to address it. Next, we have tackled the adoption of Machine Learning (ML) techniques for automating the management of TSN BNs. There, we have stressed the importance of assisting the ML agents' decisions with analytical performance models for ensuring a fully reliable policy-making. Last, we have carried out a simulation-based proof-of-concept (PoC) for assessing the capacity of ATS-based BNs for accommodating heterogeneous and sporadic data flows.

ACKNOWLEDGMENT

This work is partially supported by the European Union's Horizon 2020 research and innovation programme under the CHARITY project with grant agreement No. 101016509. It is also partially supported by the Academy of Finland Project CSN - under Grant Agreement 311654 and the 6Genesis project under Grant No. 318927.

REFERENCES

[1] I. Afolabi, T. Taleb, K. Samdanis, A. Ksentini, and H. Flinck, "Network slicing and softwarization: A survey on principles, enabling technologies, and solutions," *IEEE Communications Surveys Tutorials*, vol. 20, no. 3, pp. 2429–2453, 2018.

[2] T. Taleb, I. Afolabi, K. Samdanis, and F. Z. Yousaf, "On multi-domain network slicing orchestration architecture and federated resource control," *IEEE Network*, vol. 33, no. 5, pp. 242–252, 2019.

[3] A. Nasrallah, A. S. Thyagaturu, Z. Alharbi, C. Wang, X. Shao, M. Reisslein, and H. ElBakoury, "Ultra-low latency (ull) networks: The ieee tsn and ietf detnet standards and related 5g ull research," *IEEE Communications Surveys Tutorials*, vol. 21, no. 1, pp. 88–145, Firstquarter 2019.

[4] J. An, W. Li, F. L. Gall, E. Kovac, J. Kim, T. Taleb, and J. Song, "Eif: Toward an elastic iot fog framework for ai services," *IEEE Communications Magazine*, vol. 57, no. 5, pp. 28–33, 2019.

[5] J. Specht and S. Samii, "Urgency-based scheduler for time-sensitive switched ethernet networks," in *2016 28th Euromicro Conference on Real-Time Systems (ECRTS)*, July 2016, pp. 75–85.

[6] J.-Y. Le Boudec, "A theory of traffic regulators for deterministic networks with application to interleaved regulators," *IEEE/ACM Trans. Netw.*, vol. 26, no. 6, pp. 2721–2733, Dec. 2018. [Online]. Available: https://doi.org/10.1109/TNET.2018.2875191

[7] "IEEE Draft Standard for Local and metropolitan area networks–Bridges and Bridged Networks Amendment: Asynchronous Traffic Shaping," *IEEE P802.1Qcr/D2.1, February 2020*, pp. 1–152, 2020.

[8] J. Specht and S. Samii, "Synthesis of queue and priority assignment for asynchronous traffic shaping in switched ethernet," in *2017 IEEE Real-Time Systems Symposium (RTSS)*, Dec 2017, pp. 178–187.

[9] W. Steiner, S. S. Craciunas, and R. S. Oliver, "Traffic planning for time-sensitive communication," *IEEE Communications Standards Magazine*, vol. 2, no. 2, pp. 42–47, 2018.

[10] N. Nasser, L. Karim, and T. Taleb, "Dynamic multilevel priority packet scheduling scheme for wireless sensor network," *IEEE Transactions on Wireless Communications*, vol. 12, no. 4, pp. 1448–1459, 2013.

[11] T. Taleb, I. Afolabi, and M. Bagaa, "Orchestrating 5g network slices to support industrial internet and to shape next-generation smart factories," *IEEE Network*, vol. 33, no. 4, pp. 146–154, 2019.

[12] C. Benzaid and T. Taleb, "Ai-driven zero touch network and service management in 5g and beyond: Challenges and research directions," *IEEE Network*, vol. 34, no. 2, pp. 186–194, 2020.

[13] D. Ceccarelli and Y. Lee, "Framework for Abstraction and Control of TE Networks (ACTN)," Informational, Internet Engineering Task Force (IETF), Request for Comments 8453, August 2018.

[14] J. Prados-Garzon, T. Taleb, and M. Bagaa, "Learnet: Reinforcement learning based flow scheduling for asynchronous deterministic networks," in *2020 International Conference on Communications (ICC)*, June 2020, pp. 1–6.

[15] A. Zappone, M. Di Renzo, and M. Debbah, "Wireless networks design in the era of deep learning: Model-based, ai-based, or both?" *IEEE Transactions on Communications*, vol. 67, no. 10, pp. 7331–7376, Oct 2019.