On the Benefits of Multi-Agent Systems for Operational Expenditure Savings

Marc Ruiz[1*], Hailey Shakespear-Miles[1], Sima Barzegar[2], Andrea Sgambelluri[3], and Luis Velasco[1]
[1] *Optical Communications Group (GCO), Universitat Politècnica de Catalunya (UPC), Barcelona, Spain*
[2] *Barcelona Supercomputing Center (BSC), Barcelona, Spain*
[3] *Scuola Superiore Sant'Anna (SSSA), Pisa, Italy*
e-mail: marc.ruiz-ramirez@upc.edu

ABSTRACT
This paper evaluates the performance in terms of operational expenditures (OpEx) reduction of multi-agent systems (MAS) in charge of the autonomous control of traffic flows supporting 6G connectivity services with stringent requirements. MAS enables in near-real time adaptation of resources to support connectivity services through multiple paths. In this way, energy savings are achieved with respect to the benchmarking approach, where connectivity is statically managed and consequently, resources for guaranteeing committed requirements under any circumstance are always active and consuming energy.

Keywords: ***6G, Multi-Agent Systems, Energy Consumption Savings***

I. INTRODUCTION

6G connectivity services foreseen for the next decades will impose strict Quality of Service (QoS) requirements, such as high bandwidth, ultra-low end-to-end latency, and ultra-high reliability, making them a key challenge for next-generation network operators. To support these complex demand scenarios, networks are evolving towards the network programmability paradigm i.e., advanced programmable switches equipped with high bandwidth pluggable interfaces. These adaptive and flexible network solutions facilitate the deployment of high-speed interfaces (from 100 Gb/s to 400 Gb/s) in a modular and cost-effective manner, allowing for greater flexibility in network design and scalability [1]. However, such programmability typically comes with large energy consumption, which is especially challenging when considering the scalability of the solutions.

Among different use cases supported by network programmability, autonomous flow routing is attracting special research interest. Thus, by dynamically selecting among multiple available paths based on different criteria, individual connectivity services (flows) can be accordingly balanced in near-real time in order to guarantee strict QoS requirements while minimizing multi-objective cost functions [2]. However, for the success of this autonomous operation, programmable networks need to be fueled with distributed intelligence and in-band telemetry in order to coordinate routing decisions with target QoS requirements and desired operational objectives. To this aim, multi-agent systems (MAS) are key to guarantee such adaptive and near-real time autonomous operation of 6G connectivity services [3]. As consequence of adopting all these technologies, autonomous networks promise to reduce operational expenditures (OpEx) including energy consumption reduction.

This paper focus on evaluating the performance of using MAS for autonomous network operation of traffic flows in support of 6G connectivity services in terms of energy consumption. In particular, we aim at evaluating the energy consumption savings that a MAS-based autonomous flow routing use case provides with respect to a benchmarking static operation approach, where flows are routed following either single path routing or multi-path routing with static policies, e.g., equal cost multi-path (ECMP) strategy.

II. AUTONOMOUS FLOW ROUTING OPERATION

In this section, we summarize the autonomous flow routing operation system considered in this work, based on the work presented in [2]. In a pure SDN-based approach, route selection can be performed at flow provisioning time based on the network topology and current and expected network conditions. However, in the distributed intelligence approach sketched in Fig. 1a, a MAS deployment aims at dynamically controlling a connectivity service (traffic flow) between a source and a destination node. At the source node, the flow routing module optimizes the path selection to meet committed QoS (i.e., do not exceed a maximum delay value) while minimizing cost. Traffic is routed through different paths, with multiple sub-flows following the same route. Upon reaching the destination, e2e delay is measured, statistics computed, and relayed to participating node agents.

The scenario depicted in Fig. 1b is devoted to illustrate the challenging scenario where routing decision making is needed due to the variation of the traffic flow under-control (labelled as *f1*) and also the network dynamicity introduced by other traffic flows (*f2* and *f3*). In particular, the illustrative network scenario contains five packet nodes and three different packet flows: R1-R3 (*f1*), R1-R4 (*f2*), and R2-R3 (*f3*), along with traffic variations over time for all flows. Initially, R1 has two alternative paths for flow *f1*: R1-R2-R3 and R1-R4-R5-R3. Note that traffic can be routed via R2 (the shortest path), if acceptable delay for flow R1-R3 can be assured. However, when flow R2-R3 is established or its traffic increases, the delay for flow R1-R3 exceeds the maximum, prompting R1 to divert traffic through the alternative route via R4, that can be longer (and consequently, have higher cost) to

979-8-3315-9777-1/25/$31.00 ©2025 IEEE

mitigate delay. This adaptive routing continues as traffic and delay conditions fluctuate, with R1 adjusting routing through available paths to optimize delay/cost.

This autonomous flow routing operation can be combined with port management actions, to dynamically dimension the resources supporting the available paths. Focusing on the flow source node, this means activating/deactivating interfaces (ports), to adapt capacity to current traffic needs. Although node ports can be pre-assigned and fixed to paths, which entails a fixed deployment cost, dynamic activation/deactivation reduces OpEx costs coming from energy consumption. Thus, MAS can anticipate the need of adding/releasing ports in order to fulfil QoS in an energy-efficient way.

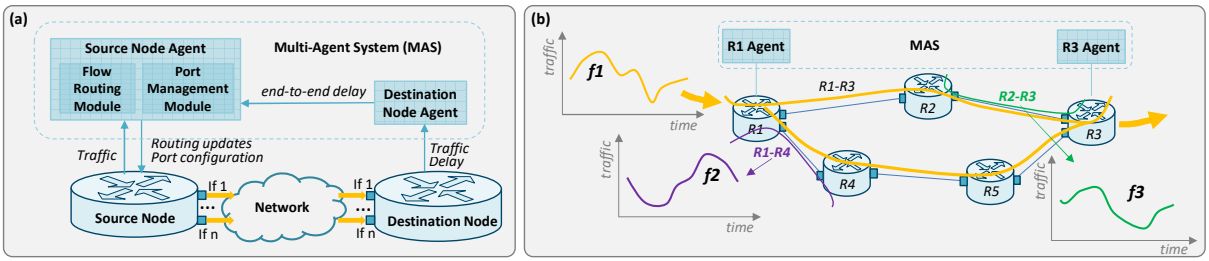

Fig. 1: MAS for autonomous flow routing (a); network scenario (b)

III. NODE ARCHITECTURE AND ENERGY CONSUMPTION MODEL

In order to support the MAS-based autonomous flow routing presented in previous section, we consider the programmable switch architecture in [4] as reference node architecture. Note that this kind of devices have shown promising energy consumption savings by adopting dedicated control plane mechanisms oriented to maximize energy efficiency [5]. The programmable switch architecture centers around the Intel Tofino Application-Specific Integrated Circuit (ASIC) model, specifically within the Edgecore Wedge100BF-65X platform. This switch utilizes a protocol-independent switch architecture (PISA), enabling programmability through the P4 language. The switch is composed of a data plane and a control plane. The data plane executes match-action pipelines defined in P4, allowing for custom packet processing rules. The control plane, managed via software entities such as an SDN controller and/or a MAS agent, is responsible for configuration and state management. Additionally, the infrastructure includes telemetry capabilities using P4-INT, which supports real-time monitoring of packets traversing the network, leading to telemetry agents integrated with the MAS infrastructure [3].

The switch supports a variety of port types and configurations such as Quad Small Form-Factor Pluggable (QSFP) ports capable of 100G operation, with a total switching capacity of 3.2 Tb/s. These ports are suitable for high-speed data transmission in metro scenarios, where the required capacity of different flows through multiple routes can be properly managed with the granularity of such ports. However new generation of Tofino-based programmable switches are supporting larger switching capacities, e.g. 12.8 Tb/s, by supporting 400G QSFP with double density (QSFP-DD) capabilities [6]. This kind of switches are more adequate for core network scenarios, where flows under control might rapidly scale to ultra-high bitrates and more stringent QoS requirements.

Assuming the programmable switch model described before, we adopt the energy consumption computation approach presented in [7] and successfully applied in [5]. Thus, the switch has a baseline consumption (*base*) that is a fixed value in the state where no ports are enabled and it is independent of data/control plane software components running in the switch. Then, at a given time, each port can be in one of the following states: *i*) disabled, as for the baseline case, where they are not contributing to increase energy consumption; *ii*) enabled, where the port adds an additional power consumption cost (*enable*) due to its port configuration (administratively up, but optical interface not yet enabled); *iii*) active, where optical interface is up and running, which entails an additional power cost (*active*) that is typically larger than *enable* cost. Finally, the energy consumption scales with the utilization of its resources, namely, by processing packets at the ports and in the ASIC. Without loss of generality, we assume a linear cost (*use*) proportional to the traffic supported at each port.

Hence, as a result of MAS decision making (flow routing + port management), a given port p at a given time t can be active (which is denoted as $x_{pt} = 1$; 0, otherwise). If active, the amount of traffic in a port (which is denoted as y_{pt}) is computed assuming load balancing among all the ports of a path. Equation (1) models the energy consumption (EC) of a given switch at a given time t as the contribution of the different parameters and variables:

$$EC(t) = base + \sum_{p \in P} \left(enable + (active - enable) \cdot x_{pt} + use \cdot y_{pt} \right) \quad (1)$$

Table 1 details the different power consumption parameters for the two configurations described above, namely, metro configuration (3.2 Tb/s switch with 100G QSFP ports) and core configuration (12.8 Tb/s switch with 400G QSFP-DD ports). These values, which are in line with the reference power consumption values in [7], as well as

product specifications in [6], will be consistently used in the results presented in next section.

Table 1: Power Consumption Parameters (in Watts)

Scenario	Sw. Capacity	100 Gb/s Ports	400 Gb/s ports	*base (switch)*	*enable (port)*	*active (port)*	*use (b/s)*
Metro	3.2 Tb/s	32	-	108W	1.8W	4.5W	1.3e-11W
Core	12.8 Tb/s	-	32	245W	3.2W	12W	2.1e-11W

IV. RESULTS

For numerical evaluation purposes, we developed a Python simulator that reproduces network scenarios as the one presented in Fig. 1b. In particular, it aims at generating traffic flows of variable bitrate between source and destination nodes according to realistic weekly patterns obtained from real operator network traffic. The traffic of each flow is routed through multiple paths, and each path is assigned to switch ports at each node, so that capacity is guaranteed. The simulator can be configured to reproduce two main approaches. On the one hand, the *static* approach assumes a fixed number of ports assigned to each path and performs ECMP to balance flow traffic among available paths. The number of ports is dimensioned to guarantee enough capacity at peak traffic, assuming perfect knowledge on future incoming traffic, which leads to tight and accurate dimensioning. This mode does not guarantee QoS requirements since no intelligent decision making is performed by any agent. On the other hand, the MAS approach follows the operation explained in Section II, i.e., telemetry is used to monitor actual flow QoS at destination node and perform routing actions and port management in the source node. According to those actions, the number of enabled/active ports assigned to every path change with time.

Fig. 2 and Fig. 3 show an example of performance of static and MAS approaches, respectively, for a scenario reproducing the network and traffic in Fig. 1, assuming metro configuration for the switches. In particular, the figures show traffic at each of the alternative paths of flow *f1* (a-b) and the power consumption at source node (c). The traffic of flows *f2* and *f3* follow different weekly patterns but similar traffic volumes than that of *f1*. It is worth mentioning that, with the current traffic loads, both operation approaches allow accomplishing QoS requirements during all the time, so both are equally valid from the service assurance perspective. As expected, the ECMP operation of static approach perfectly balances traffic among alternative paths, whereas MAS adapts routing to use only two paths when QoS assurance requires it. In this way, deactivation of ports is possible during significant time in a day. This translates to remarkable power consumption savings, which can be observed in Fig. 3c, in contrast with the flat, constant power consumption of static approach in Fig. 2c.

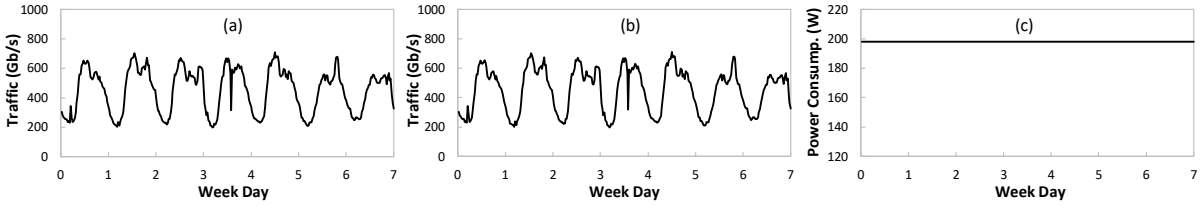

Fig. 2: Static approach: traffic through route A (a) and route B (b), and power consumption (c)

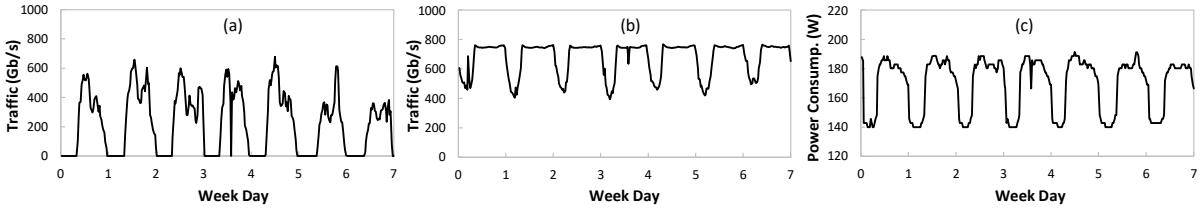

Fig. 3: MAS approach: traffic through route A (a) and route B (b), and power consumption (c)

The benefits of MAS approach in terms of energy consumption reduction sketched in Fig. 3, are exhaustively evaluated in Fig. 4. The scenario has been extended with a third route, that uses a direct bypass to avoid congestion caused by other flows. Moreover, low and high congestion scenarios have been generated by reducing and increasing, respectively, the traffic of flows *f2* and *f3* flows. The figures show the average and maximum power consumption of *f1* source node as a function of *f1* traffic volume (normalized to the maximum traffic that can be supported without violating QoS in the absence of congestion). Curves are depicted in each case only for the loads that guarantee QoS assurance. Note that, in line with the exhaustive analysis carried out in [2], the static approach using ECMP routing can support less traffic to guarantee the same QoS requirement than MAS approach. In terms of power consumption, static approach is planned in a way that alternative routes are activated only when the traffic peak for that normalized load requires it, which produces staggered power consumption as soon as new paths are activated. Moreover, weekly average and maximum traffic consumption remains similar, due to the

nature of the approach.

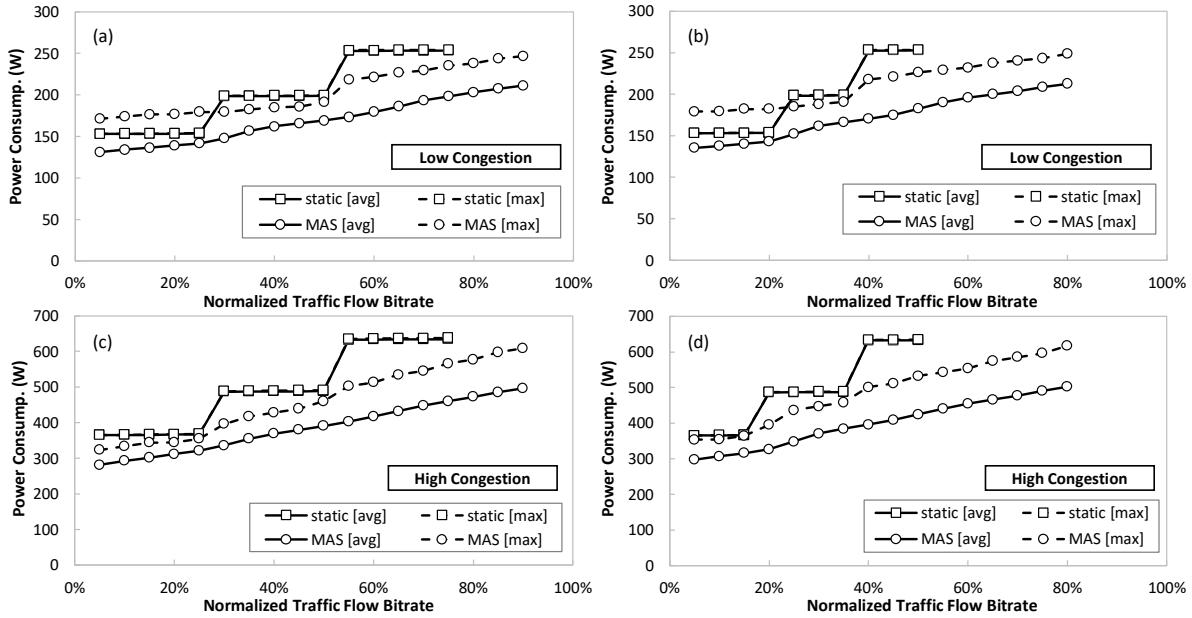

Fig. 4: Performance of static vs MAS-based operation for metro (a-b) and core (c-d) configurations

Observing the results of MAS approach, one can observe a smooth increasing of power consumption with traffic load, which allows remarkable power consumption reduction in terms of both average and maximum values. Table 2 extends the information in Fig. 4 with the relative savings with respect to the static approach for different switch configurations and congestion scenarios. The table also includes the energy efficiency computed as the amount of flow traffic that can be served with QoS guarantees given a target power consumption (200W in metro configuration and 500W in core configuration). The results show large benefits in terms of power savings and energy efficiency derived from adopting MAS infrastructure to control programmable networks in near-real time.

V. CONCLUSIONS

Adopting MAS for near-real time operation allows reducing energy consumption in programmable networks, achieving switch power savings larger than 20% under different scenarios. Indeed, energy efficiency is largely increased (even doubled) with respect to static operation, which allows us concluding that MAS-based autonomous network operation clearly reduces OpEx in 6G networks.

Table 2: MAS approach results (w.r.t static)

Config	*Congestion*	*Power Savings* mean	max	*Energy Efficiency Gain*
Metro	*Low*	19%	32%	54%
	High	21%	33%	85%
Core	*Low*	25%	36%	75%
	High	26%	37%	114%

ACKNOWLEDGEMENTS

The research leading to these results has received funding from the Smart Networks and Services Joint Undertaking under the European Union's Horizon Europe research and innovation programme under Grant Agreement No. 101096466 (DESIRE6G), and from the ICREA Institution.

REFERENCES

[1] Cisco Systems, Growing the Network with 400 Gbps Coherent Pluggable Optics, Cisco White Paper, Oct. 2022. [Online]. Available: https://www.cisco.com/c/dam/en/us/solutions/collateral/service-provider/routed-optical-networking/white-paper-sp-400g-coherent-optics.pdf (last access: 05/09/2025).

[2] S. Barzegar *et al.*, "Autonomous Flow Routing Based on Deep Reinforcement Learning," in Proc. ICTON 2024.

[3] P. Gonzalez *et al.* "Near-Real-Time 6G Service Operation Enabled by Distributed Intelligence and In-Band Telemetry," J. of Opt. Comm. and Netw., 2025.

[4] M. Groshev *et al.,* "D5.1: Preliminary experimental setup and data set collection", project deliverable, https://doi.org/10.5281/zenodo.10609158 (last access: 05/09/2025).

[5] J. A. Brito *et al.*, "Energy-Aware Edge Infrastructure Traffic Management Using Programmable Data Planes in 5G and Beyond, " MDPI Sensors, 25, 2375, 2025.

[6] Edgecore Programmable Switches specifications, https://www.edge-core.com/solution-inquiry.php?cls=6&id=95, (last access: 05/09/2025).

[7] J. Lim *et al.*, "How Much Does It Burn? Profiling the Energy Model of a Tofino Switch," ETH Zurich Library, available online.