Activity Report on ENI PoC#22: NTN, 5G SRv6 integration for TSN (Time Sensitive Network) by Artificial Intelligence

Concepts Designing on the RAN by Using the IPv6 Protocols

During this first period of activity, we investigate the IP Radio Access Network (IPRAN) design for integrating Non-Terrestrial-Networks (NTN) with the terrestrial one. Indeed, an IPRAN is a wireless access network that utilizes IP-based network layer protocols to integrate heterogeneous access technologies, particularly for back-haul scenarios. As IP-based mobile transport networks become increasingly prevalent in carrier network development, traditional solutions fail to meet future data transport needs. IPRAN emerges as a direct and widely adopted solution for wireless access, functioning as a transport network between access networks (e.g., gNB, WiFi access points) and the core network. It builds on traditional IP networks, incorporating Operations, Administration, and Maintenance (OAM) and Quality of Service (QoS) mechanisms to address wireless backhaul requirements. IPRAN provides IP connectivity between base stations and controllers, offering carriers a flexible, reliable, cost-effective backhaul network solution.

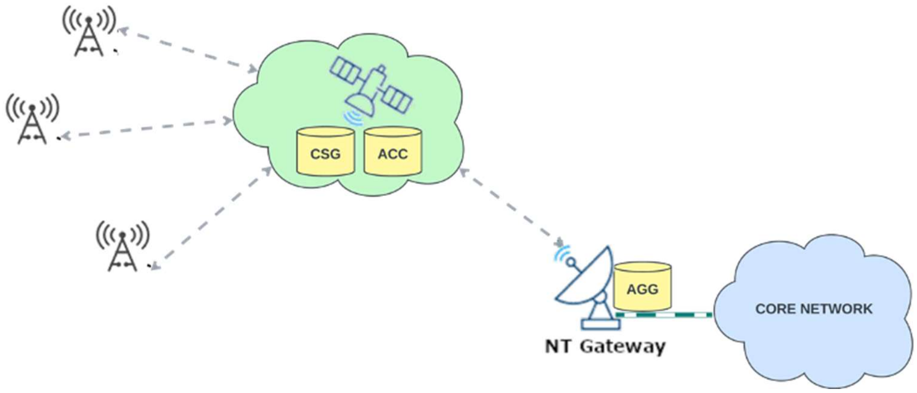


Figure 1 IP-RAN Architecture

Our IP-RAN was designed considering non-terrestrial technologies that offer enabling technologies, such as carrier aggregation, multi-access and connectivity, and multi-transmission point services when integrating with the terrestrial network. This allows the UE to communicate and leverage multiple access resources. *Figure 1* shows an example of IP-RAN adopted during the activities conducted in this PoC. In the proposed solution, non-terrestrial platforms can serve as the interface toward the CN to provide continuum connectivity. In this scenario, the IP-RAN can provide the tools for accomplishing the objective of managing diverse access resources and aligning them with user needs and technical demands such as data rate, latency, and reliability adaptability of the communication services to the changing UE demands and topology conditions, such as the changing topology that arises in using the LEO constellation.

The satellites can operate as Cell Site Gateway (CSG), providing a secure and reliable connection between gNB and the CN for mobile users. The CSG can enable reliable communications between mobile users and services devoted to network management, telemetry, and security. Figure 2 shows the protocol stack underlying the proposed IP-RAN solution for the multi-connectivity achieved through the LEO satellite. We propose to use LEO satellites for their capability to offer access functionality for wide geographical areas, maintaining the latency comparable with the constraints of the 5G technology.

Figure 2 (a) represents the protocol stack for the control plane (CP) instead of *Figure 2* (b) for the user plane (UP). In this design, we use the IP protocol on board the satellites for routing data and the control messages, such as those of the Xn interface over the Intra-Satellite-Links (ISL), for facilitating the implementation of multi-path technologies. Further, this solution allows us to manage the route paths over the LEO's links by exploiting more reliable and efficient routing protocols such as SRv6 to make this management transparent toward the 5G infrastructure even for the change of topology due to the variability of the LEO's links.

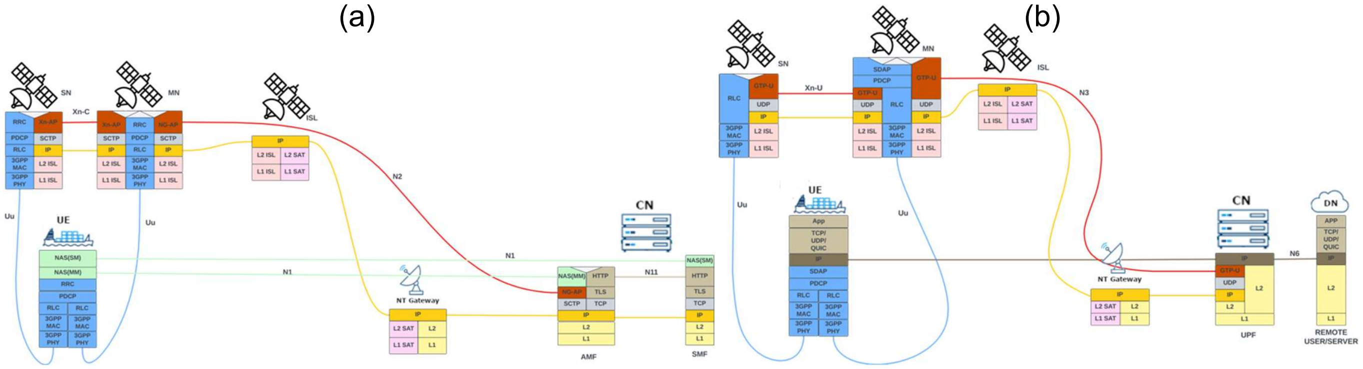


Figure 2 Protocol stack IP-RAN

○ The Segmentation Routing over IPv6 (SRv6) and IP-RAN

The SRv6 protocol was defined to facilitate the functionalities for deploying multiple services (005 ETSI GR IPE, 2022; 3GPP RAN, 2014; 3GPP RAN, 2020). SRv6 is a source protocol that acts at the data plane, and maintaining a protocol simplification allows the propagation of packet forwarding rules, called segments or policies, along the nodes composing the path. By using SRv6, it is also possible to define the segment as a function for packet processing called at a specific location in the network. By definition of segment routing in the IETF document RFC 8402 (Clarence, et al., 2018), the source node of the path is referred to as network ingress, and it is in charge of steering packets across an ordered segments list. A segment can represent any topological or service-related instruction, such as steering packets to a destination using a path that minimizes latency or processing packets with a specific QoS treatment. The network ingress is the only one that maintains a per-flow state.

In contrast, the other path nodes execute merely the per-flow explicit routing instruction associated with the received packets. The authors (Clarence, et al., 2020) define how to encode an ordered segments list as an ordered list of IPv6 addresses, exploiting new IPv6 routing headers known as Segment Routing Headers (SHRs). In SRv6, the active segment is the packet's destination address, and a pointer indicates the next active segment in SRH. The full list of ordered segments is obtained by visiting all the addresses stored in SRH. The segments list acts as a sequence of SRv6 tunnels that allow mobile, fixed, and enterprise services to be carried together, reducing operators' investment and O&M costs. The overall forwarding process of SRv6 is based on the segments list stored in the SRH, imposed by the source node containing the identifiers of the segments to be visited along the path. The length of the list is indicated in the Segment List counter (SL) field of the SHR. The IPv6 destination address contains the ID of the next node to be visited. Each node crossed decrements the SL counter and copies the next node's ID into the packet's destination address. Note that SRv6 is compliant with the SDN architecture, so we can assume that a centralized controller can compute the requested paths, assign the SIDs, and push the SID list to the ingress node. This way, the ingress node adds an SRH to each IPv6 packet without extra processing. The IETF document RFC 8986 (Clarence F. a., 2021) defines a segment as a function running at a specific location in the network. This modality of SRv6 is known as Network Programming and extends the SRv6 behavior, allowing the combination of packet forwarding operation with packet processing. An ingress node keeps on steering packets through the ordered list of segments. Each of these instructions represents a function to be called at a specific location in the network. A function is locally defined on the node where it is executed and may range from simply moving forward in the segment list to any complex user-defined behavior.

Using the SRv6 protocol, we can also design rerouting procedures needed to manage the change in topology arising in the LEO constellation by the Topology Independent Loop-Free Alternate Fast Reroute (TI-LFA FRR) technology. TI-LFA FRR is an advanced rerouting technology that responds immediately to link or node failures in the SRv6 network. This method is topology-independent, which means it does not rely on the presence of predefined alternative paths and can dynamically calculate backup paths to bypass failed nodes or links. When an event of a link failure happens, TI-LFA FRR automatically activates backup entries and adds new route information to packets by using the SID End.X, ensuring that they can be forwarded along a predefined backup path. This technology can be used in SRv6 BE (Basic Encapsulation) scenarios and SRv6 TE (Traffic Engineering) policies, where SRH information must be added to manage repair lists.

Another technology useful for efficiently managing the topology change is the Seamless Bidirectional Forwarding Detection (SBFD). SBFD allows the integration of rapid and reliable failure detection mechanisms to trigger the rerouting procedure. SBFD is designed to operate efficiently in complex network environments, significantly reducing failure detection time compared to traditional Bidirectional Forwarding Detection (BFD) methods. This is crucial for maintaining the resilience and performance of SRv6 networks, especially in TE (Traffic Engineering) scenarios where path stability is essential. SBFD's ability to detect failures quickly enables faster activation of failover mechanisms such as TI-LFA FRR, ensuring that backup paths are deployed promptly to reduce the probability of traffic disruption drastically. The strength of SBFD relies on using a simplified approach that does not require a complex state machine and better supports large-scale networks. This feature is particularly beneficial in SRv6 networks where numerous SRv6 TE Policy sessions can be configured. SBFD can handle multiple sessions simultaneously with less overhead, improving network scalability and fault management more efficiently.

○ Management and Orchestration in IP-RAN

Management and Orchestration is another tool that, jointly with those of 5G, such as slicing and SDN-based control functions, provide unprecedented flexible tools for network and service operators to optimize the QoS of the services. Machine Learning techniques further boost the orchestration and management tools (Chergui, Ksentini, Blanco, & Verikoukis, 2022), enabling more efficient generation of automatic decision-making policies. This efficiency is conditioned by the capability to monitor network parameters that carry information about the events that trigger management operations.

Monitoring parameters in large and heterogeneous networks such as the 5G-NTN considered here poses significant challenges because monitoring should be scalable, even gathering millions of parameters over variable network conditions, avoiding using excessive resources. For this reason, we propose an architecture for Operation Administration and Management (OAM), such as shown in Figure 3, where we split the monitoring of the parameters for the access, transport, and core domains. The measuring functionalities are performed independently and collected at the OAM, where machine learning algorithms can elaborate on these data to extract policies for single-domain or cross-domain management. Note that the management operations should be performed in real-time for the access and transport domains to provide adaptive policies toward the UE according to the dynamics of the active connections. Instead, management operations for the core should be performed over a long-term period to account for the core's inertia concerning changes occurring in the access and transport domain. For monitoring the core parameters, we refer to the 5G standard defined in (3GPP SA, 2022), where a framework based on the Network Data Analytics Function (NWDAF) is introduced. This framework streamlines the process of consuming metrics and exposes these metrics for further orchestration. Regarding monitoring the access and transport domain based on the NTT-IP-RAN architecture discussed above, we are investigating solutions based on the SRv6 that can guarantee the monitoring of parameters between the UE and the AGG exploiting the In-Band Telemetry (Taheri, A., J., & M. A., 2019) (Clarence, et al., 2018).

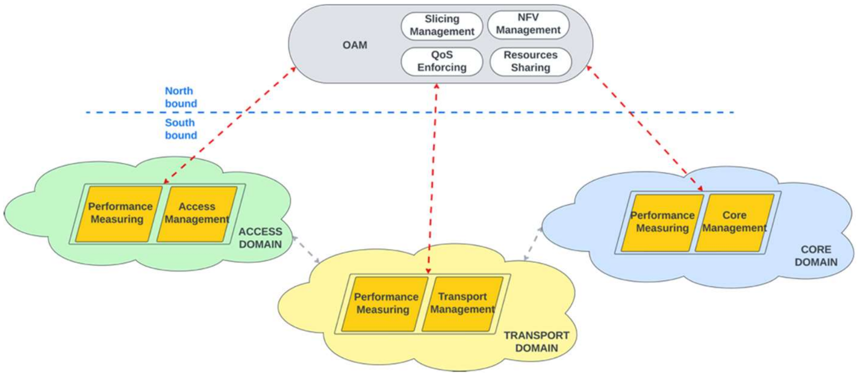
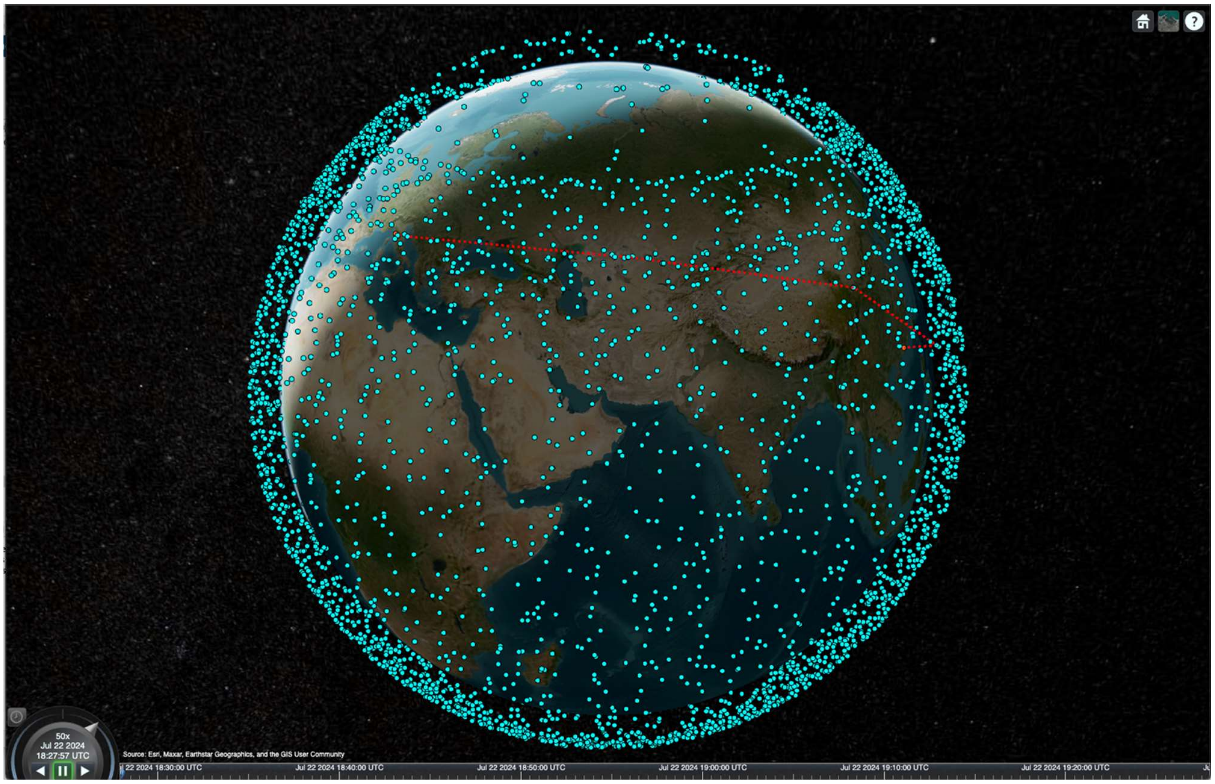


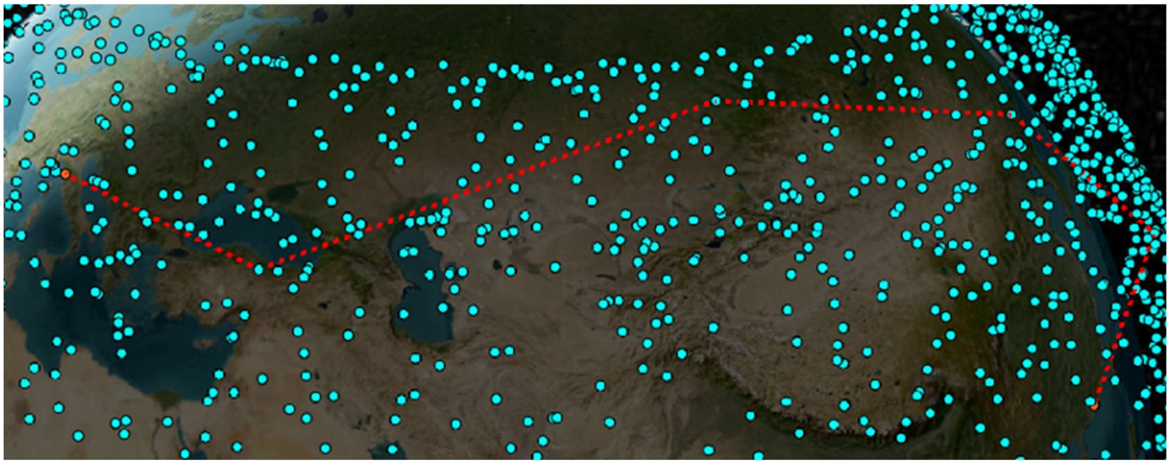
Figure 3 OAM Architecture

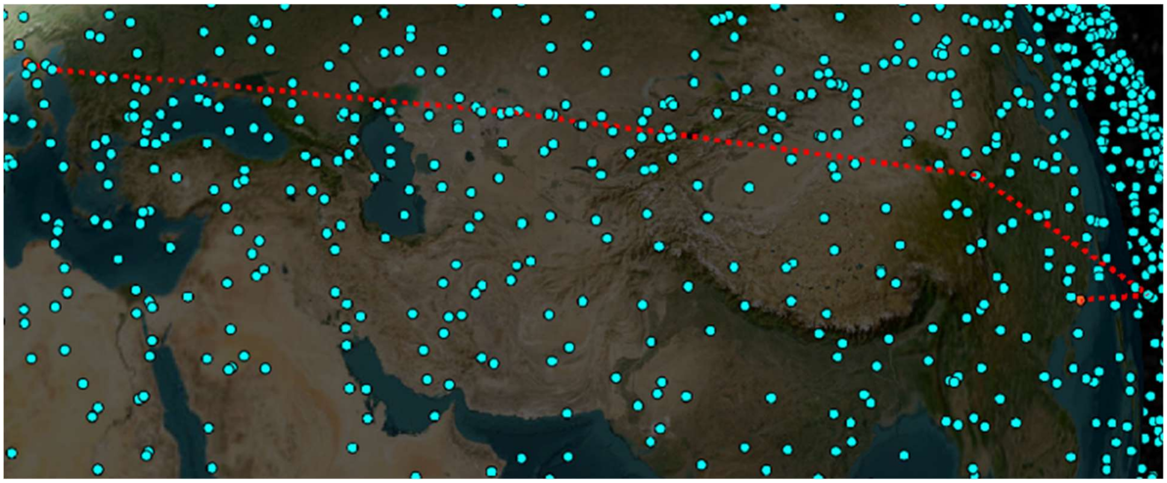
Simulations and Performance Analysis of the Designed Solutions

The NTN network performance analysis is performed using simulation based on MATLAB toolboxes. For the part concerning satellite dynamics, we used the Satellite toolbox and the displacement data of the Starlink constellation satellites. We used the Satellite communication toolbox, the Communication toolbox, and the 5G toolbox for the communication link simulation part.

We developed a simulation setup for the link budget for the link ground station-satellite and satellite-ground station. In this case, we implemented the narrowband Internet-of-Things (NB-IoT) scenario in non-terrestrial networks (NTN) using the parameter sets described in 3GPP TR 36.763. The transmitter repeats the same signal over additional periods to improve the radio coverage in NB-IoT systems. We also developed a simulation setup to analyze the link budget for the intra-satellite link based on optical communication. Optical satellite communication provides the advantage of larger bandwidth, a license-free spectrum, higher data rate, and lower power consumption compared to radio frequency-based satellite communication. Our setup's link budget calculation for uplink and downlink includes the atmospheric effects of absorption and scattering.


Source: Esri, Maxar, Earthstar Geographics, and the GIS User Community





Based on the simulated topology and link budgets in the next tasks, we plan to perform a performance analysis of SRv6-based routing protocols that can be used to handle topology changes in NTN networks efficiently to perform AI-based computational tasks.

In the following, instead, we provide a quantitative performance analysis of the communication and computational load for an edge-based infrastructure to support federated learning (FL) mechanisms for AI-based tasks for the operation of OAM. This analysis in this first step is grounded in a mathematical communication channel model between the access point at the RAN and the mobile users that can access the network using radio technologies such as 5G and WiFi, during the next activities of the PoC we will extend the result to the radio interface of the Non-Terrestrial platforms.

To accomplish the aim stated in this section, we focus on analyzing user mobility near the access point at the RAN and assessing the likelihood that a certain number of users are connected to an access point. We are interested in the performance of the link access because the connection probability impacts the data rate available for information exchange during AI tasks, thereby influencing the time required to complete these tasks. We exploit the findings to model the communication channel between the user and the access point at the RAN (Dhillon & Andrews, 2014) (Sarabjot, Harpreet, & Jeffrey G., 2013) (Jeffrey G., Abhishek K., & Harpreet S., 2016). In our study, we assume that users are mobile within a given geographical area, and that access point placement can be described by two independent Poisson point processes (PPP), with intensity varying based on environmental factors such as topology of the navigation area, wireless technology (5G, WiFi), and user speed. We calculate the load on access points, including the number of connected users and the communication channel's signal-to-noise ratio (SNR), which determine the available data rate. We then explore how this communication channel influences a cooperative learning system in an environment similar to a semi-urban scenario.

We define a two-tier heterogeneous communication model set with parameters similar to a semi-urban environment to evaluate the communication and computational load due to a federation procedure to address an AI-based task. This model features two kinds of RAN access points (AP1 and AP2): macro-cells and micro-cells, with the following scenario parameters: access point placement intensities set at λ1 = 15 and λ2 = 25 per square kilometer. The transmission powers are PAP1 = 60 dBm for macro-cells and PAP2 = 30 dBm for micro-cells, with association biases B1 = 0 dBm and B2 = 5 dBm, respectively. Mobile users are modeled with a density of λ_u = 55 users per square kilometer and P_u = 23 dBm transmission power.

The shadowing effects for each tier follow a log-normal distribution with means $\mu_1 = \mu_2 = 0$ dBm and standard deviations σ_1 = 4 dBm and σ_2 = 8 dBm. The path loss exponent is set at α = 4, with a bandwidth W = 10 MHz. The random channel gains h1 and h2 are Rayleigh distributed with an average power of unity, and the thermal noise spectral density is set at $\sigma_0 = -174$ dBm/Hz. These parameters are derived from ITUT and ETSI documentation studies for semi-urban scenarios characterized by dense radio coverage, significant neighbor interference, and environmental heterogeneity that affects signal shadowing and fading.

For the performance evaluation, we consider the load at the access point in terms of the number of connected users, varying in the interval [3-30] users per access point. This setup aims to reflect the complex dynamics of a crowded area, where the interplay between the dense presence of access points, user mobility, and varying communication loads significantly impacts the efficiency of the communication and computing infrastructure.

Using the simulation parameters introduced above, we calculated the data rates for both uplink and downlink connections across various numbers of users connected to an access point. These data rates and the corresponding load from the AI-based Tasks, as well as the load introduced by the federation procedure, are detailed in Table.1. The load for the AI tasks and the federation procedure, placed at the edge, is expressed in seconds. This way, we have metrics for evaluating the capability to provide a real-time service.

Table.1 Communication Performance

AP Load (n. users)	Data Rate (Mbps)		AI-Task Load (sec.)		FL procedure Load (sec.)	
	UL	DL	UL	DL	UL	DL
10	2.2	9.4	0.0077	5.62 10-4	3.72	0.87
20	1.8	8.5	0.0094	6.21 10-4	4.54	0.96
30	1.6	7.3	0.0106	7.23 10-4	5.11	1.12

To conduct a performance analysis of the computational load introduced by AI-based procedure, we explored integrating the transfer learning approach, such as the distillation, with federated learning. The idea is to investigate the load to transfer knowledge from a large neural network (the teacher) to a smaller, more efficient network (the student). In this teacher-student learning framework, the teacher model's probability distribution serves as a soft target to guide the training of the smaller network. This approach allows for efficient inference learning, particularly in scenarios where computational and communication resources can be constrained or we need to execute a light model maintaining high performance and accuracy, thus making the computational demands of the learning process more manageable. During the execution of the AI-based task, we assume that each user transmits an intermediate layer of the teacher model to the access point on the uplink.

Further, it returns the teacher logit on the downlink for local distillation. The teacher model used in our simulations has an intermediate layer size of 0.017 Mbits and a logit size of 0.0053 Mbits. Since all users share a model with identical parameters, the time required for each user during the execution of the AI procedure solely depends on the available data rate, as shown in the "AI Task Load" column of Table.1. As the number of connected users increases, the Signal-to-Noise and Interference Ratio (SNIR) decreases, which lowers the data rate and, consequently, extends the time needed for information exchange.

After distillation, each user begins the federated learning process by uploading its trained local model to the aggregation server. After the federation process, users download the updated federated model. Both the local and federated models have the same size of 1.022 Mbits. Again, the time required for uploading and downloading these models depends solely on the data rate, as indicated in the "FL Procedure Load" column of Table.1.

To analyze the computational load in terms of introduced delay in seconds, we employ the TensorFlow Profiler, which provides detailed insights into the execution of TensorFlow code. Using this profiler, we can assess the average time required for a single computation step. Specifically, the "Average Step Time" metric measures the time taken to complete a training step on a given device, which includes processing a batch of data and adjusting model parameters accordingly. This metric breakdown allows us to analyze how time is distributed across various operations, such as input processing, forward and backward passes, gradient updates, and other overheads during batch processing. Our analysis assumes users endowed with homogeneous constrained hardware resources, specifically a single-core CPU running at 2.00 GHz with 8 GB of RAM. Table.2 shows the average step time and the total computation time needed to complete a training epoch of the student model on each user. The table compares measurements between a centralized method and the federated one for different numbers of users connected to an access point.

Table.2 Computation time performance

N. usr. for the Fed.	Step Time	Epoch Time
Centralized	0.2676	20.60
Fed. 2 usr	0.2352	9.196
Fed. 4 usr	0.2144	4.288
Fed. 6 usr	0.1773	2.304
Fed. 8 usr	0.1875	1.875
Fed. 15 usr	0.1965	1.179
Fed. 30 usr	0.1313	0.393

The simulation results indicate that the centralized method exhibits the highest average step time. In contrast, the average step time for the FL method shows only minor variation with the number of users. The table also reveals that the epoch computation time decreases as the number of users increases. This reduction is attributed to the distributed nature of FL, which spreads the data processing load across multiple clients, enabling faster computation times for individual users.

Bibliografia

005 ETSI GR IPE. (2022). *TIPv6 Enhanced Innovation (IPE); 5G Transport over IPv6 and SRv6.* Retrieved from https://www.etsi.org/deliver/etsi_gr/IPE/001_099/005/01.01.01_60/gr_ipe005v010101p.pdf

3GPP RAN. (2014). *3GPP, Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall Description; Stage 2; (Ver. 12.0.0).* 3GPP.

3GPP RAN. (2020). *3GPP, Evolved Universal Terrestrial Radio Access (E-UTRA) and NR; Multi-Connectivity; Stage 2; (Ver. 16.0.0).* 3GPP.

3GPP SA. (2022). *Technical Specification 23.288 (Ver.17.4.0) Architecture enhancements for 5G System (5GS) to support network data analytics services.* Retrieved from https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3579

Chergui, H., Ksentini, A., Blanco, L., & Verikoukis, C. (2022). Toward Zero-Touch Management and Orchestration of Massive Deployment of Network Slices in 6G. *IEEE Wireless Communications*, 86-93.

Clarence, F. a. (2021). *ETF RFC 8986: Segment Routing over IPv6 (SRv6) Network Programming.* Retrieved from https://www.rfc-editor.org/info/rfc8986

Clarence, F., Darren, D., Stefano, P., John, L., Satoru, M., & Daniel, V. (2020). *IETF RFC 8754: IPv6 Segment Routing Header (SRH).* Retrieved from https://www.rfc-editor.org/info/rfc8754

Clarence, F., Stefano, P., Les, G., Bruno, D., Stephane, L., & Rob, S. (2018). *IETF RFC 8402: Segment Routing Architecture.* Retrieved from https://www.rfc-editor.org/rfc/pdfrfc/rfc8402

Dhillon, H. S., & Andrews, J. G. (2014). Downlink Rate Distribution in Heterogeneous Cellular Networks under Generalized Cell Selection. *IEEE Wireless Communications , Volume: 3 Issue: 1*, 42-45.

Jeffrey G., A., Abhishek K., G., & Harpreet S., D. (2016). A Primer on Cellular Network Analysis Using Stochastic Geometry. *arXiv:1604.03183* , 1-46.

Sarabjot, S., Harpreet, S. D., & Jeffrey G., A. (2013). Offloading in heterogeneous networks: Modeling, analysis, and design insights. *IEEE Transactions on Wireless, Volume: 12 Issue: 5*, 2484-2497.

Taheri, D. B., A., K., J., V., & M. A., K. (2019). IntOpt: In-band Network Telemetry optimization for NFV service chain monitoring. *IEEE Int. Conf. on Communications (ICC).*