Pervasive Monitoring and Distributed Intelligence for 6G Near Real-Time Operation

L. Velasco*, M. Ruiz and P. González
Universitat Politècnica de Catalunya
Barcelona, Spain. *luis.velasco@upc.edu

F. Paolucci
CNIT
Pisa, Italy

A. Sgambelluri and L. Valcarenghi
Scuola Superiore Sant'Anna
Pisa, Italy

C. Papagianni
University of Amsterdam
The Netherlands

***Abstract*—In-band network telemetry (INT) can provide additional insight of the network status, which can be used for distributed control loop automation. However, INT yields substantial transmission overhead that negatively affects scalability and network performance. To achieve scalable near real-time operation, we propose a solution for the next generation of mobile networks integrating: *i*) pervasive monitoring based on INT; *ii*) intelligent data aggregation at both data and control plane to deal with scalability issues; and *iii*) intelligent distributed control for near real-time service operation. A use case is presented to showcase how the above components work together to operate 6G services across different network segments, including radio access, packet transport and datacenters.**

Keywords—Pervasive Monitoring; Near real-time 6G operation.

I. INTRODUCTION

Near real-time control and operation is necessary in order to meet the stringent performance requirements of 6G services. In fact, such operation should not be based on reactive approaches, but on predictive ones able to anticipate specific events and perform proactive network adaptation. Examples include, e.g., handovers leading to additional delays and overloaded edge nodes leading to poor performance, as well as dynamic steering and priority, virtual network functions (VNF) (de)activation / replication, or slice reconfiguration. To achieve near real-time operation and control, intelligent decision making should be moved as close as possible to the data plane resources, and be complemented with pervasive monitoring to anticipate signal degradation, queue congestions, etc. In this paper, we propose a solution that includes end-to-end (e2e) in-band network telemetry (INT) from the user equipment and across the various network segments, including radio access (RAN), transport packet, and datacenter networks.

II. PROPOSED SOLUTIONS

We propose a combination of fundamental pillars to meet the requirements revealed in the previous section, namely, pervasive monitoring, intelligent telemetry data aggregation, and intelligent distributed operation.

A. Pervasive Monitoring

Near real-time operation requires a proactive telemetry system able to follow the various performance indicators at increasing rates. In traditional solutions based on centralized data lakes, cloud-hosted big data analytics often do not provide feedback to orchestrators and/or controllers with enough reaction time. Network telemetry exploiting INT and postcard technologies enables accurate monitoring, relying on distributed and federated agents. This requires novel entities, collecting measurements from heterogeneous sources, i.e., the RAN, programmable devices, the cloud, and application entities.

INT solutions provide per-packet telemetry by embedding network state in each packet (or cloned copies) and conveying the INT/postcard report. However, the massive collection and processing of telemetry data hinders scalability. For this very reason, *two-stage* P4 telemetry collectors were proposed in [1] in charge of processing and aggregating postcard telemetry reports at wire speed. However, such solution needs to be extended so that collected measurements can be processed and consumed locally by node agents, enabling high network awareness at the considered segment.

B. Intelligent telemetry data aggregation

Encoding per-hop information on a per-packet base causes packets' headers to grow linearly with every hop. This, not only wastes bandwidth, but can lead to packet fragmentation if the maximum transmission unit (MTU) is exceeded. To alleviate such overhead, the authors in [2] spread telemetry data across multiple packets within the same flow (per flow aggregation of telemetry data). As an example, a path trace can be potentially retrieved by inserting a single hop ID within each packet. The path is composed from the individual hops that have been stored in the flow packets. The reconstruction of the path takes place at a *telemetry processor*. Although this solution requires some network state to be stored in the network nodes, it allows collecting telemetry data avoiding scalability issues at the collectors, without degrading network performance.

Telemetry measurements can be additionally aggregated by telemetry processors targeting dimensionality reduction, as well as reducing data rate at the control plane [4]. Examples include compression of individual measurements and time series aggregation. In both cases, techniques based on statistics, machine learning, e.g., autoencoders (AE), and data stream mining, can be used. E.g., results in [4] show 625:1 compression ratio using AEs. In addition, intelligence entails, among others: *i*) adaptation, i.e., dynamically deciding when and how data aggregation needs to be done; *ii*) consolidating/correlating heterogeneous measurements; and *iii*) adding value to measurements, e.g., AEs can be used in both forward and backward directions to quantify relevance metrics at the latent feature space and input [5].

C. Intelligent distributed operation

Distributed decision making has been proposed for network and service operation, not only to relieve the SDN controller from fine grain tasks and increase scalability, but also as an enabler for near real-time control [3]. In such approach, agent nodes are augmented with intelligent algorithms, e.g., based on reinforcement learning that make autonomous decisions as a function of the observed conditions, collected through telemetry.

The research leading to these results has received funding from the European Commission thought the HORIZON SNS JU DESIRE6G (G.A. 101096466) and the MICINN IBON (PID2020-114135RB-I00) projects, and from ICREA.

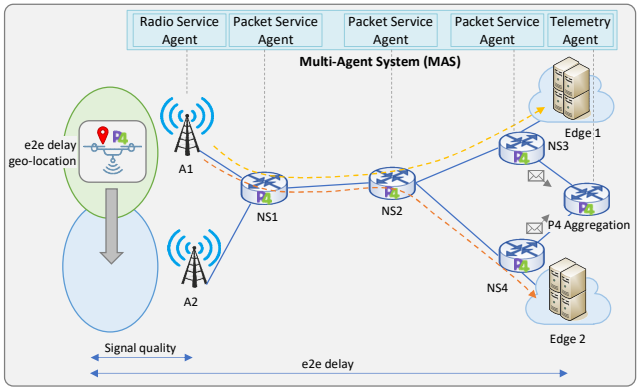

Fig. 1. Illustrative scenario supporting an e2e 6G service.

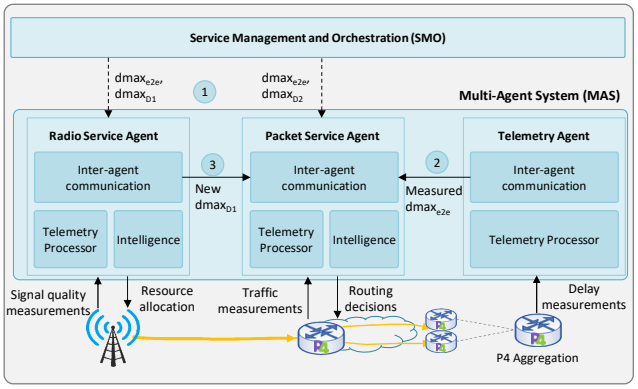

Fig. 2. Pervasive telemetry and distributed intelligence solution.

In addition, communication among agents with similar or different capabilities allows creating distributed control systems that can cooperate to achieve some common objective (named multi-agent systems -MAS), like ensuring e2e delay for services under changing conditions.

III. Illustrative Use Case

Let us present a use case that combines the solutions described in the previous section for the near real-time control of e2e 6G services. Fig. 1 presents the scenario, where a service provides connectivity between an unmanned aerial vehicle (UAV) and VNF providing computing and storage resources for real-time augmented/virtual reality (AR/VR) high quality video reconstruction. Strict e2e and segment latency and jitter requirements imposed by the application, require continuous monitoring of a large number of stream flows, to dynamically make routing and edge computing resource selections that satisfy the committed performance.

For that very reason, INT is used to extend packet headers with metadata added by the UAV and the intermediate network nodes: *i*) the UAV adds geo-location coordinates; *ii*) the RAN aggregates signal quality indicators, e.g., the wireless received signal strength indicator (RSSI) for WiFi or the channel quality indicator (CQI) for 4G/5G, and the latency experienced in the access segment; and *iii*) intra/inter-switch latencies and jitter are collected from the packet nodes. To cope with telemetry scalability issues, two-stage P4 collectors aggregate INT reports in different ways, employing both standard report aggregation and report correlation. The former option aggregates relevant metadata in a single report packet per configurable time window or number of packets, while the latter provides metadata statistics related to different flows, switches and e2e intents (e.g., maximum/average latency values).

Depending on the INT application mode, telemetry reports are either directly exported by each INT node, from their data plane to the collector switch (P4 aggregation in Fig. 1), or they are embedded into the packets along the data path. For the latter, we employ deterministic and probabilistic per-flow aggregation techniques to reduce transmission overhead thus, spreading the telemetry values across multiple packets of a flow. Finally, the collector switch strips the aggregated metadata and selectively sends them to the telemetry processor.

The collected data are consumed at the intelligent control plane, where a MAS controls the service near real-time. The MAS consists of a number of heterogeneous service agents running as part of node agents that communicate among them. Service agents might include (Fig. 2): *i*) a telemetry processor to collect from the local node and process telemetry data, including intelligent data aggregation; *ii*) an inter-agent communication module, which is used for telemetry data distribution and state and model sharing; and *iii*) technology-specific (i.e., RAN, packet, etc.) intelligence for autonomous decision making based on local and remote observations. The service management and orchestration (SMO) system provides guidelines to the MAS, while giving freedom to the MAS for operating on the resources allocated to the service.

Fig. 2 illustrates this operation; the agents receive from the SMO the resources that can be used for the connectivity service and the max e2e delay ($dmax_{e2e}$) that needs to be ensured. Based on the required performance, let us assume that each segment is assigned with an initial budget delay ($dmax_{Di}$) (labeled 1 in Fig. 2). Once in operation, e2e delay and other performance indicators are measured and results distributed among the agents in the MAS (2). For instance, in the case of the radio segment, measurements and forecasted metrics are used to enforce the resource block group adaptation with margin in time, reducing the service SLA violations. However, imagine that at some point in time, the radio segment cannot provide the committed delay in its domain. In this case, the RAN agent announces the new delay budget for its segment to the other service agents (3), so packet agents can make decisions, e.g., changing routing, to ensure the new budget delay in their domains.

IV. Conclusions

A solution enabling near real-time service operation has been proposed. INT collects heterogeneous measurements from UAVs, RAN and packet networks. Aggregation at data and control plane is used to deal with scalability issues. Service agents consume telemetry data to ensure e2e delay and cooperate among them in the case that the committed performance is not met in one or more network segments.

References

[1] F. Alhamed *et al.*, "P4 Postcard Telemetry Collector in Packet-Optical Networks," in proc. ONDM, 2022.

[2] K. Papadopoulos et al., "PFA-INT: Lightweight In-Band Network Telemetry with Per-Flow Aggregation," in proc. IEEE NFV-SDN, 2021.

[3] S. Barzegar *et al.*, "Distributed and Autonomous Flow Routing Based on Deep Reinforcement Learning," in proc. PSC, 2022.

[4] L. Velasco *et al.*, "Is intelligence the answer to deal with the 5 V's of telemetry data?," in proc. OFC, 2023.

[5] M. Ruiz *et al.*, "Deep Learning -based Real-Time Analysis of Lightpath Optical Constellations [Invited]," IEEE JOCN, vol. 14, 2022.