Latency-aware network architectures for 5G backhaul and fronthaul

D. Larrabeiti[1], G. Otero[1], J. P. Fernández-Palacios[2], L. M. Contreras[2], J. A. Hernández[1]

[1] *Univ. Carlos III of Madrid, Spain*
[2] *Telefónica Global CTO, Madrid, Spain*
dlarra@it.uc3m.es

Abstract: 5G poses important challenges regarding latency management, specially in fronthaul and backhaul traffic transport. Operators are combining standards in search of a unified architecture that features virtualization, programmability and performance control.
© 2022 The Author(s)

1. Introduction

Network function virtualization (NFV) provides flexibility, increased utilization of resources, and reliability via the service mobility features. As long as the target SLA is preserved, services like edge caching or firewall functionalities can be dynamically moved around the network [1, 2] when necessary. This is possible with the appropriate slice-capable technology like the one developed in EU H2020 PASSION project [3]. The most complex service to be virtualized in the next years is the 5G radio processing in Cloud-RAN scenarios. According to 3GPP TR38.913, 5G New Radio should support latency values below 0.5 ms UL/DL for Ultra-Reliable Low Latency Communications (URLLC). This requirement enforces a tight latency budget in the network for both fronthaul and backhaul traffic.

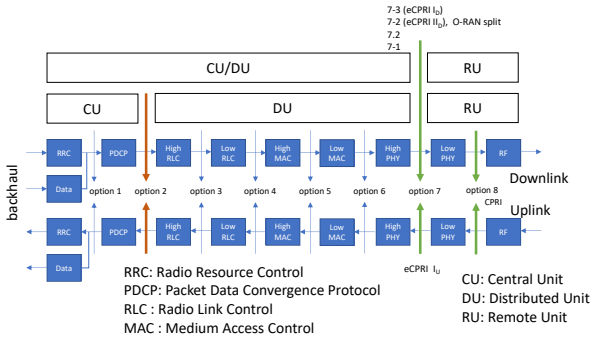

Fig. 1. Split options for the 5G New Radio signal processing tasks (3GPP).

The specific bandwidth, latency budget, and synchronization requirements in the RAN (Radio Area Network) heavily depend on the chosen functional split. Two main reference radio processing stacks with similar functional splits exist: one from 3GPP TR38.801 and another from eCPRI. eCPRI takes the 3GPP layers as reference and defines a set of intra-PHY splits, two for the downlink and one for the uplink (I_D, II_D, and I_U). These generate High Priority Fronthaul (HPF) traffic and have very strict latency requirements. In outline, intra-PHY split, has 100 μs budget according to eCPRI and IEEE 802.1CM, and 250 μs if 3GPP TR 38.801 transport requirements estimation are considered. On the other hand, midhaul (intermediate radio splits above MAC layer) and backhaul traffic (user data) have more relaxed latency constraints. This enables to push the functionality of the central processing units far from the remote radio units, increasing the sharing ratio of cloud resources. Table 1 shows estimations for the rate requirements of different 5G New Radio configurations in intra-PHY split I_U (only data plane is included). Note how the effect of the number of antenna elements require, in the worst case, Tb/s transmission rates for massive MIMO systems.

2. Dealing with fronthaul and backhaul traffic

Only dedicated fiber seems to be able to cope with the transmission requirements of HPF. However, the effective rate is proportional to the cell's load in intra-PHY split I_U. This means that there is a chance to save a lot of resources via packet multiplexing, as most cells may be idle or with low load. The first steps of fronthaul data into the network are usually via Ethernet aggregator configurations taken from IEEE802.1CM. This enables to narrow

Table 1. Requirements for OFDM transport in 5G new radio split I_U at 100% load per RF channel

Channel BW (MHz)	Δf	# Subcarriers	T_s (μs)	# Antennas	Generated Data Rate (Gbps)	Burst Size (KB)
50	15	3167	66.7	2	2.85	23.2
100	60	1584	16.7	32	91.24	185.6
200	120	1584	8.3	32	182.48	185.6
200	120	1584	8.3	256	1459.81	1485.0
400	120	3167	8.3	32	364.84	371.1
400	120	3167	8.3	256	2918.71	2969.1

down the maximum or a small percentile of the packet latency with queueing theory as in [4]. Also, given the fact that HPF can be configured with maximum priority, tighter bounds can be achieved in situations where we need to gain extra physical distance [5]. However, as LR Ethernet interface rates grow toward 400 Gb/s (IEEE 802.3CN) the effect of queuing delay becomes negligible and the use of the standard's bound is sufficient.

If the access link is based on a shared PON like GPON (only feasible for the lowest fronthaul options), the ranging mechanism of GPON's PLOAM provides information about propagation plus equalization delay, and the uplink queuing delay can be made deterministic by using GPON's circuit emulation services. However, specific engineering needs to be made for HPF, as a 125 μs TDM period makes it necessary to synchronize OFDM symbol generation and the timeslots where the whole OFDM burst needs to travel. In this sense, WDM-PON seems a better alternative. Midhaul and backhaul traffic are more likely to take a few IP hops, whose queuing delay can be estimated [5] if the same priority queuing scheme is applied. Alternatively, delay-preserving packet scheduling disciplines like Weighted Fair Queuing may be configured for aggregates of HPF traffic. With the aim of hardware reuse and cost efficiency, future networks will probably merge fronthaul and backhaul traffic. The convergence of legacy and future radio access technologies (generating backhaul and fronthaul data) in a shared network is a research topic of recent interest [6]. Studies about the network requirements and orchestration needs of such networks are paramount to support future services.

3. Delay awareness through OAM protocols

HPF shows the importance of latency in next generation multilayer networks featuring network slicing. In Figure 2, the transport data plane has two layers: a packet-switched layer (IP, carrier Ethernet, MPLS, MPLS-TP, PON) and a circuit-switched optical layer (TDM/WDM). A microwave network segment may be present.

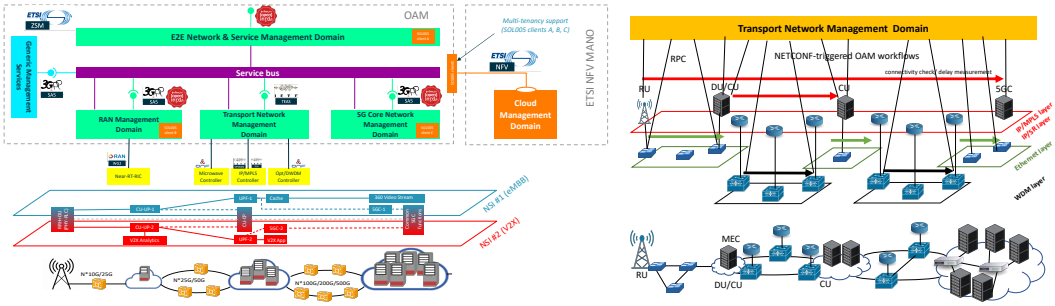

Fig. 2. End-to-end network slicing and layer-independent OAM management

At the circuit-switched optical layer, OTN ITU-T Rec. G.709/Y.1331 (06/20) has embedded support to round-trip time computation between Path Connection Monitoring End Points (and also the 6 available Tandem Connection), by means of Delay Measurement bits in the ODU overhead, with a resolution of the duration of two OTU frames. Upcoming S-BVT technologies aim at all-optical dynamic exploitation of WDM [3,7] in the MAN, bringing beyond-100G services in a cost-effective way, since most of the effects of packet aggregation/distribution latency can be prevented. Until recently, the lack of OAM interfaces at the packet switched layer for IP and security matters have impelled operators to develop ad-hoc connectivity/latency measurement tools and SNMP per-se does not feature in-band OAM flows. This has made them use the OAM utilities provided by MPLS/MPLS-TP and Ethernet. However, the endorsement by operators and manufacturers of open multilayer OAM solutions is changing this approach [8]. IETF LIME WG (Layer Independent OAM Management in the Multi-Layer Environment), ended in 2016 and produced three RFCs, currently a Proposed Standard: two YANG data models for OAM protocols connection oriented and connectionless (RFC8531, RFC8532), and a retrieval method YANG

Disclaimer: Preliminary paper, subject to publisher revision

data model for connectionless OAM (RFC8533). RFC8533 provides technology-independent RPC operations for OAM protocols that use connectionless communication, extensible with technology-specific details. In essence, the RPC model enables issuing commands to a Network Configuration Protocol (NETCONF) server to run secure OAM commands. The "connectionless-oam-methods" of module RFC8533 defines RPC 'continuity-check' ,equivalent to IP ping (RFC792, RFC4443) and MPLS LSP ping (RFC8029), and 'path-discovery' (equivalent to IP traceroute) operations. A parallel evolution supports Frame Delay and Frame Delay Variation in ITU-T rec. Y.1731/IEEE 802.1ag (as IEEE 802.1Q amendment) a performance monitoring functions. ITU-T Y.1710/Y.1711 and ITU-T G.8113 for MPLS and MPLS-TP , respectively. In 2017, IEEE, MEF, and ITU-T SG15 started a liaison toward a IEEE 802.1 CFM YANG DM, whose result is IEEE 802.1Qcp-2018, available at GitHub. Connectivity and latency supervision tools at different layers are becoming available to control entities through NETCONF or RESTCONF, as device vendors keep opening and standardizing their interfaces.

4. Network intelligence and delay

The goal for telecom operators is to provision packet-switched primary and backup paths with a maximum target latency by issuing high level connectivity requests from an orchestrator, invoking the slice managers in the RAN, in the Transport, and the 5G Core networks (Fig. 2). The service will be typically defined as a chain of VNFs, making use of MEC (Multi-Access Edge Computing) capabilities. As noted above, different network entities can supply latency information and have mechanisms to enforce some sort of latency budget. Furthermore, the network should automatically react to changes detected in the underlying infrastructure to make sure that the end-to-end network slicing SLA is preserved. The implicit lack of predictability of backup routes latencies after network failures needs special attention, especially in the packet switching layer. Network managers should leverage the intelligence of SDN and perform exhaustive *what-if* bandwidth and latency analysis on their databases in order to program SLA-preserving policies.

5. Conclusions

Delay and bandwidth management is especially relevant in 5G. Even though it is one of the most challenging latency-constrained use case, C-RAN can be considered yet another case of network slicing in the more general 5G end-to-end slicing framework providing a general solution to industry verticals and multi-operator network sharing. The use of SDN and the current trend for open multilayer OAM interfaces may solve integrated end-to-end and per-segment performance monitoring in the short term. At provisioning time, the underlying technologies should be capable of enforcing a maximum delay and perform delay-aware actions commanded via YANG/NETCONF standard APIs. Finally, orchestrators must be able to predict the delay of primary and backup alternatives before failures.

Acknowledgements

Work funded by the EU H2020 projects PASSION (780326), Int5Gent (957403) and ES ACHILLES (AEI/10.13039/501100011033) projects.

References

1. G. Otero, D. Larrabeiti, J. A. Hernández, P. Reviriego, J. P. Fernández-Palacios, V. López, M. S. Moreolo, J. M. Fabrega, L. Nadal, and R. Martinez, "Optical Interconnection of CDN Caches with Tb/s Sliceable Bandwidth-Variable Transceivers featuring Dynamic Restoration," European Conference on Netw. and Comm. (EuCNC), 2019.
2. G. Otero, J. P. Fernández-Palacios, M. S. Moreolo, D. Larrabeiti, J. A. Hernández, J. M. Fabrega, and R. Martínez, "Scaling Edge Computing through S-BVT and Pb/s Switching Devices in Large Dense Urban Metro Networks [Invited]," 2020 22nd International Conference on Transparent Optical Networks (ICTON), Italy, 2020.
3. M. Svaluto Moreolo, J. M. Fabrega, L. Nadal, R. Martínez, R. Casellas, J. Vílchez, R. Muñoz, R. Vilalta, A. Gatto, P. Parolari, P. Boffi, C. Neumeyr, D. Larrabeiti, G. Otero, and J. P. Fernández-Palacios, "Programmable VCSEL-based photonic system architecture for future agile Tb/s metro networks," J. Opt. Commun. Netw. 13, 2021.
4. G. O. Pérez, J. A. Hernández and D. Larrabeiti, "Fronthaul network modeling and dimensioning meeting ultra-low latency requirements for 5G," in J. Opt. Commun. Netw., 2018.
5. G. Otero Pérez, D. Larrabeiti López and J. A. Hernández, "5G New Radio Fronthaul Network Design for eCPRI-IEEE 802.1CM and Extreme Latency Percentiles," in IEEE Access, 2019.
6. G. Pérez, A. Ebrahimzadeh, M. Maier, J. A. Hernández, D. L. López, and M. F. Veiga, "Decentralized Coordination of Converged Tactile Internet and MEC Services in H-CRAN Fiber Wireless Networks," in J. Lightwave Technol., 2020.
7. D. Larrabeiti, G. Otero, J. A. Hernández, P. Reviriego, J. Femández-Palacios, V. López, M. Svaluto Moreolo, R. Martínez, Josep M. Fabrega, "Tradeoffs in optical packet and circuit transport of fronthaul traffic: the time for SBVT?," 2020 International Conference on Optical Network Design and Modeling (ONDM), 2020.
8. Telecom Infra Project (https://telecominfraproject.com/)

Disclaimer: Preliminary paper, subject to publisher revision