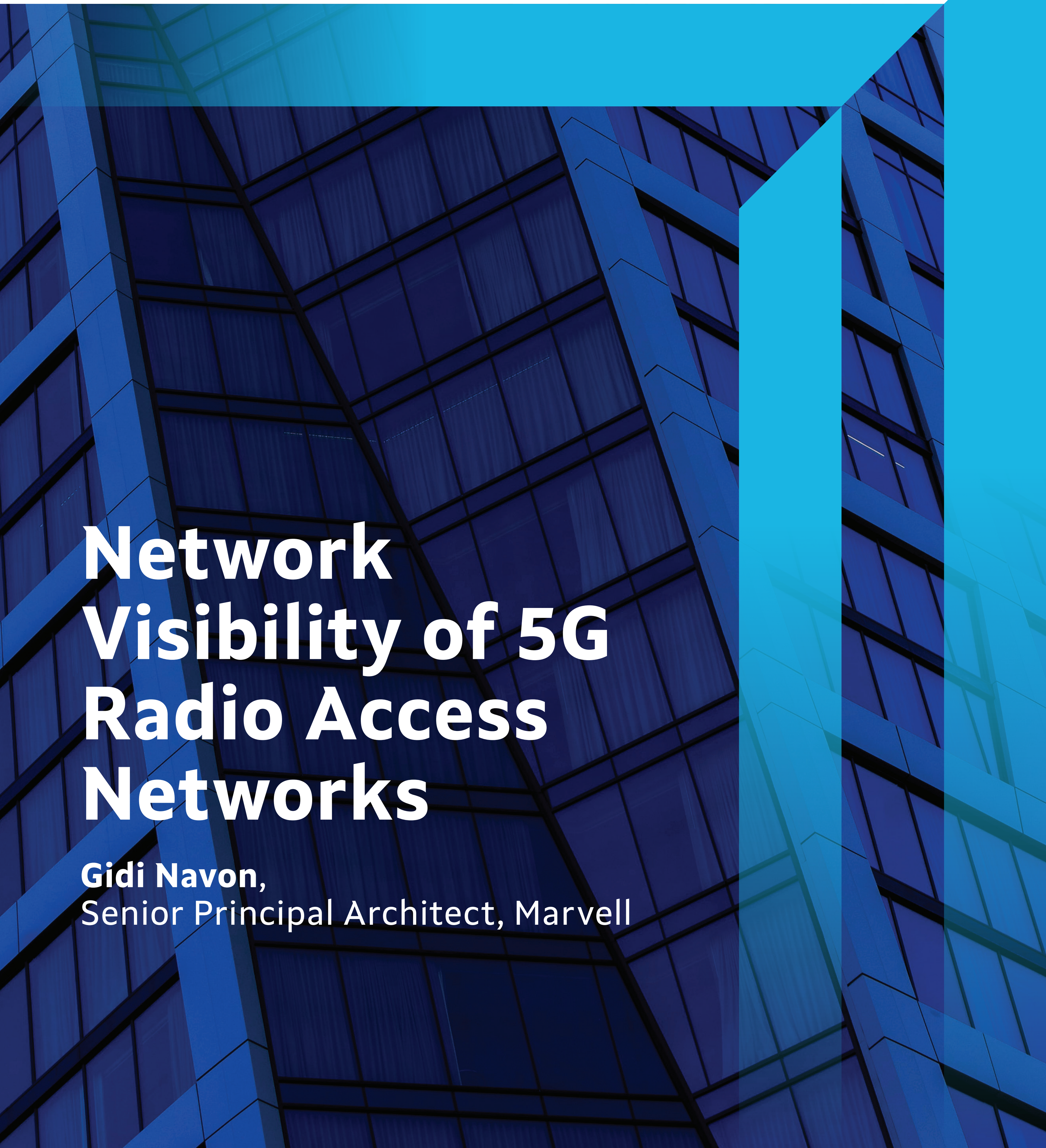

MARVELL

Network Visibility of 5G Radio Access Networks

Gidi Navon,
Senior Principal Architect, Marvell

The Radio Access Network (RAN) is dramatically changing with the introduction of 5G networks and this, in turn, is driving home the importance of network visibility. Visibility tools are essential for mobile network operators to guarantee the smooth operation of the network and for providing mission-critical applications to their customers. In this whitepaper, we will demonstrate how Marvell's Prestera® switches equipped with TrackIQ visibility tools are evolving to address the unique needs of such networks.

The Changing RAN

The RAN is the portion of a mobile system that spans from the cell tower to the mobile core network. Until recently, it was built from vendor-developed interfaces like CPRI (Common Public Radio Interface) and typically delivered as an end-to-end system by one RAN vendor in each contiguous geographic area. Lately, with the introduction of 5G services, the RAN is undergoing several changes as shown in Figure 1 below:

- In centralized RAN architectures (known as C-RAN), baseband processing resources, called distributed units (DUs), are pooled in a macro base station hub or far edge data center and dynamically shared with remote radio units (RUs).
- The segment between the RU and the DU, called fronthaul, has transformed from using an interface called CPRI to enhanced CPRI (eCPRI), which runs over standard Ethernet packets and sometimes over UDP/IP packets.
- The move to packet-based fronthaul segments created in some deployment scenarios a switched network that aggregated many RUs into just a few DUs, reducing the number of fiber links and the number of interfaces on the DU.
- In the context of Cloudification and Virtualization of the RAN, known as Cloud-RAN and vRAN, 5G software building blocks are provided over virtual machines (VMs) or containers, and on top of commodity servers equipped with O-RAN compliance accelerators or on cloud native solutions.
- Disaggregation between software and hardware, in which full solutions previously provided by a single 5G vendor (Ericsson, Huawei, Nokia, ZTE and Samsung) can now be provided by multiple players each specializing on one aspect of the solution.

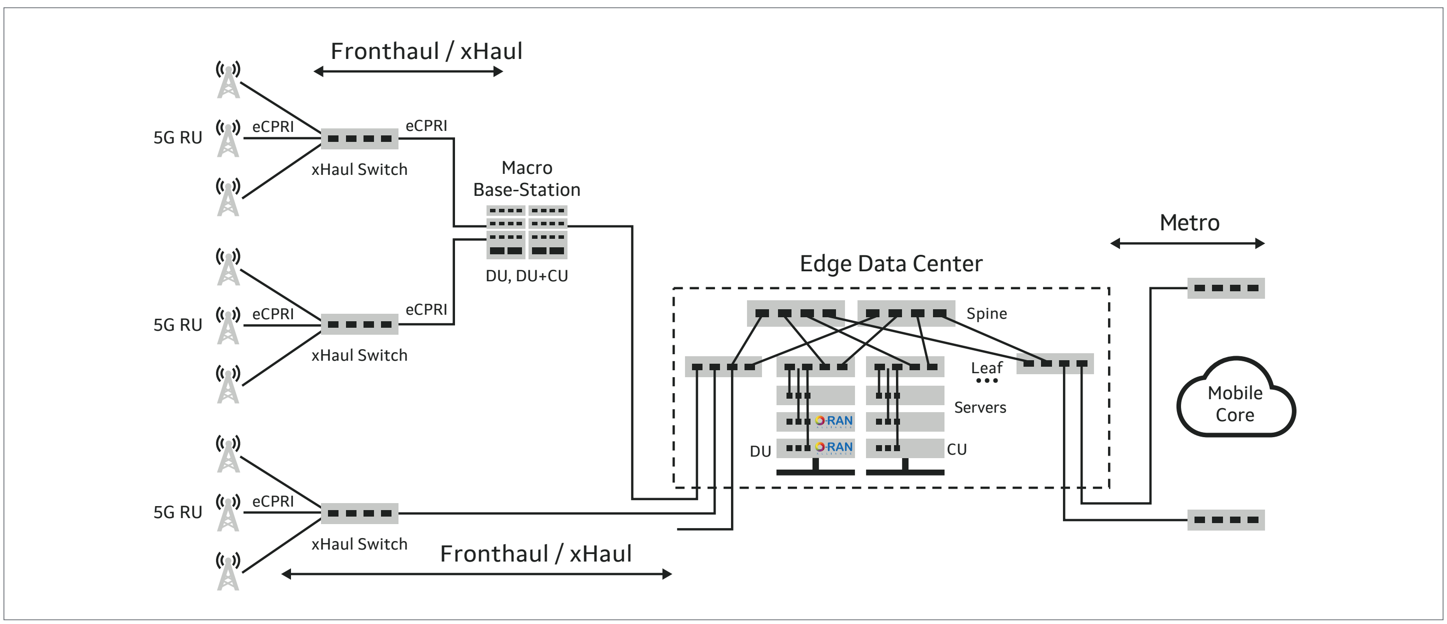

Figure 1: The Changing RAN

The above trends have a direct impact on the management and visibility tools used by MSOs to run an efficient and reliable network and monetize their investment by providing new advanced services, such as ultra-reliable low latency communications (URLLC).

Monitoring Latency and Jitter

A key task that visibility tools are expected to monitor in the RAN is latency. There are several latency requirements for each segment of the RAN.

The typical one-way frame latency budget for fronthaul is around 75usec, out of which more than half can be taken by propagation delay over fiber. The max latency per node thus needs to be kept very low, and this can only be guaranteed if queuing delay is at a minimum, with only a few packets at max waiting for transmission.

It is not enough to periodically check the latency of this segment; it is crucial to continuously check the latency of all packets, and check for any gradual or sudden increase in latency.

On the other extreme, it makes no sense to overwhelm a collector with the latency of each and every packet, but instead provide intelligent statistics on the latency distribution on the minimum, average and peak latency (providing an indication on the jitter added by each node). It is also important to provide statistics on trend analysis to anticipate problems before they arise, as well as on anomaly alerts that may affect the performance of the mobile network.

End-to-end latency is not enough. To identify the factors contributing to the latency, operators need to track the per hop latency, and also the path packets traversed in the network, making sure routing and load balancing decisions are optimal. Tracking latency on a per-flow basis usually provides the best granularity operators can relate to. A specific flow carries one type of application, for example an eCPRI flow, that has a known latency target. Each flow traverses the network via one route, thus after identifying and verifying this path, what is left is monitoring the latency on each hop of this path.

Marvell's per-flow visibility and telemetry tools are best suited to provide the above for the Radio Access Network. The tools utilize new technologies called In-band Telemetry (INT) or In-situ OAM (IOAM) shown in Figure 2. With these technologies, information about the path the packet traversed, and the time it spent in each node, is added to the user's packets as additional metadata, until it reaches its final destination or until it reaches the last hop of its journey:

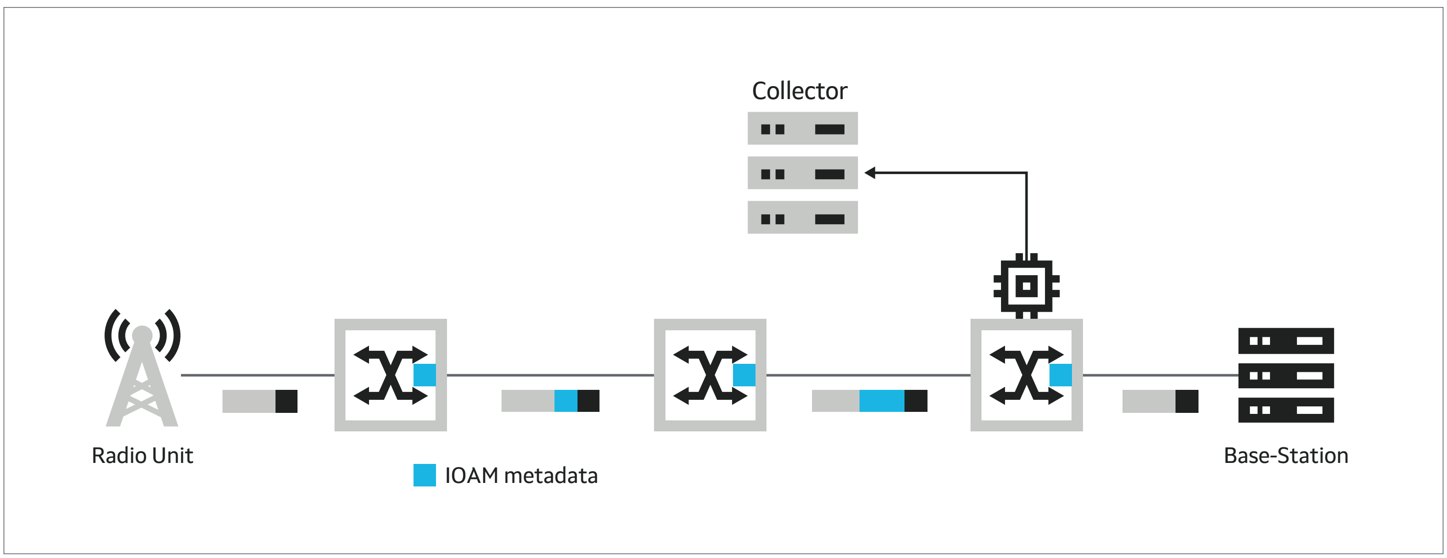

Figure 2: In-band Telemetry

The collector receives digested, per-flow information, and is able to present the nodes in the network the flow has traversed, with additional information on the minimum, average and peak latency at each of these nodes. (See Figure 3 below.) There are several methods to provide the per-flow information to the collector, including IP Flow Information Export (IPFIX) that has the ability to include this additional metadata.

This information is gathered from the last node in the path, without the need to collect information from multiple nodes, and without the need to attempt to correlate the information into a coherent view. Precision Time Protocol (PTP) implemented in the network can also provide an accurate wall-clock timestamp in the IOAM or INT packets, and provide valuable information on the propagation delay in the network.

A video demonstration of the flow-tracking capabilities based on IOAM can also be viewed here.

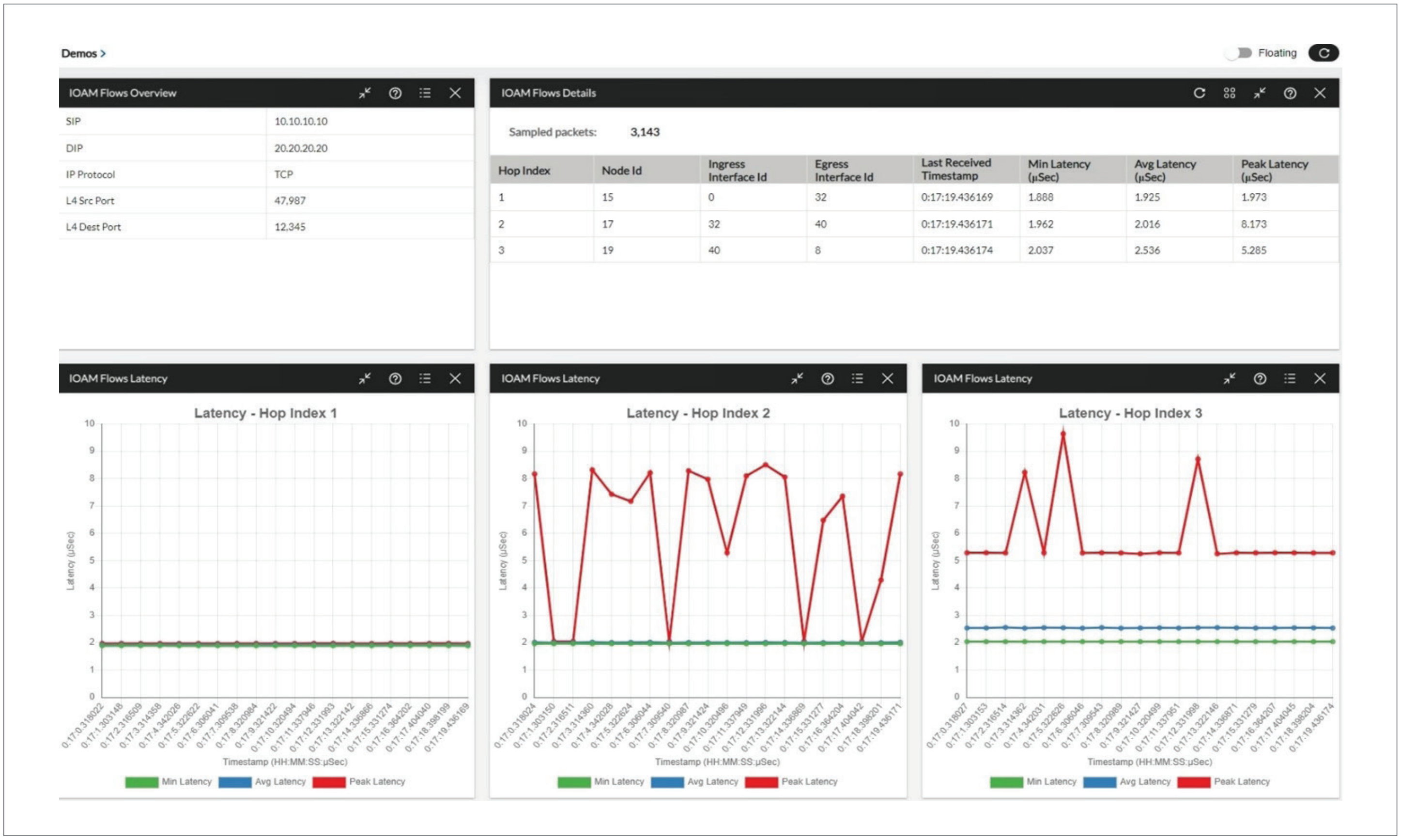

Figure 3: Flow-Tracking based on IOAM

Some mobile network operators prefer a passive monitoring technique, which continuously tracks the latency of all queues, without adding metadata to user packets or adding additional OAM packets to the network. Monitoring the latency of queues instead of the latency of all flows provides an aggregated view on the differentiated services provided by the operator and dramatically reduces the amount of monitored entities. This is critical for a nationwide mobile network with thousands of network elements.

Marvell's queue latency monitoring approach further reduces the amount of data that reaches the centralized controller by continuously analyzing the data and placing the latency information into latency bins. Figure 4 below shows a histogram of the latency on a specific port & queue. The Y-axis is the percentage of time (inside a specific second) the queue latency was in each of the latency bins shown on the X-axis. This is a live graph, in which each color represents a specific second during the day. In this example, the last six seconds are shown.

Let's examine the 'red' second (the one on the left of each latency bin): It shows that for 12% of this second, the queue latency was between 0-10usec, 15% of this second latency was between 10 and 20usec, and so on. Such data can be sent to an analytic server for anomaly detection, trend analysis, catastrophic event detection and more.

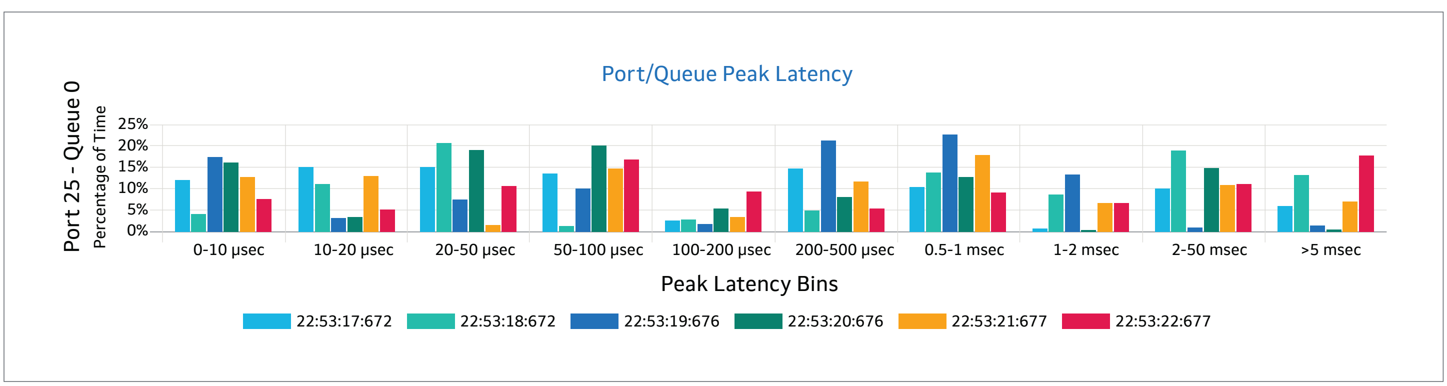

Figure 4: Peak Latency Monitoring shown in Histogram

Resource Monitoring

Monitoring latency is a critical way to identify problems in the network that result in latency increase. However, if measured latency is high, it is already too late, as the radio networks have already started to degrade. The fronthaul network, in particular, is sensitive to even a small increase in latency. Therefore, mobile operators need to ensure the fronthaul segment is below the point of congestion thus achieving extremely low latencies.

Visibility tools for Radio Access Networks need to measure the utilization of ports, making sure links never get congested. More precisely, they need to make sure the rate of the high priority queues carrying the latency sensitive traffic (such as eCPRI user plane data) is well below the allocated resources for such a traffic class.

A common mistake is measuring rates on long intervals. Imagine a traffic scenario over a 100GbE link, as shown in Figure 5, with quiet intervals and busy intervals. Checking the rate over long intervals of seconds will only reveal the average port utilization of 25%, giving the false impression that the network has high margins, without noticing the peak rate. The peak rate, which is close to 100%, can easily lead to egress queue congestion, resulting in buffer buildup and higher latencies.

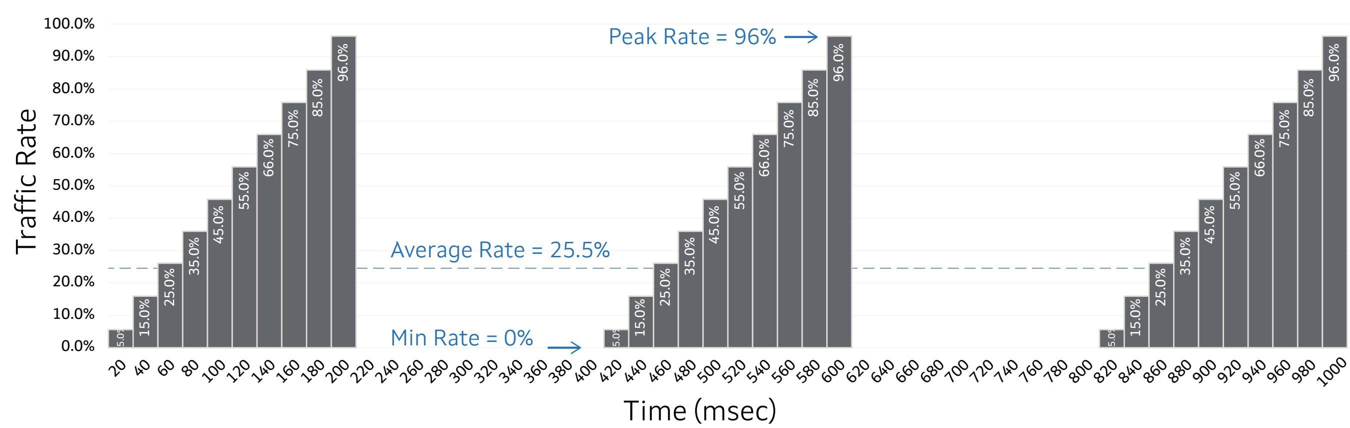

Figure 5: Port Utilization Traffic Scenario

Understanding port utilization is key for network planning and upgrade decisions from, for example, a RAN based on 10GbE links to a radio network based on 25GbE or 100GbE fiber links. Marvell's resource monitoring tool continuously monitors the rates of all queues in the system, and provides the percentage of time, inside a second, that the port was between different port-utilization bins. In addition, the tool provides the minimum, average and peak queue rates in very short time intervals.

Figure 5 shows a graph of the minimum, average and peak rates (expressed as a percentage of the port's speed) for a specific queue over time. While the samples are slow (every 1 sec), the underlying measurements are very fast measuring the correct values. Note that slow monitoring techniques would not have detected the correct peak rate and would have created the false impression that peak rate is much lower.

In addition, the fast-changing rates inside each second are placed in rate histograms that reveal the percentage of time, inside each second, a queue was transmitting at a specific rate interval. Figure 6 shows such a histogram for a specific port. The vast amount of data points (continuous measurement of rates in the entire network) is dramatically reduced, allowing continuous monitoring of the entire network and providing extremely valuable information.

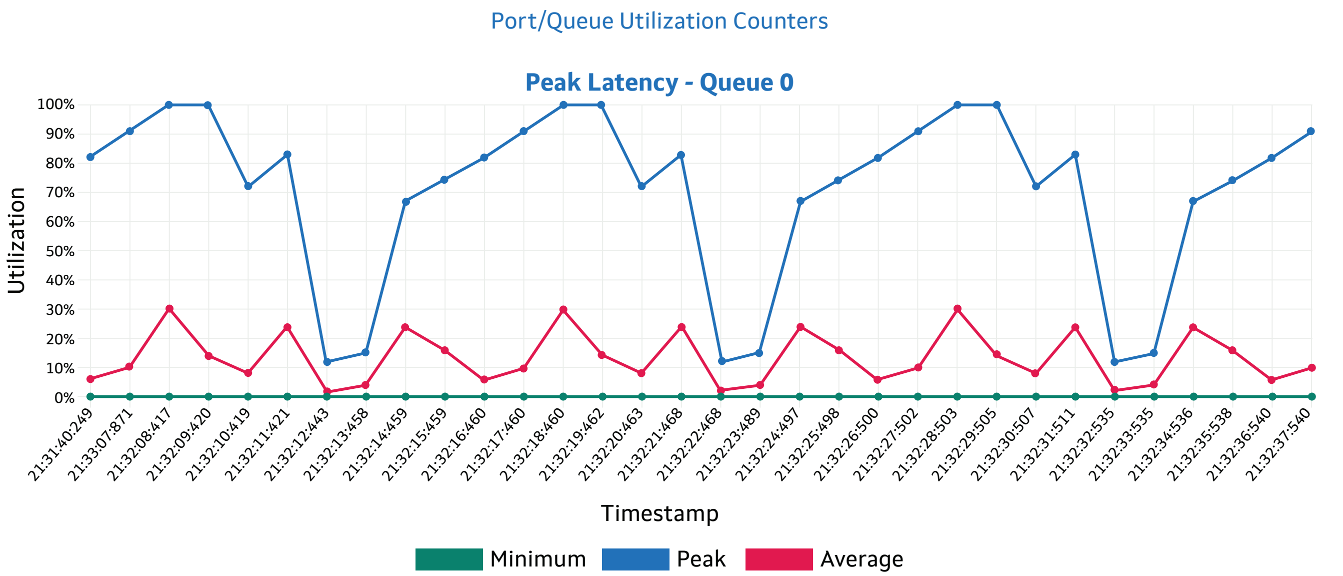

Figure 6: Monitoring Minimum, Average and Peak Resource Usage

Open RAN Monitoring

Cloud RAN, vRAN and the proliferation of players in the Open RAN create new challenges for network visibility tools. Many functions are now deployed as virtual functions over virtual machines (VMs) or containers. The Digital Unit (DU) functions, which include complex baseband processing, can be implemented and accelerated using O-RAN compliant Network Interface Cards based on, for example, Marvell's OCTEON Fusion®-O device. Virtualizing and running these real time functions on cloud servers is a challenging task that can consume most of the latency budget. As shown in Figure 7, such Cloud RAN solutions are expected to be deployed in Edge Data Centers, in proximity to the cell towers.

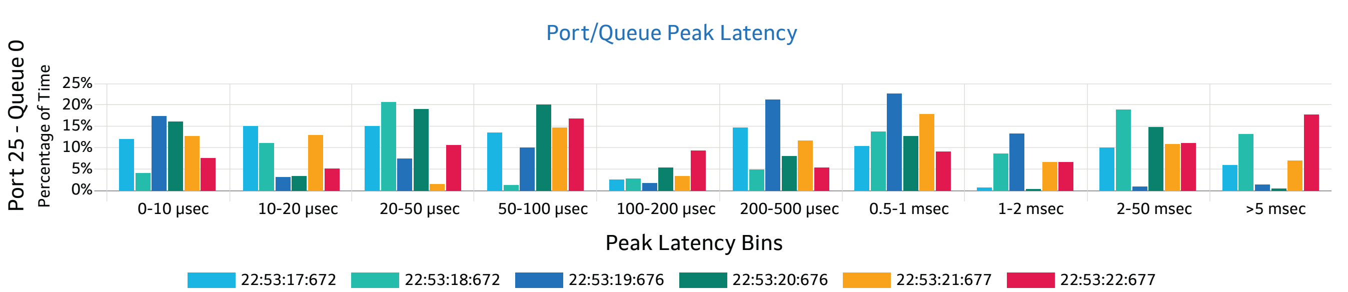

Figure 7: Resource Monitoring shown in Utilization Histogram

Like other data centers, Edge Data Centers are built in a Leaf-Spine CLOS architecture. Using Marvell's Prestera Switches in the fronthaul Radio Access Network and in the Edge Data Center provide seamless visibility to the latency and the utilization of the network, including the utilization of the links towards the servers. It is critical to monitor and make sure that the load is being spread evenly between the servers and that no congestion is ever reached.

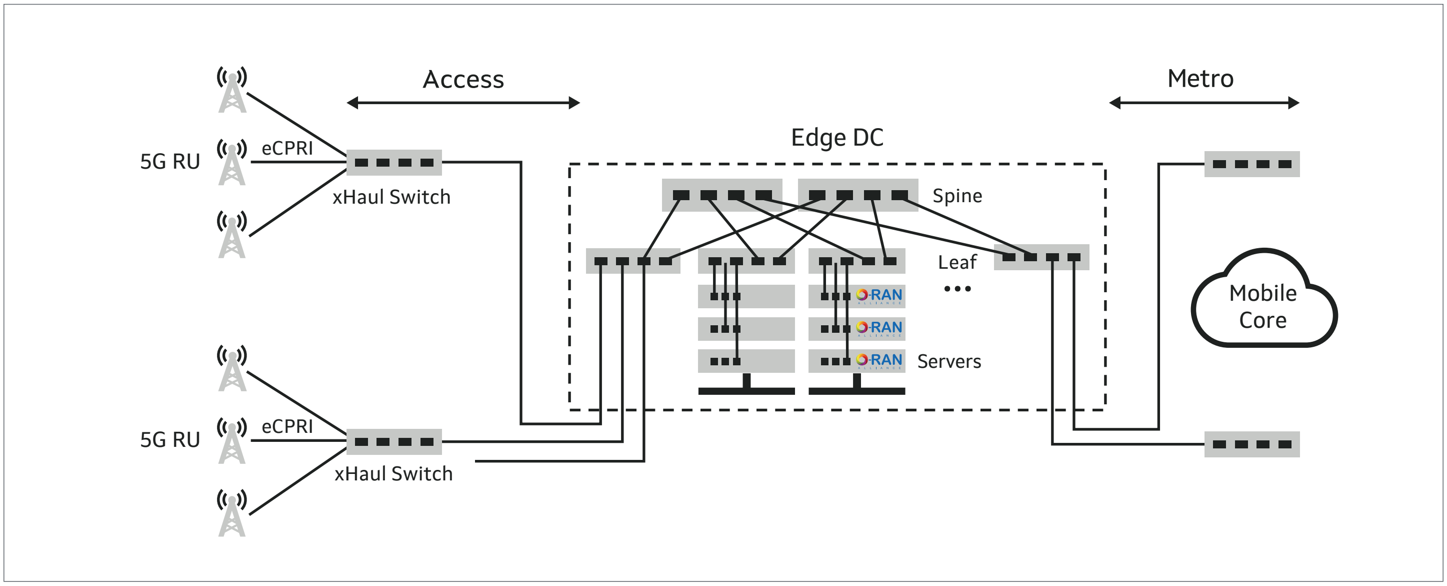

Figure 8: Cloud RAN and vRAN

Summary

5G Radio Access Networks are now being widely deployed, making it essential to accommodate their low latency, high bandwidth connectivity that is enabling new industrial and enterprise use cases. Marvell's Prestera switches equipped with TrackIQ visibility tools ensure the smooth operation of the network, allowing mobile operators to provide mission-critical, low-latency applications to their customers.