White Paper

Engineers focused on Open RAN deployment
MNOs, system integrators working on private 5G

Determining the Right Hardware for Open RAN Deployment

Tests by Aspire Technology* compare performance of 2nd Generation Intel® Xeon® Scalable processors and 3rd Generation Intel® Xeon® Scalable processors for Open RAN 5G centralized units and distributed units.

Overview

With the disaggregation of hardware and software, and the delivery of radio access network (RAN) baseband software as virtualized network functions, the underlying features and performance of the RAN hardware infrastructure is critical to providing the capacity needed for user-dense mobile services. For mobile network operators (MNOs) and original equipment manufacturers (OEMs) alike, understanding these capabilities is key to the selection of the right hardware to maximize network capacity, efficiency, stability, and performance with the attractive cost efficiency that is promised by Open RAN solutions.

Traditionally, RAN baseband functionality is delivered from a cell site using tightly integrated proprietary hardware and software from a relatively small number of incumbent OEMs. Through Open RAN, the industry is now opening up RAN interfaces to allow for a more competitive and diverse ecosystem.

Key to this evolution is the standardization of open interfaces and decoupling of hardware and software, the baseband application software can now be deployed on commercial-off-the-shelf (COTS) hardware. In the Open RAN architecture, layer 1, layer 2 and layer 3 functionality in a baseband unit is split across distributed units (DU) that supports layer 1 and layer 2, and centralized units (CU) that supports layer 3. Further, the interface between the DU and RU has also been standardized to a new open fronthaul interface relying on evolved Common Public Radio Interface (eCPRI), enabling compliant open DUs (called O-DUs) to interoperate with compliant open radio units (O-RUs). This enables flexible and modular deployment of baseband software and hardware and software to be upgraded independent of each other.

The 3rd Generation Partnership Project (3GPP [1]) has defined several architectural splits. The focus of the commentary in this paper is the 7.2 split, which realizes the layer 1 function in the baseband stack as two separate network functions; the low-PHY component of layer 1 is processed by the open radio unit (O-RU, as defined by O-RAN Alliance[2]) and the high-PHY component is processed by the open distributed unit (O-DU, as defined by the O-RAN Alliance).

This split achieves a balance across the open fronthaul interface, with respect to bandwidth and latency requirements. The software-based RAN stack allows for virtualization and/or containerization of the DU and increased deployment flexibility for the Open RAN solution.

It is important to note that the workload characteristics and performance requirements on the CU and DU differ. The DU terminates the eCPRI fronthaul connection towards the O-RU, and is bound by stringent latency and bandwidth requirements. The DU implements much more computationally intensive functionality, whereas the CU terminates the less time-critical layer 3 functionality, and thus may support traffic from multiple DUs.

Table of Contents

One way to implement the Open RAN architecture includes hardware infrastructure powered by 2nd Generation Intel® Xeon® Scalable and 3rd Generation Intel® Xeon® Scalable processors, allowing for high performance CU and DU network elements.

Aspire Technology, an independent telecommunications company and an Intel® Network Builders partner, tested both CPU families in a like-for-like test environment within its Open Networking Lab in Dublin, Ireland. The objective was to independently compare the performance of each family of processors in the context of a 5G mobile telecommunications environment, specifically focused on the impact of performance of the CU and DU operations, and the impact these have on the overall cloud and Open RAN solutions.

Finding the Right Processor for 5G Open RAN

The scope of this testing activity is to compare the performance of 2nd Generation Intel Xeon Scalable processor servers with 3rd Generation Intel Xeon Scalable processor-based servers when executing Open RAN software, specifically when executing the CU and DU elements of the 5G RAN.

For this investigation, the 2nd Generation Intel Xeon Scalable processor-based server will be referred to as "baseline"; the 3rd Generation Intel Xeon Scalable processor-based server will be referred to as "DUT1".

Test Configuration and Setup

Intel's FlexRAN™ software reference architecture enables deployment of software-based Open RAN LTE and 5G Base Stations (eNB/gNB) in several varieties.

For this CPU benchmarking exercise, we deployed a 5G standalone (SA) gNB bare-metal solution on the baseline and DUT1 infrastructure, each in two scenarios:

- **Physical abstraction layer (PAL) scenario:** Simulated user equipment (UE) using Radisys' PAL UE Sim*. FlexRAN was not utilized in this scenario (see Figure 1).
- **OTA scenario:** Over the air using an n78 O-RU and commercial UE (see Figure 2).

Both deployments were connected to a Radisys 5G SA core network, primarily using the access mobility management function (AMF), session mobility management function (SMF) and user plane function (UPF).

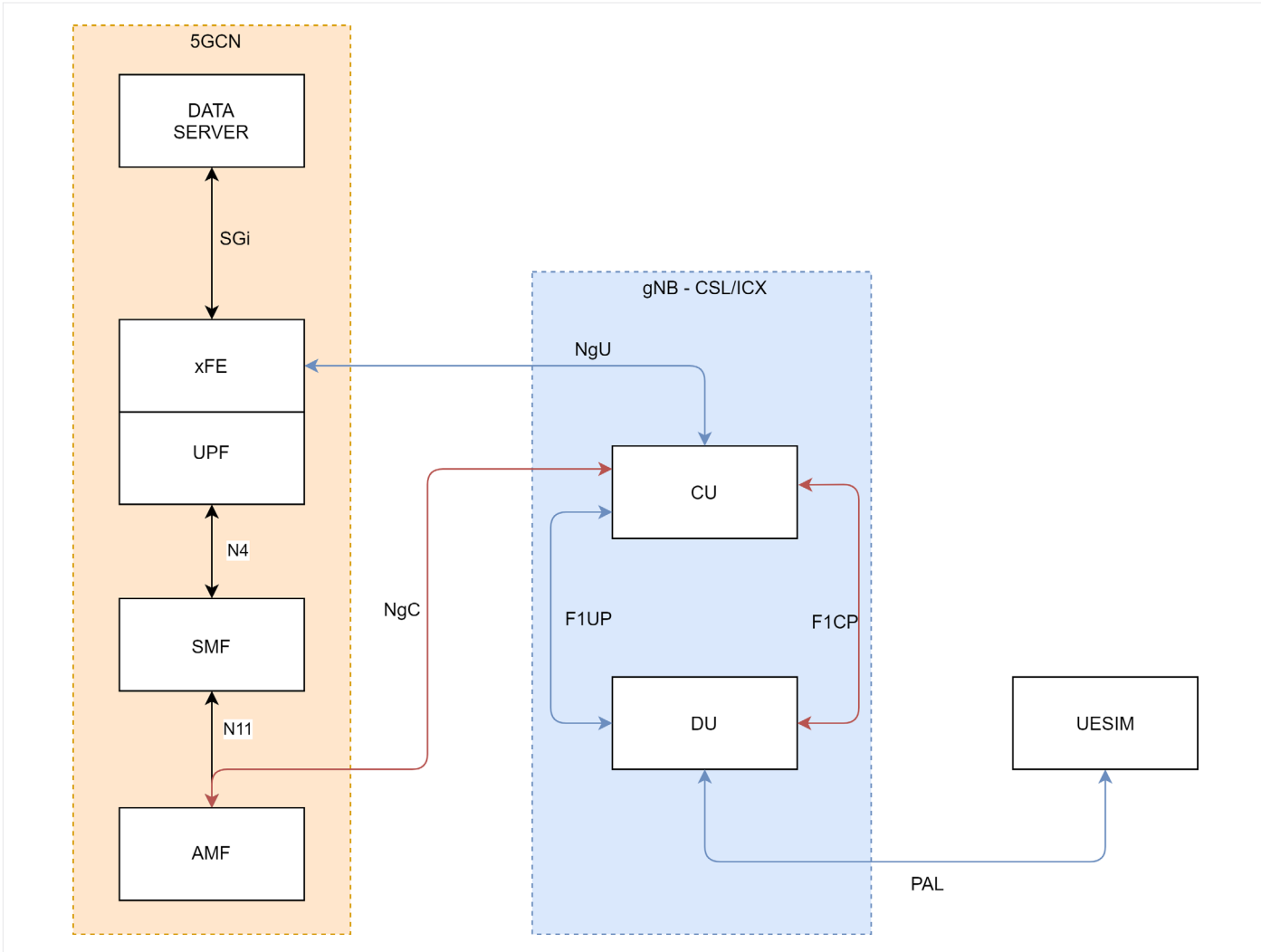


Figure 1. End-to-end 5G SA PAL network architecture.

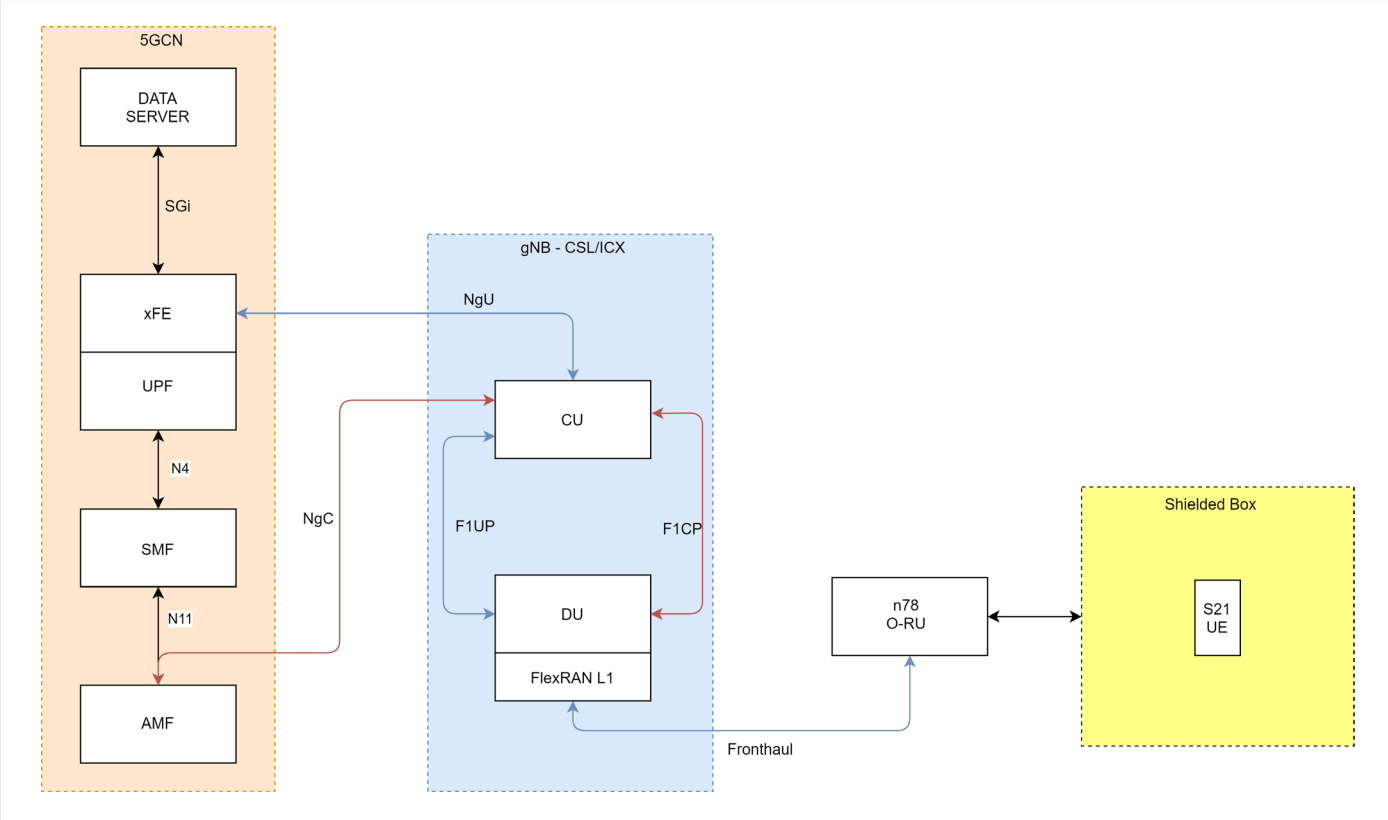


Figure 2. End-to-end 5G SA OTA network architecture.

The RAN configuration for each is as follows:

OTA - 1 cell, n78 band TDD, 100MHz channel, mu1 (30KHz spacing), up to 256QAM DL, 64QAM UL, 2x2 MIMO, DDDS SF

PAL - 1-6 cells, n256 band TDD, 100MHz channel, mu1 (30KHz spacing), up to 256QAM DL, 64QAM UL, 2x2 MIMO, DDDS SF

The exact hardware configurations are described below:

- **2nd Generation Intel Xeon Scalable-based servers (baseline):** Dell PowerEdge* R740xd with dual Intel® Xeon® Platinum 8260M processors with 24 cores at 2.4GHz; Intel® QuickAssist Technology (Intel® QAT) 8970, ACC100 Silicom Lisbon card (P3iMB-M-P1)
- **3rd Generation Intel Xeon Scalable processor-based servers (DUT1):** Intel® Server M50CYP reference board featuring dual Intel® Xeon® Platinum 8360Y processors with 36 cores running at 2.4GHz, Intel® QAT 8970, ACC100 Silicom Lisbon card (P3iMB-M-P1). 54 logical cores were disabled in BIOS to facilitate a better comparison with baseline CPU.

Both servers under test had 6x 64GB DDR4 2666MHz RAM, TSC core frequency of 2.5GHz, uncore frequency of 1.7GHz.

In terms of software configuration, each server was running FlexRAN v20.11, Radisys 5G Core Network v2.3.3 and Radisys 5G gNB v2.2 (for PAL) and 5G gNB v2.3 (for OTA).

The CPU load was generated on the gNB servers, by transferring uplink and downlink user plane traffic at various data rates in different CPU, UE and cell configurations, to understand the impact of high and low data rates, high and low number of UEs and limited CPU availability for a given thread.

Metrics were collected for both the CU (layer 3) and DU (layer 2/layer 1) to further understand the performance delta between the families of CPUs.

In order to achieve parity of performance, CPU pinning was implemented to guarantee a like-for-like test configuration in both PAL and OTA scenarios. Intel® Hyper-Threading Technology was enabled in BIOS.

Results Overview

Comparison was done of the CPU utilization on baseline and DUT1 under the same test configurations and scenarios. The PAL and OTA test results show the CPU utilization of DUT1 is lower than baseline, demonstrating the improved efficiency of DUT1.

In the PAL test scenario, DUT1 CPU utilization for layer 2 threads is lower by between 8% and 43%, and for layer 3 is lower by 23.7% to 30.2% depending on the number of cells, UEs and data rates. Generally, the higher the data rate the greater the CPU utilization reduction on DUT1 versus baseline for both DU (layer 2 and layer 1) and CU (layer 3). This was especially noticeable when testing with moderately low throughput and a high number of UEs.

In the OTA test scenario (one cell and one UE), DUT1 CPU utilization for layer 1 BBU threads is lower by 23.8%, layer 2 threads is lower by 11.2% and for layer 3 threads is lower by 8.5%. The layer 1-related CPU utilization efficiency gains (layer 1 BBU threads) with DUT1 are approximately the same (22%-25%) for all data rates tested. However, for some low-level functions, the efficiency gains for DUT1 are greatest at high data rates (72% more efficient at 800Mbps).

PAL Testing Details

Two different setups were used when testing in the PAL scenario:

1. **Basic CPU allocation with different UE data rates:** gNB vendor-recommended core allocation for basic test configuration.
2. **Minimized CPU allocation with different UE data rates:** to determine subscriber capacity limits using the minimum CPU allocation required to run the gNB functionality.

For layer 2 and layer 3, the higher the UE data rate the more efficient the CPU utilization (per Mbps transferred).

Results are generally consistent across test scenarios with greater efficiency gains on DU specific CPU cores and at higher data rates.

The number of subscribers served on DUT1 was higher by 25%-40%, where DUT1 was again most efficient at the higher data rates.

Basic CPU allocation with medium-to-high UE data rates

In a six cell configuration, 23 simulated UEs were attached to the gNB, and various download data rates (200 - 350Mbps) were used to generate CPU load on the CU and DU network functions. In the 350Mbps test case, the CU CPU became saturated when the 24th UE was added (limited by basic CPU allocation for CU performance threads – this could be extended). In order to have a like-for-like comparison across test cases, the CPU utilization percentage was recorded with 23 UEs transferring data for all data rates used in these results.

This configuration was used to understand the performance when using medium-to-high UE data rates.

For the basic configuration PAL scenario with medium to high data rates, it is noticeable that the rate of increase in CPU utilization for DU-allocated cores becomes greater when increasing the data rate. For CU-allocated cores the utilization is more linear, though there is a slight CPU utilization rate change when increasing the user throughput from 200 Mbps to 250 Mbps.

The results observed (see Figure 3) on both baseline and DUT1 CPUs showed the DUT1 CPU was more efficient running gNB applications.

DU-Assigned Cores

- Across all DU-assigned cores, there was an average of 42% lower CPU utilization on DUT1 compared to baseline (see Figure 3).

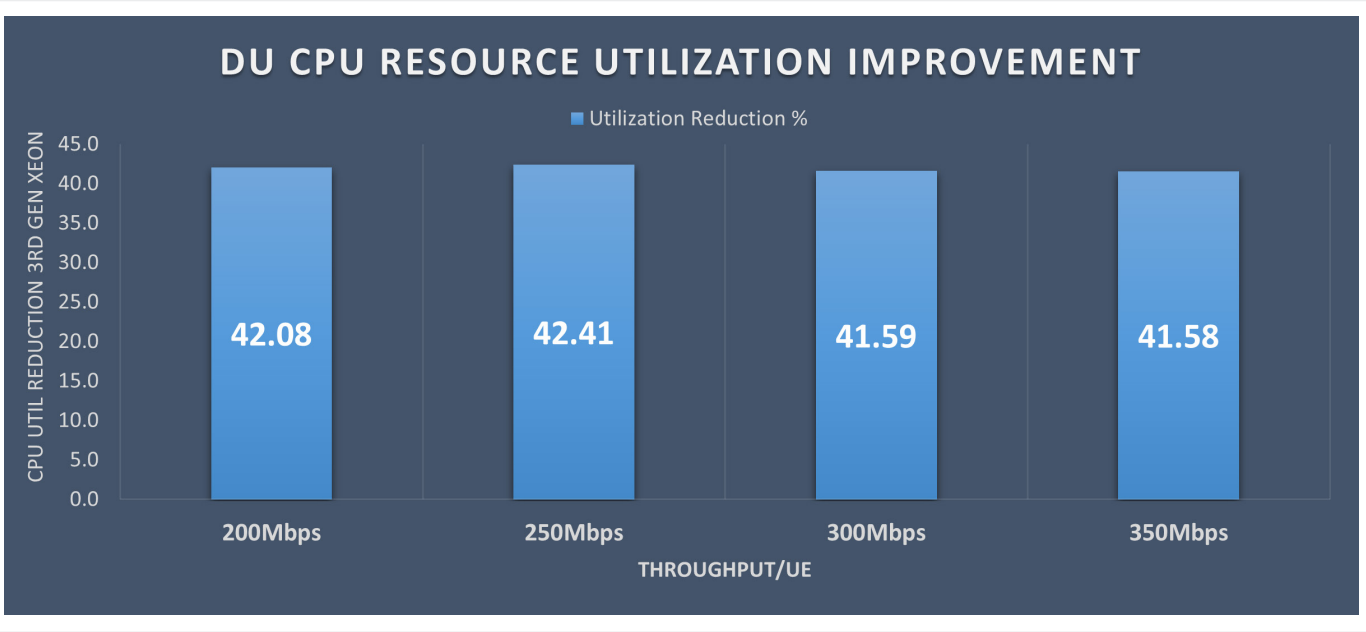
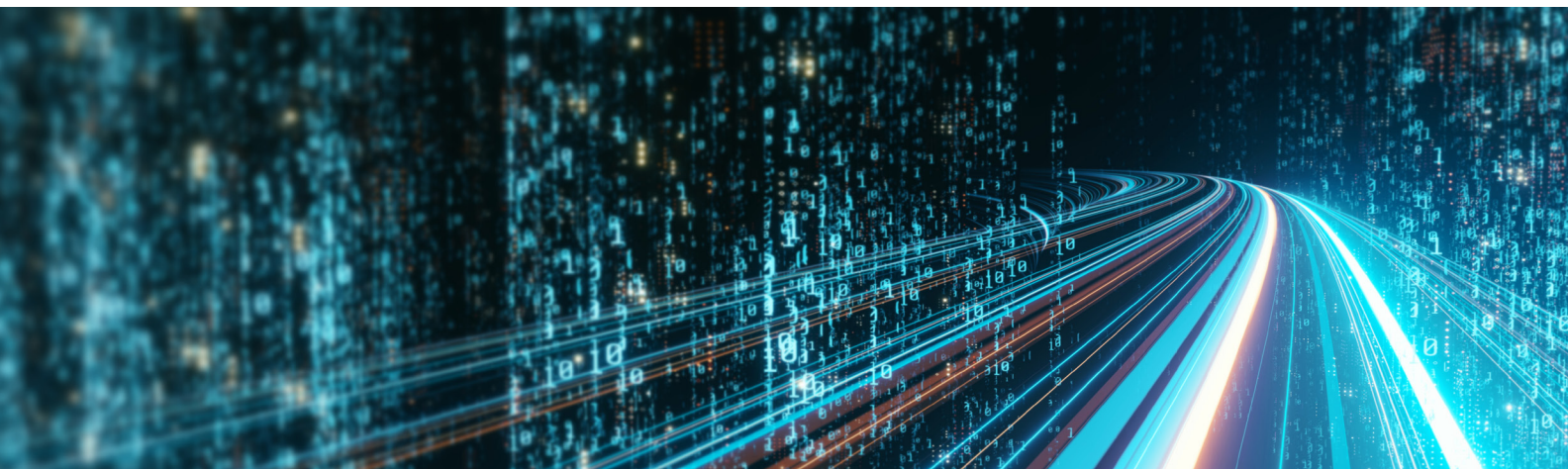


Figure 3. Average DU CPU utilization reduction in DUT1 vs. baseline for a six-cell configuration with 23 UEs (higher is better).



- For the cores assigned to the DU worker thread, the most heavily utilized DU thread, there was an average of 32% lower CPU utilization on DUT1 than the baseline (see Figure 4).

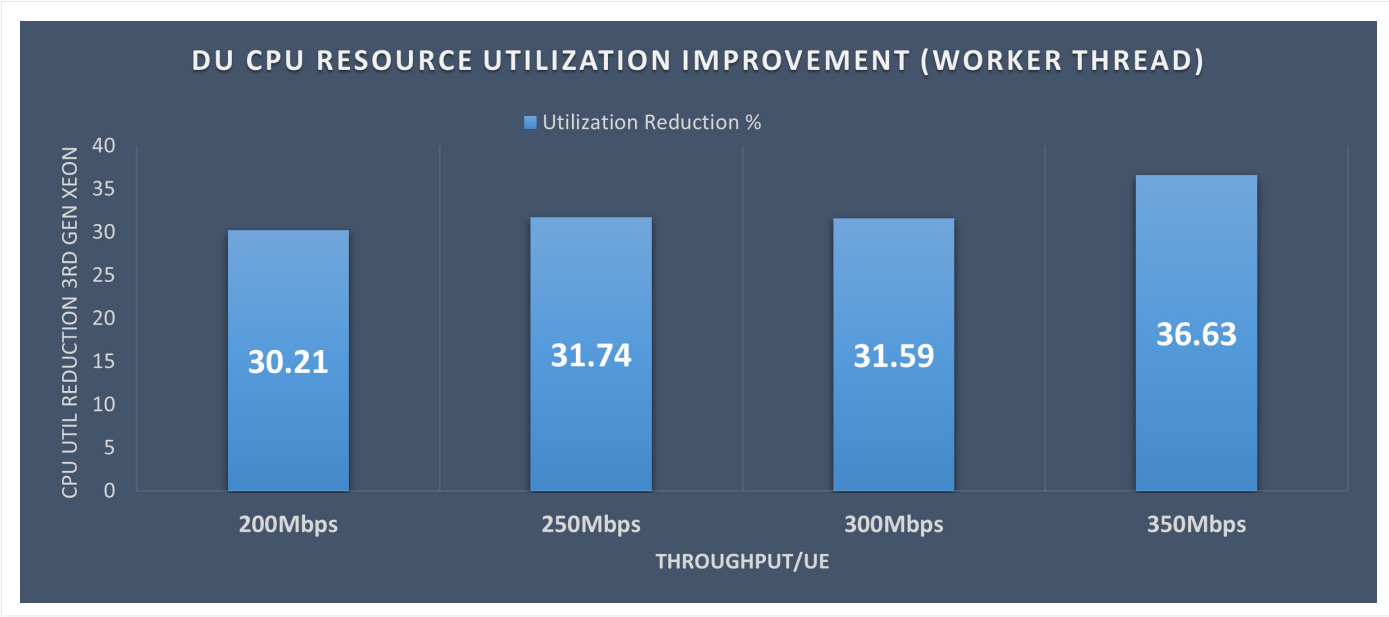


Figure 4. DU worker thread CPU utilization reduction in DUT1 vs. baseline for a six-cell configuration with 23 UEs (higher is better).

CU-Assigned Cores

- For all CU-assigned cores, there was an average of 37% lower CPU utilization on DUT1 than the baseline (see Figure 5).

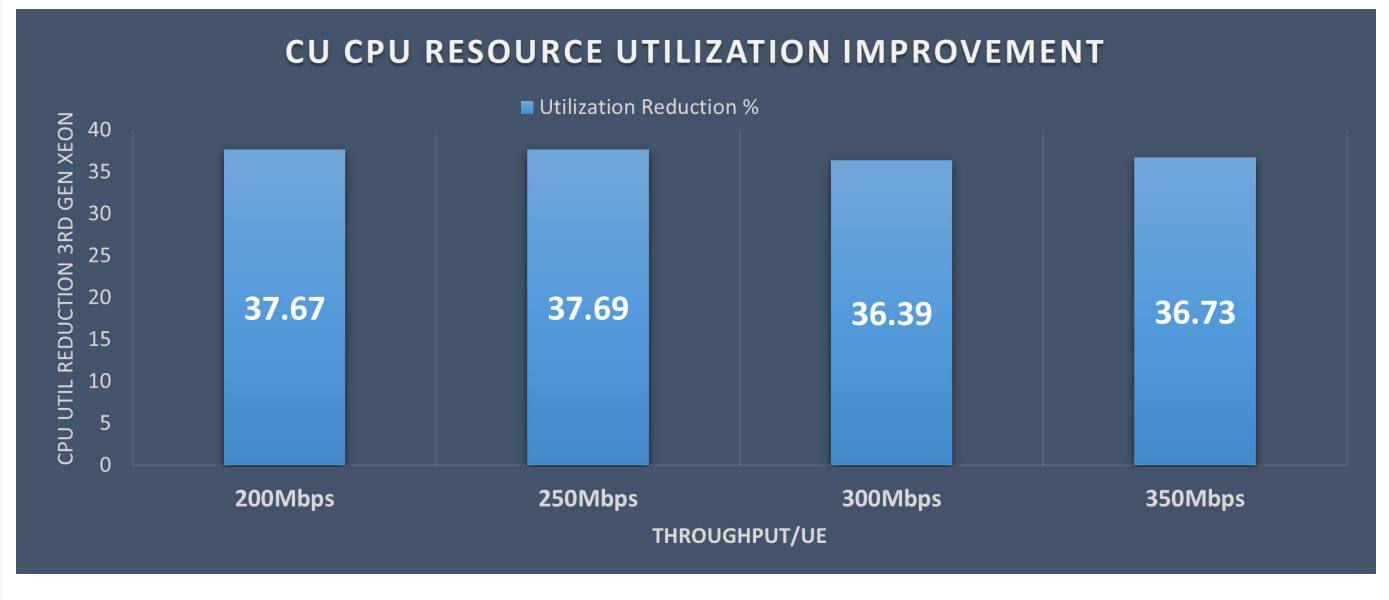


Figure 5. Average CU CPU utilization reduction in DUT1 vs. baseline for a six-cell configuration with 23 UEs (higher is better).

- For the cores allocated to the CU downlink thread, the most heavily utilized CU thread, there was an average of 29% lower CPU utilization on DUT1 than the baseline (see Figure 6).

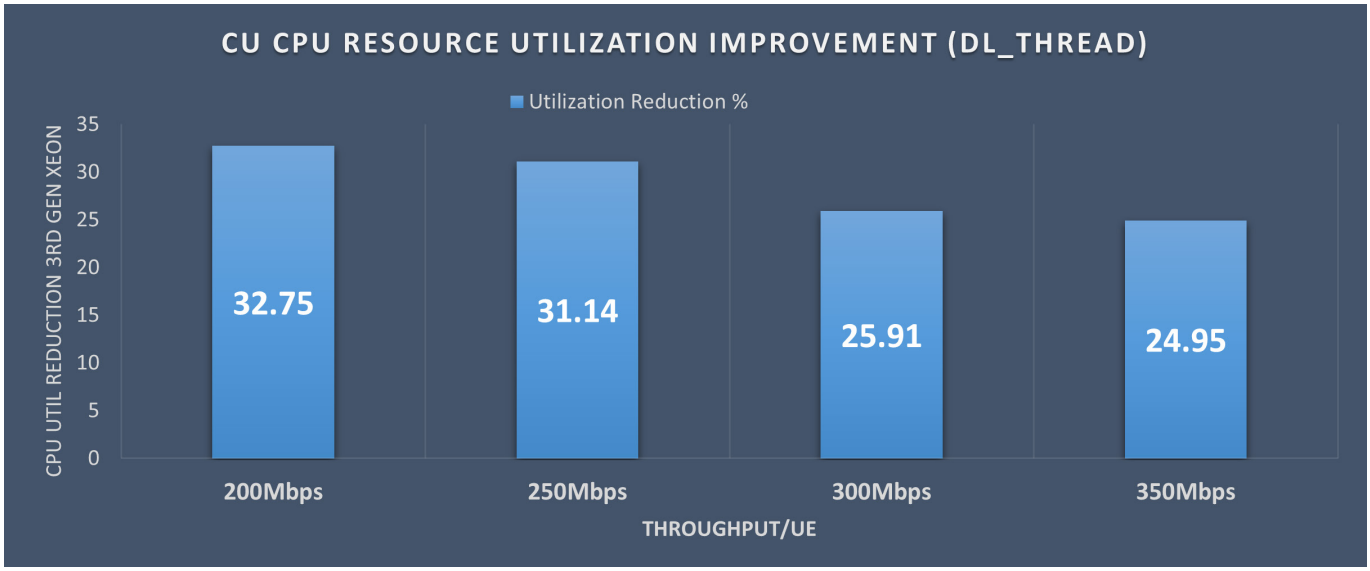


Figure 6. CU downlink thread CPU utilization reduction in DUT1 vs. baseline for a six-cell configuration with 23 UEs (higher is better).

- For both DU and CU core utilization, DUT1 became more efficient in all tested scenarios.

Normal CPU allocation with low data rates and high number of UEs

In a six-cell configuration, 60 simulated UEs were attached to the gNB and various low downlink data rates (10-160 Mbps) were used to generate CPU load on the CU and DU network functions.

This configuration was used to understand the performance when using low data rates and higher number of UEs.

The higher the data rate, the greater the CPU utilization reduction on DUT1 versus baseline for both DU (layer 2 and layer 1) and CU (layer 3).

These were the results observed across baseline and DUT1:

- Across all DU-assigned cores, CPU utilization is reduced comparing DUT1 vs. baseline. As shown in Figure 7, these values are scaling with load:
  - 12.71% at 10Mbps
  - 33.8% at 100Mbps
  - 39.13% at 160Mbps

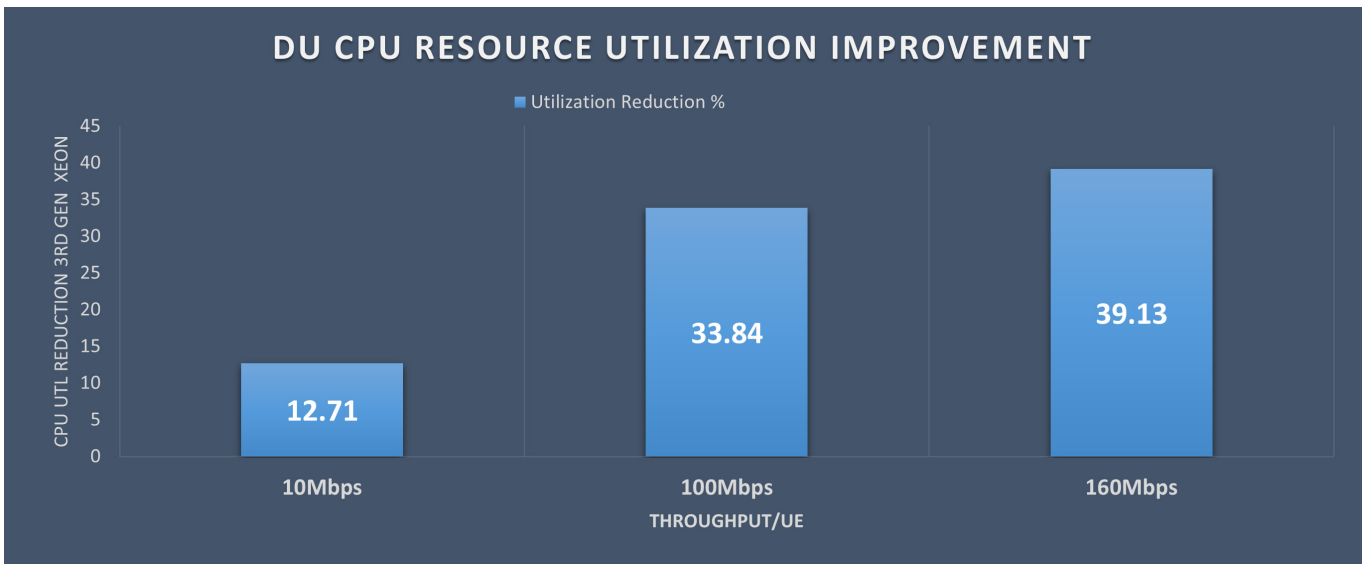


Figure 7. Average DU CPU utilization reduction in DUT1 vs. baseline in a six-cell configuration with 60 UEs (higher is better).

- Specifically for the cores assigned to the DU worker thread, CPU utilization is again reduced comparing DUT1 vs. baseline. As shown in Figure 8, these values are scaling with load:
  - 7.56% at 10Mbps
  - 27.76% at 100Mbps
  - 32.24% at 160Mbps

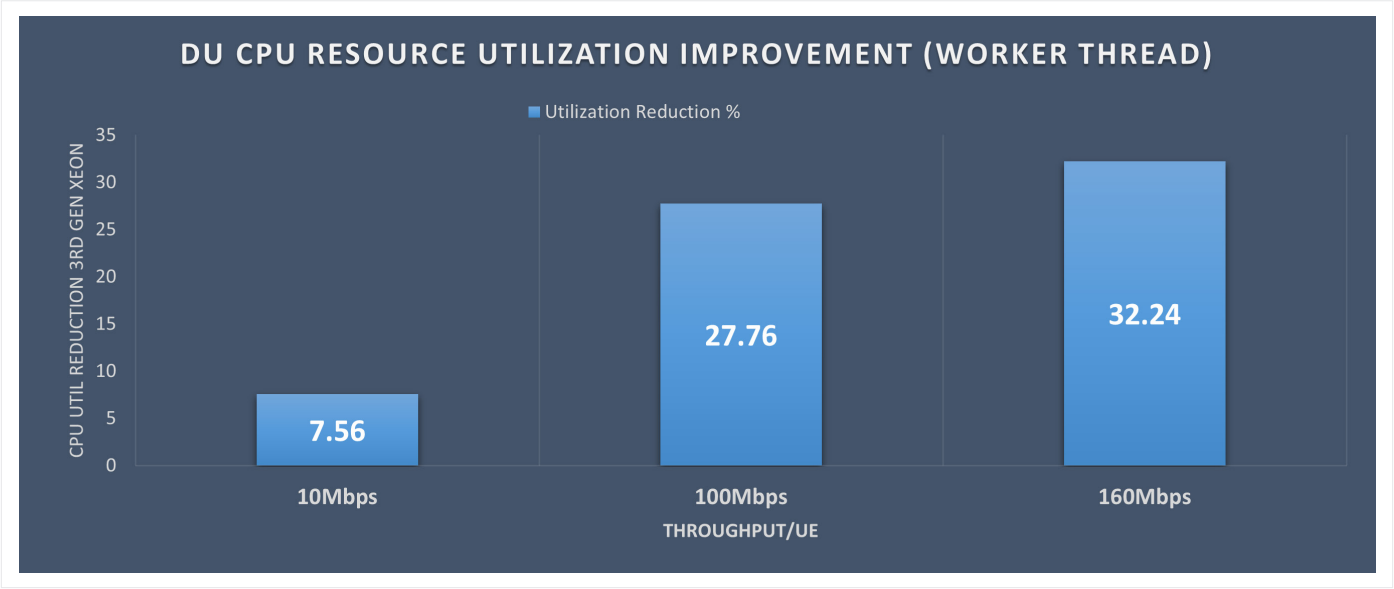


Figure 8. DU Worker Thread CPU Utilization reduction in DUT1 vs. baseline in a six-cell configuration with 60 UEs (higher is better).

- Across all CU-assigned cores, an average of 25% lower CPU utilization was observed on DUT1 vs. baseline (see Figure 9).

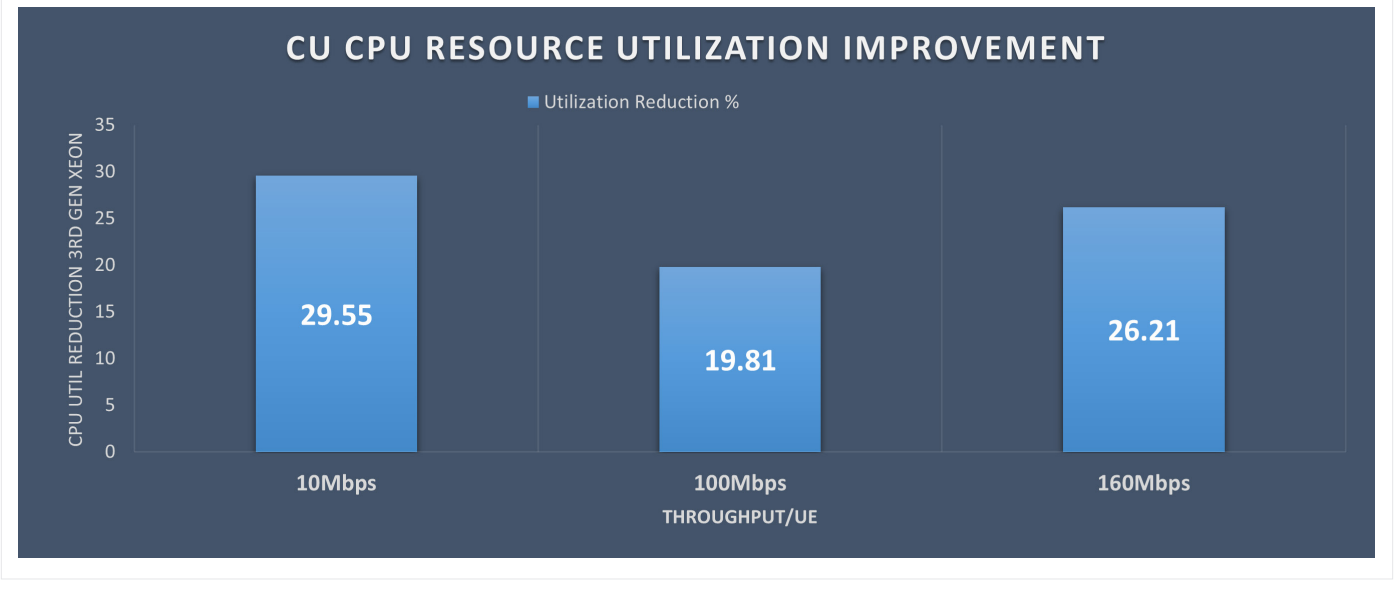


Figure 9. Average CU CPU utilization reduction in DUT1 vs. baseline in a six-cell configuration with 60 UEs (higher is better).

- Specifically for the cores assigned to CU downlink thread, an average of 25% lower CPU utilization was recorded on DUT1 vs. baseline (see Figure 10).

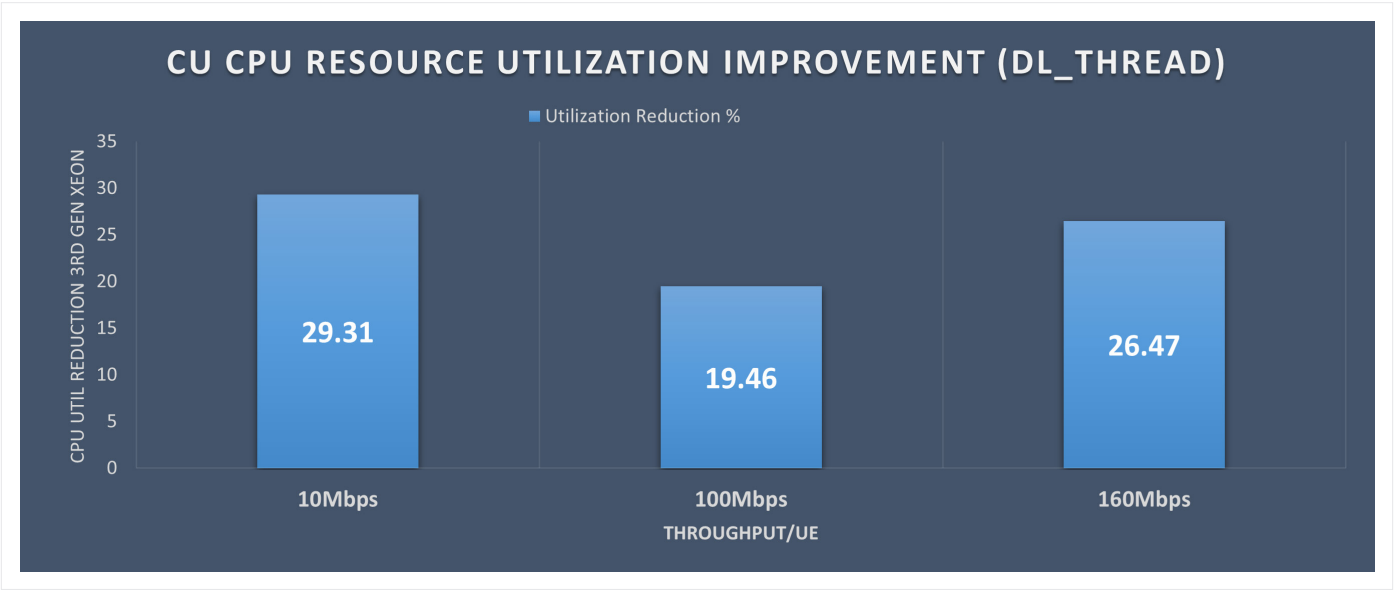


Figure 10. CU downlink thread CPU utilization reduction in DUT1 vs. baseline in a six-cell configuration with 60 UEs (higher is better).

COMPONENT	CPU UTILIZATION DELTA
DU ALL Cores	33.33%
CU ALL Cores	23.75%

Figure 11. Average CPU utilization reduction in DUT1 vs. baseline running DU and CU workloads (higher is better).

Minimized CPU allocation with different UE data rates

A single CPU core was allocated for CU downlink threads to determine the maximum number of UEs that could be served per CPU at specific UE data rates. The configuration involved three cells and up to 48 simulated UEs, used to better understand subscriber capacity of the gNB handled by the DUT1 and baseline. When saturation was reached on the CU DL thread, the test case was stopped. The number of subscribers served on DUT1 was higher by 25%-40%, where DUT1 was again most efficient at the higher data rates. (see Figure 12).

RATE PER UE	DELTA SERVED # UES
750 Mbps	40.00%
140 Mbps	29.60%
120 Mbps	25.80%
100 Mbps	33.30%

Figure 12. Subscriber capacity increase in DUT1 vs. baseline (higher is better).

OTA Testing

The deltas between baseline and DUT1 for layer 1 logical core utilization in OTA testing are in the range of 20.7% - 72.5%. While DUT1 is observably more efficient on layer 2 and layer 3, the scale of this is difficult to determine as testing was limited to 1 UE in this setup.

In a single cell configuration, one Samsung S21 UE was attached to the gNB and various downlink data rates ranging between 400 and 800Mbps and an uplink of 20Mbps. Data was transmitted to generate CPU load on the CU and DU network functions, as well as the FlexRAN layer 1. These were the results observed across baseline and DUT1 CPUs (see Figure 13):

- Across all DU assigned cores, CPU utilization is reduced comparing DUT1 vs. baseline. The largest reduction was observed on UL test cases, with utilization up to 51% reduced on DUT1.
- Across all CU-assigned cores, an average of 10.5% lower CPU utilization on DUT1 vs. baseline was recorded.
- Specifically for the cores assigned to CU downlink thread, an average of 8.1% lower CPU utilization was observed on DUT1 vs. baseline.
- Specifically for the cores assigned to layer 1 BBU, CPU utilization is reduced comparing DUT1 vs. baseline. DUT1 is 22%-25% more efficient than baseline. The efficiency gains with DUT1 are similar for all data rates.
- Regarding the number of logical cores required for FlexRAN layer 1 BBU tasks, there is a large reduction on DUT1 vs. baseline. This extends to 72% reduction for downlink, and 65% reduction for uplink, in the most user plane-intensive test case (800Mbps). This logical core utilization reduction ranges from 28.2% - 72.5% on downlink to 20.7% - 72.49% for uplink.
- Regarding the channel utilization of FlexRAN layer 1 BBU tasks PUSCH, PDSCH, PRACH, OTHERS and PDSCH FEC are all more efficient on DUT1 vs. baseline. There is an average reduction on DUT1 vs. baseline of 27.6%.

COMPONENT	CPU UTILIZATION DELTA %
CU ALL Cores	8.49%
DU ALL Cores	11.20%
L1_BBU	23.82%
L1 DL Logical Cores	37.28%
L1 UL Logical Cores	42.49%

Figure 13. CPU utilization reduction in DUT1 vs. baseline (higher is better).

Conclusions

One of the main motivations of Cloud RAN, and more specifically Open RAN, is to develop a solution and an ecosystem where both CAPEX and OPEX are reduced. This has been clearly communicated to the industry with CAPEX ranked one and OPEX two, in a recent GSMA survey conducted by Aspire for the most important considerations when adopting Open RAN (view a webinar on the report here).

One critical element to this is the advancement of all hardware components within the supply chain that make up the vertically and horizontally integrated RAN. The results seen from the performance benchmarking follow this path with clear reduction in required CPU utilization, to deliver the same unit of RAN functionality. It is also clear that more significant gains are seen in the DU and this was expected due to higher computational activity within this network function, in the processing of layer 1 and layer 2 traffic than in layer 3 traffic processed in the CU. Given also the relatively low load on the CU and the performance gains seen from a DU as a direct relationship to computational activity, it is expected that the gains seen for the CU will improve in a more loaded environment.

When we translate the results to CAPEX and OPEX, there are a number of areas we can map to directly. From a CAPEX perspective, it is clear that less infrastructure is required in the number of servers needed to deploy CU and DU network functions. This could broadly be mapped to a 40% reduction in compute requirements for DU network functions and a 25% reduction in compute requirements, for the CU network function, under certain distribution models. In a pooling environment, this represents a significant saving to the operator. This pooling is realized through placement of the CU/DU network functions at either the edge or regional data centers and moving from a cell site DU deployment.

From an OPEX perspective, the significant savings here are a result of energy efficiency. CPU reduction percentages can be mapped conservatively to energy reduction, given that the data center environmental energy footprint should also reduce accordingly. Given the emphasis on energy within cloud and Open RAN, this represents again an incremental reduction in the carbon footprint for a next Generation release of hardware. It should also be pointed out that 3rd Generation Intel Xeon Scalable CPU has additional energy saving functionality over its predecessor, so it is expected that implementing these

features on top of the measured energy gains, will yield a significant carbon footprint reduction for 3rd Generation Intel Xeon Scalable CPU technology over 2nd Generation Intel Xeon Scalable CPUs. This additional carbon footprint reduction will be independently evaluated by Aspire and exact savings further reported in a subsequent white paper.

Finally, in terms of capacity, and the results show that like for like, 3rd Generation Intel Xeon Scalable CPU-based servers will significantly increase the capacity of the RAN, in terms not only of user data handling, but also allowing for increased subscribers and additional cell carrier addition.

The availability of 3rd Generation Intel Xeon Scalable processors is great news for the cloud and Open RAN ecosystem, bringing clear and unambiguous benefits to the industry, targeting the very topics that are on top of operator's agendas in adopting the technology, CAPEX and OPEX. It also addresses the other hot topic for both MNOs and other industries alike, that of energy efficiency where the CPU will make a significant impact on the carbon footprint of the RAN from a server perspective. Finally, from a capacity perspective, a like-for-like server deployed as a network function host will provide notable capacity improvements.

Aspire Technology ULC is an independent telecommunications company headquartered in Dublin, Ireland. Aspire provides software and services to perfect and optimize the deployment, performance and operations of customer systems and networks. More recently we have expanded our services to include open networks and have invested strongly in our open network's lab. For further information please visit www.aspiretechnology.com.

Radisys is a global leader in open telecom solutions and services. Its disaggregated platforms and integration services leverage open reference architectures and standards combined with open software and hardware, enabling service providers to drive open digital transformation. Radisys offers an end-to-end solutions portfolio from digital end points, to disaggregated and open access and core solutions, to immersive digital applications and engagement platforms. For more information, visit www.radisys.com.

Further Study

Some topics that could be considered for further study:

- Testing with additional numbers of UEs in OTA and PAL environments
- Utilizing additional bandwidth on user plane interfaces (NgU/Fronthaul) – 25/40/100Gbps NICs
- Using varied radio configurations – E.g., Massive MIMO, UL/DL Modulation, alternative Radio Bands/Bandwidths, TDD/FDD, slot formats
- Comparison using alternative gNB vendors for layer 2/layer 3 software
- Comparison between gNB deployed as a VNF/CNF/PNF
- Multi-sector configurations where UE mobility is utilized in testing

Learn More

Aspire Technologies

Intel® Network Builders

Intel® Xeon® Scalable processors

3GPP

ORAN Alliance

Radisys

0922/TM/HO9/PDF ♻Please Recycle 351928-001US