5 ways to reduce power use in the RAN

Table of contents

The RAN has huge potential for energy savings

Energy efficiency and sustainability are top priorities for mobile network operators.[1] No wonder when electricity prices have increased by 300% since 2016.[2] The radio access network (RAN) offers huge potential for savings, with more than 70% of the network's power used there.[3]

Network operators can pursue several strategies to reduce power bills and improve sustainability. This paper will outline how the compute resources in virtualized RAN (vRAN) can be optimized to reduce power requirements.

84%

Of network operators rate energy efficiency and sustainability as a top priority[1]

300%

Increase in electricity prices in major markets since 2016[2]

70%+

Of network operators' total network power usage is consumed by the RAN[3]

Helping you to improve sustainability

Intel plans to achieve net-zero greenhouse gas emissions in our global operations by 2040.

"The impact of climate change is an urgent global threat," said Intel chief executive officer Pat Gelsinger (pictured, right). "Intel is in a unique position to make a difference not only in our own operations, but in a way that makes it easier for customers, partners, and our whole value chain to take meaningful action too."

ABI Research evaluated more than 80 telco network infrastructure and equipment vendors for their sustainable impact and implementation. Intel ranked #1 among all chipset and component providers and #1 among all Open RAN vendors.[4]

1. Matching processor resources to vRAN demand to save power

Advanced general-purpose processors (GPPs) are increasingly used for demanding virtualized RAN (vRAN) workloads. The first stage in optimizing power consumption in the RAN is to select processors that deliver the right level of performance.

The physical layer processing in vRAN demands deterministic, low-latency, and high-performance computing resources. Capacity enhancement measures such as Massive Multiple-Input Multiple-Output (massive MIMO) and beamforming further increase the compute requirements.

Network operators must use processors that deliver the compute performance required while helping them optimize their power costs.

With Intel® processors, network operators can right-size processors to match the capacity, compute, and power requirements of their vRAN sites while using a single software base.

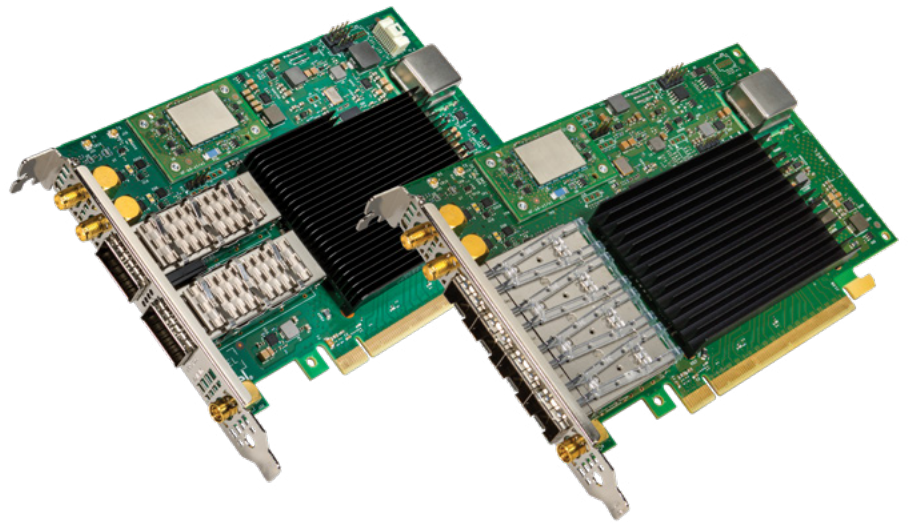
Intel's compute and Ethernet portfolio includes:

- **4th Generation Intel® Xeon® Scalable processors**, which are suitable for medium- to high-capacity vRAN deployments. The Intel® Advanced Vector Extensions (Intel® AVX) for vRAN instruction set architecture helps to achieve up to 2x capacity gains compared with the previous generation.[5]

  The processors' Intel® vRAN Boost feature integrates vRAN acceleration into the CPU, delivering additional compute power savings of approximately 20%.[5] These advances combine to more than double the performance per watt versus the previous processor generation.

- **Intel® Xeon® D-1700 processors**, with up to 10 processor cores. These processors are suitable for lower-capacity vRAN implementations, such as small-cell indoor and outdoor sites.

- **Timing-enhanced Intel® Ethernet 800 Series Network Adapters**, which deliver the network timing accuracy and synchronization required for 5G vRAN solutions. Support for IEEE 1588 Precision Time Protocol and Synchronous Ethernet, an optional, integrated GNSS receiver, and extended holdover allow network operators to replace dedicated timing hardware with a standards-based Ethernet adapter to simplify deployments and reduce power consumption.



1. Matching processor resources to vRAN demand to save power

2. P-States lower processor frequency to save power

GPPs are designed with multiple discrete compute entities called cores. The multicore architecture enables more efficient simultaneous processing of multiple tasks, thereby increasing performance.

The cores in Intel® Xeon® processors can be individually controlled to save power by changing frequency and using lower power levels.

Some workloads, such as packet processing, need compute resources continuously, but the workload's intensity can vary. Workloads such as these can be optimized using performance states called P-states, which change the core frequency and voltage.

Processor telemetry exposes information about the utilization of the cores, which can be used to adjust their frequencies in response to the load on them. If a core is underutilized, its frequency can be lowered, reducing the power and throughput capacity but without affecting RAN performance.

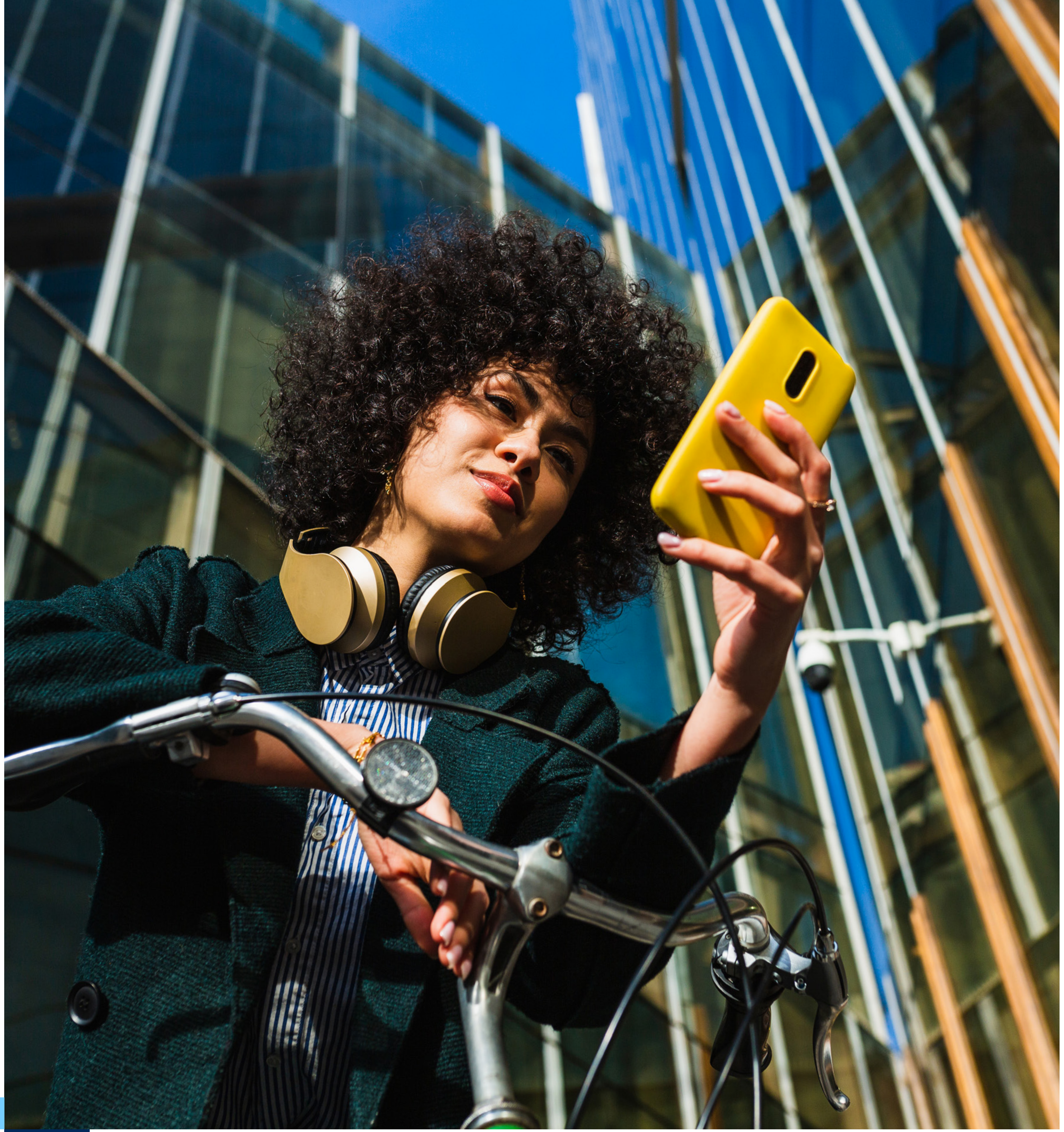
up to 20%

Power reduction seen by Vodafone in proof of concept using P-states, C0/C1 states, and uncore frequency scaling[6]

3. C-States enable processor microsleeps to save power

Some workloads in the RAN pipeline are 'bursty,' characterized by widely variable peaks and troughs in demand at unpredictable intervals. One example is the Layer 1 (L1) physical layer processing. Even in times of high traffic, there will be small periods (measured in microseconds) where demand will be lower.

During the troughs, idle cores can be put into a low-power state to save power.



This is possible thanks to C-state power management supported by the processor. C0 is the active state where the processor core is executing instructions. C1 is a state where the core is not working but can return to C0 with a latency as low as 1 microsecond. Higher C-state numbers, such as C6, put the core into a deeper sleep, saving more power but taking longer to return to C0.

The near-instant exit time from C1 makes it suitable throughout the vRAN processing pipeline. Sleep states might only last 10 microseconds, but the combined power savings can be significant. Deutsche Telecom, for example, saw power savings of approximately 30% in a proof of concept.[7]

The Intel® FlexRAN™ reference architecture is a complete L1 implementation that helps vendors develop RAN solutions optimized for Intel® processors. It enables intelligent scheduling of application workloads on CPU cores so the Linux kernel can put unneeded cores into a lower-power C-state (C1 or C6) for the next Transmission Time Interval (TTI). Metrics such as throughput and the number of user equipment devices can help the software determine how many cores will be required in the next TTI.

The telemetry in Intel processors provides information about the utilization of the cores, memory subsystem, and other processor resources. When cores are underutilized, it can be possible to consolidate their workloads into fewer cores and put the resulting unused cores into a C6 state.

Optimizing the RAN software to take advantage of C1 and C6 sleep states avoids having cores running on full power when not required, even for extremely short periods.

30%

Power reduction seen by Deutsche Telekom in proof of concept using C-states[7]

20%

Power reduction in low-traffic period seen by Ericsson in proof of concept using C-states[8]

10%

Power reduction in high-traffic period seen by Ericsson in proof of concept using C-states[8]

4. Pooling workloads on servers or cores to save power

Network infrastructure is typically dimensioned for peak usage. However, the number of users and type of traffic can vary dramatically, depending on the time of day or geography.

Consider a pool of cells served by a pool of distributed unit (DU) servers. In standard DU server pooling, each cell cluster is statically connected to a specific DU instance during site provisioning. Every DU server remains active all the time to process traffic from any user from any of the cells in the cluster it serves. General purpose processors allow for the pooling of cells within a single DU server or sharing the DU server resources between multiple cells (see Figure 1).

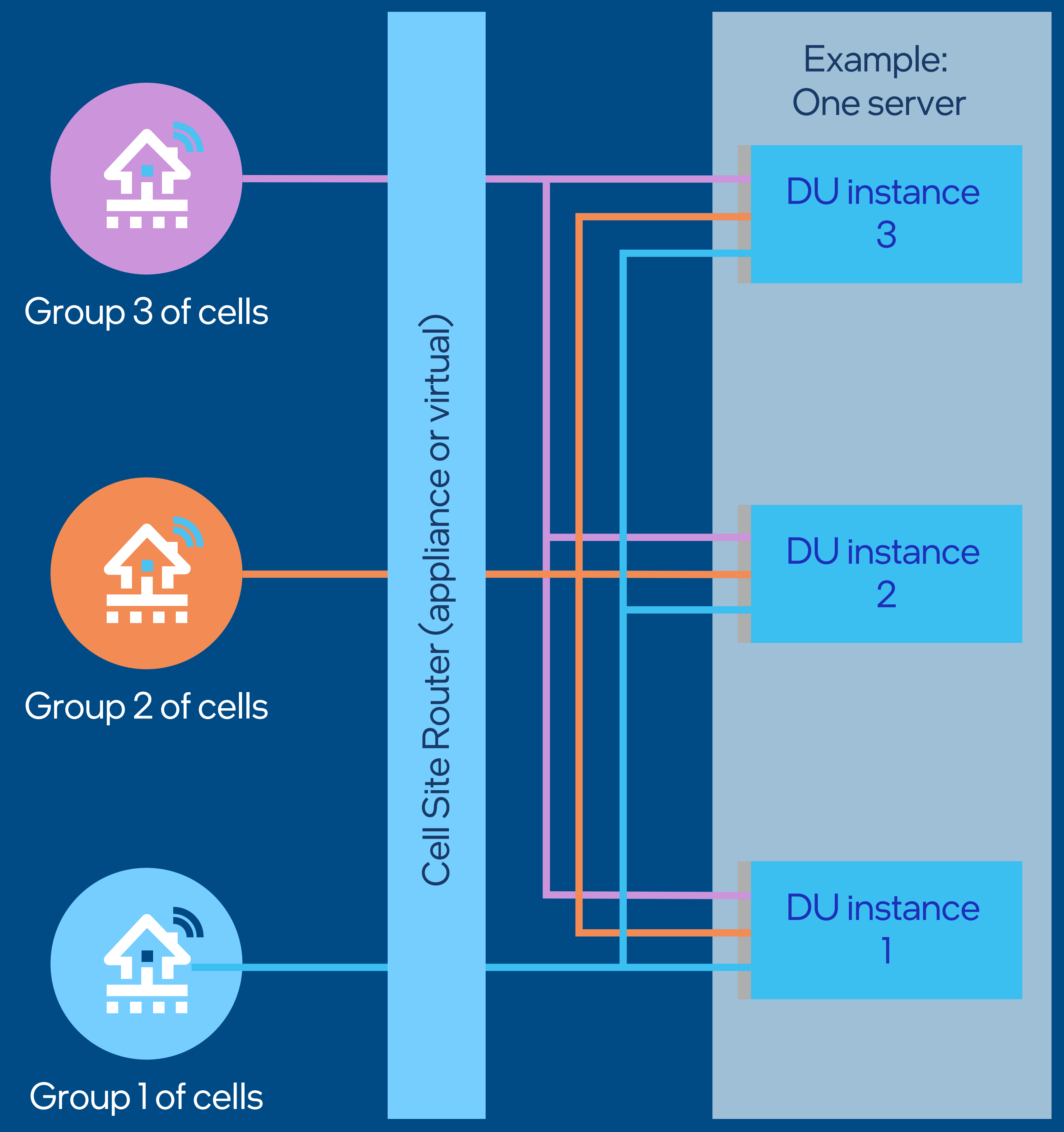


Figure 1: Pooling at a site with a single server

However, when the number of users and traffic decrease, the compute power of all the active DU servers is not needed. This presents an opportunity to keep active only the servers needed to meet current traffic needs. The remaining servers can be turned off, reducing power consumption and related costs.

Advanced DU pooling technology enables new users to be dynamically connected to any DU server in the pool. Existing users from any cell can also be migrated to any DU server. This technology can also be scaled to apply to cores in a CPU. In this case, the processing is consolidated into a group of cores, leaving unused cores for other functions. They can be put in a sleep state if they are not needed.

5. Using AI to reduce power consumption

Artificial intelligence (AI) brings new opportunities to optimize power consumption in the RAN by predicting network load and enabling more dynamic resource allocation. AI software can run inside the virtualized RAN, on the DU, centralized unit (CU), or the near-real-time or non-real-time RAN Intelligence Controller (RIC), depending on the application's latency tolerance.

AI software running on the near real-time RIC, the non-real-time RIC or CU can, for example, save power by dynamically modulating processor frequency using P-states. Software on the DU could put cores into power-saving microsleeps using C-states.

One promising idea is to use AI to learn traffic patterns and intelligently reduce power consumption using radio sleep modes. An AI application can instruct the radio to “sleep” an individual carrier, an entire radio, or even reduce the number of transmitters in a MIMO configuration—all without affecting the performance and user experience—by helping the RAN to ensure users are on active carriers.

As Figure 2 shows, the remote radio head consumes 67% of the energy in the RAN, so there is significant potential for savings.[9]

Capgemini and Intel worked together to create an application for the non-real-time RIC that uses AI to reduce the energy used by the radio unit. It’s an rApp, a microservices-based app that runs on the non-real-time RIC for use cases with a latency tolerance greater than 1 second. Capgemini tested the solution across twenty 4G and 5G radio sites with 150 cells. During the low traffic period, it switched off up to 120 cells.[10]

The algorithm uses closed-loop automation to improve the accuracy of the energy-saving decisions over time. The rApp uses Capgemini’s NetAnticipate platform and Intel AI software, including BigDL Chronos and Intel® oneAPI software. The rApp runs on 3rd generation Intel® Xeon® Scalable processors and Intel supported Capgemini in optimizing the solution.

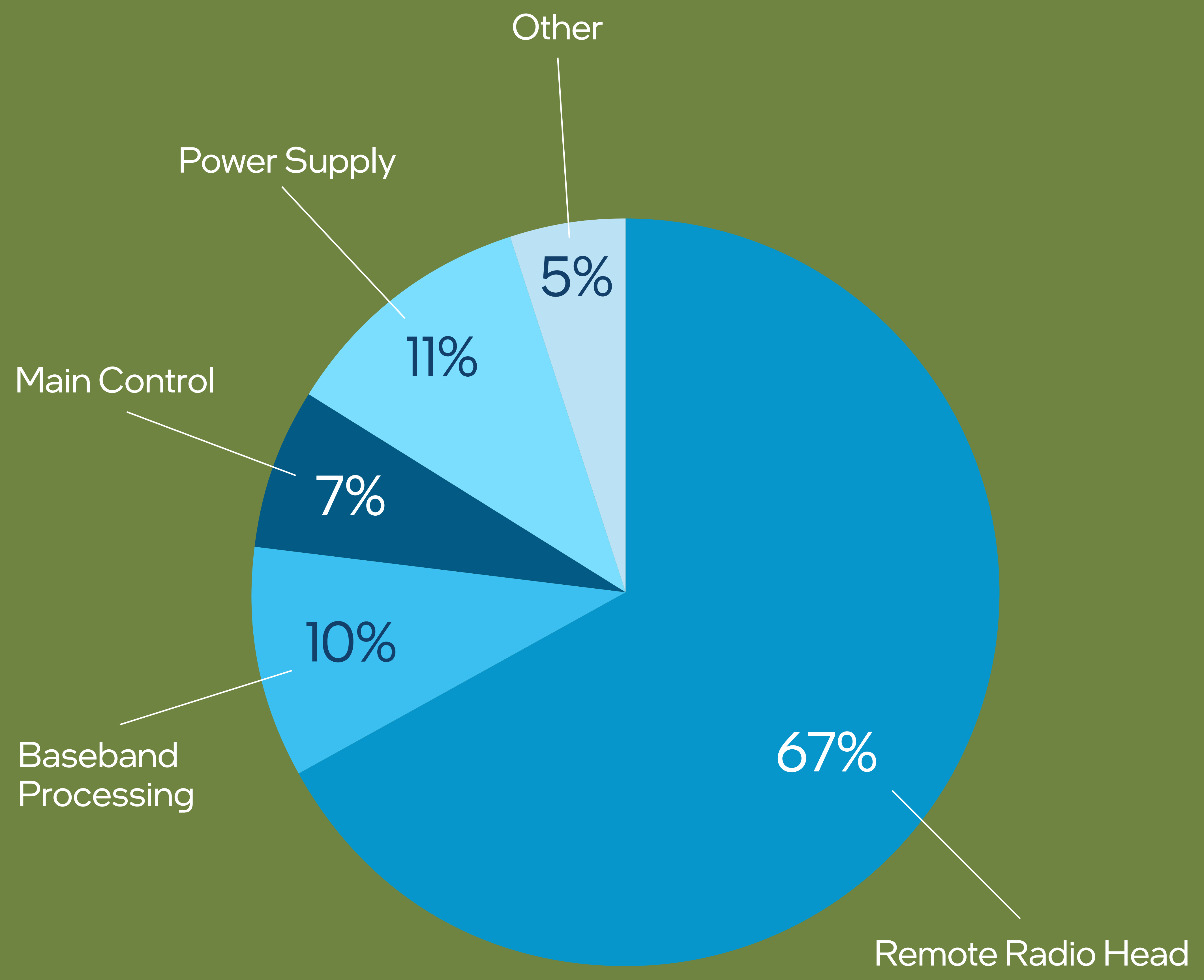


Figure 2: Power consumption in a base station without air conditioning[7]

Conclusion

As network operators seek to improve their sustainability and reduce power costs, Intel and the mobile ecosystem provide creative options to save energy in the RAN.

Network operators can choose processors that are right-sized to the vRAN they are serving to reduce energy requirements. vRAN workloads can be pooled on processing resources to enable unneeded servers to be powered down.

At the processor core level, C-States and P-States enable power savings by matching power consumption to the workload demand on the processor. Intel® FlexRAN™ reference software helps network operators and vendors access processor telemetry information, control power states in the RAN, and enable sophisticated AI-based power-saving applications.

Learn more

- Intel commits to net-zero greenhouse gas emissions in its global operations by 2040
- Intel: Driving the Future of Sustainable Computing
- Fact sheet: 4th Gen Intel® Xeon® Scalable processors with Intel® vRAN Boost
- High-efficiency compute for mid-range vRAN deployments
- First Intel structured ASIC for 5G, AI, cloud, and edge announced Intel® eASIC™ N5X devices
- How to make the business case for open and virtualized RAN
- Unleashing the potential of AI in the RAN

1. The Green Generation: Bridging 5G and 6G, GSMA Intelligence, November 2022
2. Electricity Market Report, International Energy Agency, July 2022
3. Going Green: Benchmarking the Energy Efficiency of Mobile, GSMA Intelligence, June 2021
4. Intel: Driving the Future of Sustainable Computing, ABI Research, 2023
5. See configuration information at intel.com/performanceindex
6. Dynamic power management to achieve energy savings in multi-vendor Open RAN systems, Vodafone, Intel, Keysight, Radisys, Wind River, MWC Barcelona 2023
7. DT's Open RAN Journey, Tecknexus.com, July 2022, p49-53
8. Embracing energy efficiency in Open RAN architecture, Ericsson, 13 July 2023
9. A holistic study of power consumption and energy savings strategies for open vRAN systems, Mavenir and Intel, February 2023
10. AI Enabled Energy Savings in Open RAN, Capgemini and Intel, MWC Barcelona 2023