A Holistic Study of Power Consumption and Energy Savings Strategies for Open vRAN Systems

WHITE PAPER

February 2023

TABLE OF CONTENTS

EXECUTIVE SUMMARY

As mobile traffic increases, mobile networks are transforming into more software-driven, virtualized, flexible, intelligent and energy efficient systems. These trends have spurred significant changes in the core network design with the advent of Software-Defined Networking (SDN) and Network Functions Virtualization (NFV), which have enabled building more agile, less expensive network entities. However, the radio access network has largely remained unchanged until recently, even though much of the CAPEX and OPEX in building and managing networks results from the Radio Access Network RAN. Traditionally, RAN components such as radio transceivers and baseband were implemented on proprietary hardware, and these components typically used vendor-specific protocols for communications. The software and interfaces between different RAN components were customized for optimal performance of the proprietary hardware.

Disaggregated base station architecture with open interoperable interfaces, introduced as Open RAN, enables vendor diversity by ensuring interoperability of various network components. Decoupling the hardware and software, virtualizing the hardware, and hosting the software that manages and orchestrates networks in the cloud results in supply chain diversity and solution flexibility leading to increased competition and innovation. These Open vRAN solutions address capacity, scalability, flexibility, and coverage issues, while supporting open and interoperable network interfaces. Using the innovative concept of RAN intelligent controller (RIC), AI/ML enabled solutions on near-Real Time RIC and non-Real Time RIC platforms will optimize network energy savings, and various user-level and cell-level performance metrics. 80% of traffic is carried on 20% of the sites[1,2]. Traffic moves through the network and with AI and RIC, some of the radios and interstitial cell sites could be turned off when there is little traffic on them. This is accomplished by moving users to other bands or cell sites through load balancing. In a study conducted in 2020/2021 with 31 operators[1], for base-stations without air conditioning, 67% of energy consumption is in the radio, and only 10% of the energy is consumed in the BBU. Mechanisms to conserve power on the radio, such as turning it off when not needed, using more efficient power amplifiers, and using massive MIMO antennas will bring the most benefit on energy savings.

Mobile network operators have been eager to find solutions to reduce the energy consumption of their networks either by using the latest radio access technology or by optimizing the use of active and passive components in their network. While wide-scale deployment of 5G systems has led to increased energy consumption due to densification of the network and the use of additional radios in new frequency bands, 5G technology has been more energy efficient than its predecessors by offering improved spectrum efficiency achieved through massive MIMO and low-overhead radio protocol design.

Massive MIMO, as one of the key enablers of 5G new radio, has the potential to provide significant improvements in spectral efficiency and energy efficiency. Massive MIMO can increase network coverage, capacity, and user throughput using multi-antenna and multi-user MIMO techniques such as beamforming, beam steering, and spatial multiplexing. The power consumption of a 5G base station using massive MIMO is dominated by the power consumption of the radio units whose power amplifier(s) consume most of the energy, thus determining the energy efficiency of the radio unit. Higher power amplifier efficiency will contribute to improving the overall energy efficiency. Application of the power amplifier in the linear regions of the power amplifier result in higher efficiency.

Decoupling of software from hardware in the mobile core network with open interfaces has been happening for several years. The 5G service-based architecture is a cloud-native concept that works in principle like a cloud service running in a data center. Hence, data centers are becoming more important as cloud services are growing. It is anticipated that this trend will continue, as 5G will address more vertical industries and enterprise businesses. Virtualization and cloudification of 5G RAN has been gaining a lot of interest recently. The virtualization of RAN means that the baseband functions such as L1, L2, L3 and transport processing are processed as software by general purpose processors such as Intel x86 processors using a Commercial Off-The-Shelf (COTS) computing platform.

This white paper examines power consumption and energy efficiency in cloud-native Open RAN systems. The focus is on the implementation aspects of cloud-native virtualized Open RAN systems powered by Intel x86-based commercial servers and containerized and highly optimized RAN software developed by Mavenir and Intel. Strategies and technologies that are incorporated or are under development to optimize the overall system power consumption and to ensure energy efficiency across various network components and the network are described.

1. General Aspects of Network Energy Savings

Energy consumption of a mobile network is one of the key concerns of a network operator because it leads to an increase in the OPEX with an adverse impact on green-house gas emissions and the environment. Mobile network operators have always been eager to find optimal solutions to reduce the energy consumption of their networks either by using the latest radio access technology or by optimizing the use of the active and passive components in their network. While wide-scale deployment of 5G systems has led to increased energy consumption due to densification of the network and the use of additional radios in new frequency bands, 5G technology has been more energy efficient than its predecessors. 5G offers improved spectrum efficiency which is achieved through use of new technologies such as massive MIMO and low-overhead radio protocol design. To support the growing use of cellular connectivity in the 5G era, while reducing energy consumption and CO_2 emissions on a per-bit basis, the mobile networks need to become more energy efficient[1].

A study was conducted in 2020/2021 with 31 global operators[1]. The largest percentage of power consumed in a site is related to the main equipment consisting of the radio unit, baseband, and main control elements, as shown in Figure 1[1]. This accounts for approximately 50% of the total site power consumption of which 40% is consumed by the radio unit (RRH) noted as radio processing. The second major consumer is the air conditioning, accounting for 40% of total power, particularly in the United States. In several countries there is no air conditioning used on sites, and the baseband units are air cooled. The radio unit would therefore represent a higher percentage of the overall power consumption in a base station. In the radio unit, the power amplifier of the radio unit consumes most of the power (59%). These energy consumption percentages may vary depending on the RAN equipment power efficiency, the technology of air conditioning units, and the location of the base station etc.

Breakdown of power consumption of network, base station, and radio unit[1]

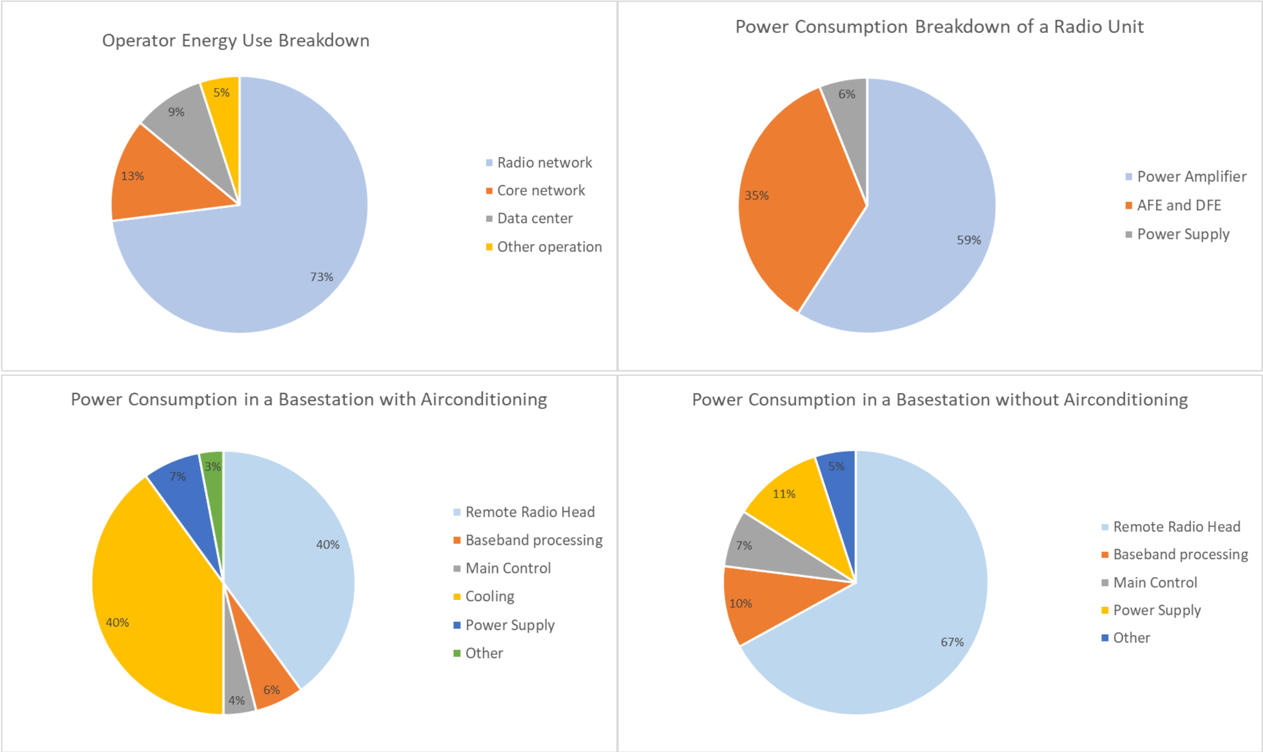


Figure 1

Open RAN is gaining momentum as an alternative to the traditional RAN. Open RAN enables operators to use equipment from different vendors, while ensuring interoperability between them. With its virtualization capabilities, an Open vRAN system allows the servers to be shut down during off-peak times or be dedicated to other applications to reduce idle mode power consumption. According to a survey conducted by the NGMN Alliance (see Figure 2), 80% of cell sites typically carry 20% of total traffic only and the busiest 10% of cell sites carry 50% of traffic[1,2]. Therefore, sharing the RAN resources among cell sites can reduce and make efficient use of computing resources with significant power savings. Moreover, there is an increased competition within Open RAN among radio manufacturers to design products with lower complexity and lower power consumption.

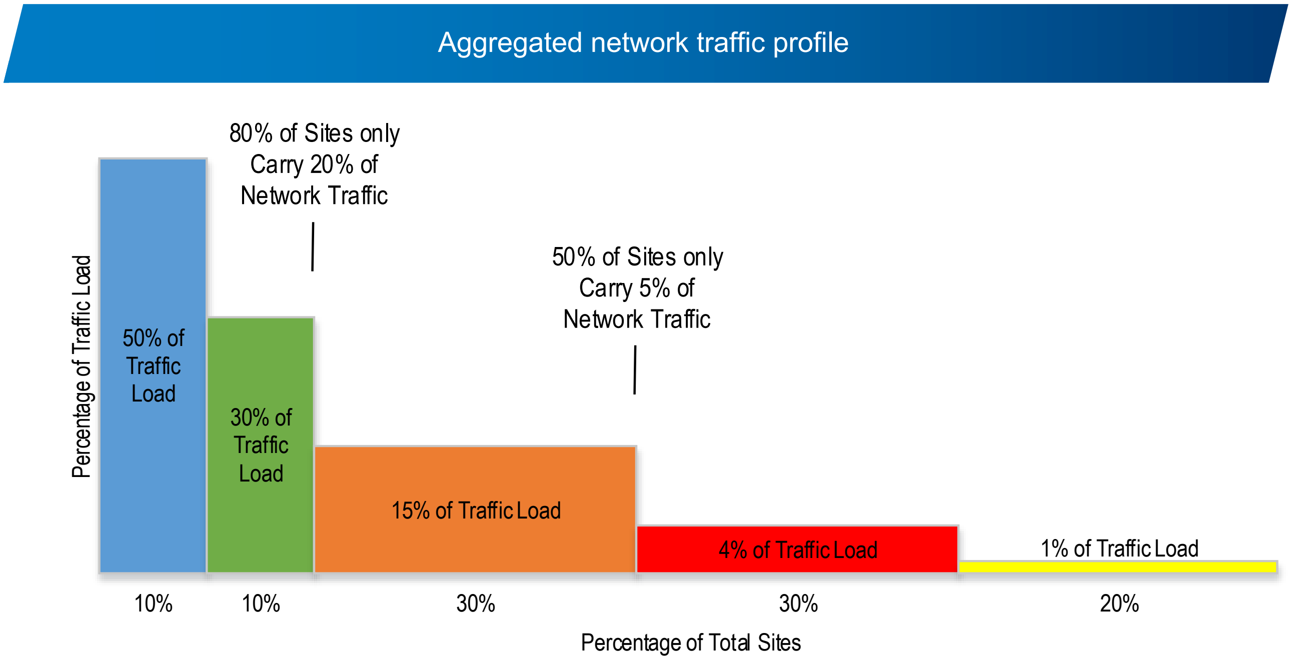


Figure 2

Traditional base stations handled baseband processing locally, thereby, collectively consuming more power. The radio unit is by far the most significant contributor to the total RAN power consumption power in a base station. Improvement in energy efficiency is possible using modularization within the system to allow dynamic shutdown of functionalities that are not in use or not required to remain active or on standby for the purpose of synchronization and signaling. This will result in a high degree of proportionality between instantaneous energy consumption and network load.

Despite improvements in energy efficiency, the RAN continues to consume more power than any other part of the network. New technologies such as massive MIMO and wider operation bands and bandwidths allow us to communicate more data using less power (bits/J). The exponential growth of user traffic means that the information and communication technology industry is expected to consume more power in the future, despite the recent advances in energy efficiency. Network upgrade is an important step toward minimizing power consumption. Some estimates show that upgrading from 4G to 5G RAN hardware could save as much as 90% of the energy needed per bit[2].

Energy consumption in mobile networks is dominated by the radio access portion of the network[2]. Unlike its predecessor, 5G NR uses a low-overhead (lean) design that minimizes always-on transmissions to enhance network energy efficiency. There are several factors that make a 5G RAN more energy efficient than LTE. These include:

> In contrast to LTE, the reference signals in NR are transmitted only when necessary.

> For NR, the timing of the SS Block can be set by the network operator. Typical value of SSB transmission is 20ms but it can be set between 5 and 160ms (5, 10, 20, 40, 80 and 160). During this time set by the operator, a number of SS Blocks will be transmitted in different directions during a 5ms Period. The 20ms SS-block periodicity is four times longer than the corresponding 5ms periodicity of LTE PSS/SSS transmission. The longer SS-block period was selected to allow for enhanced NR network energy performance and in general to follow the ultra-lean design paradigm.

> Carrier aggregation and massive MIMO in radio units, with wider carriers and more antennas, will lead to an increase in the compute/processing requirements, making energy-efficient processor design more critical.

> The use of resource pooling for 5G baseband processing can lead to both cost and power efficient use of resources.

Figure 3 shows another example comparing the distribution of power consumption across various components of 4G and 5G mobile networks[2]. It shows that RAN consumes most of the power in an operator's network. Although RAN power consumption is reduced in 5G, it is still over 50% of the total power consumption. Another trend is the rise in data center power consumption in 5G[16]. With many of the core network services moving to the cloud in 5G, a reduction in the energy consumption of core network elements from 4G to 5G and an increase in data center energy consumption as the core services move to data centers in 5G is observed.

Breakdown of power consumption in a 4G/5G mobile networks

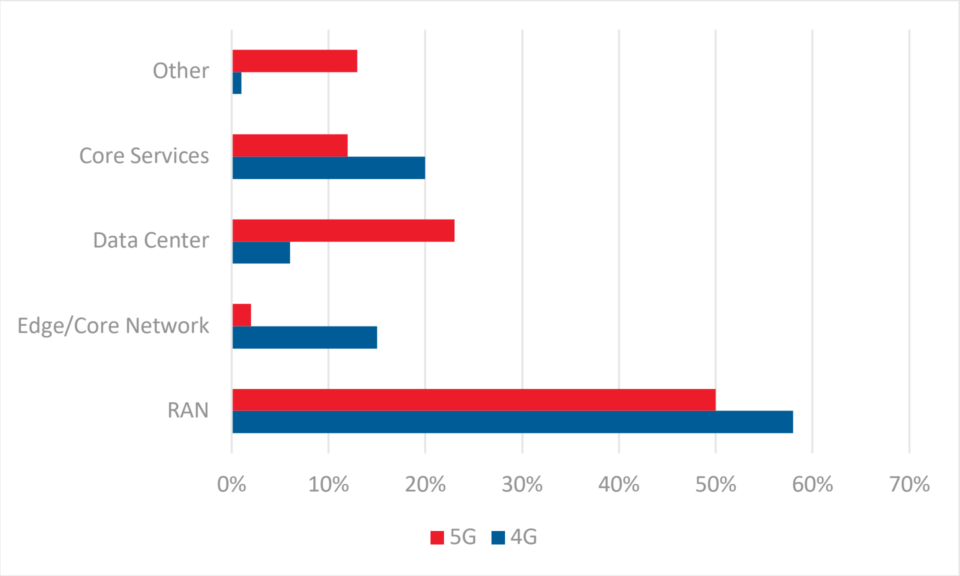


Figure 3

2. Open vRAN Architecture

Open vRAN deployments are built with disaggregated, virtualized and software-based components, connected through open and standardized interfaces, and interoperable with different vendors. Disaggregation and virtualization, based on cloud-native principles, enable flexible deployments. This increases the resiliency and reconfigurability of the RAN. Open and standardized interfaces also allow operators to onboard different equipment vendors, which opens up the RAN ecosystem to smaller players. Open interfaces and software-defined protocol stacks enable the integration of intelligent, data-driven closed-loop control for the RAN. The O-RAN specifications implement these principles on top of 3GPP LTE and NR RANs. ORAN supports the 3GPP NR 7.2 split for base stations. The 7.2 split disaggregates base station functionalities into a Central Unit (CU), a Distributed Unit (DU), and a Radio Unit (RU). It connects them to intelligent controllers through open interfaces that can consume appropriate data from the RAN and deploy control actions and policies to the RAN.

The O-RAN architecture supports the RAN intelligent Controllers (RICs) that perform management and control of the network at near-real-time (near-RT, response time to control loop actions is >10ms and <1sec) and non-real-time (non-RT, response time to control loop actions is >1sec) time scales. These entities house xAPPs and rAPPs respectively that permit closed loop optimization based on Operator driven Intent, and data collected from the network.

Open RAN architecture with RAN Intelligent Controller (RIC) platforms

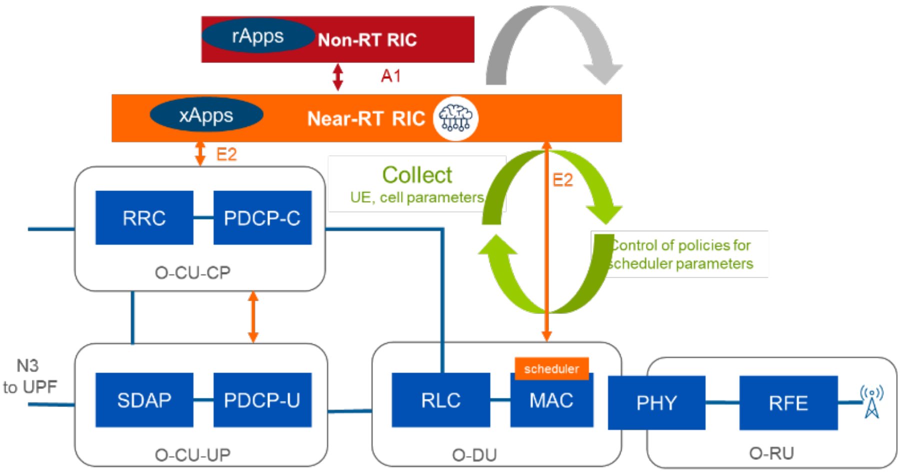


Figure 4

Network operators are increasingly looking into building intelligence into their networks to maximize their return on investment and minimize operational costs. This is also the goal of the O-RAN Alliance and the RAN intelligent controller entity. The non-RT and near-RT RIC allow combining platform telemetry through O2 interface with application telemetry through the O1 and E2 interfaces, shown in Figure A1 in Appendix A.

The xAPPs and rAPPs executing in the near-RT RIC and non-RT RIC, respectively, leverage AI/ML techniques to implement various use cases. These include maximizing spectral efficiency and use of deployed radio resources with active antenna management and orchestration, enhancing real-time response, dynamic adjustment of resource allocation as network conditions change, closed-loop automation for best configuration settings or configuration to meet the given workload requirement with minimal resources, faster new user provisioning, predictive analytics, network anomaly detection and intervention at scale, security anomalies and intrusion detection and automated application of corrections, and automated cell deployment. Application telemetry for many of these use cases include dynamic resource and network performance information from the RAN stack, including L1. The KPIs and resource information may be at the antenna level, cell-level, or user-level.

The application of near-RT RIC and non-RT RIC to optimize energy and exploit energy savings opportunities are addressed in the forthcoming sections.

Appendix A describes the details of the O-RAN architecture. Appendix B describes the details of virtualization and containerization of RAN software. Appendix C describes energy efficiency considerations in Cloud Computing.

3. RAN Power Consumption and Power Efficient Solutions

A radio access network comprises of several cell sites with site infrastructure equipment and base station equipment[1]. The site infrastructure has power supplies and cooling systems equipment. The RAN equipment is the base station comprising of a baseband unit and one or more radio units. From the base station, data is transmitted over the air interface to the user devices distributed across the cell. Figure 5 depicts the energy usage at a cell site in different stages from the main AC power supply to the radio unit, where the energy efficiency at each stage is defined as follows[1]:

- The power consumption in the site infrastructure is measured from the AC main supply to the DC power supply for the base station. The energy efficiency of the site infrastructure can be measured by dividing the DC input power of the base station by the AC input power of the site. This measure is the inverse of the power usage effectiveness, which is commonly used for data centers.
- The power consumption in the base station is measured from the DC power input to the cabinet-top power output of the base station antenna. The power efficiency of a base station can be measured by dividing the cabinet-top power by the DC input power of the base station.
- Air interface is the link from the output of the antenna on the top of the cabinet to the radio receiver of the user device. The energy efficiency of the air interface can be measured by dividing the service provided by the base station (e.g., number of bits delivered to the user, coverage, or the number of users served by the base station) by the output power at the top of the cabinet.

Typical RAN site, and energy flow from main AC input to reception at the user terminal[1]

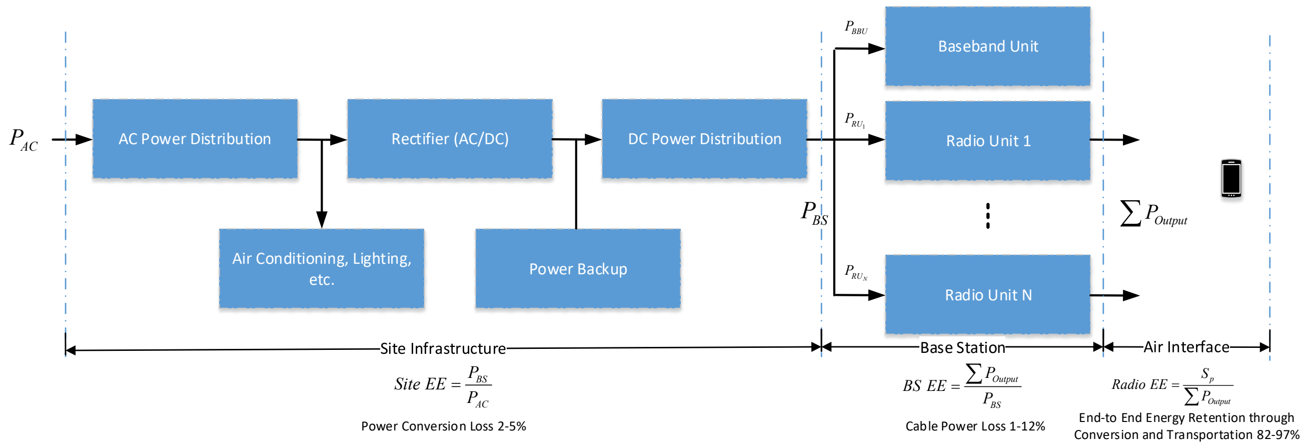

P_{AC}
AC Power Distribution
Rectifier (AC/DC)
DC Power Distribution
Air Conditioning, Lighting, etc.
Power Backup
P_{BS}
P_{BBU}
Baseband Unit
P_{RU_1}
Radio Unit 1
P_{RU_N}
Radio Unit N
$\sum P_{Output}$
Site Infrastructure
Base Station
Air Interface
$Site\ EE = \frac{P_{BS}}{P_{AC}}$
Power Conversion Loss 2-5%
$BS\ EE = \frac{\sum P_{Output}}{P_{BS}}$
Cable Power Loss 1-12%
$Radio\ EE = \frac{S_p}{\sum P_{Output}}$
End-to End Energy Retention through Conversion and Transportation 82-97%

Figure 5

The overall energy efficiency is defined by these three factors: power efficiency of the site infrastructure, power efficiency of the base station equipment, and energy efficiency of the air interface. The overall energy efficiency is obtained as the product of the afore-mentioned energy efficiencies. When summed up over a period of time, the energy performance is obtained, as measured in bits/Joule.

3.1 Current Standards Work on going for Energy efficiency

Both 3GPP and O-RAN are actively working on standardizing energy efficiency interfaces and designs. This section lists the major on-going work. ETSI specification “ETSI ES 203 228 V1.3.1 (2020-10) Environmental Engineering (EE); Assessment of mobile network energy efficiency”[17] provides a view on metrics for energy efficiency assessment.

3.1.1 TR38.864 Network energy savings

TR 38.864 studies various techniques to improve energy savings for NR:

- Time domain techniques – including UE wake up signal (WUS) for gNB
- Frequency domain techniques – including multi-carrier energy savings enhancements
- Spatial domain techniques – including adaptation of spatial elements
- Power domain techniques – including adaptation of transmission power of signals and channels
- High layer aspect techniques – including cell selection/reselection

3.1.2 TR37.817 enhancement for Data Collection for NR and EN-DC

TR 37.817 describes AI enabled RAN intelligence mechanisms to support network energy savings, load balancing, and mobility optimization use cases:

- It describes AI functionalities (i.e., data collection, model training, model inference), where model training function may reside in OAM or RAN nodes, and model inference function resides in RAN nodes, as shown in Figure 6.

AI enabled RAN intelligence to support network energy savings, load balancing, and mobility optimization use cases

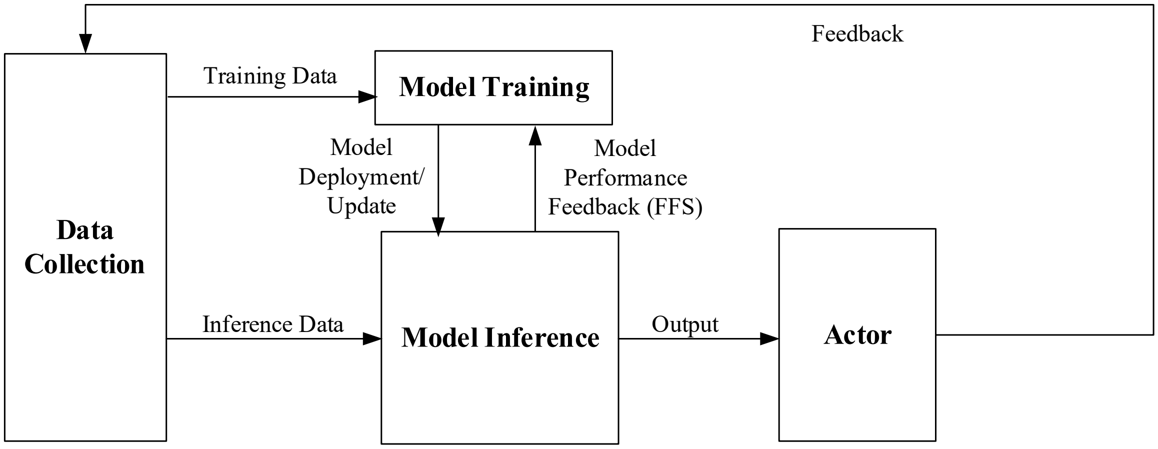


Figure 6: Source: 3GPP TS 38.817

3.1.3 TS 28.310 Energy efficiency of 5G

- TS 28.310 specifies concepts, use cases, requirements and solutions for the energy efficiency assessment and optimization for energy savings of 5G networks:
    - Transition from peak to off-peak hours: switch off NR capacity cell, while using gNB or eNB cells to provide coverages.
    - Transition from off-peak to peak hours: switch on NR capacity cell.

Transitioning mechanisms specified in TS 28.310

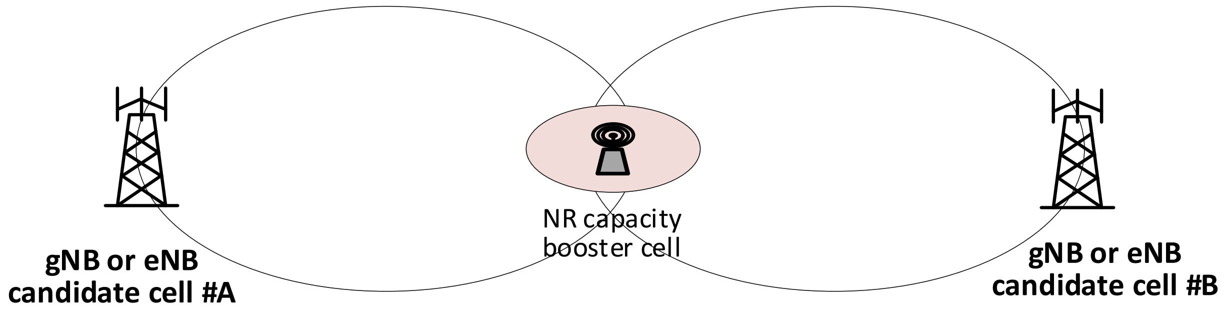


Figure 7: Source: 3GPP TS28.310

3.1.4 O-RAN WG1 Network Energy Savings Use Cases

The following use cases have been documented in O-RAN Working Group 1 - the detailed technical reports and standardizations are in WG4 and WG6.

- Carrier and cell switch off/on
- RF channel reconfiguration
- Advanced sleep mode
- O-cloud resource energy saving mode

3.2 Energy Efficiency KPIs

Energy efficiency is an end-to-end system requirement involving all components of the wireless network. The overall objective is for Open RAN products to be more energy efficient than traditional RAN while permitting cloudification, virtualization and disaggregation of the network. While RUs represent the largest fraction of the overall power consumption in a RAN, energy usage of the hardware infrastructure supporting other functions such as DU and CU also need to be considered.

It is recommended that RUs be designed with energy efficient power amplifiers and transceivers and have the ability to switch off most of the transmitter elements and digital frontend when the network is under-loaded. Commercial-off-the-shelf servers using general purpose processors to support CU/DU functions should be designed with power efficient hardware, including low-voltage modules. To ensure optimal energy efficiency, each HW element of the network is examined and appropriate KPIs are defined[6].

The energy efficiency KPI is commonly defined as the data volume (in kbits) divided by energy consumption (in kWh) of the considered network elements. The unit of this KPI is bits/Joule[6]. These are calculated differently for integrated and aggregated CU and DU (also referred to as a gNB) and disaggregated CU's and DU's..

Two metrics (data energy efficiency, coverage energy efficiency) are used to effectively benchmark the deployment and operation performance of a wireless network in terms of energy efficiency. Appendix D details these metrics and the Energy Efficiency KPIs for non-disaggregated and disaggregated CU and DU's.

4. 5G Advanced RAN Power Consumption Optimization Techniques

5G systems by design are expected to improve energy efficiency. More efficient power amplifiers have been developed and used in 5G radios, renewable energy sources for powering on-grid and off-grid sites, including solar power, are starting to be widely adopted. Moreover, new generation of batteries are becoming an integral part of any 5G site to enhance energy management, and liquid cooling is being implemented to reduce the need for air conditioning. 3GPP NR specifications have enabled several new techniques that can help to improve energy efficiency at network level which include the following:

- Lean carrier design
- Improved sleep modes
- BWP and RF carrier management
- Massive MIMO and beamforming-based operation
- Artificial Intelligence and Machine Learning based RAN and core optimizations

The following sections describe the energy savings features that were specified as part of 5G NR and analyze the impact of each feature on the overall energy efficiency of the radio access network. All these features are currently being studied in O-RAN as part of the network energy saving initiative[5].

4.1 Symbol-Level Sleep Mode

Feature description

Micro-DTX (discontinuous transmission) aims to save energy by disabling the power amplifiers when there are no data to transmit. This feature sometimes is referred to as symbol muting.

5G NR has an ultra-lean design, which minimizes always-on transmissions to enhance network energy efficiency and reduce interference. 4G has only one type of reference signal design - a 'one size fits all' downlink reference signal design called CRS (Cell-Specific RS). In contrast to the setup in LTE, the reference signals in NR are transmitted only when necessary. NR downlink reference signals are used for specific roles and can be flexibly adapted for different deployment scenarios. Hence 5G is more efficient than 4G.

In principle, the feature can work even when a single blank symbol is available. However, considering the time for the PA and RF circuitry to transition between ON and OFF states and to stabilize RF output following transition, very short transmission blanking durations may not be practical. Thus, the best energy saving results can be achieved, if the transmission is not disabled and enabled for consecutive symbols.

The DU scheduler can help achieve the best results, if it schedules the data to be transmitted in bursts, i.e., postpone transmission of non-critical and non-delay-sensitive symbols to later slots to prolong the transmission blanking periods. It is also advised to spread the data in the frequency domain, rather than in the time domain. These schemes are known as low-energy scheduler solutions, which need to be incorporated in DUs. Figure 8 shows a representation of symbol level power shutdown.

Symbol level power shutdown

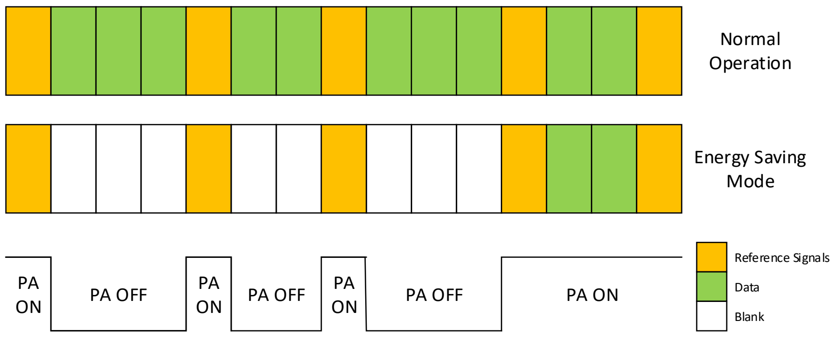


Figure 8

In symbol-level transmission blanking, when a DU identifies some downlink symbols with no data, the DU turns off the PA(s) and some analog components, thereby reducing the power consumption of the base station with no impact on the user experience. By adjusting the number of SSB beams when cells have no traffic or light traffic in a specified period of time, one can increase the proportion of symbols for which symbol-level power saving can be used. During automatic beam adjustment, base station needs to adjust the transmit power of common channels to ensure coverage is not adversely impacted. When the load is relatively low, up to 30% reduction of power consumption can be achieved through symbol-level shutdown.

Energy savings impact of various DTX schemes[18]

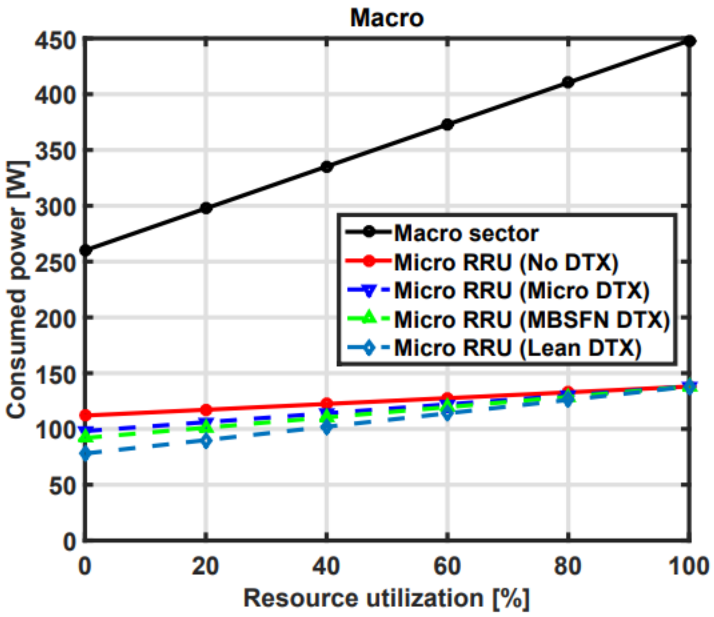


Figure 9: The energy savings impact of various DTX schemes on average cell power consumption as a function of traffic load[18]

The RU control circuitry can prepare in advance to disable more components of the transmit chain than just the PA, therefore saving relatively more power.

Another feature that has some similarity with micro-DTX is the PA bias adjustment, where the PA operating point is changed by changing the PA bias to a setting that is more suitable for the conditions (e.g., reduction of required RF power). To save power, micro-DTX typically switches the PA bias voltage from the nominal operational voltage of -50V to a lower voltage idle setting (e.g., negative 5-8V).

The PA bias adjustment is similar to the micro-DTX feature by way of setting of the normal operation voltage to the value other than default, which is then used by the micro-DTX to toggle between the PA normal and idle settings. Figure 9 shows the results of a study on the effect of various DTX schemes on average cell power as a function of cell spectral efficiency.

Analysis of impact of Open RAN with RIC versus proprietary RAN

These aspects equally apply to Open RAN with RIC and proprietary RAN architectures.

4.2 Advanced Sleep Modes

Feature description

Power consumption of a radio base station varies with cell load. As the cell load increases, the radio power amplifiers consume the largest fraction of the total power consumption. However, in low traffic conditions, the RU power consumption is mainly attributed to the digital frontend modules. In no traffic conditions and in between control/signaling transmission, the gNB/eNB components consume energy even though there is no need for transmission. Thus, the opportunity arises to reduce unnecessary radio unit energy consumption by proactively deactivating components when they remain unused/under-utilized for transmission.

Shutting down components of a radio unit, and discontinuous transmission was used in LTE as an energy saving scheme on the base station side. The power amplifier is shut down at symbol level for at least the duration of a symbol and for multiple repetitions as long as there is no control or data to transmit to the users. In LTE, transmission of the cell-specific reference signals irrespective of traffic load contributes to a considerable system overhead and limits the ability to use DTX. These signals are continuously and periodically transmitted over the entire cell coverage.

The beamforming and potential multiplexing gains in 5G NR allow the use of reduced transmit power to reach a given distance and/or to meet a given user QoS requirement, with the potential benefits of interference mitigation and energy savings. They also allow for larger data rates given a fixed transmit power, which enables a larger BS deactivation time, and thus further energy savings. 5G NR was designed based on a lean design principle, in which control signals are not continuously transmitted in every radio frame, but transmitted on demand and more sparsely, based on traffic requirements. This lean design allows more efficient operation of the mMIMO, and facilitates gNB (de)activation, including sophisticated shutdown mechanisms at symbol, carrier, or channel (antenna) level, which can significantly reduce energy consumption.

In a 5G system, there are more opportunities to activate energy savings features such as shutting down gNB hardware components. An NR carrier allows the activation of energy savings features for multiple granularities of time in between the sparse control signaling and when cell load is low. In the case of shutting down gNB components, the longer this cell inactivity, the more gNB components can be shut down across time, depending on the time needed for their wake up and stabilization. The adaptive deactivation of gNB components for multiple granularities of time will enable sleep mode of different levels.

Cells may enter multiple sleep states/modes depending on the radio hardware (antenna, power amplifier, etc.) deactivation tradeoff between power consumption and wake-up time specific to each sleeping mode where the deeper the sleep state, the less power consumption, and a longer wake-up time, which impacts user experience in terms of user throughput and packet delay. Both 4G LTE and 5G NR systems support this mechanism, although options and parameters differ.

5G NR allows for configurable signaling periodicities, which enables more effective sleep modes ranging from deactivation of some components of the base station for several micro-seconds to switching off almost all of them for one second or more. The definition of the sleep modes currently under consideration in O-RAN for the purpose of radio energy savings is provided in Table 1. The actual power saving values by the base station invoking the various sleep states/modes will depend based on network scenarios and traffic models.

Table 1: Definitions of advanced sleep modes

Sleep Mode	Time Duration (millisecond)	Description (Use Case)
Micro Sleep Mode (SM1)	[OFDM symbol duration to 1ms]	The gNB does not need to operate TX/RX within the next few OFDM symbols ($T_{\text{OFDM-Symbol}} < T \leq 1\text{ms}$). The RF transceivers can be turned off for the prescribed period. Note: No need for additional specification as existing specifications can cover implementation of the sleep mode automatically when there is no transmission.
Light Sleep Mode (SM2)	~ [5 - 10]	The gNB does not need to operate TX/RX within the next 5ms≤T≤10ms. The RF transceivers and additional hardware components can be turned off for the prescribed period. The power consumption is typically lower compared to that of micro sleep mode. This sleep mode can be implemented via transmission blanking when there is no SSB in downlink or PRACH or other uplink signals to be transmitted/received.
Deep Sleep Mode (SM3)	~ [50 - 100]	The gNB does not need to operate TX/RX within the next 50ms≤T≤100ms. Additional hardware components can be turned off for the prescribed period. Note that some hardware components such as timing circuitry must be kept at standby or on, allowing fast transition to Active State. The power consumption in this mode is typically lower compared to that of light sleep mode. This sleep mode requires coordination with other neighboring cells since user experience would be affected if cell is turned off for over 50ms.
Hibernate Sleep Mode (SM4)	~ [1000]	The gNB does not need to operate TX/RX within the next T ms (T ~ [1000] ms). Most hardware components can be turned off except some hardware components such as timing circuitry which must be kept at standby or on, allowing fast transition to Active State. The power consumption in this mode is typically much lower compared to that of deep sleep mode. Sleep durations greater than 1000ms are not precluded.

Analysis of impact of Open RAN with RIC versus proprietary RAN

RIC based intelligence can be used to optimize the energy consumption across multiple CU/DU and RU modules, (versus single proprietary RAN node) by maximizing sleep cycles and by intelligent scheduling which minimizes the number of symbols/slots to be scheduled.

Conclusion of above comparison with respect to Open RAN impact

Open RAN with RIC will support energy efficiency by identifying opportunities for cores to move to sleep states (C1, C6 more and more), as described in Section 7.3.1, based on telemetry data obtained from the RAN via the E2 interface. This can significantly improve the energy efficiency of Open RAN systems.

4.3 BW/BWP Adaptation and Carrier Management (RF Carrier Activation/Deactivation)

Feature description

A key performance requirement of 5G is to support downlink and uplink peak data rates in the excess of 20 and 10Gbits/s, respectively. To achieve this requirement, wide system bandwidths are being supported by the new radio. The UEs are required to support channel bandwidth up to 100MHz for sub-6GHz bands and 400MHz for above 6GHz bands, which is much wider than the bandwidth of 20MHz in LTE. Since UE capabilities vary, it is often limited to the bandwidths less than the maximum bandwidth supported in the specification. Moreover, high UE power consumption can be a major issue if UE is required to perform transmission or reception in a wide bandwidth all the time regardless of how much the actual traffic load is.

To reduce UE power consumption and to guarantee the data transmission rate, the concept of bandwidth part (BWP) was adopted by 3GPP. A BWP consists of several continuous physical resource blocks (PRB) with specific numerology. For each serving cell, there are at most three BWPs that can be configured for downlink or uplink namely initial BWP, first active BWP, and default BWP. When a large data packet needs to be transmitted, the UE can be instructed to activate a BWP with a wider bandwidth. Otherwise, the UE can be informed to switch to a BWP with a narrower bandwidth to save power (see Figure 10). BWP switching framework in NR is beneficial for developing an energy efficient network.

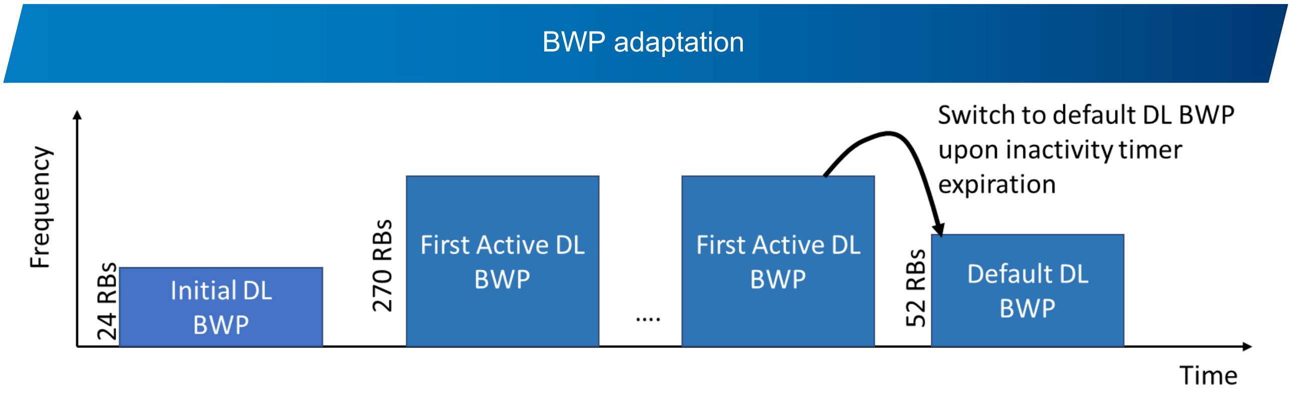
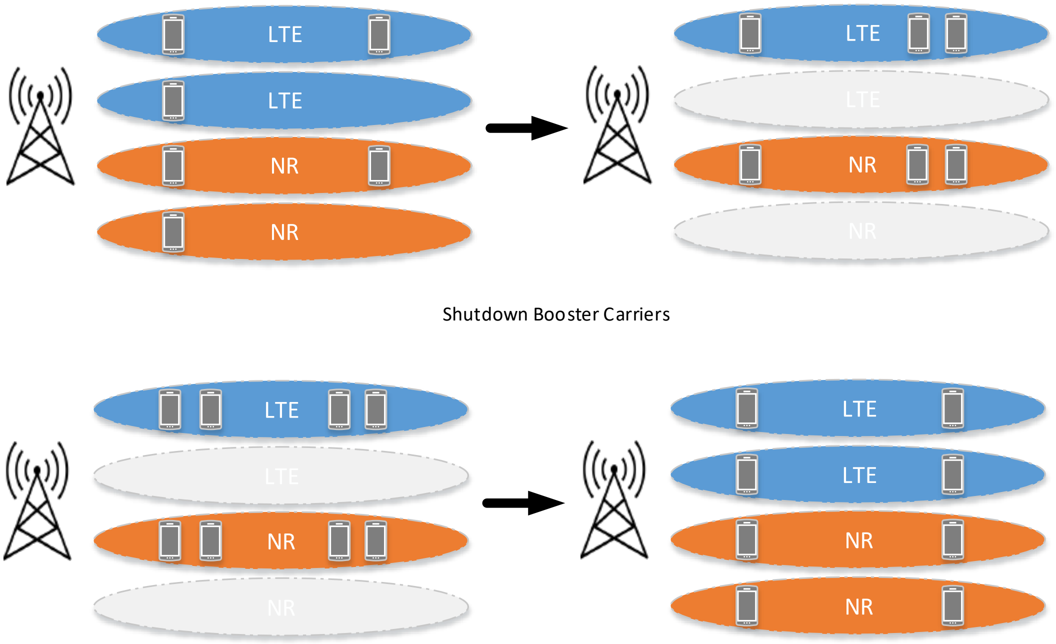


Figure 10

When the load of the entire base station is low during off-peak hours, the base station energy consumption can be reduced by retaining only the coverage-layer cells and shutting down the capacity-layer cells, as indicated in Figure 11.

RF carrier shutdown



Shutdown Booster Carriers

Recovery Booster Carriers

Figure 11

To improve system-level energy efficiency with a multi-carrier system, the number of active RF carriers must be dynamically managed. When the load of the entire base station is low during off-peak hours, the BS energy consumption can be reduced by moving UEs from lightly loaded cells onto other cells with similar coverage. These cells can absorb the slight increase in traffic without impacting user experience, and the offloaded cells can be shut down as indicated in Figure 11. Likewise, as congestion in a cell grows due to sudden surge in traffic, new cells can be powered on, and traffic can be load-balanced across all the cells to help improve user experience across all the cells.

Another approach to inter-BS energy savings is cell zooming, which is not based on carrier shutdown. It adapts the transmission power by reducing the cell coverage of lightly loaded cells, while simultaneously increasing the area covered by neighboring cells (see Figure 12). When using this mechanism, the network topology changes should be carefully handled to avoid service outage. A data-driven approach may be used to optimize the cell zooming mechanism.

Cell zooming procedure

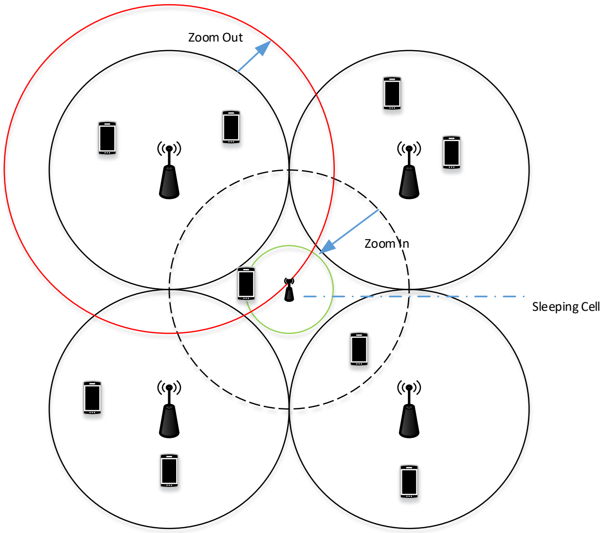


Figure 12

Analysis of impact of Open RAN with RIC versus non-Open RAN

The RF carrier shutdown feature (typically hosted by the SMO and non-RT RIC in O-RAN architecture) periodically checks the service load of multiple carriers and if the service load is below a specified threshold, the capacity-layers are shut off (see Figure 12). The UEs served by those carriers can camp on or access the services from the carrier providing basic coverage. When the load of the carrier providing basic coverage is higher than a specified threshold, the base station turns on the carriers that have been shut down for service provisioning. When shutting down a carrier, it is important to ensure that basic coverage is maintained.

RF carriers can be shut down by non-RT RIC rAPPs more intelligently using information from telemetry data from E2 interface with O-RAN. Additional information from this data may permit more opportunities for shutdown, and over more granular time scales with minimal impact to user experience.

Conclusion of above comparison with respect to Open RAN impact

O-RAN can provide more differentiation versus proprietary RAN through access to additional data the provides valuable insights about the RAN and the UE performance.

4.4 RF Transceiver and Antenna Module Switching

Feature description

RF channel (or transceiver) shutdown is another power saving scheme where a multi-transceiver base station with 64/32 channels mutes some of the RF channels with low traffic condition, thereby reducing the power consumption of the base station. Turning half of the channels off will cause a coverage loss of up to 6dB due to reduced antenna gain and transmitted power, directly causing a significant drop in the throughput of cell edge users.

Analysis of impact of Open RAN with RIC versus non-Open RAN

Coordinated with traffic forecast, via AI/ML prediction engine, certain mMIMO antenna elements (RF channels) can be deactivated when the PRB usage and the number of RRC connected users are below certain thresholds. The antenna array scaling would impact the azimuth and elevation radiation patterns and antenna sidelobes, adversely affecting the coverage of the base station. Therefore, deactivation of RF transceivers and their associated sub-arrays must be carefully done considering the potential impacts on the coverage. Figure 13 shows an example antenna array scaling from 64TRX to 4TRX based on the traffic load of the cell. After the channel is shut down, since there will be a degradation to the total transmit power and antenna gain, the power spectral density of the remaining channels should be increased to ensure that the cell coverage is not affected.

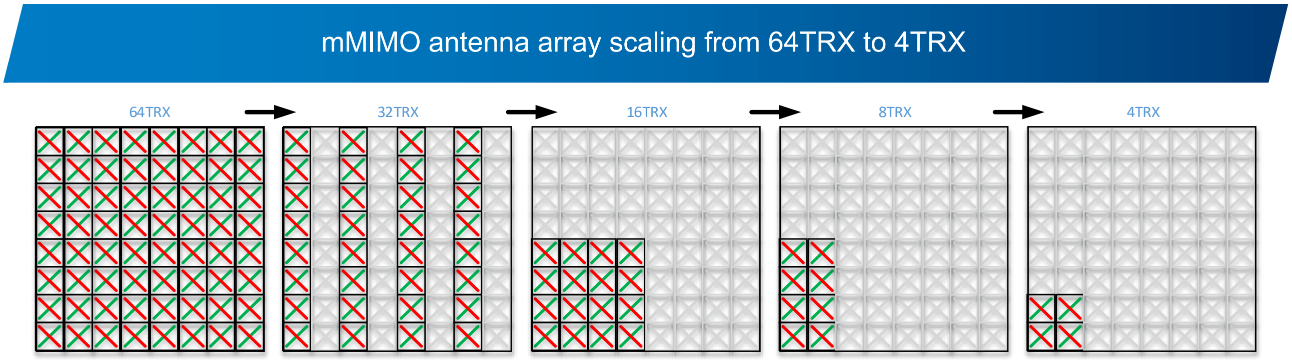


Figure 13

Conclusion of above comparison with respect to Open RAN impact

The AI/ML schemes can be hosted on the rAPP to facilitate energy savings. Open vRAN architecture permits the ability to collect the necessary data needed for intelligent RF channel shutdown.

4.5 Intelligent Scheduling Impact on Power Consumption

Feature description

To ensure that the radio unit power amplifiers can shut down in the low load conditions, the DU scheduler must take certain considerations into account when scheduling the active UEs in a cell. As shown in Figure 14, some scheduling decisions can result in short duration of blank TTIs (i.e., a transmission time interval with blank PRBs) or no opportunity for the RU PA(s) to shut down and conserve power. To prolong the PA shutdown duration, the UEs should be grouped and scheduled intelligently to allow more consecutive blank TTIs. Since every PA requires certain time to transit from active state to standby state (with minimal static power consumption) and vice versa, very short blank transmission intervals would practically result in no possibility for the RU PA(s) power shutdown and less effective energy savings strategy.

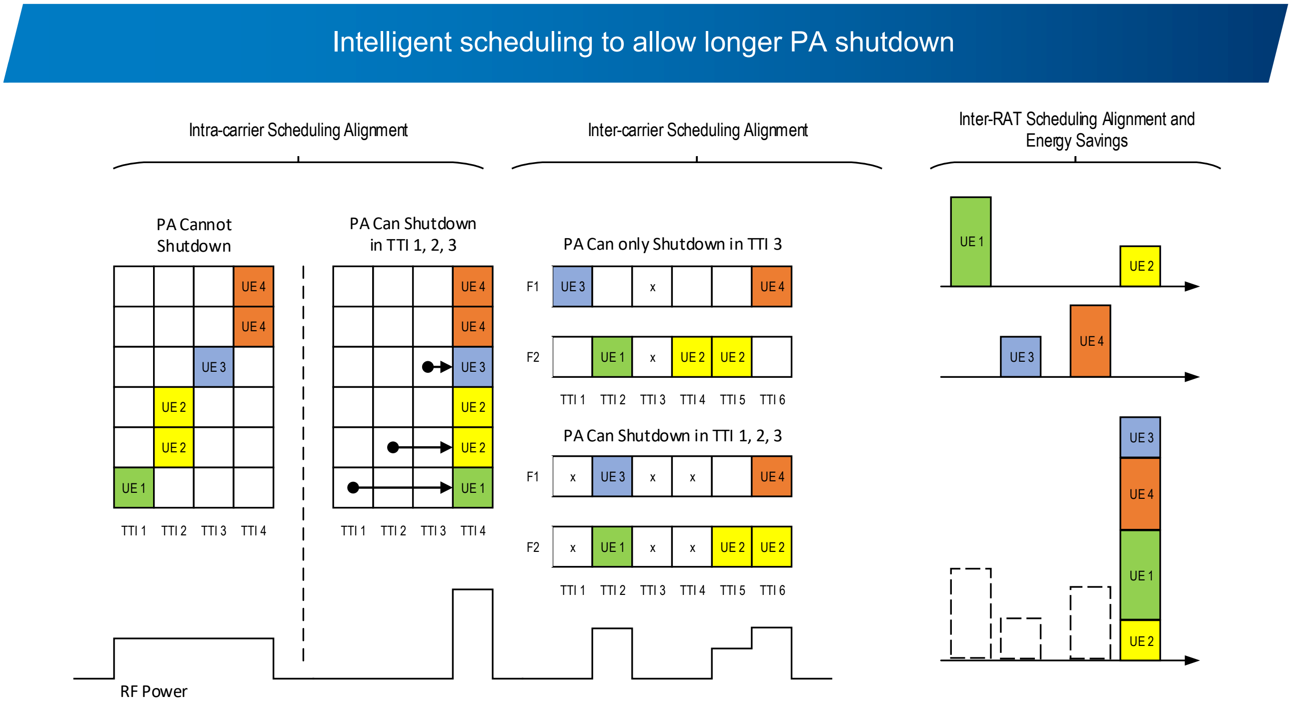

Intelligent scheduling to allow longer PA shutdown
Intra-carrier Scheduling Alignment
Inter-carrier Scheduling Alignment
Inter-RAT Scheduling Alignment and Energy Savings
PA Cannot Shutdown
PA Can Shutdown in TTI 1, 2, 3
PA Can only Shutdown in TTI 3
PA Can Shutdown in TTI 1, 2, 3
UE 1
UE 2
UE 3
UE 4
F1
F2
x
TTI 1
TTI 2
TTI 3
TTI 4
TTI 5
TTI 6
RF Power

Figure 14

Analysis of impact of Open RAN with RIC versus proprietary RAN

With Open vRAN, AI/ML based scheduling algorithms may be trained and adopted to optimize radio resource allocation and power saving in a DU based on instantaneous or statistical network conditions. Furthermore, near-RT RIC can provide energy savings guidance to DU based on an overall view of the radio network to optimize resource utilization and power consumption.

Conclusion of above comparison with respect to Open RAN impact

Open RAN xAPPs can further optimize energy efficiency on the DU compared to non-Open RAN solutions using AI/ML based scheduling algorithms that use RAN data from a single DU and from across a cluster of DUs and use that data to make predictions of different variables. This permits more opportunities for power savings.

4.6 Booster Cells Switching ON/OFF

Feature description

Switching off under-loaded cells during network operation without affecting the user experience (call drops, QoS degradation, etc.) is one way to achieve RAN energy efficiency. A typical energy savings scenario is realized when capacity booster cells are deployed under the umbrella of cells providing basic coverage and the capacity booster cells are switched off to enter dormant mode when its capacity is no longer needed and reactivated on a need basis. In a practical network deployment, the energy savings can be divided into centralized and distributed energy savings. For the distributed energy savings, the NR capacity booster cell may decide to switch off when it detects that its traffic load is below a certain threshold, and its coverage can be provided by the basic coverage providing cell, as indicated in Figure 15. The coverage providing cell decides to reactivate the NR capacity booster cell when it detects additional capacity is needed.

Analysis of impact of Open vRAN with RIC versus proprietary RAN

For centralized energy savings, a centralized entity, such as the orchestration and management entity, the non-RT RIC in O-RAN architecture, collects the traffic load performance measurements from the NR capacity booster cell and coverage providing cells, and may request an NR capacity booster cell to switch off when its traffic is below certain threshold.

gNB capacity booster cell(s) switching for energy savings

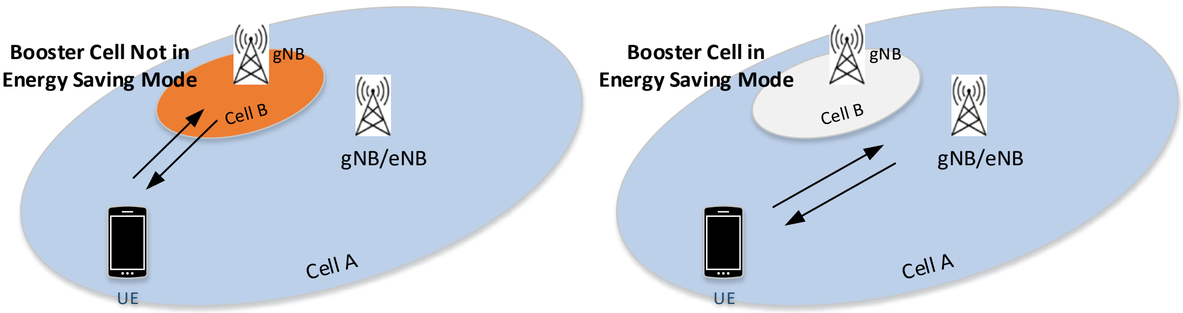


Figure 15

If a gNB is deployed with CU/DU disaggregated architecture, the F1 interface is enhanced to support the cell reactivation procedure and cell status exchange. When the booster gNB with CU/DU split decides to switch off cell(s) to the dormant state, the decision is typically made by the gNB-DU based on cell load information or by the OAM entity (non-RT RIC in O-RAN architecture). Before the cell in the gNB-DU enters into the dormant mode, the gNB-DU will send the gNB-DU configuration update message to the gNB-CU to indicate that the gNB-DU will switch off the cell subsequently sometime later.

During the switch-off period, the gNB-CU offloads the UE to a neighboring cell and simultaneously will not admit any incoming UE to this cell being switched-off. After the cell at gNB-DU enters into the dormant mode, the gNB-DU sends a new gNB-DU configuration update message to inform the inactive status of this cell to the gNB-CU. The gNB-CU needs to inform the updated cell status to the coverage provider node. When the gNB-CU receives the cell activation request/EN-DC cell activation request from a coverage provider node over the Xn or X2 interface, or the gNB-CU decides to activate the dormant cell by itself, it will trigger the gNB-CU configuration update message to the gNB-DU with a list of the cells to be activated.

Conclusion of above comparison with respect to Open RAN impact

Open RAN can further optimize energy efficiency on the DU compared to non-Open RAN solutions. Additional telemetry data for individual radios and clusters of cells is communicated to rAPPs, via the E2 interface. rAPPs would use that information and take into account the loads on each cell, and trigger load balancing between cells before switching-off power for various DUs.

4.7 Massive MIMO and Energy Efficiency

Feature description

Using extensive beamforming and spatial multiplexing capabilities, massive MIMO not only can significantly improve network capacity, but also reduce the transmit power required at the BS to achieve a target rate, given a frequency band of operation and coverage area due to increased directionality gain. However, running such larger number of antennas at the BS, together with the increased signal processing required to handle the larger capacity in a mMIMO cell, also increases the energy consumption of the BS.

Figure 16 shows the achievable energy efficiency for different values of K (the number of multiplexed UEs), when varying the number M of antennas based on a study described in[8]. These results confirm that, for a given value of K, the number of multiplexed UEs, increasing the number of antennas M increases the energy efficiency up to a maximum value, where the user throughput gains due to increasing number of BS antennas is not sufficient anymore to offset the cost incurred by its associated power consumption increase. Therefore, as shown in Figure 16, deploying hundreds of BS antennas to serve a large number of UEs is the optimal solution from an energy efficiency perspective, confirming the energy efficiency potentials of mMIMO.

Energy efficiency for a given number of antennas[8]

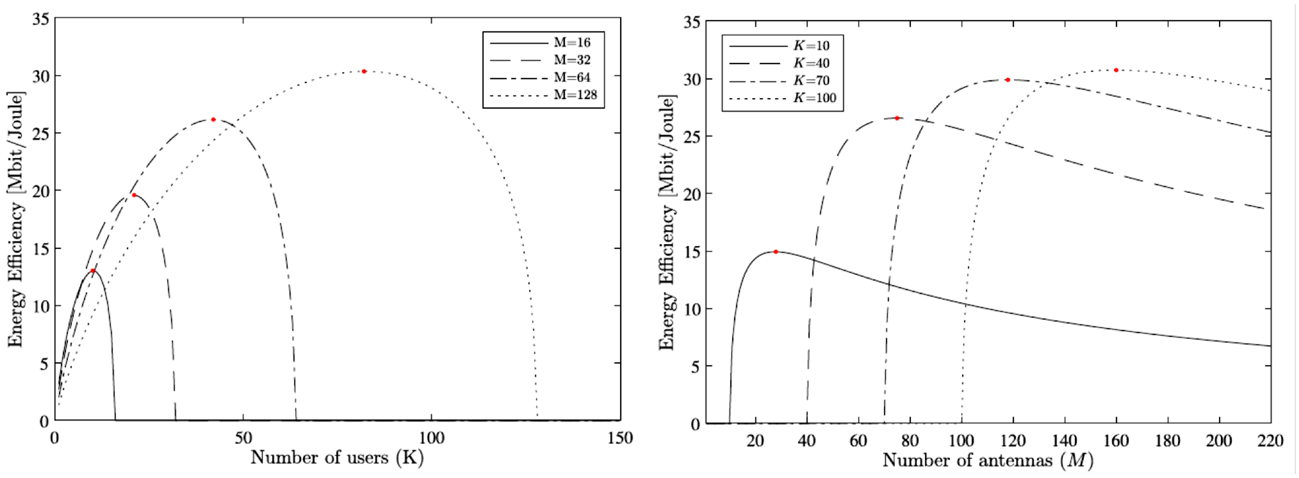


Figure 16: Energy efficiency with K multiplexed UEs for a given number of antennas M[8]

5G energy consumption depends on several factors, such as radio configuration, hardware, and traffic load, and RAN which is responsible for over 50% of the total power consumption in the system. Within the mMIMO systems, the main contributors to power consumption are power amplifiers (PAs), baseband processing modules, digital frontend modules, and RF transceivers, whose breakdown is shown in Table 2[10].

Table 2: mMIMO contributions to overall power consumption under various loading[10]

Component	100% Traffic	30% Traffic	0% Traffic
Power Amplifier	58%	36%	15%
Power Module	5%	5%	5%
RF Transceiver	16%	25%	34%
Baseband and DFE	21%	34%	46%

The PA requires a substantial amount of energy, representing almost 60% of a mMIMO radio power consumption with high traffic loads. A typical mMIMO 5G base station has 32 or 64 transceivers. For a user moving through a cell, the transmit power per mMIMO PA is much smaller. This permits the PA's to operate in the linear region of the PA and result in higher efficiency of massive MIMO systems.

Analysis of impact of Open RAN with RIC versus non-Open RAN

These principles apply to Open RAN with RIC and non-Open RAN solutions.

4.8 RRM and Load Balancing

Feature description

Radio resource management (RRM) is a collection of algorithms used to optimize radio resource and spectrum utilization and to improve cell capacity and throughput. Some of its functions are scheduling and optimization. With non-Open RAN solutions, the RAN control algorithms were previously not accessible to the operators and were proprietary to the RAN vendors, which caused a vendor lock-in scenario and reduced opportunities for innovation.

Analysis of impact of Open RAN with RIC versus proprietary RAN

In O-RAN architecture, the RIC will abstract the CU and DU functions from the rAPPs and xAPPs so that third-party developers can create innovative new solutions for RRM without worrying about the number or complexity of the multi-vendor network elements. Developers will have access to the network elements through open APIs. The abstraction of the xAPPs and rAPPs developed by different suppliers also benefits the RAN component suppliers. The operator community can partner with developers and make their networks' open APIs available to them to speed up development time and ensure diversity in the supply chain and differentiating applications from the developers.

Conclusion of above comparison with respect to Open RAN impact

RAN telemetry and additional UE specific messaging is much more easily available with Open RAN, through the open interfaces. The rAPPs will utilize these RAN-based datasets together with other datasets that are external to the network, such as user mobility patterns, to train AI/ML algorithms for RRM and load balancing.

4.9 Beam Forming

Feature description

5G NR is a beam-based radio access technology. Beamforming improves user throughput and facilitates interference management, thus, making efficient use of power. There are two major approaches to beamforming in wireless networks namely, adaptive beamforming and grid-of-beams (GoB) beamforming. Adaptive beamforming is based on the uplink reference signals, by which BS can estimate the radio channel and perform beamforming procedures, aimed at e.g., SNR maximization, or inter-user interference suppression. This approach requires SRS management procedures and accurate channel estimation. New beamforming weights are computed very often, thus increasing the signal processing overhead.

Another option is to create a static set of beams as a grid of beams. In this approach, each beam is associated with a unique synchronization signal block. Instead of a complicated procedure of channel estimation from uplink reference signals on the BSs side, the UE measures reference signal received power (RSRP) related to SSB in the downlink and reports it back to the BS. Based on such reports, the user is associated with one of the static beams.

Analysis of impact of Open RAN with RIC versus non-Open RAN

While aggregated mobility KPIs/PMs are used for slow adaptations (e.g., 5 min), individual failure reports are used for faster adaptations (e.g., 100ms) or for AI/ML based analysis of failure patterns. Measurement reporting periodicity and measurement type are configurable on a per cell basis considering the tradeoff between gain in mobility performance and associated signaling overhead. Mobility KPIs and failure events are forwarded from the CU to the near-RT RIC, and the near-RT RIC configures the cell handover offsets and measurement reporting in the CUs.

Open RAN near-RT RIC permits intelligent beam forming solutions for Operator specific energy savings optimizations. AI/ML based models can be used i) to build beam groups, ii) to decide on cell individual measurement configuration, iii) to detect changes in mobility characteristics, iv) to group the UEs in UE groups, and v) to calculate the optimal cell individual offsets. In case of dynamic beam pattern optimization, relevant mMIMO beam pattern information must be available at the Near-RT RIC, e.g., mobility reports might indicate a specific beam pattern.

Conclusion of above comparison with respect to Open RAN impact

Open RAN near-RT RIC intelligent beam forming solutions, using real-time data over E2 and processed with AI/ML algorithms, has the potential to unlock significant energy savings for Open RAN architected systems.

5. Radio Unit Power Saving Features

In O-RAN architecture, a radio unit is used to convert radio signals send to and received from the antenna system into a digital baseband signal, which can be connected to the DU over the O-RAN fronthaul interface. The high-level functional diagram of an Open RAN RU shown in Figure 18 consists of the following blocks:

- Synchronization and Fronthaul Transport
- Lower PHY Layer Baseband Processing
- Digital Front End (DFE) Processing
- RF Front End (RFFE) Processing

Some of processing blocks shown in Figure 17, such as digital frontend and RF frontend, are replicated when supporting multiple RF carriers and/or multiple transmit/receive antennas.

Reference architecture of an Open RAN RU

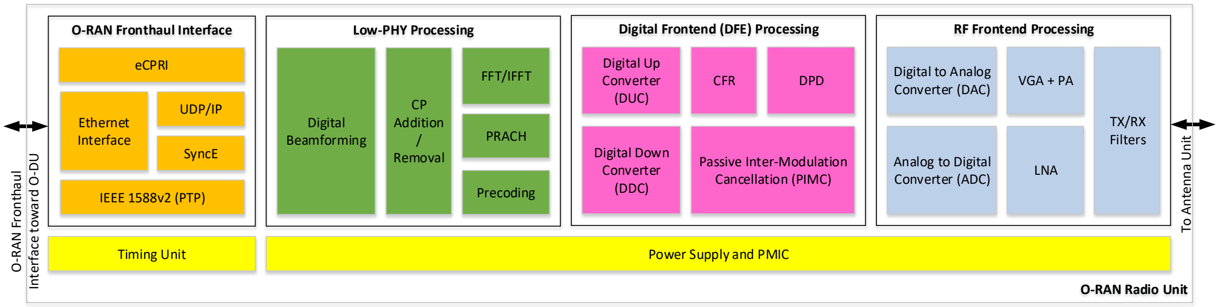

O-RAN Fronthaul Interface
eCPRI
Ethernet Interface
UDP/IP
SyncE
IEEE 1588v2 (PTP)
Low-PHY Processing
Digital Beamforming
CP Addition / Removal
FFT/IFFT
PRACH
Precoding
Digital Frontend (DFE) Processing
Digital Up Converter (DUC)
CFR
DPD
Digital Down Converter (DDC)
Passive Inter-Modulation Cancellation (PIMC)
RF Frontend Processing
Digital to Analog Converter (DAC)
VGA + PA
Analog to Digital Converter (ADC)
LNA
TX/RX Filters
O-RAN Fronthaul Interface toward O-DU
To Antenna Unit
Timing Unit
Power Supply and PMIC
O-RAN Radio Unit

Figure 17

The digital frontend comprises specialized blocks for the transmit and receive paths. The TX path contains a spectrum shaping filter and a digital upconverter (DUC) to convert the signal to the desired carrier frequency. In addition, it contains two fundamental blocks namely digital pre-distortion (DPD), and crest factor reduction (CFR). The CFR reduces the peak-to-average power ratio (PAPR) of the 4G/5G OFDM signals by clipping the signal peaks that create highest distortion. The DPD compensates the power amplifier distortion in the RF frontend to improve the PA linearity and power efficiency. Both CFR and DPD improve the energy efficiency of the RU. Minimization of the PA power consumption is a source of continuous improvement and innovation because PAs represent a large fraction of the overall power consumption in the RAN.

Energy savings gain from carrier and cell switch ON/OFF in a RU would be equivalent to power consumption of the RU hardware components that can be shut down or put into energy savings mode during low traffic load. In the RU reference diagram shown in Figure 17, the main energy savings gains would be attained from shutting down RF and digital processing units (depending on the antenna configuration), while O-RAN fronthaul processing, timing and power units and some other components would still be fully or partially functioning, and the power consumption would stay more or less consistent during the switching off period.

Applying various energy savings schemes such as micro-DTX, cell shutdown, and other advanced sleep modes in an RU depends on several factors including deployment scenario, traffic pattern/load, user distribution, antenna type, radio hardware, temperature, etc. The reason that total power consumption in cell sleep/shutdown and deep sleep cases does not vary with cell load is that these features can only be enacted by external operation and management control when the cell load is below a certain threshold. The power consumption in the latter cases is due to dormant fronthaul, timing, and power modules of the RU.

When the traffic load is low, a 64TRX mMIMO system may be scaled down to 32/16/8/4TRX such that some of the RF components in the RU can be switched off to save energy. Energy savings gains from RF channel reconfiguration would be equivalent to power consumption of the RU hardware components that can be shut down or put into energy savings mode when the number of active antennas is reduced. There is energy saving gains from reducing the number of RF channels from 4TRX to 2TRX in low load conditions.

6. CU/DU Processor Power Management and Energy Efficiency Considerations

While most of the power consumption in operator’s network occurs in the radio access network (RAN), operators also deploy their core networks, OSS/BSS support systems, and service data centers on servers using general purpose processors (GPP). With the advancement of server platforms, most wireless and wireline network workloads (virtualized/containerized functions) can be processed on COTS servers, while at the same time, power saving methods can be exploited for energy efficiency without compromising the stringent performance requirements of the network.

The goal of an energy efficient network can be considered as maximizing network capacity and user throughput while minimizing power consumption, i.e., an optimal tradeoff between energy and performance. By properly configuring the peripherals such as memory, storage, network and accelerator cards, COTS servers can satisfy the requirements of baseband processing in wireless access, user-plane data forwarding as well as control-plane message processing.

The energy efficiency is increasing with each new generation of processors. However, there is a greater opportunity to implement power saving features and become more energy efficient. COTS servers provide a range of management functions and APIs to enable data center or site management software to monitor and react to the utilization of each individual server in a deployment, allowing power-aware scheduling and optimization of power consumption across the deployment. Automation is one of the key contributors to achieving an energy efficient network. Consolidating workloads onto fewer servers, optimizing the cooling approach for each deployment location, and selecting the optimum placement in the network for a workload would enable significant reduction in overall energy consumption. Network workloads require fine-grained and fast response to traffic variation.

Advanced telemetry combined with the CPU power management features described below allows power saving techniques to be extended across the network. New generation of GPPs, which can be used across multiple segments of wireless networks, have various features and capabilities that can be used to improve energy efficiency. The new CPUs have had increased number of core counts and provide more compute power than previous generations. Given that not all resources may be used all the time, it is important to save power when processing capability is not required. This section examines the techniques that can be used to save power and improve energy efficiency in Intel x86 CPUs.

CPU power management modes may include the control of certain C-state (Power States) and P-State (Performance State) of the CPUs. C-state x, Cx, means one or more subsystems of the CPU is at idle. The states are numbered starting from zero and are denoted as C0, C1, … and P0, P1…, where the higher the number is, the more power is saved. C0 means no power saving, so everything is powered on. P0 means maximum performance, thus maximum frequency, voltage, and power are used. The basic C-states [defined by Advanced Configuration and Power Interface (ACPI)] are as follows[7]:

- C0: Active, CPU/Core is executing instructions. P-states are relevant here, CPU/Core may be operating at its maximum performance (thus at P0) or at a lower performance/power (thus at anything other than P0).
- C1: Halt, nothing is being executed, but it can return to C0 instantly. Since it is not working (but halted), P-states are not relevant for C1 or any C-states other than C0.
- C2: Stop-Clock, similar to C1 but it takes longer time to go back to C0.
- C3: Sleep, it can go back to C0, but it will take considerably longer time.

Modern CPUs have several C-states. As an example, Intel® Xeon® E3-1200 v5 family has C0, C1, C1E (C1 Enhanced), C2, C3, C6, C7 and C8, where C1 and C1E are CC states (Core C-States) only, and C2 is only a PC-state (Package C-state). All others are both a CC-state and a PC-state. There are also thread level C-states because of Intel® Hyper-Threading feature. However, the individual threads can only request C-states, but power saving action only takes place when the core enters in that C-state.

7. Energy Efficiency in Applications / Network Functions

There are different application designs that affect energy consumption. An application that implements a run to completion model and sleeps will allow energy savings when there is no work to do. Another different approach is a polling workload that is constantly polling I/O or other resources on the O-Cloud. There are different architectural approaches to a 5G software application but the majority of the workloads that a 5G software application executes can be categorized as follows:

> Bursty workloads in which CPU usage peaks and troughs at different random intervals. This is similar to the network traffic seen on 5G RAN networks.

> Sustained workloads with varying intensity in which CPU usage is a relatively constant value and varies periodically based on time. This is similar to packet processing workloads or polling-based application designs.

The type of energy savings that can be employed by the application developer will ultimately be based on the SW architecture design and the workload type (bursty or sustained). Possible approaches include leveraging the C-states and P-states that are available in GPP Cloud platforms to save power.

Table 3: Example power savings technologies

Workload Type	Description	Example Power Saving Technology
Bursty Workloads	Workloads that have peaks and troughs at different random intervals. Example includes L1 physical layer processing.	Leverage light sleep C-states and deep sleep C-states for quick transitions between peak and average and core consolidation during low traffic periods.
Sustained Workloads with varying intensity	Workloads with relatively consistent load that varies over time. Examples includes packet processing, or polling threads.	Leverage P-states to change core frequency and voltage

8. Intel® FlexRAN™ Reference Architecture

The FlexRAN™ reference architecture from Intel serves as a blueprint to facilitate development of Open RAN solutions, assisting equipment manufacturers and operators to reduce time, effort, and cost. It enables fast development of software-based LTE and 5G NR base stations that can be instantiated on any wireless network node. RAN virtualization presents several significant challenges as the processing and timing requirements are very stringent for the physical layer and fronthaul functions. These functions are critical because they impact many aspects of RAN capacity and coverage. Intel® Xeon® processor family provide high performance cores and instructions, such as Intel® AVX for vRAN signal processing instructions, which are useful for software-based L1 processing.

8.1 FlexRAN™ Reference Software

The software developed for one Intel® Xeon® CPU generation is reusable in the next generations[11]. FlexRAN™ reference software uses a distributed scheduler (not the same as L2 scheduler) framework to map tasks to a set of available CPU cores. Each of the cores can independently queue/dequeue tasks to/from a shared set of priority queues. The queues are protected using atomic operations and the scheduler itself has a low overhead. Using this framework, one can divide the L1 pipeline into independent tasks that can be scheduled in parallel to make use of all available cores. Some of the tasks are dynamically split into finer tasks based on number of user, layers, and/or antennas to reduce processing latency.

The block diagram in Figure 18 shows the FlexRAN^{TM} reference software, which takes radio signals from the RF frontend and provides real-time physical layer processing on servers empowered by Intel® Xeon® scalable processors. The FlexRAN^{TM} reference architecture performs the entire 4G and/or 5G L3, L2, and L1 processing. The FlexRAN^{TM} SDK provides optimized signal processing libraries for x86 processors. Its task controller facilitates scaling across multiple processors, and the data plane development kit (DPDK) delivers platform services such as high-throughput packet forwarding and memory management. The FlexRAN^{TM} reference architecture is designed to run on various operating systems supporting NFV and containerized network functions (CNF). Intel® Xeon® scalable processors are designed for cloud-native and virtualized networks such as macro-cell or indoor deployments. High performance Intel® Xeon® processors are ideal for equipment with space and power constraints, like multi-access edge or distributed computing nodes deployed closer to the radio antenna where they may be exposed to severe environmental conditions.

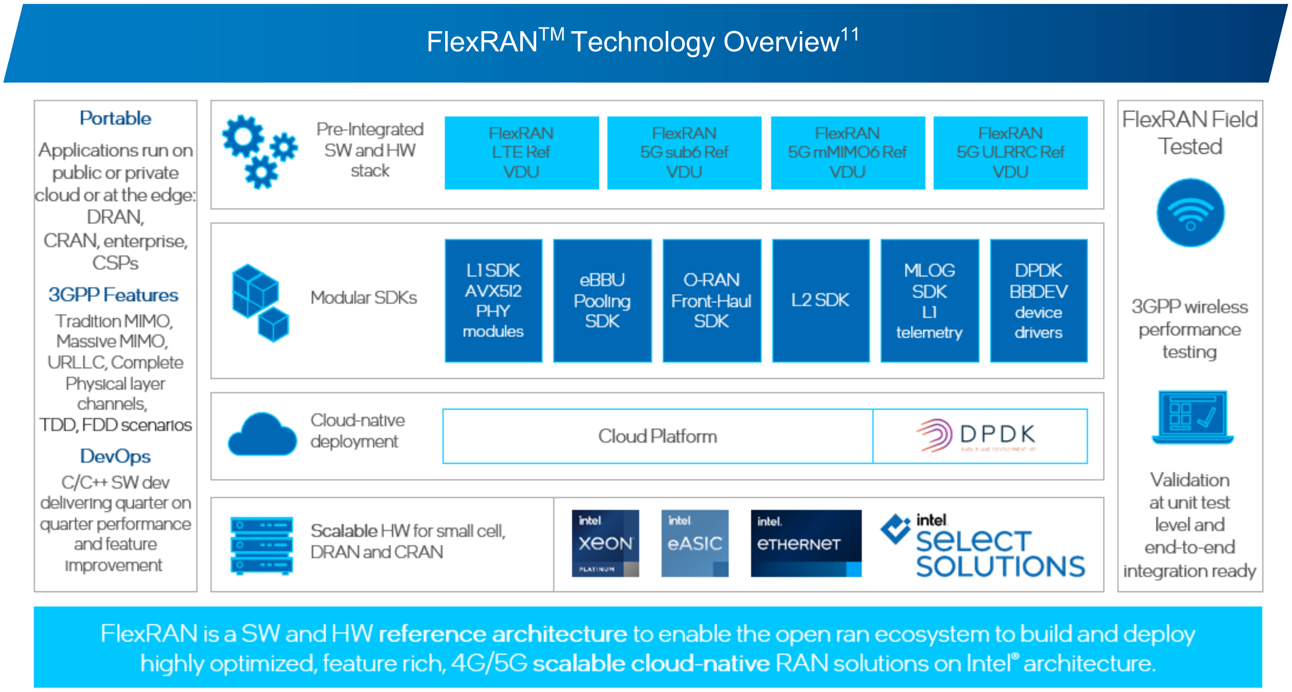

FlexRAN™ Technology Overview[11]
Portable
Applications run on public or private cloud or at the edge: DRAN, CRAN, enterprise, CSPs
3GPP Features
Tradition MIMO, Massive MIMO, URLLC, Complete Physical layer channels, TDD, FDD scenarios
DevOps
C/C++ SW dev delivering quarter on quarter performance and feature improvement
Pre-Integrated SW and HW stack
FlexRAN LTE Ref VDU
FlexRAN 5G sub6 Ref VDU
FlexRAN 5G mMIMO6 Ref VDU
FlexRAN 5G ULRRC Ref VDU
Modular SDKs
L1 SDK AVX512 PHY modules
eBBU Pooling SDK
O-RAN Front-Haul SDK
L2 SDK
MLOG SDK L1 telemetry
DPDK BBDEV device drivers
Cloud-native deployment
Cloud Platform
DPDK
Scalable HW for small cell, DRAN and CRAN
intel xeon PLATINUM
intel eASIC
intel. ethernet
intel select solutions
FlexRAN Field Tested
3GPP wireless performance testing
Validation at unit test level and end-to-end integration ready
FlexRAN is a SW and HW reference architecture to enable the open ran ecosystem to build and deploy highly optimized, feature rich, 4G/5G scalable cloud-native RAN solutions on Intel® architecture.

Figure 18

8.2 Integrating vRAN Acceleration

L1 forward error correction (FEC) is a compute-intensive 4G and 5G workload. FEC resolves data transmission errors over noisy channels. FEC techniques detect and correct a limited number of errors in 4G or 5G data without the need for retransmission. Initially, FEC was implemented in PCIe add-in cards to accelerate processing in hardware. The 4th Generation Intel® Xeon® processor family will include a SKU with Intel® vRAN Boost, a feature that eliminates the need for an external accelerator card by fully integrating vRAN acceleration directly into the Intel® Xeon® processor. Intel® vRAN Boost helps reduce operators' component requirements, which translates to reduced system complexity and can provide energy efficiency advantages. More information on the energy savings can be found in this blog post titled "Meeting Future Data Demand with vRAN Processing Capacity, Energy Efficiency"[15]. These accelerators natively support deployment in a virtualized environment (container or VM) based on single root I/O virtualization (SR-IOV) and Intel® virtualization technology for directed I/O (VT-d). The SR-IOV and VT-d are exposed to L1 software using a generic, light-weight interface within DPDK known as BBDEV. These accelerators significantly increase the cell capacity and performance of Xeon® processors running L1 workloads.

8.3 Opportunities for Energy Efficiency in Virtualized L1 Applications

8.3.1 Application Sleeps

Like any other application, the virtualized L1 application has busy and idle periods within each TTI. With the recent 3rd Gen Intel® Xeon® Scalable processor, the latency of switching between CPU processor idle states has been reduced to 1us, this enhancement allows real time applications like virtualized L1 applications take advantage of standard GPP energy savings techniques. During these idle periods, an application should sleep (through usleep)) or yield the processor to another tasks (shedyield)), if the Linux Kernel scheduler finds there is no work to do on the CPU core it will enter into a C-state. If the sleep duration is too small, then the OS may not send the core to C1 state and may just keep it active in C0 state. For longer sleeps OS may enter C1 states, resulting in some power savings during this period. Sleeps of 10 micro-seconds, have been shown to the deliver good optimizations for power savings versus real time performance. The real time performance of C1 state switching allows C1 state to be used throughout the virtualized RAN workloads in DU and CU.

With this low sleep period, (10us) the active cores are not allowed to go to C6 state, a deep sleep state that delivers more power savings compared to C1 state. C6 state has a higher exit latency and so needs more thought in how to use in real time workloads like DU. Intel® FlexRAN™ software has demonstrated intelligent scheduling of CPU cores used by the application to allow the Linux Kernel schedule cores into C6 state using sleep durations equal to the slot time (500us). This technique can be used on a dynamic basis on each TTI based on some high-level L1/L2 statistic like throughput, resource block utilization, users/TTI, etc. to make decisions on how many cores need to be active for the next TTI. Since the application decides when to shut down or bring back cores to active state, some coordination in suspend/resume operation might be necessary. This would maximize the opportunity for the physical core to enter a C-state.

8.3.2 Pooling

In a virtualized RAN, applications are designed as microservices and the infrastructure that orchestrates these microservices use Kubernetes. Network infrastructure is typically dimensioned for peak usage. However, the number of users and usage patterns can vary dramatically over a given period of time. The number of users is typically much higher during the day than at night. The usage pattern of applications such as video, voice calls, text messages, interactive gaming, etc. vary over the course of a day. Some of these variations are predictable and some are not. Moreover, the processing requirements for workloads can vary considerably depending on the traffic patterns. For example, while the L1 workload is heavy during high-activity periods when users consume more data, the L2 workload depends on the number of simultaneous users.

There are additional functions in a DU, such as analytics collection and processing, AI/ML inferencing models and others that need to be run from time to time which are flexible in terms of when they can be executed. Pooling enables dynamic allocation of the CPU cores to each workload. The hardware resources may be within a server blade or across server blades. The pooling benefits are maximized when the same type of CPU core is used for all L1/L2 functions. The pooling can be implemented within a server, i.e., pooling the CPU cores, memory, and I/O for multiple cells, or implemented within a rack of servers. These are not mutually exclusive and can be leveraged concurrently to improve power consumption depending on the deployment scenario.

As shown in Figure 19, in a distributed computing scenario, there is a single server per site where resource pooling can be implemented, whereas in a centralized computing scenario, there are more than one server in a centralized location such as edge cloud and pooling can be achieved within the server and within the rack of servers. Since the compute resources scale with the network traffic, the power consumption should be measured over a long period of time as opposed to reporting peak or low load power consumption. Figure 19 represents an Indoor or Omni cells with1 cell per site. A similar analysis would apply to a macro configuration with 3 cells per site.

Compute resource pooling within a server and across a rack of servers

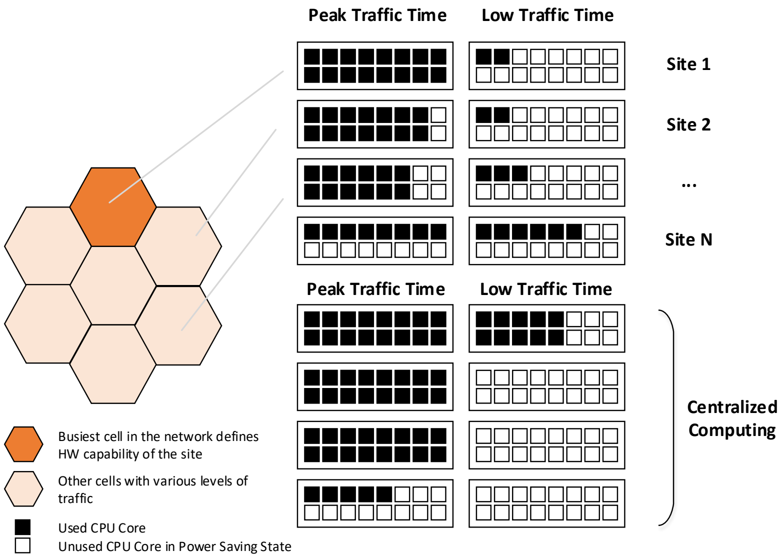

Peak Traffic Time
Low Traffic Time
Site 1
Site 2
...
Site N
Peak Traffic Time
Low Traffic Time
Centralized Computing
Busiest cell in the network defines HW capability of the site
Other cells with various levels of traffic
Used CPU Core
Unused CPU Core in Power Saving State

Figure 19

Pooling may be achieved through centralization of multiple cells per site to a single DU instance running in the cloud platform. More aggressive types of BBU resource pooling would allow a looser coupling of the RU to the DU network function. The RU is a fixed physical network function and cannot be changed. Therefore, it is the DU network function that must support cloud-native design principles such as scaling, zero touch automation, orchestration, and management etc. Pooling minimizes the total number of deployed CPU cores and servers and enables regular maintenance to be conducted without interrupting traffic. It makes it possible to monitor server health and do preemptive maintenance by moving traffic to other servers.

RIC based intelligence can be used to optimize the energy consumption (across CU/DU and RU modules, by maximizing sleep cycles, by intelligent scheduling which minimizes the no of symbols/slots to be scheduled). This in turn brings in energy efficiency by pushing ores to Sleep states (C1, C6, and more).

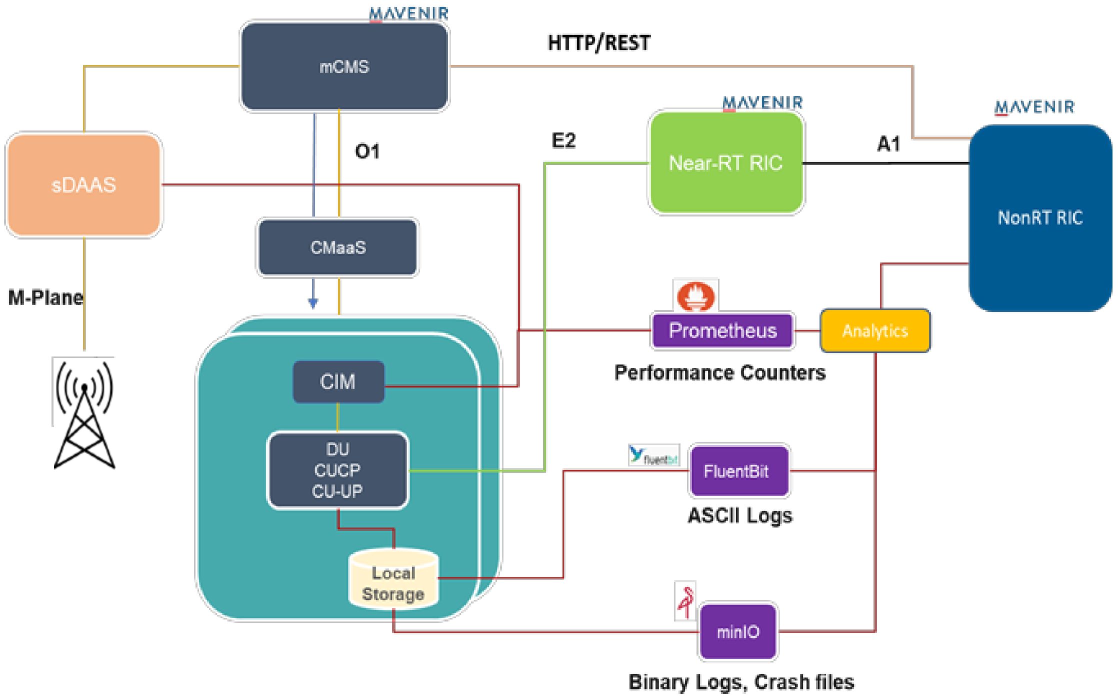
9. Mavenir Open RAN and RIC support for Power Efficient Implementations

While many of the energy savings features discussed earlier have been or are being implemented in Mavenir RAN and Radio products, there are ongoing efforts in 3GPP and O-RAN to define advanced energy efficiency measures that are expected to make Open RAN products even more energy efficient. Most of these energy efficient features exploit AI/ML based techniques to collect network performance analytics, train the machine learning models, and to optimally configure the system parameters to ensure maximal energy efficiency. The application of AI/ML algorithms within near-RT RIC and non-RT RIC to energy savings is described for each of the features, in Section 4. The following section describes the architecture aspects of near-RT RIC and non-RT RIC to enact the closed loop optimizations needed to achieve energy savings objectives.

9.1 Near-RT and non-RT RIC Role in Power Savings

RAN Intelligent controllers (RIC) in an Open RAN system can help optimize energy efficiency and performance across the entire operator network, as detailed in Section 4. In a large-scale Open RAN network architecture, shown in Figure 20, an energy-savings rAPP running on a non-RT RIC platform can analyze and regulate power consumption at various network nodes. The setup consists of energy savings rAPP and service management and orchestration O1 configuration and performance management (CM/PM) services. The energy savings rAPP running on non-RT RIC platform is used to continuously monitor the cell load utilization, using the O1 interface and according to standardized performance metrics. Once the energy savings rAPP detects low traffic load on a cell, it would check the coverage and load levels of the cells against others nearby in the cluster. The rAPP would then trigger energy saving actions on the cells based on operator-defined criteria via the O1 interface. The rAPP would then continue to monitor the load on the neighboring cells and activate the cell's energy-saving mode if the activity level exceeds certain threshold. The non-RT RIC can provide O1 CM/PM services to the rAPP via an R1 interface. The rAPP is able to consume the services to gather the performance metrics and modify the configuration of the E2 nodes over the O1 interface. The RIC and its framework services interact with the SMO O1 CM/PM services to collect the performance metrics and modify the E2 nodes.

Mavenir network performance and energy consumption optimization framework



Mavenir Centralized Management System (mCMS)
Smart Deployment as a Service (SDaaS)
Container Interface Module (CIM)

Configuration Management as a Service(CMaaS)
Prometheus is an open-source SW for monitoring and alerting functionality in cloud-native environments including Kubernetes.

Figure 20

Mavenir non-RT RIC is a containerized application that uses advanced machine learning algorithms to optimize network performance and train ML models using long-term RAN data for dynamic and adaptive policy and control. The non-RT RIC is responsible for setting high-level declarative policies and intents, sending configuration recommendations, and use-case-specific prediction and enrichment information via rAPPs to the near-RT RIC over the A1 interface. The non-RT RIC is hosted in a service management and orchestration framework, typically deployed in a centralized cloud, which is responsible for RAN fault, configuration, accounting, performance, and security (FCAPS) operation and orchestration of platform infrastructure resources.

Mavenir further provides O-RAN compliant near-RT RIC platform with an AI-powered extensible application (xAPP), which can control the traffic steering functionality of 5G RAN, a key feature that is responsible for managing the connectivity and mobility of users in the network. A comparison against a SON-based RAN handover algorithm shows improvement in mobility overhead KPI via reducing the number of handovers by approximately 50% and an increase in the throughput KPI by over 20% for cell-edge UEs, which can help make Open RAN operation more energy efficient.

9.2 Role of AI/ML in Network Energy Savings

The diverse and stringent service requirements of 5G networks, together with their increasing complexity are making traditional approaches to network operation and optimization less adequate. To bridge this gap, and to provide 5G networks with the intelligence required to find optimum operation points, equipment vendors and operators have started to equip their products with ML-based functionalities. Using network measurements, supervised and unsupervised ML tools are being extensively used to model 5G network behavior and subsequently to make decisions and/or predictions of complex scenarios. This is particularly relevant to energy efficiency. Minimizing 5G energy consumption is an end-to-end network problem, which highly depends on complex BS and UE distributions, varying traffic demands and wireless channels as well as hidden network trade-offs. Therefore, understanding and predicting UE behavior and service requirements over time are critical to optimize the network operation and configuration of mMIMO, RF carriers, sleep modes, etc. The current trend is to replace rule-based heuristics and associated thresholds with, e.g., optimal parameters configured through the knowledge acquired by machine learning models.

Given the dynamic nature of wireless networks, and lack of network measurements for all network procedures in all possible configurations, reinforcement learning (RL) is being widely explored to optimize 5G network performance and energy efficiency. Shutting down network elements is a combinatorial problem with several variables. RL agents may be used to let the network interact with the environment and learn optimum resource shutdown policies to minimize the total network energy consumption. It is expected that current promising machine learning models will be improved over time, while the wireless industry collects and makes available larger data sets related to different problems.

Machine learning techniques are being used to collect network analytics, learn intelligently from data and optimize the performance of the network. It is believed that virtualization technology improves the energy efficiency and resource utilization, resulting in significant energy savings. To achieve energy efficient virtualization and network optimization, AI/ML models can further improve energy efficiency through load sharing and consolidation and traffic steering. Likewise, energy consumption in the data centers can be minimized by intelligent resource allocation and management through ML-based approaches. The opportunities to apply AI/ML models based on specific use cases have been addressed in Section 4.

10. Conclusions

This white paper outlined the key available technologies, methods and strategies that a service provider can adopt, in building and deploying networks, to change their sustainability index and the impact of their carbon credits. It discussed the current state of energy consumption in 4G/5G wireless networks and energy saving principles and holistically reviewed the practical schemes such as RF carrier shutdown, channel shutdown and symbol shutdown as well energy saving techniques and technologies that can be used in the cloud-native RAN deployments such as CPU core utilization and resource pooling strategies to mitigate 5G network energy consumption. Enhanced technologies, e.g, deep sleep and symbol aggregation shutdown, which have been developed in 5G era are detailed. Open RAN with RAN Intelligent Controller (RIC) offers the granular ability to optimize RAN energy consumption by having additional access to RAN telemetry data, and deriving insights to take appropriate actions in a timely manner. The application of software based virtualized RAN provides unique energy saving opportunities and needs to be applied to all software nodes.

REFERENCES

[1] Green Future Networks – Network Energy Efficiency by NGMN Alliance, December 2021 https://www.ngmn.org/wp-content/uploads/211009-GFN-Network-Energy-Efficiency-1.0.pdf

[2] Green Future Networks – Sustainability Challenges and Initiatives in Mobile Networks by NGMN Alliance, December 2021 https://www.ngmn.org/wp-content/uploads/210719_NGMN_GFN_Sustainability-Challenges-and-Initiatives_v1.0.pdf

[3] O-RAN Architecture Description v08.00, October 2022.

[4] ETSI ES 203 228 V1.3.1, Environmental Engineering (EE) Assessment of mobile network energy efficiency, 2020.

[5] Considerations, Best Practices and Requirements for a Virtualized Mobile Network, GSMA, 2020

[6] 3GPP TS 28.310 V17.4.0, Management and Orchestration, Energy Efficiency of 5G, September 2022.

[7] https://metebalci.com/blog/a-minimum-complete-tutorial-of-cpu-power-management-c-states-and-p-states/

[8] Intel WP, Benefits of Virtualizing the Layer 1 in a RAN Stack, 2022 https://networkbuilders.intel.com/solutionslibrary/benefits-of-virtualizing-the-layer-1-in-a-ran-stack

[9] Intel, FlexRAN Power Demo Enhanced Power Savings Features 2022 https://www.intel.com/content/www/us/en/events/mobile-world-congress.html

[10] Leveraging Power Management Technology for vRAN Technology Guide https://cdrdv2.intel.com/v1/dl/getContent/736642

[11] https://networkbuilders.intel.com/solutionslibrary/power-management-technology-overview-technology-guide

[12] https://builders.intel.com/docs/networkbuilders/power-management-enhanced-power-management-for-low-latency-workloads-technology-guide-1617438252.pdf

[13] https://spectrum.ieee.org/3d-cmos

[14] Georgia Tech Electronics and Micro-System Lab (GEMS)

[15] https://community.intel.com/t5/Blogs/Tech-Innovation/Edge-5G/Meeting-Future-Data-Demand-with-vRAN-Processing-Capacity-Energy/post/1415977

[16] https://www.jrseco.com/how-much-power-does-5g-consume/#:~:text=Data%2Ddriven%20energy%20consumption&text=Telecom%20providers%20expect%20their%20energy,a%20U.S.%20network%20service%20provider.

[17] https://www.etsi.org/deliver/etsi_es/203200_203299/203228/01.03.01_60/es_203228v010301p.pdf

[18] https://www.diva-portal.org/smash/get/diva2:826087/FULLTEXT01.pdf
Energy Efficiency of Heterogeneous LTE Networks, Henrik Forssell

[19] https://www.etsi.org/deliver/etsi_ts/128300_128399/128310/16.01.00_60/ts_128310v160100p.pdf
Management and orchestration; Energy efficiency of 5G (3GPP TS 28.310 version 16.1.0 Release 16)

Intel technologies may require enabled hardware, software or service activation. No product or component can be absolutely secure. Costs and results may vary.

About Mavenir

Mavenir is building the future of networks and pioneering advanced technology, focusing on the vision of a single, software-based automated network that runs on any cloud. As the industry's only end-to-end, cloud-native network software provider, Mavenir is transforming the way the world connects, accelerating software network transformation for 250+ Communications Service Providers in over 120 countries, which serve more than 50% of the world’s subscribers.

APPENDIX

A: Open RAN Architecture: Components, Interfaces and Use Cases

Traditional radio access networks comprised base stations with integrated functions including baseband processing unit, cabling and connectors, RF processing unit, and the antenna subsystem, provided as a single package by few major OEMs. In contrast, the disaggregated RAN architecture splits the base station into three distinct parts: a centralized unit (CU), a distributed unit (DU), and a radio unit (RU). Combined with the notion of open interfaces, this would allow the operators to choose different hardware based on the needs of a particular network node. The virtualized CU and DU can be provided by different vendors, and the operators can choose the amount of processing power and additional functionalities that they require based on their deployment scenario. It must be noted that the disaggregated RAN architecture was originally defined by 3GPP to address 5G requirements. A key objective of Open RAN was to create multi-vendor RAN solutions that allow for decoupling of hardware and software, open interfaces and virtualization, hosting software that manages and orchestrates networks in the cloud, resulting in supply chain diversity, solution flexibility, and new capabilities leading to increased competition and innovation.

The O-RAN Alliance, established initially by prominent network operators, is the primary organization that is defining open interfaces of Open RAN architecture and complement 3GPP RAN specifications to ensure full interoperability between network elements made by different vendors. The O-RAN Alliance has been organized in various technical working groups, each focusing on a specific area, specifying the required elements of the disaggregated base station model, which includes hardware and software architectures of CU, DU, and RU and the open interfaces as well as cloudification and orchestration mechanisms for management and control of the virtualized functions. O-RAN defines the RAN architecture with a focus on open interfaces between the logical nodes and physical partitions of the RAN functions. In some deployment scenarios, e.g., a pico-cell or macro-cell, the physical layer functions are split between the DU and RU. O-RAN has defined an open fronthaul interface which is adopted in the split architecture. The DU contains the higher physical layer functions, while the RU hosts the lower physical layer functions[3].

A high-level view of the O-RAN architecture is illustrated in Figure A1. It shows that four key interfaces namely, A1, O1, open fronthaul M-plane and O2, connect service management and orchestration (SMO) framework to O-RAN network functions and the O-cloud. It further illustrates that the O-RAN network functions can be virtualized network functions (VNFs); i.e., VMs or containers, running on the O-cloud and/or physical network functions (PNFs) utilizing customized or white-box hardware. All O-RAN network functions are expected to support the O1 interface when interfacing the SMO framework. Within the logical architecture of O-RAN, the radio side includes near-RT RAN intelligent controller (RIC), O-CU-CP, O-CU-UP, O-DU, and O-RU entities.

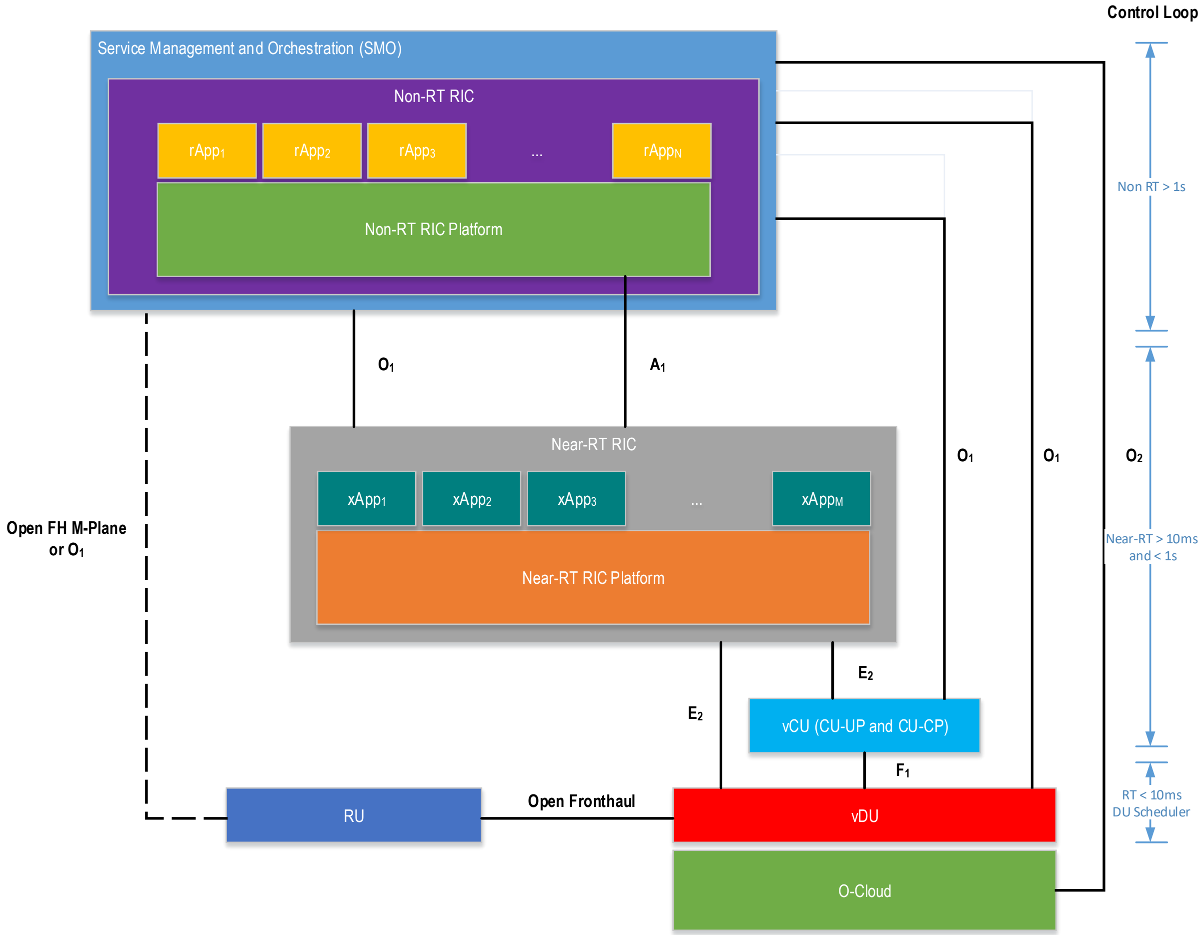


Figure A1: Components and interfaces of O-RAN architecture[3]

The O-cloud is a cloud computing platform comprising a collection of physical infrastructure nodes that meet O-RAN requirements to host the relevant O-RAN functions (such as near-RT RIC, O-CU-CP, O-CU-UP, and O-DU), the supporting software components (such as operating system, virtual machine monitor, container runtime, etc.) and the appropriate management and orchestration functions. The management side includes SMO framework containing a non-RT RIC function. The O-RU terminates the open fronthaul M-plane interface towards the O-DU and SMO. O-RAN near-real-time RIC is a logical function that enables near-real-time control and optimization of RAN elements and resources via fine-grained data collection and actions over E2 interface. It may include AI/ML workflow together with model training, inference, and updates[3].

The use of AI/ML tools is growing in importance and will allow processing the data faster to aid the non-RT RIC decision-making processes. The non-RT RIC is connected to the near-RT RIC via the A1 interface, while the SMO is connected to both the RAN's CU and DU components using the O1 interface (see Figure A1). The rAPPs connect to the non-RT RIC via the R1 interface and over standard APIs. This enables the use of rAPPs that are independent of the RIC implementation (i.e., supplied by third parties). Some of the use cases include policy, configuration, load balancing, cloud management, slice management, anomaly detection, fault prediction, and energy saving, and these are built as applications on the non-RT RIC platform. The non-RT RIC manages and optimizes RAN performance in control loops that last more than one second.

The O-RAN RIC architecture divides the necessary functions into service management and orchestration (SMO), non-RT, near-RT RIC, and the applications. Separating the non-RT RIC and the near-RT RIC decouples the high- and low-latency closed loop control functions. This addresses the MNOs' requirements for an open ecosystem where components from a variety of vendors can be integrated. Moreover, the cloud-native and distributed nature of 5G networks, including edge computing, will only aggravate the situation and manual configuration and optimization will no longer be manageable or cost efficient. Similar to the non-RT RIC, the near-RT RIC also utilizes open APIs to connect to xAPPs applications. These applications are part of the network function layer and operate in control loops lasting between 10ms and 1s. Near-RT RIC interacts with the non-RT RIC via the A1 interface, and it is connected to the RAN components' CU and DU via the E2 interface. This is an important interface given the need for significant alignment between several vendors, such as RIC platform and xAPPs developers, and both RAN hardware and software vendors, to ensure the RAN performance is not degraded. The xAPPs developed by platform developers or third-party developers control use cases such as radio bearer management, load balancing, and interference mitigation. Other more complex use cases such as massive MIMO beamforming optimization are also under development.

In O-RAN network architecture, open fronthaul interface is defined as the one-to-one or one-to-many link(s) between the O-DU and the O-RU(s). The previous generations of cellular systems used CPRI as the interface between the BBU and the remote radio units. While simple in design, CPRI required significant transport bandwidth proportional to the bandwidth of the baseband signal and the number of antennas. This disadvantage posed a significant challenge to the introduction of 5G services that rely on much larger bandwidths and increased number of antennas. Known as eCPRI, this packet-based transport technology significantly reduces the fronthaul bandwidth, but it also presents some new challenges. It exposes some of the disadvantages of packet-based transport such a its inherent packet delay variation. Furthermore, eCPRI is not a synchronous technology and relies on synchronization technologies such as precision time protocol (PTP) and optionally synchronous Ethernet (SyncE). Open fronthaul interface facilitates the use of standardized multi-vendor interfaces, which paves the path to successful interoperability between O-DU and O-RU (see Figure A2).

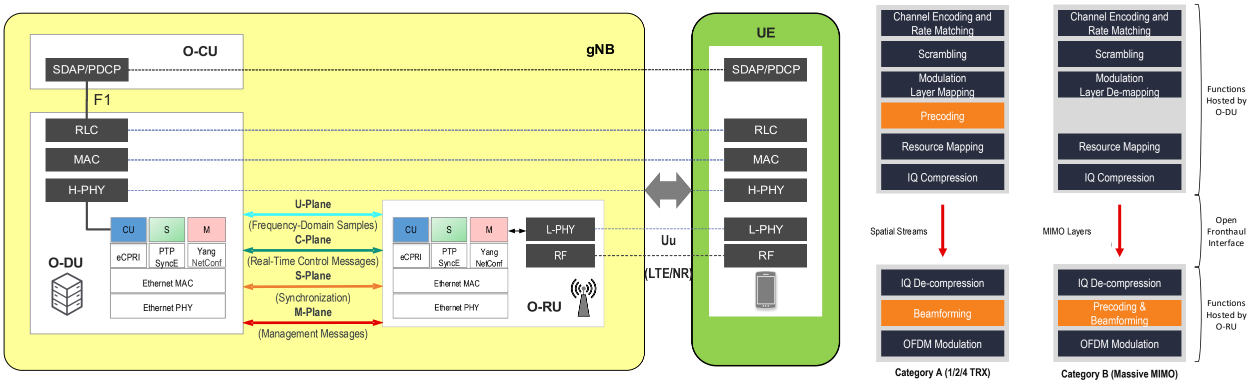


Figure A2: Disaggregated base station components, protocols, and O-RAN fronthaul composition[3]

B: Virtualization and Containerization of RAN Software

Containers and virtual machines use commercial off-the-shelf (COTS) hardware to achieve flexibility and scalability as compared to purpose-built hardware. Virtualization uses a common software infrastructure that enables the operation of multiple virtual machines, each having its own operating system on a single physical server (see Figure B1). It can be considered as a system with multiple computers each having its own operating system running on a single physical server with a hypervisor. The function of a hypervisor is to orchestrate and assign resources to these virtual machines. In contrast, containerization in the cloud-native domain provides improved development-to-delivery cycles. In containerization scheme, a single operating system instance can support multiple containers, each running within its own execution environment. Thus, it enables the deployment of multiple applications using the same operating system on a server. In containerization, microservices are critical to running applications in the cloud native environment. Microservices-based architecture splits applications into multiple services that are loosely coupled and can be developed, deployed, and maintained independently. Whereas a monolithic architecture has a single instance where all the application's functions reside.

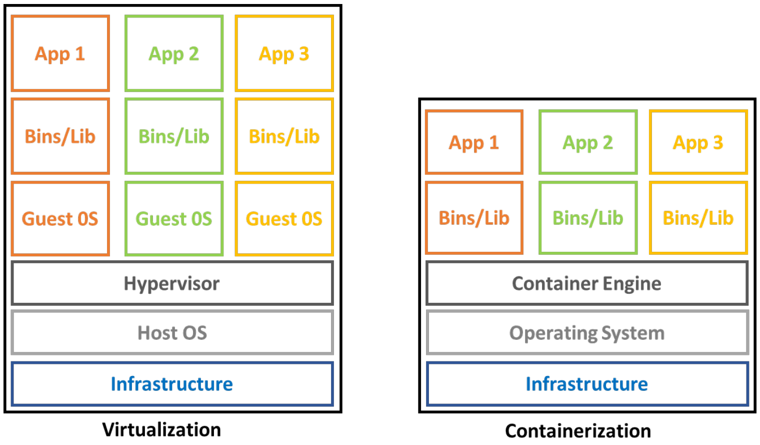


Figure B1: Illustration of virtualization and containerization

The choice of containers versus virtual machines depends on the environment requirements and the following factors:

- Different operating system support on a single server: All containers support a single operating system, which is typically Linux-based. Containers based on a different operating system will require a different host. This will increase the number of host servers. Virtualization can support multiple operating systems by using hypervisors on a single server. Thus, virtualization is preferable when the application requires a different operating system.
- Startup time: Containers start immediately as they do not have a separate operating system. The operating system for a container is on a server, which is always up and running. Whereas, in case of virtual machines, each virtual machine has its own operating system, and the start of virtual machines starts the operating system, increasing overall startup time.
- CPU utilization: CPU utilization is very high in virtual machines because of the number of operating systems used. Multiple containers use one common server operating system, which reduces the overall CPU utilization in a containerized environment.

- Image sharing: The size of the image in a container is small compared to virtual machines, as containers do not have an operating system. The operating system in a virtual machine increases its size and makes the transfer of virtual machines to a host very difficult. Containers, on the other hand, are portable because of their small size.
- Development cycle: Containers use microservice-based architecture which provides greater agility by limiting dependencies on other parts of the system. This shortens the development cycle compared to monolithic architectures, where changes to be implemented in a small part of the application requires the entire application to be rebuilt.

B.1 O-Cloud Components

Decoupling of software from hardware in the ICT industry and more specifically the mobile core network has been happening for several years. The 5G service-based architecture is cloud-native concept, in the sense that it works in principle similarly to a cloud service running in a data center. Thus, data centers are becoming more important as cloud services are growing. It is anticipated that this trend will continue, as 5G will address more vertical industries and enterprise businesses. Virtualization and cloudification of 5G RAN has been gaining a lot of interest recently. The virtualization of RAN means that the baseband functions such as L1, L2, L3 and transport processing are processed as software by general purpose processors (GPP) such as x86 processors using a commercial off-the-shelf (COTS) computing platform.

From the point of view of O-RAN, an O-cloud is a collection of O-cloud resource pools at one or more location and the software to manage nodes and deployments hosted on them. An O-cloud will include functionality to support both deployment-plane and management services. It provides a single logical reference point for all O-cloud resource pools within the O-cloud boundary. The cloud platform is a set of hardware and software components that provide cloud computing capabilities to execute RAN network functions. The cloud platform hardware includes computing, networking and storage components, and may also include various acceleration technologies required by the RAN network functions to meet their performance objectives. The cloud platform software exposes open and well-defined APIs that enable the management of the entire life cycle for network functions. The cloud platform software is decoupled from the cloud platform hardware (i.e., it can be sourced from different vendors). An example of a cloud platform is OpenStack or a Kubernetes deployment on a set of COTS servers, interconnected by a spine/leaf networking fabric.

There is an important relationship between specific virtualized RAN functions and the hardware that is needed to meet performance requirements and to efficiently support the functionality. As a result, a hardware or cloud platform combination that can support an O-CU function might not be appropriate to adequately support an O-DU function. When RAN functions are differently configured for each specific deployment scenario, certain aspects must be taken into consideration. For example, any RAN function that involves real-time movement of user traffic will require the cloud platform to control for delay and jitter, which may in turn require features such as real-time operating systems, avoidance of frequent interrupts, CPU pinning, etc. An O-cloud includes functionality to support both user-plane and management services.

The O-cloud provides a single logical reference point for all O-cloud resource pools within the O-cloud boundary, where an O-cloud resource pool is a collection of O-cloud nodes with homogeneous profiles in one location which can be used for either management services or user-plane functions. The allocation of NF deployment to a resource pool is determined by the SMO. An O-cloud node is a collection of hardware, i.e., CPUs, memory, storage, NICs, accelerators, BIOS, BMC, etc., and can be thought of as a server. Each O-cloud node will support one or more roles, which are defined as the functionalities that a given node may support including compute, storage, networking for the user-plane related functions, as well as optional acceleration functions and the appropriate management services[4].

The O2 interface is a collection of services and their associated interfaces that are provided by the O-cloud platform to the SMO. The services are categorized into two logical groups:

- Infrastructure Management Services (IMS), which include the subset of O2 functions that are responsible for deploying and managing cloud infrastructure.
- Deployment Management Services (DMS), which include the subset of O2 functions that are responsible for managing the lifecycle of virtualized/containerized deployments on the cloud infrastructure.

An O-cloud resource pool comprises one or more O-cloud nodes, each with one or more network cards and optionally, one or more accelerator cards. If the resource pool contains multiple compute nodes, it may also include an O-cloud network fabric that interconnect the O-cloud nodes. The O-cloud network fabric may provide connectivity between the servers, to the O-RU through an O-RAN 7-2x compliant fronthaul transport and to a regional O-cloud through a backhaul transport. The O-cloud network fabric may also be shared across multiple O-cloud resource pools or across multiple O-cloud instances (via separate network domains). The O-cloud network fabric is managed by the infrastructure management services.

In implementation of any logical network functionality, decisions need to be made regarding which logical functions are mapped to which cloud platforms or which functions need to be collocated with other logical functions. In the context of O-cloud, each specific mapping is regarded as a deployment scenario. Figure B2 shows the prominent O-cloud deployment scenarios. Scenario A is an edge cloud which centralizes near-RT RIC, virtual O-CU, and O-DU functions to support very dense deployments (e.g., dense urban areas) that provide a high-capacity fronthaul network. This type of deployment requires edge clouds with substantial hardware acceleration capabilities. Scenario B separates the virtual O-CU and O-DU functions from the near-RT RIC, which can be placed in a regional cloud and uses E2 interface for interaction with O-CUs and O-DUs. This allows near-RT RIC to have a global view for optimization. In scenario C virtual O-CU network functions are collocated with the near-RT RIC in a regional cloud. The regional cloud and edge cloud(s) must, in this case, satisfy the latency requirements of 3GPP-defined F1 interface. This scenario enables deployment in locations with limited fronthaul capacity and a number of O-RUs. There are two additional variations of this scenario, C.1 and C.2, to support specific needs of network slices. Scenario D is a replica of scenario C in which O-DU functions are not virtualized in an O-cloud, but rather supported by an O-RAN-capable physical NF. Scenario E is a replica of scenario C in which the O-RU functions are virtualized into a common O-cloud, in addition to the O-DU functions. Scenario F is a replica of scenario E in which O-DU and O-RU functions are virtualized into separate O-clouds[4].

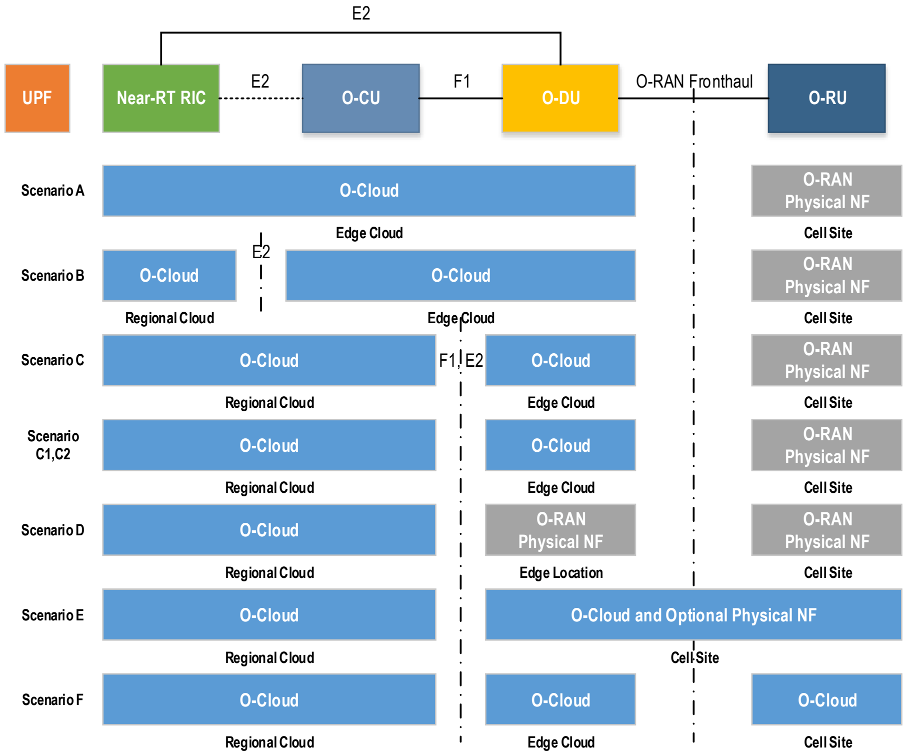


Figure B2: Deployment scenarios[4]

C: Energy Efficiency Considerations in Cloud Computing

Mobile edge computing in cellular networks provides processing resources (compute and storage) for applications with networking close to the end users, typically within or at the boundary of operator networks. Edge computing can also be placed at enterprise premises. The edge infrastructure can be managed or hosted by communication service providers or public cloud service providers. Several use cases require various applications to be deployed at different sites. In such scenarios, a distributed cloud is useful which can be seen as a computing environment for applications over multiple sites, including connectivity managed as one solution. The main benefits that the edge solutions provide include low latency, high bandwidth, device processing and data offload as well as trusted computing and storage.

To take advantage of the emergence of 5G, cloud providers have expanded their hybrid and edge offerings, partnered with telcos, and built or acquired 5G-specific services. AWS Outposts, Google Anthos and Microsoft Azure Stack are hybrid cloud platforms and services that facilitate 5G MEC use cases. The O-cloud, i.e., the O-RAN cloudification and orchestration platform, as we described in the previous section, facilitates flexible deployment options and service provisioning models of O-RAN virtualized network elements in telco clouds. The O-cloud is the cloud computing platform comprising a collection of physical infrastructure nodes that can host the relevant O-RAN functions, the supporting software components and the appropriate management and orchestration functions.

O-cloud energy efficiency can be achieved by reducing the power consumption of various components without degrading the network performance. The O-cloud components' power consumption can be optimized through actions such as adaptive shutdown of hardware, scaling up/down network functions, and optimization of CPU/GPU/FPGA power usage, memory usage, CPU/GPU frequency, etc. depending on network operational conditions. The non-RT RIC can configure changes towards the O-cloud using multi-dimensional data (e.g., traffic load data over E2 nodes, load over O-cloud in terms of compute/storage). When an O-cloud node is operating at low loading condition, the deployed network functions or its microservices can be relocated or shut down to free up the O-cloud nodes. Moving virtualized network functions within O-cloud and releasing nodes are among the infrastructure management services (IMS) internal functionalities which might be triggered by the SMO. Forcing nodes to an idle/dormant state may also be realized by blocking certain nodes and not assigning any new workload to them.

Idle O-cloud nodes can be shut down to reduce energy consumption during light traffic periods. To do that, the non-RT RIC subscribes to O2 data via SMO, which includes configuration of O-cloud nodes (i.e., Kubernetes cluster, resource pools, application containers, etc.). The non-RT RIC provides guidance to SMO federated O-cloud orchestration and management (FOCOM) function. The FOCOM monitors O-cloud and E2 node resources based on O1/O2 data and requests shut down via O2-ims. The O-cloud IMS will estimate if sufficient O-cloud resources are remaining or are available to serve expected requests. After execution or rejection of respective requests, the O-cloud IMS will communicate its actions via O2-ims.

An O-cloud node is a collection of processing units (CPUs/GPUs), memory, storage, NICs, hardware accelerators, BIOS, board management controllers (BMCs), etc., and can be viewed as a server or computing hardware, such as a physical or virtual machines. Programs running on nodes are packaged as containerized or virtualized microservices. One or more microservices form a virtualized or containerized network function (VNF/CNF). An NF may comprise functionalities supporting one or multiple cells or parts of one or multiple cells. As shown in Figure B3, the O-cloud Node 1 has one NF microservice deployed which can be relocated to the O-cloud Node 2, making O-cloud Node 1 idle. After relocation of the NF microservice, the idle O-cloud Node 1 can be shut down to conserve energy[5].

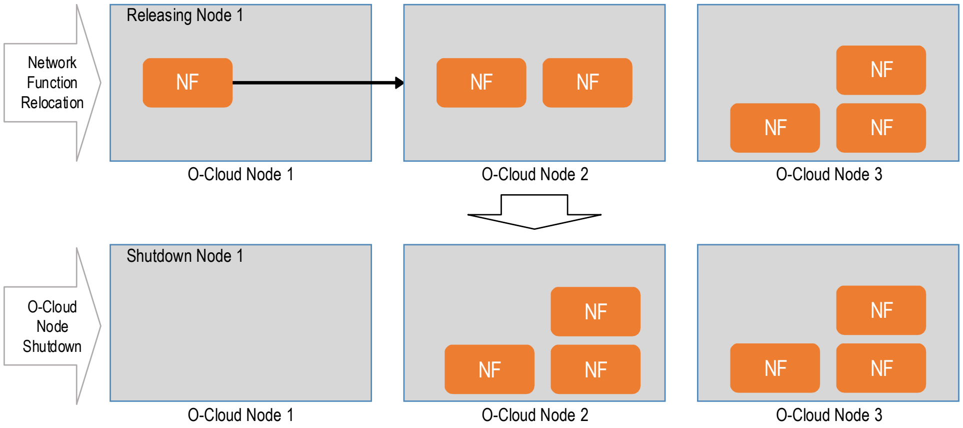
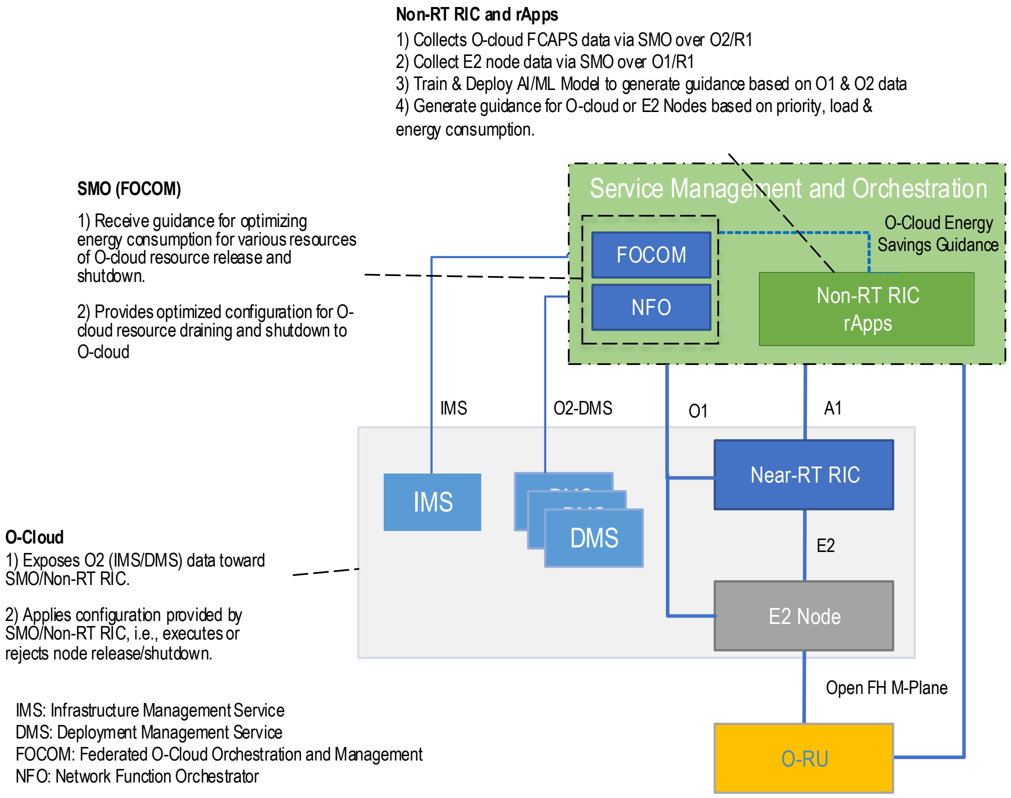


Figure B3: Energy saving in O-cloud by O-cloud NF relocation and node shutdown[5]

Non-RT RIC can use O1 and O2 data to provide guidance for shutdown decision. For example, O-cloud resource utilization metrics such as CPU, memory, and storage utilization can be analyzed by using O-cloud monitoring service telemetry. Such data can be correlated with RAN related load and energy consumption information obtained per network function or entity via the O1 interfaces. In O-RAN cloud-native network architecture, the FOCOM is responsible for accounting and asset management of the resources in the cloud. The NFO is responsible for orchestrating a group of network functions as a composition of NF deployments in the O-cloud. The IMS is responsible for management of the O-cloud resources and the software which is used to manage those resources whereas the DMS is responsible for management of NF deployments into the O-cloud. In this deployment, decision making for O-cloud node shutdown configuration/guidance, including AI/ML model training and inference, lies with the SMO/non-RT RIC entity. The overall procedure is illustrated in Figure B4[5].



Non-RT RIC and rApps

1) Collects O-cloud FCAPS data via SMO over O2/R1
2) Collect E2 node data via SMO over O1/R1
3) Train & Deploy AI/ML Model to generate guidance based on O1 & O2 data
4) Generate guidance for O-cloud or E2 Nodes based on priority, load & energy consumption.

SMO (FOCOM)

1) Receive guidance for optimizing energy consumption for various resources of O-cloud resource release and shutdown.

2) Provides optimized configuration for O-cloud resource draining and shutdown to O-cloud

O-Cloud

1) Exposes O2 (IMS/DMS) data toward SMO/Non-RT RIC.

2) Applies configuration provided by SMO/Non-RT RIC, i.e., executes or rejects node release/shutdown.

IMS: Infrastructure Management Service
DMS: Deployment Management Service
FOCOM: Federated O-Cloud Orchestration and Management
NFO: Network Function Orchestrator

Figure B4: Cloud resource energy saving using non-RT RIC[5]

The O-cloud energy saving strategy/guidance (shown in Figure 9) describes a method in which CPU active and idle state power saving features (P-states and C-states of x86 CPUs) used in the O-cloud node should be utilized in order to drive them into low-power modes when one or multiple CPU cores are operating at idle or low loading condition. This would create a method of O-cloud energy saving in which a preferred CPU power management mode (a P- and/or C-state) can be configured by SMO/non-RT RIC.

The energy saving through CPU power management modes may correspond to different CPU energy saving states (e.g., CPU frequency, voltage, certain sleep modes, etc.) that can be externally controlled. The O-cloud might for instance be configured with a range (or utilization factor) of permissible CPU power management modes. The O-cloud is then allowed to do fast adaptations of the CPU power management modes autonomously (e.g., based on instantaneous load of one or multiple CPUs) within that range. Alternatively, the O-cloud might be configured with a maximum O-cloud energy saving CPU state. The O-cloud is then allowed to autonomously select among CPU power management modes up to the maximum CPU power management modes. The energy savings can be maximized using these strategies, while still limiting the impact on QoS/user experience (e.g., potential latency impact on user-plane traffic). The operator is allowed to adjust the O-cloud energy saving gains versus O-cloud performance based on a tradeoff between the two.

At the O-cloud node, the power consumption can be controlled in different ways. The C-states reflect the capability of an idle processor to turn off unused components to save power. The decision to invoke a particular C-state can be made by examining the NF deployment distribution on the O-cloud instance, consolidating the deployments on the limited set of O-cloud nodes based on O2 telemetry analysis such as CPU utilization, memory utilization, etc. Note that the C-state can be changed at the CPU level, or at the individual core level of the CPU[5].

D: Energy Efficiency KPIs

D.1 Energy Efficiency for gNBs[19]

D.1.1 Energy Efficiency for non-disaggregated gNBs

$$EE = \frac{\sum_{Samples} DRB.PdcpSduVolumeUL + DRB.PdcpSduVolumeDL}{\sum_{Samples} PEE.Energy}$$

Non-disaggregated gNBs are gNBs without CU/DU split. The parameters are interpreted as follows.

1) $DRB.PdcpSduVolumeUL$ is a measure of data volume of PDCP service data unit of a DRB in the uplink, delivered by the PDCP layer to SDAP layer

2) $DRB.PdcpSduVolumeDL$ is the measured data volume of PDCP SDU of a DRB in the downlink, delivered to the PDCP layer.

The total data volume (in kbits) is obtained by measuring the uplink and downlink PDCP SDU bits of all DRBs of the non-split gNBs over the measurement period.

PEE.Energy measurement provides the energy consumed (in kilowatt-hours) by the subject gNB.

D.1.2 Energy Efficiency for dis-aggregated gNBs

$$EE = \frac{\sum_{Samples}\left[\begin{array}{l}(F1uPdcpSduVolumeUL + XnuPdcpSduVolumeUL + \\ X2uPdcpSduVolumeUL\ ) + (F1uPdcpSduVolumeDL + XnuPdcpSduVolumeDL + \\ X2uPdcpSduVolumeDL)\end{array}\right]}{\sum_{Samples} PEE.Energy}$$

For gNBs with CU/DU split, the parameters are interpreted as follows.

1) $F1uPdcpSduVolumeUL$ is the measured data volume of PDCP SDU in the uplink delivered to gNB-CU-UP from gNB-DU via F1-U interface

2) $XnuPdcpSduVolumeUL$ is the uplink data volume received from an external gNB-CU-UP via Xn-U interface

3) $X2uPdcpSduVolumeUL$ is the uplink data volume received from an external eNB via X2-U interface.

4) $F1uPdcpSduVolumeDL$ is the measured data volume of PDCP SDU in the downlink delivered by gNB-CU-UP to gNB-DU via F1-U.

5) $XnuPdcpSduVolumeDL$ is the downlink data volume delivered to external gNB-CU-UP via Xn-U interface, and

6) $X2uPdcpSduVolumeDL$ is the downlink data volume delivered to an external eNB via X2-U interface.

The total data volume (in kbits) is obtained by measuring the amount of uplink and downlink PDCP SDU bits of all interfaces (F1-U, Xn-U and X2-U) of the split gNBs over the measurement period.

The energy consumption (in kWh) is obtained by measuring the power, energy and environmental (PEE) parameters of the considered network elements over the same period of time.

Several PEE parameters have already been defined in 3GPP in accordance with ETSI[6] which are applicable to 5G a physical network function, i.e., power (average, minimum, maximum), energy consumption, temperature (average, minimum, maximum), voltage, current, and humidity[7].

D.2 Energy Efficiency metrics

To evaluate the effect of energy efficiency mechanisms on radio access network power consumption, some energy efficiency metrics need to be defined. These metrics must be comprehensive, reliable, and widely accepted to allow comparisons. They must capture both the energy consumed by the system under study as well as the performance measured at network level such as coverage, capacity, and delay.

To achieve these goals, ETSI environmental engineering technical committee, the ITU-T SG5, and the 3GPP RAN have specified metrics to assess mobile network energy efficiency under different operating conditions. One of the goals of 5G networks is to enhance energy efficiency through a tradeoff between the system capacity and the power consumption. Hence, definition of network load-dependent metrics is key for the next generation of green communication networks[4,6].

D.2.1 Data Energy Efficiency

The first metric represents the data energy efficiency metric of a mobile network expressed in bits/Joule.

It is defined as the amount of data transmitted per each unit of energy. For a given network coverage area $_A$ the data energy efficiency is defined as:

$EE_{DA} = DV_A / EC_A$[Bits/Joule]

where the data energy efficiency metric $EE_{DA}$is the ratio between the overall data volume DV_A transmitted/received by the network over a given coverage area and the energy EC_Aneeded by network equipment installed in this area for transmitting/receiving this data volume.

D.2.2 Coverage Energy Efficiency

Another standard energy efficiency metric expressed in m^2/Joule represents the coverage energy efficiency of the network. It is defined as the area which can be covered with the wireless signal, using one Joule of consumed energy. For a given network coverage area $_A$, the coverage energy efficiency is expressed as:

$EE_{CA} = S_A / EC_A$ [m^2/Joule]

where the coverage energy efficiency metric $EE_{CA}$is calculated as the ratio between the overall size of the area $S_A$covered by the network and the total energy EC_Aconsumed by the network equipment allocated to serve the analyzed area[6,7].

The total energy consumption of the network to serve the coverage area can include the sum of the energies consumed by the RAN equipment, transmission equipment (e.g., wired or wireless backhauling equipment) and other supplementary equipment (e.g., air-conditioning, power backup, etc.).

Intel technologies may require enabled hardware, software or service activation.

No product or component can be absolutely secure.

Your costs and results may vary.