SMaRT-5G Project

Sustainable Mobile and RAN Transformation (SMaRT)

Authors:

Sarat Puthenpura, ONF

Timon Sloane, ONF

Marcin Dryjański, Rimedo Labs

Murali Ranganathan, META/TIP

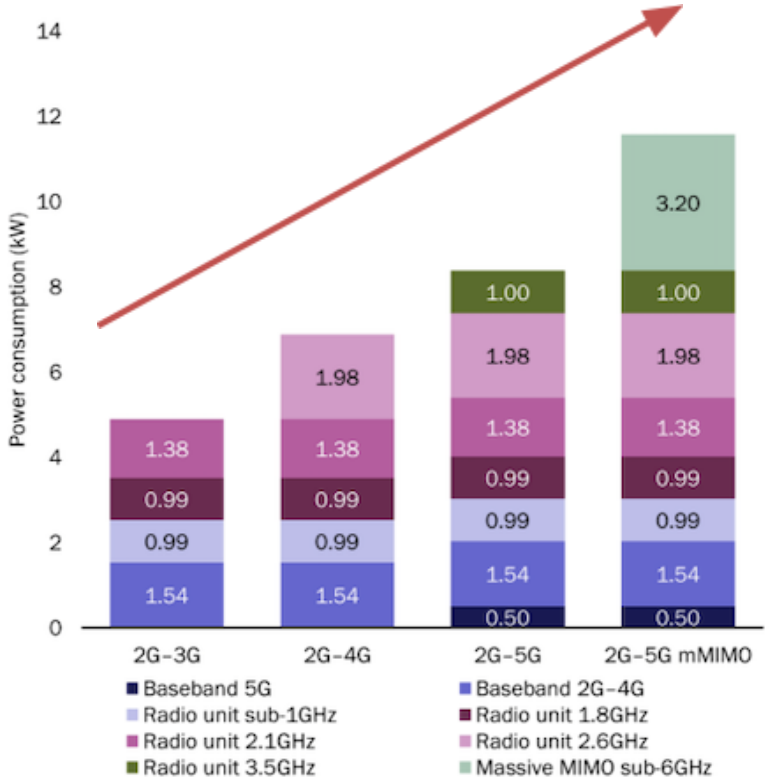
Vaibhav Singh, Ashutosh Tiwari and Christian Maciocco, Intel

N. K. Shankaranarayanan, Rutgers WINLAB

Table of Contents

Introduction

Energy costs for telecom operators can be as high as 10 percent of the total operating expenditure. This is especially true as 5G is being rolled out by operators. Even though the 5G-NR standards are more energy efficient than their predecessors (for example, energy consumption per unit of data - watt/bit is much less for 5G than 4G), the overall power consumption of 5G base stations can be very high. For example, a report from Huawei indicates that 5G base stations could consume about 1.7x more energy than its 4G counterpart [1]. With most networks delaying shutdown of legacy technologies (2G-4G), the overall energy consumption keeps increasing. **Figure 1** is an excellent summary of the maximum power consumption of a base station that supports multiple mobile generations by component (source: Analysys Mason), showing a steep upward trend.



Source: Analysys Mason

Figure 1 – Power consumption by various generations of base stations

Due to radio propagation characteristics, the use of new spectrum bands in which 5G operates will increase the density of mobile sites. As a matter of fact, 5G requires higher RAN densification to offer the same coverage albeit with higher capacities and throughput.

Additionally, massive MIMO and higher MIMO modes, which are highly beneficial to support mobile broadband at high frequency bands (FR2), require more power. Furthermore, with mobile edge computing the number of data centers will increase, and computationally demanding use cases will further intensify energy usage. A third aspect of 5G operations is the higher operating bandwidth for each channel, which has increased from 20 MHz (LTE) to as high as 400 MHz. The sub-division of larger bandwidths into Bandwidth Parts (BWPs), which is aimed at reducing power consumption of devices, will cause incremental energy consumption for the base stations. Thus overall, the energy consumption in mobile networks is expected to go up significantly as 5G is rolled out.

This rapidly growing energy use is creating two significant stresses for telecom operators:

1. Operators are under significant pressure to reduce their carbon footprint as telecom already accounts for 2 to 3 percent of total global energy demand, as per a recent McKinsey report [2].
2. Energy costs for operators are significant - for example, for one major US operator, the energy cost for running the RAN is reportedly on the order of $1B USD annually [3].

All these aspects are why mobile operators are under intense pressure to deploy intelligent solutions to optimize the energy consumption of their mobile networks.

Energy Saving Approaches – A quick look at Return on Investment (ROI)

A recent public report from AT&T indicated that the company spends about $1.6B annually in energy costs [3]. It is also estimated that RAN accounts for 73 percent of this cost, meaning that RAN energy costs are about $1.1B annually. Even if we achieve a mere 1 percent savings in RAN energy, the annual savings translate to $11M. Studies reveal that we can achieve energy savings on the order of 10 to 15 percent by leveraging a variety of approaches, which would further increase the financial attractiveness of optimizing power utilization on mobile networks. Below is a simple illustration of this point.

One of the approaches to saving energy in RAN is to turn off cells or frequency layers (aka carriers) which are not needed from a capacity point of view during low load hours (for example, from midnight to 6:00 am in a business district). We will be looking into this use case in detail in subsequent sections – we will just consider the basic financials for now. For example, let us assume that an MNO has 100,000 sites each with 3 sectors on average with three carriers (cells) in each sector. If two carriers can be shut down (put to sleep) during off

hours for 5 hours on average, then the cell sleep hours are 3x2x5 = 30 hrs per site per day. This implies that the total sleep hours per year for the entire network is 30 hrs x 100,000 sites x 365 days, which is about 1.1 billion hours (which is about 14 percent of the total cell hours). When a cell is put to sleep mode, its power amplifier is shut down. A power amplifier has a power consumption of 200 Watts (0.2 KW) on average. If we take on average 10 cents per Kilowatt hour as energy cost (in many countries, this value is much higher than this; for example, in some European countries, it is as high as 30 cents) the annual cost saving is about $22M. Note that this is just for one use case under SMaRT 5G, which has several similar and enhanced use cases under its umbrella.

There is also a study by the O-RAN Alliance which shows similar energy saving estimates for the use case mentioned above (cell on/off) [4].

If we look globally, it is estimated that Telcos spends about $25B annually in energy costs, out of which about $18.25B is for just the RAN [5]. Going with the 14 percent savings as we estimated above (that models just the cell on/off approach), this amounts to a savings of $2.55B per year globally. With some conventional calculations, this translates to a savings of 265 billion pounds of carbon dioxide every year [6].

ONF SMaRT-5G Initiative

ONF's Sustainable Mobile and RAN Transformation 5G (**SMaRT-5G**) project is a collaborative effort to develop and demonstrate an ML (Machine Learning)-driven, intelligent energy savings solution for mobile networks. Proof of Concept (PoC) implementations are planned to demonstrate progressively advanced energy savings techniques. Implementations will be undertaken both on open source RAN stacks (to make tools available to researchers) as well as on commercial-grade RAN configurations (in order to accelerate solutions that can be adopted by operators).

This initiative is structured as a phased series of PoCs designed to enable MNOs to start using the results from each PoC on both open RAN and traditional RAN architectures, thus supporting both brownfield and greenfield networks and providing the fastest possible route to impact.

SMaRT-5G is exploring two major approaches for optimizing the overall power consumption:

1. Optimizing RAN power consumption
2. Optimizing compute utilization for both RAN and Core

The ultimate strategy is to coordinate both these aspects in a holistic manner and to reach optimal operating conditions.

RAN Energy Optimization – Approaches

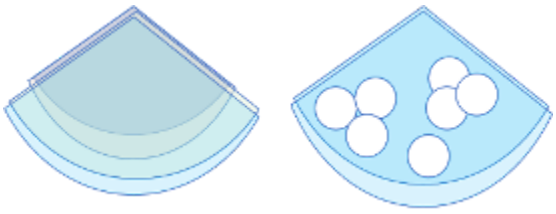
We now look into various approaches to save energy in RAN and some implementation details around them.

Cell On/Off Approaches

The first undertaking (PoC) in SMaRT-5G is to demonstrate optimized cell on/off capabilities, which involves turning carrier frequencies on and off in a mobile network. Optionally, if peak throughput is not required during low traffic conditions, Carrier Aggregation can be disabled to further reduce energy usage and provide more freedom for cell switch-off decisions.

Referring to **Figure 2**, there are two scenarios possible for cell on/off:

1. Keeping larger footprint (low band) cells on for coverage (coverage carriers), while switching on/off high band cells (capacity carriers) based on load and traffic pattern.
2. If small cell clusters are also operating within the coverage area of macros cells, switch them on/off based on load (interference is another side aspect to consider but that is beyond the scope of this paper).



Overlay - Macro **Macro + Small Cells**

Figure 2 – Overlay carriers vs small cells under the coverage areas of macro cells (different color shading denotes the coverage of different carriers)

Figure 3 is a graphical depiction of the concept. The blue area represents the capacity that can be put to "sleep" and to be "awakened" on an as-needed basis, based on traffic (capacity cells). The dark green area represents the capacity that should be made available all the time

to ensure coverage and to satisfy the minimum acceptable network performance level (coverage cells).

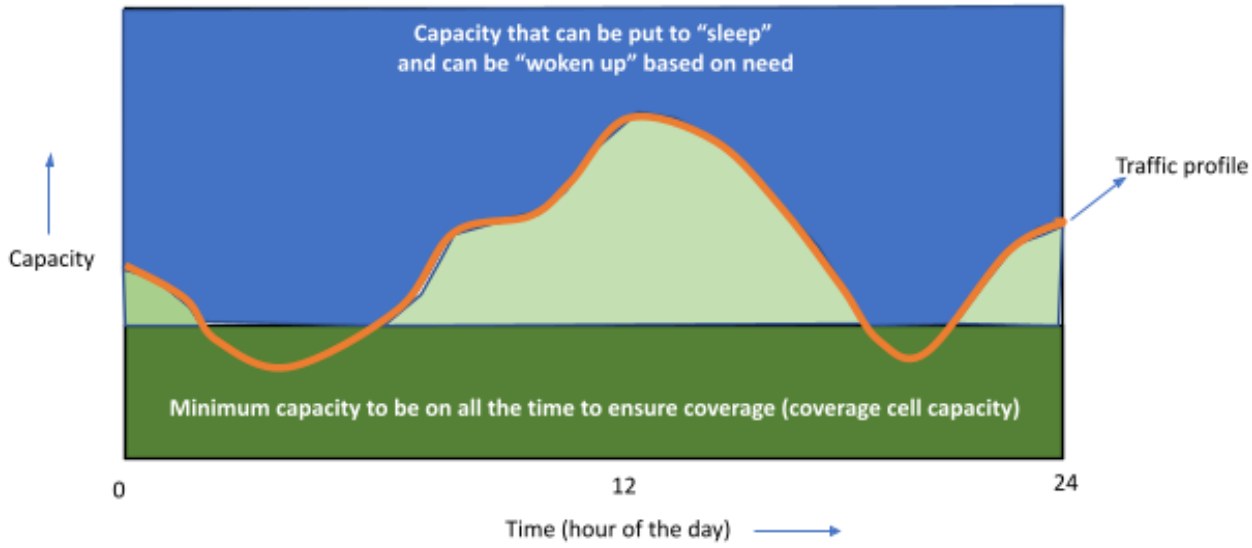


Figure 3 – Coverage and capacity cells

As shown, the sleep/wake-up capability should enable RAN capacity to closely follow demand (load) to optimize energy consumption while ensuring the needed performance levels. Network performance levels are often measured by Key Performance Indicators (KPIs) and minimum acceptable KPI values are set by MNOs. For energy management use cases, KPIs are generally evaluated at cell level (not at the UE level).

Three commonly used KPIs to monitor network performance are:

KPI Name	Description
Accessibility	Call or session setup success rate.
Retainability	Call/session drop rate.
Integrity	Throughput and latency. (For this use case, throughput is often considered to be critical)

A quick note on KPI calculations. Conventionally, cell level KPIs are calculated from performance counters that are collected from each cell, periodically (typically 5/15/30 minutes interval). Counters track cell level events and get updated when an event occurs (e.g., session drop, handover attempt). For traditional RAN, counters are reported at specific

intervals by the EMS (typically as xml files). These xml files are then parsed to extract the counter values and based on a given formula by the RAN vendor, various KPIs are calculated.

Next, we will look more closely at the conventional macro cell scenario.

As an example, consider a sector with five cells (carriers), two 700 MHz, one 850 MHz, and two 1900 MHz. A strategy would be to keep one 700 MHz cell on all the time (as the coverage cell) since it has the best propagation characteristics among the carriers, and to consider the other carriers as capacity cells.

If one needs to further extend energy savings, an entire sector may be put to sleep, or even more, the whole site may be put to sleep. This is possible in a dense urban setting when coverage compensation can be achieved by uptilting the antennas of nearby sites or increasing the transmit power. Though transmit power has an immediate impact on the cell boundary, most MNOs make maximum use of available power, leaving very little power headroom to increase coverage in the direction of a neighboring cell sector, where carriers are put to sleep. Therefore, antenna tilt adjustment may be the only viable solution for coverage compensation. Tilt adjustments are often done to manipulate vertical patterns, but newer antenna technology enables many complex adjustments of the coverage pattern as needed.

Figure 4 is a graphical depiction of how coverage compensation is done when an entire sector of a site is put to sleep and the lost coverage is compensated by adjusting the RF footprint of the sector of a nearby site facing the sleeping sector.

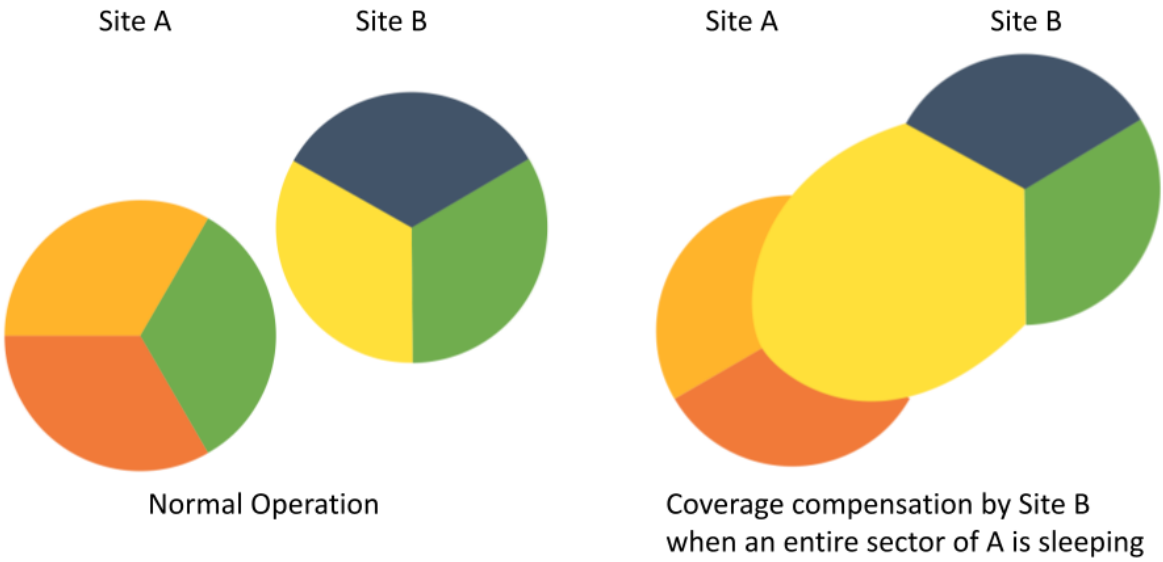


Figure 4 – Coverage compensation of sleeping cells

It is noteworthy that to compensate by another cell for the switched-off cell, we may need higher power, which is a tradeoff and needs to be taken into account. That is, the compensating sector may consume more power than normal operating conditions to provide coverage in the switched-off sector. Therefore, switching off cells in such a setup may not come without a price to pay.

Let us now look into how cell on/off concepts can be applied when small cells are involved. To begin with, consider the spatial and temporal traffic as shown in **Figure 5**, which depicts a sector with macro-cells (with overlay) and clusters of small cells within the sector's coverage area. The figure also shows the traffic distribution with different levels of peaks (p) happening at different times (t).

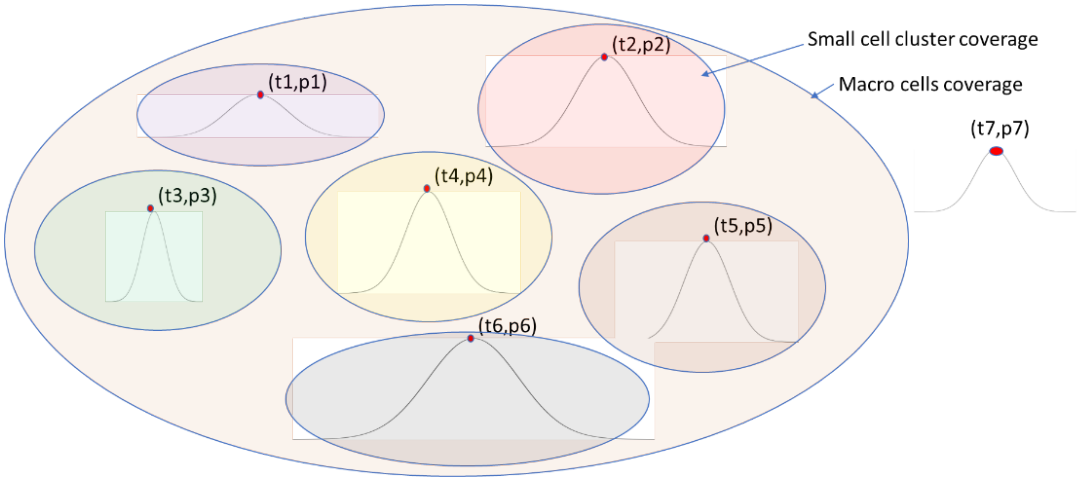


Figure 5 – Small cells and macro-cells

The following points are important to note:

1. Some macro-cells (especially low band) in the sector could serve as coverage cells, and some of the macro-cells (mid-band and high band) could be capacity cells, as was described previously.
2. The small cells are almost invariably capacity cells. The peak traffic of each small cell sector often happens at different times. For example, the peak traffic time of the green cluster may happen at 8:00 am on a weekday (say if it is at a train station) and that of the yellow cluster may happen at 11:00 am on the same day (assuming it is in a business district). This means, cells in the green cluster need to be turned on say from 7:00 am to 9:00 am on a weekday, whereas the cells in the yellow cluster may be turned on only from 10:00 am to noon on the same days. We will later see this

operation is highly effective using a disaggregated RAN architecture with virtualization of some of the RAN functions. This also means the associated core capacity can be turned on/off accordingly, and a cloud-based architecture for core is highly suitable for this purpose. Both these aspects play a key role in energy saving.

When a decision is made to switch-off a cell, even under low load conditions, some users may still be occupying the cell. This is a highly likely scenario since load balancing schemes during idle and connected states ensure that UEs are evenly distributed across all available carriers (IMSI based hashing for carrier allocation is a typically followed technique). A graceful switch-off of a cell would be needed, meaning that the cell is not switched off until there are no active users on that cell. It is also necessary to remove the cell that is due to be shut off from the idle mode and connected mode access and neighbor lists so that UEs do not try to continue to detect that cell. Graceful cell switch-off can be achieved in two ways:

1. By relying on features that are implemented in the RAN to move users to other cells (using techniques such as "Redirection at Release" or by adjusting handover thresholds to make the cell less favorable) and then locking the cell to other users (aka "soft lock"). In this RAN feature-based approach, the cell on/off criteria will be configured through feature parameters.
2. By the energy management application modifying search (sIntraSearch, sInterSearch) and handover thresholds to make the cell to be switched off less favorable to UEs.

While configuring the system to undergo cell switch-off and switch-on, two approaches are normally used, load-based and time-based:

1. In the load-based approach, load thresholds are set (based on PRB utilization, RRC connections, or other suitable load metrics) below which the cell will be a target for switch-off, and a threshold when exceeded will be used as a target for switching the cell back on.
2. In the time-based approach, based on historic traffic load patterns, the energy optimization system may simply set a time schedule for cell on/off.

It is advantageous to combine the two approaches on a single solution, which intuitively yields good results.

Let us now see how a disaggregated RAN is very effective in energy management. This can be illustrated with the help of **Figure 6.**

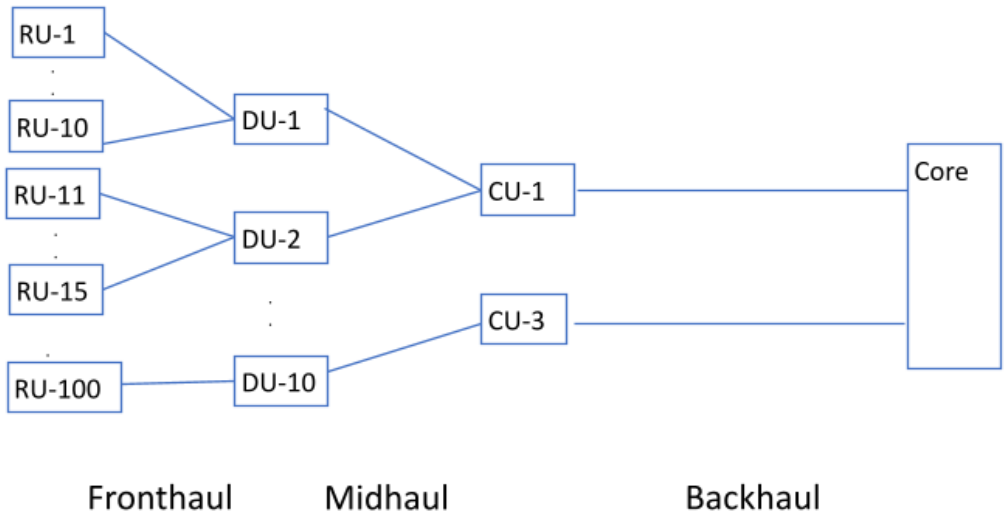


Figure 6 – Disaggregated RAN example

In this example, we have 100 RUs (Radio Units), 10 DUs (Distributed Units) with an average of 10 RUs per DU and 2 CUs (Central Units) with on average 5 DU per CU.

If we consider on an average 10 RUs per cluster and at a given time only 5 clusters are needed, then only 5 DUs and one CU is needed. Since DUs and CUs are VNFs, one can spin down those VNFs and associated cloud resources may be put to sleep. This means additional energy savings beyond the savings related to RUs. Furthermore, the corresponding core capacity (VNFs) and the associated cloud resources can also be put to sleep, giving more energy saving.

MIMO Sleep and RF Channel Switch On/Off

Let us look at two scenarios.

1. MIMO sleep mode: Consider for example 8×8 MIMO. During low traffic, the radio/antenna configuration can be changed to 4x4 or 2x2 MIMO. This would save energy. The important aspect to note is that this strategy could meet the user needs and is not reducing the carrier bandwidth, hence this could be an easier scenario to be considered by MNOs compared to cell on/off, especially by those who are overly concerned about network performance.
2. M-MIMO RF deep sleep mode: In the low traffic case, the M-MIMO panel can be turned off (put in deep sleep) partially or fully, progressively (turn RF channels off).

With RF deep sleep mode, almost all components in the Active Antenna Unit (AAU) can be shut down, and power reduction would be up to 75 percent. The efficiency of the on/off process of course depends on the M-MIMO architecture. One could consider Tx muting and PRB blanking in this context, with consideration to interference management as well, but a detailed dissemination of these aspects is beyond the scope of this paper.

Figure 7 depicts a summary of the concept under the two scenarios:

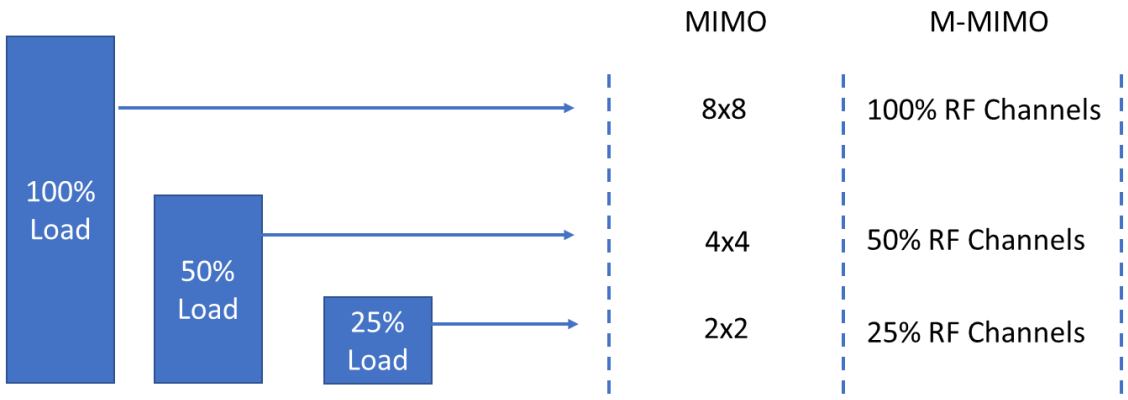


Figure 7 – MIMO and M-MIMO management to save energy

It is to be noted that impact on coverage should be compensated when MIMO configurations are manipulated. We have three options for this:

Option 1: Increase RF power but it reduces possible energy saving. Also, most MNOs make maximum use of available power, leaving very little power headroom to play with.

Option 2: Beam shape management (see **Figure 8**). The base station configuration and associated antenna patterns are adjusted to ensure coverage.

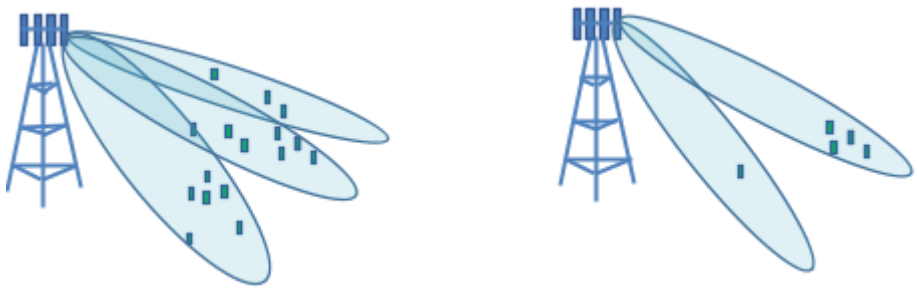


Figure 8 – Beam management to ensure coverage

Option 3: Uptilt nearby site antennas to compensate loss coverage (just like the cell on/off coverage compensation as shown in Figure 4).

More details on this topic can be seen in the article from Rimedo Labs [20].

Advanced Sleep Modes (ASM)

ASM corresponds to a gradual deactivation of the different components of the base station. Different types of sleep levels can be considered according to the transition time of each component, that is, the time needed to shut it down then wake it up again. This can be done at symbol level, sub-frame level, frame level or at the node (RU) level. This concept led to the classification of the different components of the base station into four sleep mode categories, SM1 through SM4 [8,18].

Essentially:

- ASMs switch off certain O-RU components and reduce energy consumption.
- The deeper (longer) the sleep mode level, the longer is its wake-up time (see **Table 1** below).

Sleep Mode	Deactivation Duration	Minimum Sleep Duration	Activation Duration
SM_1	35.5µS	71µS	35.5µS
SM_2	0.5ms	1.0ms	0.5ms
SM_3	5 ms	10 ms	5 ms
SM_4	0.5 s	1.0 s	0.5 s

Table 1 – Advanced Sleep Mode Characteristics

SM1: Represents the shortest sleep mode which needs a transition time of 71µs (OFDM symbol). The power amplifier and some components of the digital baseband and the analog front-end (both in Rx and Tx) are deactivated.

SM2: Corresponds to a longer sleep mode, which needs 1ms as a transition time (1sub-frame or TTI) and in which more components of the analog front-end are put to sleep and woken up compared to SM1.

SM3: The power amplifier and all the components of the digital baseband as well as almost all the components of the analog front end (except the clock generator) are put to sleep. These components need 10 ms (one frame) for going to sleep and then to wake up.

SM4: This corresponds to the entire base station going into a standby mode and almost its entire component set (except those needed for the sleep/wakeup functionalities) are put to sleep and woken up. This needs 1s transition time (hence often referred to as the deep sleep mode).

Implementation of RAN Energy Saving Solutions

Almost all RAN vendors have implemented an energy saving solution for the traditional RAN as a Distributed Self Organizing Network (DSON) solution for the basic cell on/off approach. **Figure 9** is a simplified depiction of DSON based solutions that are commonly used to implement a basic cell on/off solution to save RAN energy consumption.

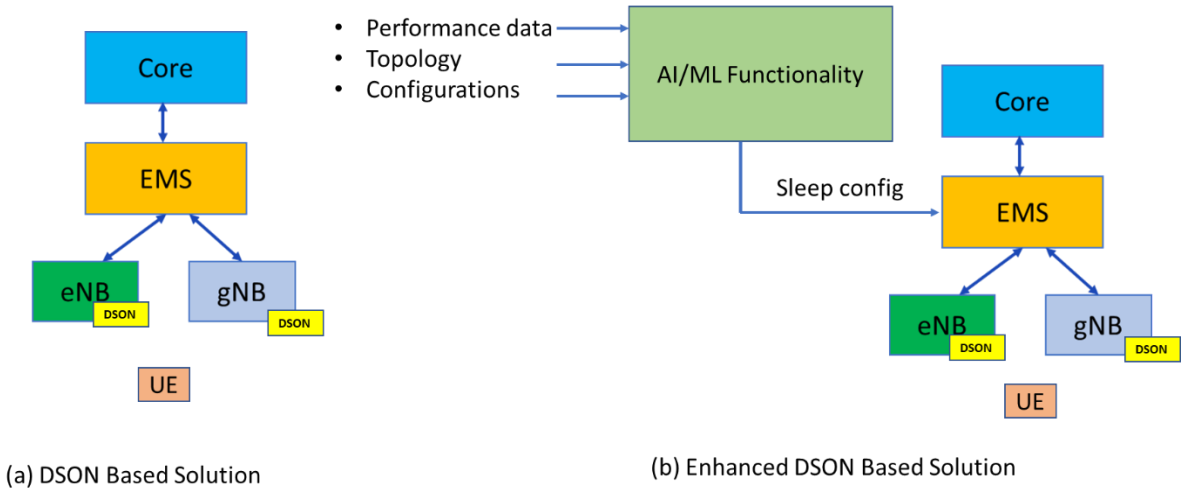


Figure 9 – Traditional RAN energy saving solutions via cell on/off

Though DSON implementations vary from vendor to vendor, on a high level it works as follows (refer to the basic solution shown in Figure 9).

First of all, the MNO has to classify cells as either coverage cells (cells that are to be on all the time to ensure coverage) or capacity cells (cells that are candidates for sleep and wake up based on load). This is often referred to as the cell qualification process and done as a separate function by MNOs. Once that is done, the MNO has to determine load thresholds. There are two sets of thresholds: (1) the load thresholds for the capacity cells below which it would be appropriate for the capacity cells to go to sleep, and (2) the load thresholds for the active cells above which the sleeping (capacity) cells should be awakened. Often these thresholds are specific values of RRC connections and PRB utilization. The DSON monitors load offered to the eNB or gNB in terms of the number of RRC connections and PRB utilization and ensures that appropriate sleep and wakeup actions are performed.

When the traffic load falls below the threshold (both RRC connections and PRB utilization) in a particular capacity cell, the cell is put to sleep mode (essentially, the power amplifier is turned off). In a sophisticated implementation, all active users in the cell that is slated to sleep are proactively transferred to other active cells (either capacity or coverage cells).

Now assume that only the coverage cells are active, and all capacity cells are sleeping. When the load (RRC connections or PRB utilization) exceeds the thresholds, all capacity cells are awakened simultaneously, not one by one. This is because the DSON feature is not aware of the rate in which the traffic increases, and to safeguard performance the entire capacity is made available at once.

One lingering problem is the latency involved in waking up sleeping cells. Vendors have made good progress in this area but may not be up to the total satisfaction of MNOs. To circumvent this issue, operators often set very conservative thresholds, but that, of course, reduces the energy saving opportunity.

Another issue is the customization of the load thresholds per cell, as the load characteristics among sites vary significantly and AI/ML-based approaches are sometimes used to address this issue since setting universal, conservative sleep/wakeup thresholds results in diminishing energy savings.

In addition, it is important to take into account that the energy saving algorithms cannot work in a vacuum. The decision to switch off the entire base station needs to be coordinated with other features. In general, it is required to provide energy-saving-aware traffic steering mechanisms, which will move the users out of the cell before it is switched off. This requires coordination and a holistic approach to energy saving to avoid instability in the network.

Gaps in Current DSON Solutions for Energy Savings

While the DSON operation looks straightforward, it has many shortcomings. These are summarized below.

- DSON solutions operate essentially in an open loop, with no inherent feedback on network performance. Cell sleep/wake up decisions are purely based on load – RRC connections and PRB utilization. While this is a logical approach, MNOs are often worried about network performance and the ability of DSON to respond quickly to sudden increase in network load; cell sleeping activity takes capacity away from the RAN, while the load is unpredictable, and cell wake up time could be of the order of a couple of minutes. Therefore, many MNOs are not fully confident to turn on the DSON feature all the time, thus they limit the activation of this feature to the network maintenance hours (say midnight to 4:00 am). However, this reduces cell sleep hours and corresponding energy saving.
- Manual setting of cell sleep/wake up thresholds. As mentioned before, the MNOs are tasked with setting thresholds and this needs to be done as a careful balance between energy savings and network performance. Therefore, MNOs are often forced to set conservative, universal sleep/wakeup thresholds, reducing sleep hours. Few MNOs have implemented AI/ML solutions to customize the thresholds on a site/sector basis as shown in Figure 9, (b) by themselves or via their vendors as custom solutions.
- Site/cell qualification process is managed separately and often in a laborious manner. It is to be noted that not all capacity cells are candidates for sleep. For example, if some capacity cells are used for dynamic spectrum sharing (DSS) or licensed assisted access (LAA), they are often dropped out of the sleep candidate list. Many such conditions exist for MNOs and those conditions often vary from time to time. Site/cell qualification has to be well integrated with the DSON solution, which is often not simple to do.
- In a network there are many automation functions (like other centralized SON type capabilities) that may be active concurrently, MNOs face difficulty in coordinating these other functions with energy saving functions. For example, if traffic off-loading is active (i.e., moving traffic from “hot” sites to less busy neighbor sites) in a cluster of sites, this function needs to be well coordinated with energy saving (e.g., priority needs to be specified among potentially conflicting actions). This can be quite complex.

- Note that sleep/wakeup decisions are done by DSON on an individual sector/site basis. However, coordination among multiple sites in a cluster would yield better energy savings via holistic traffic management among them.

Using AI/ML for Energy Saving Solution

In general, AI/ML-based approaches support time-based energy saving strategies and associated solutions. Currently, AI/ML-based intelligent energy savings leveraging data from the RAN is being discussed in 3GPP and O-RAN standards [4] [9] [10]. While messaging between an intelligent controller and different RAN components (functionality and inputs for energy saving AI/ML models) are undergoing standardization discussions, the details behind energy saving AI/ML algorithms are out of scope of standardization.

Estimating future load and cell QoS correctly is essential for designing intelligent energy saving solutions. For example, if the predicted load is lower than the actual load, switching off a cell (or other energy saving action), may cause an outage situation as the cell is switched off while its actual and immediate future load is high. On the other hand, in the case of load overprediction, cells may not be switched off when the actual future load is low resulting in less energy savings.

Thus, there is an innate tradeoff between seeking maximum energy savings and minimization of service quality degradation. Therefore, in the context of AI/ML, there is a need to incorporate MNO's preferences in future load prediction along with QoS requirements. MNOs may emphasize service quality over high levels of energy saving in their network based on several factors, such as customer contracts, geographical location, time of day, type of traffic, etc. For example, in an urban downtown area during the day operators would place higher weight on avoiding underprediction situations which could cause service quality degradation, while ensuring sufficient energy savings. In contrast, at nighttime during low traffic hours, the operator may choose to place higher weight on energy saving for these cells.

We further observe that each cell in a network may have different load distribution characteristics, different energy saving priorities, and different degree of traffic imbalances [11], needing the ML training on a per cell level based on the corresponding data characteristics and MNO preference. Note that some of the existing energy saving solutions were designed under the assumption that the load is balanced and evenly distributed under different load regimes. Recently, training imbalance and fairness have received high

attention in the ML community for different fields such as computer vision, but it has not been sufficiently addressed in intelligent RAN energy saving approaches.

Listed below are some of the design requirements for an AI/ML based solution for energy savings:

1. We should consider the overhead such as power consumption, compute resource requirements etc., in addition to the performance while determining the right AI/ML model to use for energy saving solutions. This can aid the selection among available sophisticated models such as neural networks, graph-based models, attention-based models.
2. The model training should accommodate bias in the network field data rather than assuming well-balanced loading among cells.
3. The model should be able to take input from the operator to prioritize energy saving vs cell QoS.
4. The model should take into account different distributions of data across different cells while coming up with an appropriate energy saving model. At one extreme, we could have an AI/ML model trained on a per-cell basis, and at the other extreme, there could be one common model trained on data from thousands of cells. We need to identify the right model training approach based on data availability, computing resources etc.
5. The model for final deployment should be trained and evaluated on real field data as the simulated traffic patterns may not capture sudden peaks, or valleys in the actual traffic. Simulated data could be used for developing a coarse model, initially.
6. The model should be easy to re-train based on load distribution shifts in the network.
7. Model results and predictions should be made available (e.g., RestAPIs or shared database) so that they can be reused by other applications or services. This avoids redundancy and saves computing and storage resources.

It is to be noted that a real-time check of traffic shift is often necessary (and reacting accordingly) if a high weight is given to QoS.

Advantages of O-RAN Architecture for Energy Savings

With these points in mind, let us examine how O-RAN architecture is highly suitable to implement RAN energy saving solutions. To begin with, **Figure 10** shows the skeleton O-RAN architecture for energy saving solutions:

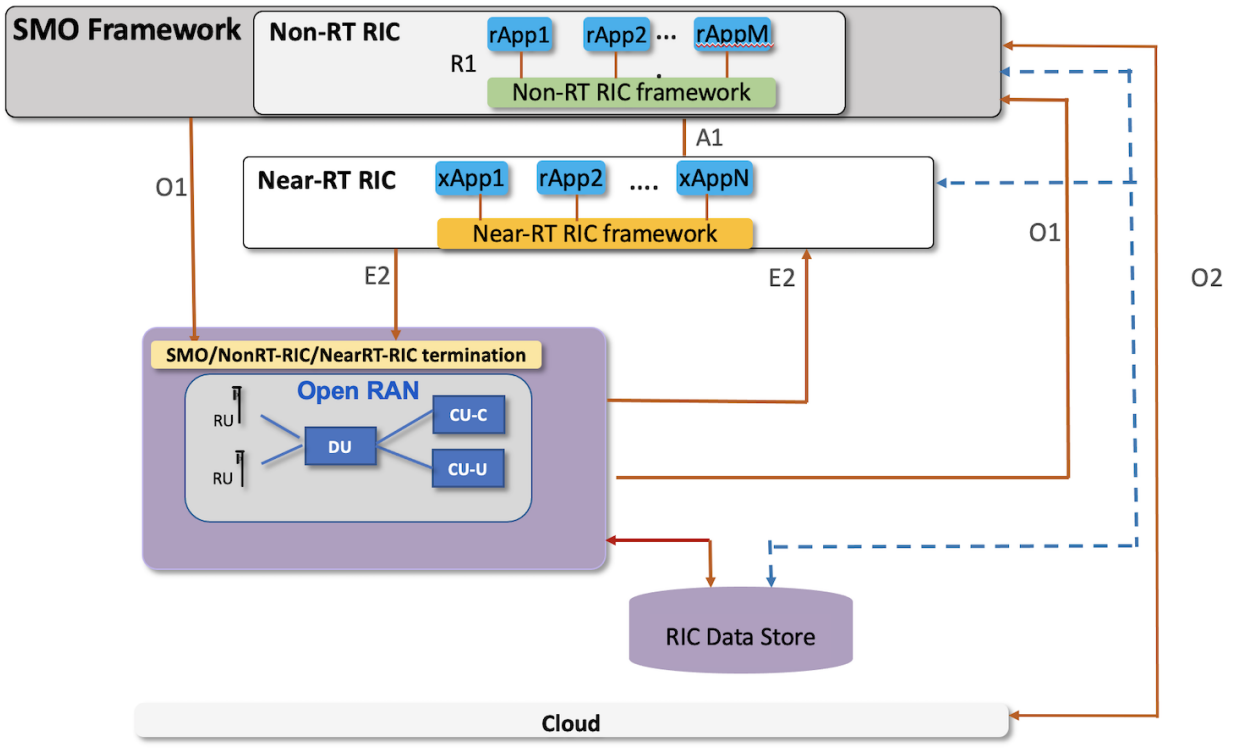


Figure 10 - O-RAN architecture for energy saving solutions

The Non-RT RIC/rApps and Near-RT RIC/xApps providing ML-based intelligent network control is the heart of the solution. In general, Non-RT RIC/rApps implement control loops with latency one second or more while Near-RT RIC/xApps implement sub-second control loops. While O-RAN does not specify a data architecture, Figure 10 shows a RIC Data Store to enable PoC implementations.

Some more details on the solution can be found in the O-RAN specification O-RAN.WG1.NESUC-R003-v01.00 [4]. Also, the Rimedo Labs article is a good resource for further information [20].

Cell on/off and RF Channel/MIMO on/off can be implemented with Non-RT RIC/rApps, whereas ASMs would need to be implemented in Near-RT RIC/xApps to achieve the necessary performance/latency. In any case, an O-RAN/RIC-based solution for energy saving presents the following advantages. First, we will look at it from an architectural point of view.

1. RIC has visibility to multiple radio units, hence holistic sleep/wake-up decisions can be made.

2. Policy driven sleep/wake-up strategy makes management of this capability easier and efficient. Operators can quickly change sleep strategy without changing any software – they just need to change the policy specifications.
3. RIC construct provides flexibility and control to make the RAN more easily programmable.
4. SMO/Non-RT RIC is a very effective platform for AI/ML based sleep policy creation. As a matter of fact, the sleep strategy could be automatically adjusted without any human intervention, in an adaptive mode.
5. Coordination of other automation capabilities can be done as policy coordination and conflict resolution. For example, dynamic spectrum sharing (DSS) and energy saving have conflicting goals and coordination between them should be more tractable in a well-structured policy driven RIC environment.
6. Future proof solution – Non-RT RIC/rApps are being designed to work with traditional RAN (by leveraging an O1 adapter) as well as with Open/Disaggregated RAN. The hope is to support rApp RAN automation/evolution without needing any "rip and replace."

Additional information can be obtained from the Rimedo Labs whitepaper [7].

Operationally a RIC-based solution has the following advantages:

- An inherently closed loop solution. Cell sleep/wakeup decisions are based on KPIs, providing a better handle on network performance to MNOs. Therefore, the feature can be turned on all the time, yielding improved energy savings.
- MNOs can strike a good balance between energy savings and network performance based on operating conditions, with full confidence due to the closed loop capability.
- Optimal sleep thresholds (PRB utilization and RRC connections) on a sector/cell basis can be implemented using AI/ML techniques without any human intervention, saving operation costs.
- Policy-based cell qualification process is easier to manage and can be done without any human intervention. This can also be integrated with the cell sleep/wakeup process, which translates to substantial efficiency in operation and associated cost reduction.
- RIC has visibility across sites in a cluster, making coordinated sleep/wakeup decisions feasible and efficient in conjunction with other network management functions (such as load balancing), thus reducing network operational complexities via automation.
- The RIC app vision allows best-in-class solution providers to develop innovative algorithms without having to deal with the complexities of network interfaces.

However, using O-RAN architecture for SMaRT-5G poses the following challenges, which are common for O-RAN adoption in general:

1. O-RAN standards are slowly maturing.
2. O-RAN architecture-based RAN is not very widely deployed and could take time to reach the needed critical mass.

Therefore, a possible strategy for making it possible for SMaRT-5G to have a rapid time-to-impact for operators, could be as follows.

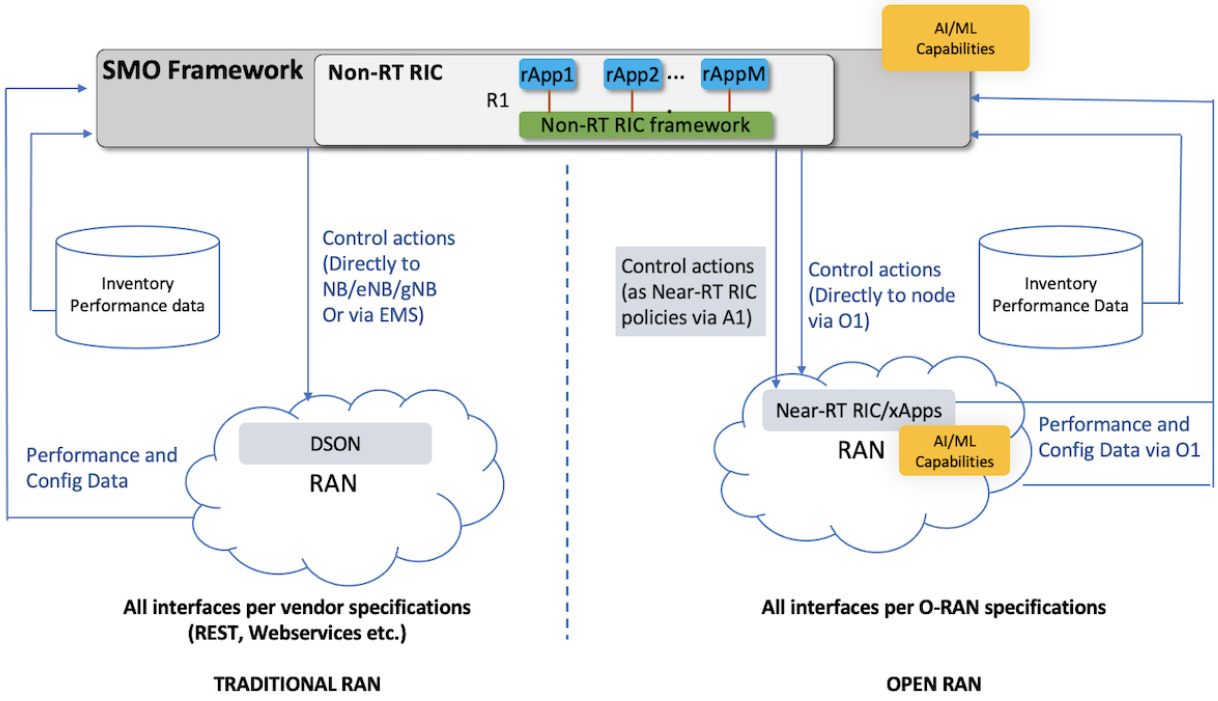


Figure 11 – RAN energy saving using O-RAN Architecture to support both traditional and open RAN

If an operator is early in the O-RAN adoption cycle (or even has no O-RAN adoption), they can develop custom interfaces between SMO and EMS or directly between SMO and the nodes (depending on the level of support from the RAN vendor) to reap the benefits of energy saving rApps right away by using the rApps in conjunction with the DSON from their vendors (see **Figure 11**). As MNOs evolve more towards O-RAN and introduce Near-RT RIC into their architecture, the energy saving benefits can be enhanced progressively by addressing more use cases such as ASM. The value provided by SMaRT-5G could be a great incentive for operators to accelerate adoption of O-RAN standards.

Note that except for ASM, the rest of the RAN energy saving approaches discussed in this paper can be implemented via Non-RT RIC/rApps. These rApps can work with DSON for traditional RAN as well as with xApps for open RAN.

xApps/rApps Features for RAN Energy Savings

In this section, we outline some basic requirements of RIC-based energy saving solutions based on xApps and rApps.

1. The implementation of the energy saving solution should be a cloud native application, since the SMO/Non-RT RIC and Near-RT RIC are cloud platforms. This provides lightweight architecture, high availability, and easy scalability and migration among other benefits.
2. The cell state suggestion model and associated functions should be separated from the RAN control part of the solution. This enables the operator to use a different model for better or more accurate state suggestions without having to change the RAN control part. This would promote functionality reuse in situations where the same kind of model may be required for a different RAN application.
3. The implementation should store topology information of the governed cells efficiently and accurately. The data structure used should represent the neighbor information along with various features necessary for the AI/ML model to make accurate suggestions. The implementation should be able to store, modify and delete a node's information quickly and correctly.
4. The RAN should provide all the useful KPIs per cell to effectively train and apply the AI/ML model for accurate suggestions. The RAN should comply with the ongoing work to specify KPIs for Energy Efficiency [4,10]. Performance metrics, QoS metrics and energy metrics are must-haves.
5. The energy saving solution should meet QoS requirements and ensure minimal impact on the QoS when a cell energy saving action is taken. For example, a graceful shut-off as described before is essential for this purpose.
6. It should be possible to adapt the solution for various RAN implementation scenarios with minimal or no changes.

Energy Savings in Mobile Core and Compute Power Optimization

When the RAN element configurations are modified to save energy based on network traffic as described in previous sections, the 5G Mobile Core capacity and the RAN/Core cloud resources can also be adjusted accordingly for energy savings. Assuming cloud-native implementations, the number of processor cores assigned to various tasks can be adjusted by scaling in or out of the cloud-native workload. Both reactive and predictive scale-out/in can be considered, using AI/ML to automate the predictive scenarios.

As RAN and Mobile Core workloads are virtualized and containerized to become CPU loads, there is an opportunity to dynamically adjust the compute and hence the power footprint to better match the real-time and projected needs of the mobile network. As the network traffic changes, the Mobile Core user plane (UPF) is most directly affected and is the focus of attention for energy savings. **Figure 12** illustrates this concept. We are considering two levels of compute optimization – (i) systems level which is scaling in/out of cloud resources for RAN and Core, and (ii) CPU level, which involves CPU optimization of resources supporting both RAN and Core.

The systems level compute energy savings is achieved by dynamic release/suspension of cloud resources via the FOCOM and NFO cloud orchestration mechanisms specified by the O-RAN Alliance [10]. The information about the resources which can be scaled out is readily available if the underlying topology of the cloud-native network functions is known. The specifications for the interactions within the SMO is ongoing work in the O-RAN Alliance [10] and there is increased interest in O-RAN SC and ONAP for a collaborative solution on O-Cloud management [19].

At the CPU level, for CPU optimization, the open standard Advanced Configuration and Power Interface (ACPI) provides a mechanism for CPU power management. ACPI specifications for power management are divided into two categories or states: P states and C states.

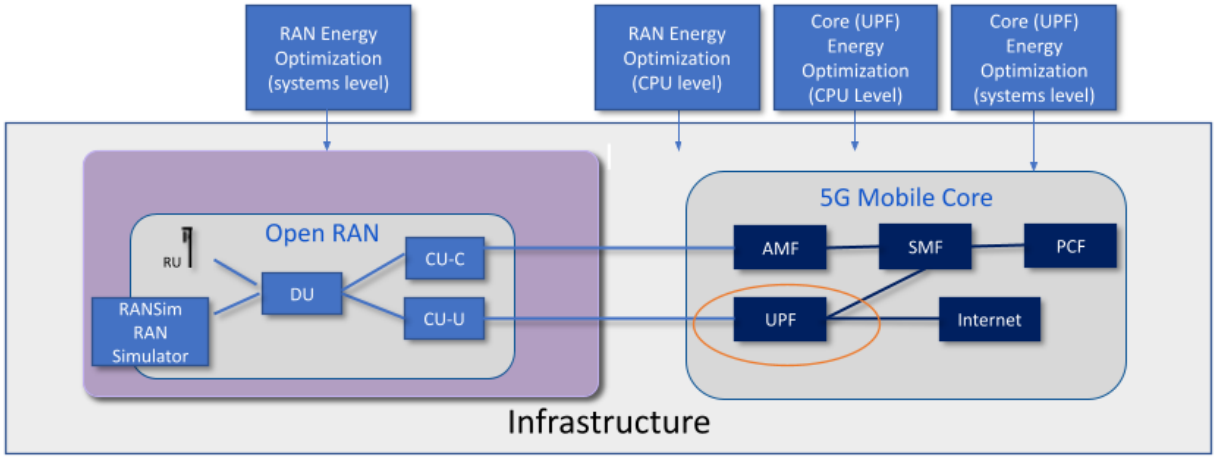


Figure 12 – Cloud resources and CPU power optimization for RAN and Core energy saving

Power performance states (P states) provide a way to scale the frequency and voltage of the processor to optimize power consumption (given that CPU power consumption varies linearly with the frequency and quadratically with the voltage). Note that this adjusts the entire CPU, and a closed loop can dynamically adjust the various CPUs throughout the mobile infrastructure to seek to balance capacity with demand.

Processor idle sleep states (C states) put selected functions of the CPU to sleep. Different processors support different numbers of C-states in which various parts of the CPU are turned off. Since the states available are CPU dependent, target systems should be carefully selected with pre-determination of the impact of various adjustments for various types of mobile workloads (RAN and Mobile Core are expected to behave differently for selected changes).

Different platforms offer a variety of hardware features to optimize. These features allow for precise tuning of e.g., core/non-core and base/turbo frequencies for specific CPU cores. As the configuration search space for setting up a system for performance (efficiency, throughput, latency, etc.) is large, we can use ML techniques to tune the CPU core's configurations statically and dynamically for the desired targets. There may be a need for optimization solutions to handle workload scheduling with advanced dynamic resource allocation technologies for improved resource utilization, reduced power consumption and reduced cost of ownership.

To enable more autonomy at the node level, modern AI/ML techniques are needed which support applications and resource owners. Algorithms such as Bayesian Optimization, Reinforcement Learning, Deep learning, etc., can be used to statically and dynamically allocate shared hardware resources (for example, cache line, memory bandwidth, and CPU

states) for different applications and improve application performance, enhance server utilization, and reduce energy costs.

Most recent work for resource allocation optimization can be divided into two general categories: search-based methods and reinforcement learning-based methods. Bayesian optimization-based search methods have been gaining attention for resource allocation [12][13][14][15]. Among reinforcement learning-based methods, deep Q-learning-based approaches are becoming popular [16] [17].

Figure 13 is an architecture for a Non-RT RIC/rApp based solution which encompasses the end-to-end RAN and Mobile Core energy management, which includes CPU power optimization.

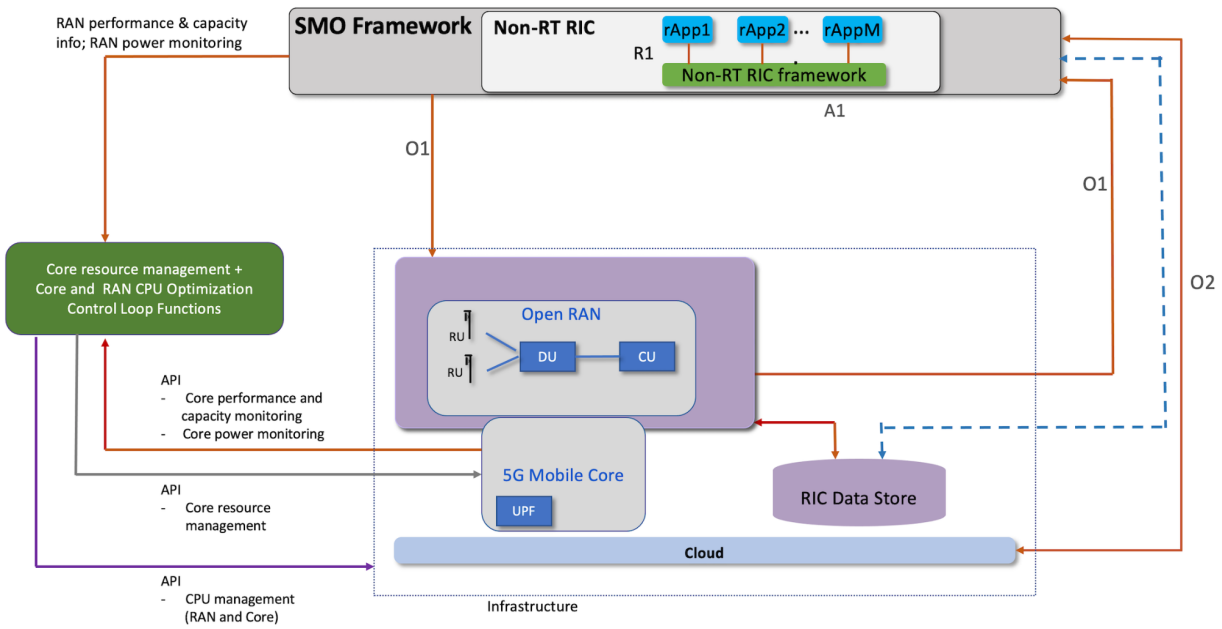


Figure 13 – End-to-end RAN and Core Energy Management

Comparing Figure 10 (RAN energy management architecture) and Figure 13, note the following enhancements:

1. Introduction of new control loops to perform Mobile Core performance/resource management, CPU management for both RAN and Core. These are not under the purview of the SMO which does however include the O2-based scale-in/scale-out management for the O-Cloud.

2. APIs to support these control loops, including External APIs to/from the SMO so that the rApps/Non-RT RIC/SMO can provide the network state information to the green box in Figure 13.

SMaRT-5G PoC Implementations

When Proof-of-Concept (PoC) implementations are designed to experiment with various concepts for energy savings under the auspices of SmaRT-5G, the following aspects are important to consider:

1. Follow open architecture: O-RAN and other relevant open standards like 3GPP and ACPI.
2. Use a phased approach aligned with MNO open network adoption so that they can progressively benefit from SmaRT-5G.
3. Support for both greenfield and brownfield networks.
4. Demonstrate PoC solutions both in open source and commercial configurations:
    a. Leverage exemplar open source software where possible for demonstrating capabilities, including SD-RAN, SD-Core, Aether, ONAP and contributions from the O-RAN Software Community (OSC). Open source implementations can enable a broad community of researchers to continue to advance the work.
    b. Selectively use closed-source commercial components and associated collaborations to prototype configurations that can be rapidly consumed by operators in production settings. The goal of the project is to create a rapid path to impact, and as such it is an explicit goal to create a commercial ecosystem to advance the technology.

If we look at the adoption of O-RAN standards for the RIC, the most mature interface is O1 and the early adoption among operators is around SMO, Non-RT RIC and rApps. This is because other entities like Near-RT RIC, xApps and supporting interfaces like A1 and E2 have a heavy dependency on vendor support.

Given these considerations, a series of PoC phases can be considered as shown in **Figure 14**.

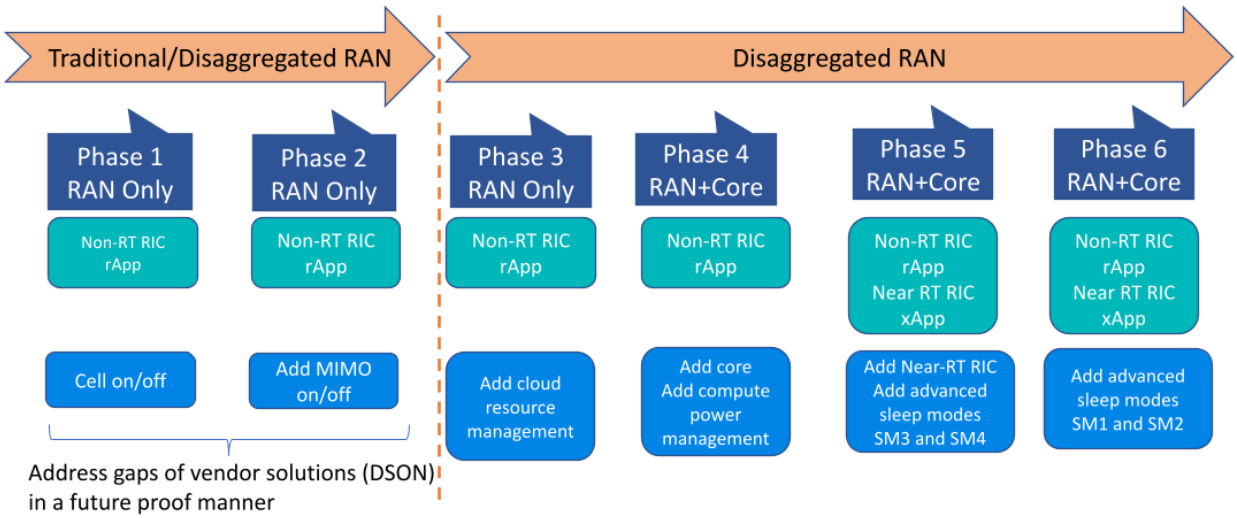


Figure 14 – SMaRT-5G PoCs: various phases

As can be seen, initially only RAN is considered (but for both traditional and disaggregated RAN). The PoC program starts with cell on/off, then MIMO on/off is added. Subsequently cloud resource optimization for RAN and then Mobile Core along with CPU power management is considered. Until this stage, the PoC would need only SMO/non-RT RIC. Finally, Near-RT RIC is introduced and ASM scenarios are considered.

The target architecture for the PoC is essentially as shown in Figure 13, with Near-RT RIC and associated interfaces added to support various ASMs. This is shown in **Figure 15**.

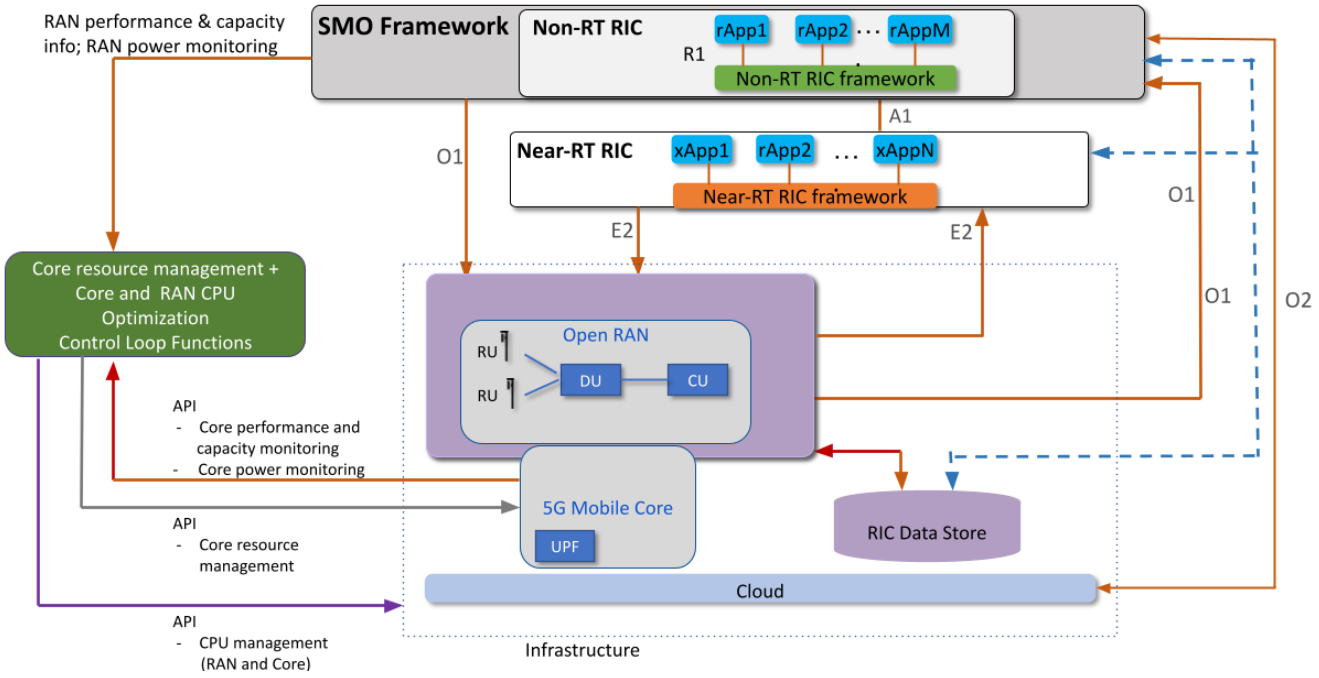


Figure 15 – PoC Target Architecture

RAN Simulator

Another important component that is essential for the SMaRT-5G PoCs is a RAN Simulator. A RAN simulator is needed to address the following:

- A simulator helps in testing the applications at scale. A "toy network" with a few network elements may not be sufficient for this purpose.
- It is high risk to try new ideas and technologies on live networks. A simulator should give ample faith to MNOs so they can test out new technologies before they are field trialed. This implies that a good simulator should be capable of showing actual network behavior as much as possible as it is controlled by the technology/algorithms being advanced in each PoC.
- The simulator should be able to encompass a full range of "what-if" scenarios including unusual circumstances, commonly known as "corner cases".

As an illustration, **Figure 16** is an adaptation of Figure 10, with a RAN simulator incorporated in the architecture.

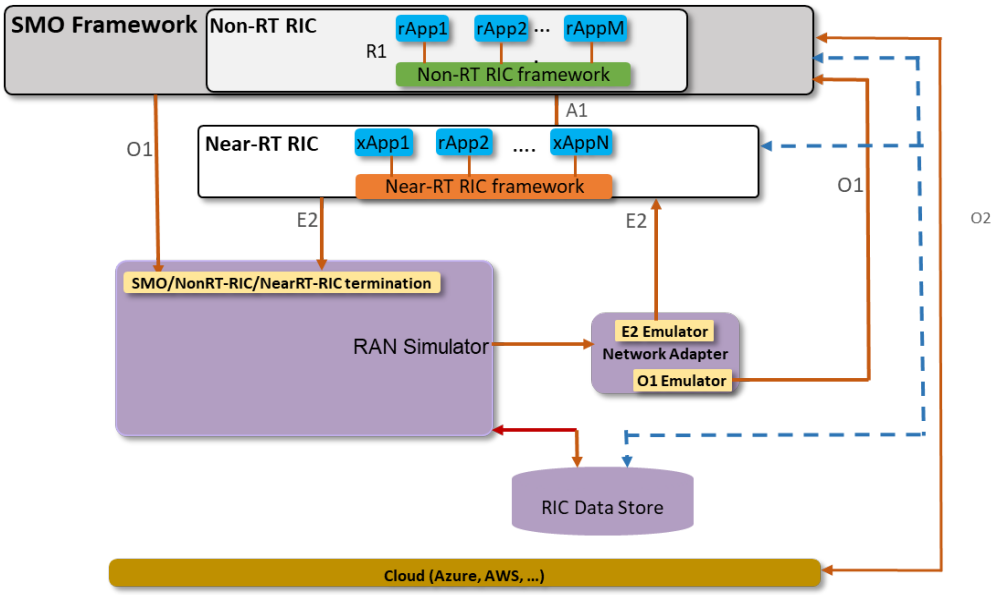


Figure 16 – Application of a RAN Simulator for energy saving studies

RAN Simulator can leverage open source work in O-RAN SC and the use of NS3, but other options can be considered as well including commercial options. For the PoC, the simulator should specially address the following capabilities (at least for Phases 1 and 2):

- Cell site configuration data
  - Sector information, carrier frequencies (UL/DL), Transmit power, azimuth, location coordinates, etc.
- Neighbor list and associated details
- RRC attempts, success and failures
  - Traffic volumes, UL/DL
  - PRB utilization, UL/DL
- Active and idle UEs
- Additional data to calculate performance KPIs (accessibility, retainability, throughput)
- MIMO configuration support (needed for Phase 2)
- Energy consumption KPIs (can also be estimated)
- Handover stats and KPIs
  - Incoming/outgoing handover attempts/failures

Bootstrapping the Open Source Version of the PoC

Let's consider how the PoCs (Figure 14) can be bootstrapped on an open source platform (note – commercial platforms are also being pursued in parallel, with the intention to demonstrate the sustainability use case on both open and commercial-grade platforms).

Phase 1 is a pure rApp play. Also, this phase does not require a Non-RT RIC since an A1 interface is not needed for this phase. Therefore, we can use the SMO based on lightweight ONAP/OSC components from the O-RAN software community.

As the first step, develop an Event Generator which can manually trigger VES events within the SMO framework as shown in **Figure 17a**. These events are triggers which can be used by rApp to create specific actions.

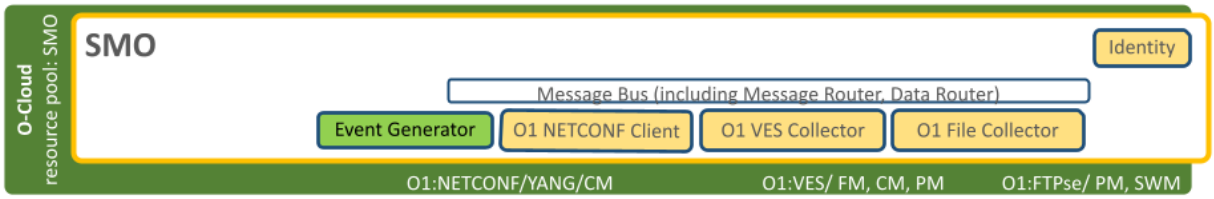


Figure 17a - Bootstrapping Phase 1 of the PoC

The next step is to develop an O1 simulator (O-DU and O-RU) and expose O1 YANG models (for configuration management (CM) as depicted in Figure 17a. For performance management (PM), an O1 simulator exposing sample real network data would be sufficient to start with, and it can be file based.

The next step is to develop a cell on/off rApp (just a microservice to start with), consuming the PM data and triggering CM to be performed by the O1 NETCONF Client.

Strictly speaking, without an R1 interface, it is non-compliant to call this application an rApp. But note that it is the first step towards a full-fledged rApp, and an R1 interface will be developed later.

Once this setup is working, an NS3 simulator (or a similar capability) with O1 interface to encompass the steps above can be used to carry out the PoC as shown in **Figure 17b**. Note that a manual event trigger is not needed at this point as the simulator would address that aspect.

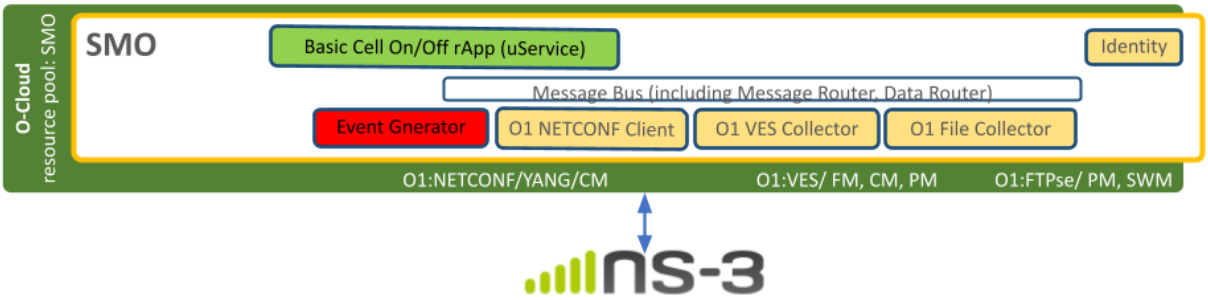


Figure 17b - Bootstrapping Phase 1 of the PoC with a simulator added

How MNOs Can Get Immediate Benefit from the PoC

The way in which the phases of the PoC are constructed is such that MNOs can benefit from each stage of the PoC in accordance with the level of O-RAN adoption they have progressively, in tune with the network evolution from a brownfield scenario (now) to a fully open RAN scenario (as shown in Figure 11). As an illustration see **Figure 18**:

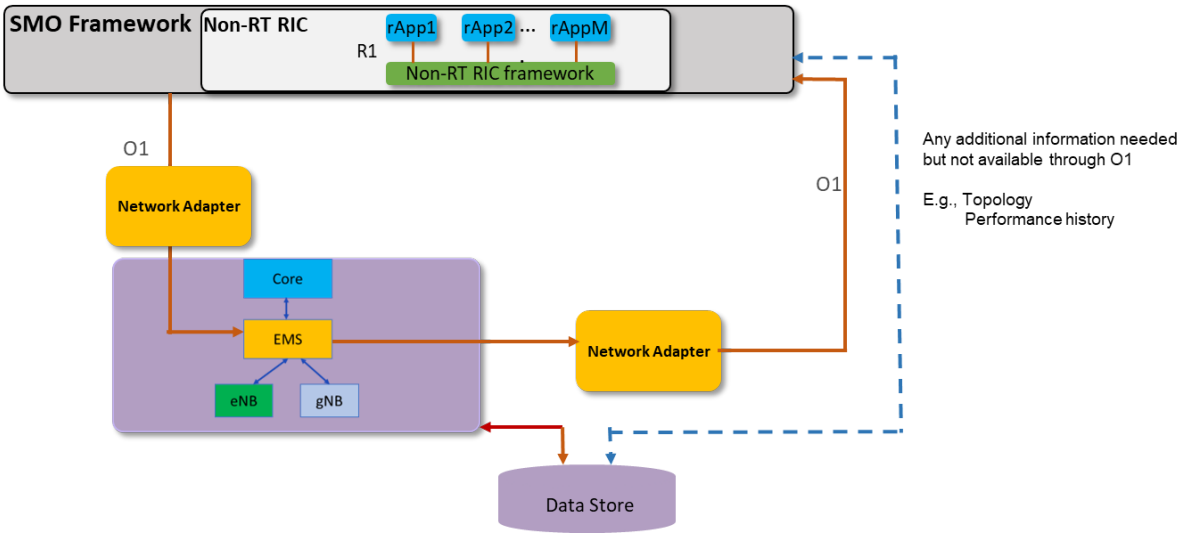


Figure 18 – Cell on/off and MIMO on/off using O-RAN architecture for brownfield networks

In Figure 18, the rApp controls the DSON in brownfield networks via EMS. For this MNOs need to develop two adaptors – one to gather PM data from the EMS and to push this data to SMO via O1 and the other one to implement parameter changes in eNB/gNB via EMS

through O1. Of course, the adaptors are vendor dependent and could be using REST or Webservice interfaces with the EMS.

If/when the MNO has an O-RAN compliant network, the adaptors can be dropped and direct O1 connectivity can be used to implement energy saving solutions.

Conclusion

RAN energy savings is an important use case for mobile network operators, and for the world. Open RAN provides efficient technological possibilities to address this important problem. ONF's Sustainable Mobile and RAN Transformation 5G (SMaRT-5G) project is a collaborative effort to develop and demonstrate an ML-driven, intelligent energy saving solution for traditional and open RAN mobile networks. Given the slow adoption of open RAN standards, it is important to devise solutions that are equally applicable to traditional RAN as well as open RAN to have impact now and enhanced impact as open RAN adoption progresses. To achieve the energy savings objectives while fostering open RAN adoption, technology trials are essential among operators, system integrators, application developers as well as platform providers. ONF presents an excellent industry forum to promote such an ecosystem, and ONF is pursuing this agenda in collaboration with industry partners in the context of the SMaRT-5G project, implementing successively more powerful PoCs on both open source and commercial RAN stacks.

References

1. D. Chen. “5G Power: Creating a green grid that slashes costs, emissions & energy use.” July 2020, Huawei. https://www.huawei.com/en/huaweitech/publication/89/5g-power-green-grid-slashes-costs-emissions-energy-use#:~:text=During%20service%20troughs%2C%20the%20power,efficiency%20by%20around%209%20percent.

2. R. Lee, D. Pinner, K. Somers, and S. Tunuguntla. “The case for committing to greener telecom networks.” February 2020, McKinsey & Company. https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/the-case-for-committing-to-greener-telecom-networks

3. M. Dano. “How AT&T’s network chief hopes to cut a $1.6B electricity bill.” September 2022, Light Reading. https://www.lightreading.com/climate-change/how-atandts-network-chief-hopes-to-cut-$16b-electricity-bill/d/d-id/780237

4. O-RAN Work Group 1 (Use Cases and Overall Architecture) Network Energy Saving Use Cases Technical Report (O-RAN.WG1.NESUC-R003-v02.00). O-RAN Alliance, 2023. https://orandownloadsweb.azurewebsites.net/specifications

5. “Ericsson Breaking the Energy Curve Report 2022: 5G network success can be achieved in an energy-efficient way.” October 2022, Ericsson. https://www.ericsson.com/en/news/2022/10/ericsson-publishes-breaking-the-energy-curve-report-2022

6. “Going green: can 5G be energy efficient?” September 2021, GSMA Intelligence. https://www.gsmaintelligence.com/event/going-green-can-5g-be-energy-efficient/

7. M. Hoffmann, M. Dryjanski. “The O-RAN Whitepaper 2023: Energy Efficiency in O-RAN.” Rimedo Labs, 2023. https://mailchi.mp/0afa75992d20/the-o-ran-whitepaper-2023-energy-efficiency

8. Z. Umar. “Base Station Sleep Modes to Trade-Off Energy Saving and Performance in 5G Networks.” March 2020. Politecnico di Torino. https://webthesis.biblio.polito.it/14459/1/tesi.pdf

9. “Management and Orchestration: Study on new aspects of Energy Efficiency (EE) for 5G.” 3GPP TR 28.813. April 2020, 3GPP. https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3743

10. O-RAN Alliance Specifications. See especially ongoing work for the MVP-C Energy Savings Feature in WG1, WG2, WG3, SuFG and AI/ML Framework specifications. https://orandownloadsweb.azurewebsites.net/specifications

11. V. Singh, M. Gupta and C. Maciocco. "Intelligent RAN Power Saving using Balanced Model Training in Cellular Networks." 2022 20th International Symposium on Modeling and Optimization in Mobile, Ad hoc, and Wireless Networks (WiOpt). Torino, Italy. 2022, pp. 357-364. https://ieeexplore.ieee.org/document/9930603

12. T. Patel and D. Tiwari. "CLITE: Efficient and QoS-aware co-location of multiple latency-critical jobs for warehouse scale computers." International Symposium on High-Performance Computer Architecture (HPCA), 2020. https://ieeexplore.ieee.org/document/9065583

13. P. Mercati, B. Li, M. A. Ergin, C. Tai, M. Kishinevsky, B. Serafimov, S. Ravisundar, E. Walsh and T. Long. "MOBO-NFV: Automated tuning of a network function virtualization system using multi-objective bayesian optimization." 2021 IFIP/IEEE International Symposium on Integrated Network Management (IM), 2021. https://ieeexplore.ieee.org/document/9463999

14. J. Sydir, B. Li, P. Mercati, C. Tai, R. Iyer, M. Kishinevsky and B. Serafimov. "DPM-NFV: Dynamic Power Management Framework for 5G User Plane Function using Bayesian Optimization." IEEE Global Communications Conference, 2022. https://ieeexplore.ieee.org/document/10001394

15. Q. Li, B. Li, P. Mercati, R. Illikkal, C. Tai, M. Kishinevsky and C. Kozyrakis. "RAMBO: Resource Allocation for Microservices using Bayesian Optimization." IEEE Computer Architecture Letters, 2021. https://ieeexplore.ieee.org/abstract/document/9380428

16. B. Li, Y. Want, R. Wang, C. Tai, R. Iyer, Z. Zhou, A. Herdrich, T. Zhang, A. Haj-Ali, I. Stoica and K. Asanovic. "RLDRM: Closed loop dynamic cache allocation with deep reinforcement learning for network function virtualization." IEEE Conference on Network Softwarization (NetSoft), 2020. https://ieeexplore.ieee.org/document/9165471

17. R. Nishtala, V. Petrucci, P. Carpenter and M. Sjalander. "Twig: Multi-agent task management for colocated latency-critical cloud services." International Symposium on High-Performance Computer Architecture (HPCA), 2020. https://ieeexplore.ieee.org/document/9065442

18. Fatma Ezzahra Salem, Tijani Chahed, Eitan Altman, Azeddine Gati, and Zwi Altman. “Optimal Policies of Advanced Sleep Modes for Energy-Efficient 5G networks.” Cornell University, 2019. https://arxiv.org/abs/1909.09011

19. J. Keeney, S. Mudiganti, T. Perala, N.K. Shankar, M. Skorupski. “ONAP and O-RAN SC.” O-RAN Software Community, March 2023. https://wiki.o-ran-sc.org/download/attachments/3604609/ONAP_OSC_Areas_20230321.pdf?api=v2

20. M. Hoffmann. “O-RAN Network Energy Saving: RF Channel Switching.” Rimedo Labs, February 2023. https://rimedolabs.com/blog/o-ran-network-energy-saving-rf-channel-switching/

About ONF

The Open Networking Foundation (ONF) is an operator-driven, community-led nonprofit consortium fostering and democratizing innovation in software-defined programmable networks. Through ecosystem building, advocacy, research, and education, ONF is accelerating the state-of-the-art in open networking and catalyzing creation and adoption of open disaggregated solutions leveraging open source software.