SMaRT 5G Conceptual Overview

Sustainable Mobile and RAN Transformation (SMaRT)

Open Networking Foundation

Authors
T. Sloane, L. Peterson, S. Ananmalay, ONF
J. Kaustubh, R. Savoor, AT&T
[C. Coletti, P. Lédl, Deutsche Telekom]
M. Dryjański, Rimedo Labs
C. Maciocco, Intel

October 2022 - Version 0.9

Executive Summary

Mobile infrastructure is the source of a concerning fraction of the world's total power consumption. Energy costs for telecom operators are substantial, accounting for upwards of 5% to 7% of operating expenditures[1],with the mobile radios access network (RAN) making up a significant component of this expense[2]. And energy costs are growing as a result of exponential traffic growth and new 5G services, with each 5G site requiring two to three times more power than its 4G-equivalent.

In addition to operating costs, the threat of climate change is compounding the growing urgency of reducing the carbon footprint of mobile networks. With telecom operators already accounting for 2 to 3 percent of total global energy demand,[3] telco operators and vendors are under ever increasing pressure to take demonstrable steps towards minimizing their environmental impact.

There is a clear and pressing need to optimize power utilization in mobile networks to balance the economic value these networks bring with the cost and environmental impact they impose. There is also a significant ROI opportunity if operators can reduce their power consumption, creating additional incentive and urgency for this work.

---

[1] 2018 numbers from 2020 McKinsey study: https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/the-case-for-committing-to-greener-telecom-networks

[2]https://www.researchgate.net/figure/Energy-consumption-composition-of-a-mobile-operator-43-Power-aware-routing-in-sensor_fig5_272864748

[3] https://www.gsma.com/futurenetworks/wiki/energy-efficiency-2/

The mobile network, by its very nature, is dynamic. Traffic demands are constantly changing due to weather, holidays, public events, remote work, day of the week, time of day, and seasonal changes. There is ample opportunity to build optimizations into the mobile network both to dynamically respond to changing demands, but also to forecast shifts in demand to proactively tune the network ahead of changing traffic conditions. And by dynamically tuning the network, there is an opportunity to optimize power consumption throughout the mobile infrastructure.

The move towards disaggregating and opening the mobile network (through both open RAN and the cloud-native mobile core) has opened the door to rapid innovation. As more interfaces are exposed there is more and more opportunity to select/adjust/modify/control various components of the network to optimize for various outcomes. The goal of this project is to experiment with various techniques to dynamically modify the mobile infrastructure, including by leveraging AI/ML control, to dynamically balance power consumption against user quality-of-experience (QoE) in an end-to-end functioning deployment (lab or otherwise).

This project being launched by ONF is exceptionally well aligned and squarely targeted at operator and industry priorities, including O-RAN's[4] top 4 of 8 high priority items, including: Energy Efficiency, Traffic Steering, QoE Optimization, and QoS Based Resource Optimization.

This project will create an end-to-end open source platform ideally suited to help advance operators call for a focus on Energy Efficiency.[5] APIs will be defined and implemented to enable software-defined control of dynamic power tuning, open source will be seeded into the industry to enable vendors to more easily and rapidly incorporate power efficiency optimizations into their product offerings, and the open platform will further empower the broader community to continue to advance this essential work.

---

[4] Open RAN Technical Priority document:
https://assets-global.website-files.com/60b4ffd4ca081979751b5ed2/623dac2e83e05152be1b6021_Open%2BRAN%2BTechnical%2BPriority%2BDocument%2B-%2BRel2%2B-%2BFV.xlsx

[5] OPEN RAN TECHNICAL PRIORITIES, Focus on Energy Efficiency:
https://assets-global.website-files.com/60b4ffd4ca081979751b5ed2/624153341a4f88926a6ed698_Open%2BRAN%2BTechnical%2BPriorities%2B-%2BEnergy%2BEfficiency%2B-%2BFV.pdf

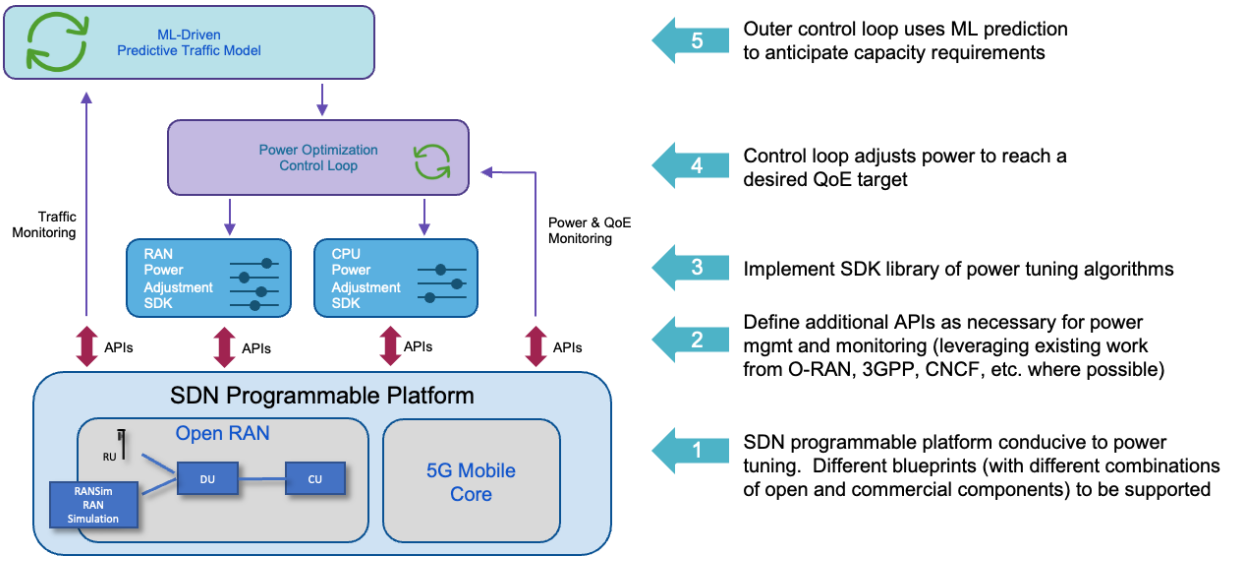
Table of Contents

Architecture

The plan is to assemble an open SDN-programmable 5G mobile platform that allows for dynamic control of the power consumption of the end-to-end 5G network (including both RAN and core).

We will then create a control loop that accepts traffic capacity and a QoE as target goals, and adjusts the power to try to match the goal, tracking the result in real-time for feedback. We call this the inner control loop.

To supplement this, an outer control loop will be created using ML models that learn from historical traffic patterns to predict capacity needs. Anticipated capacity requirements can then be provided as inputs into the inner control loop in order to dynamically adjust power ahead of changes in demand in order to predictively maintain the desired QoE.

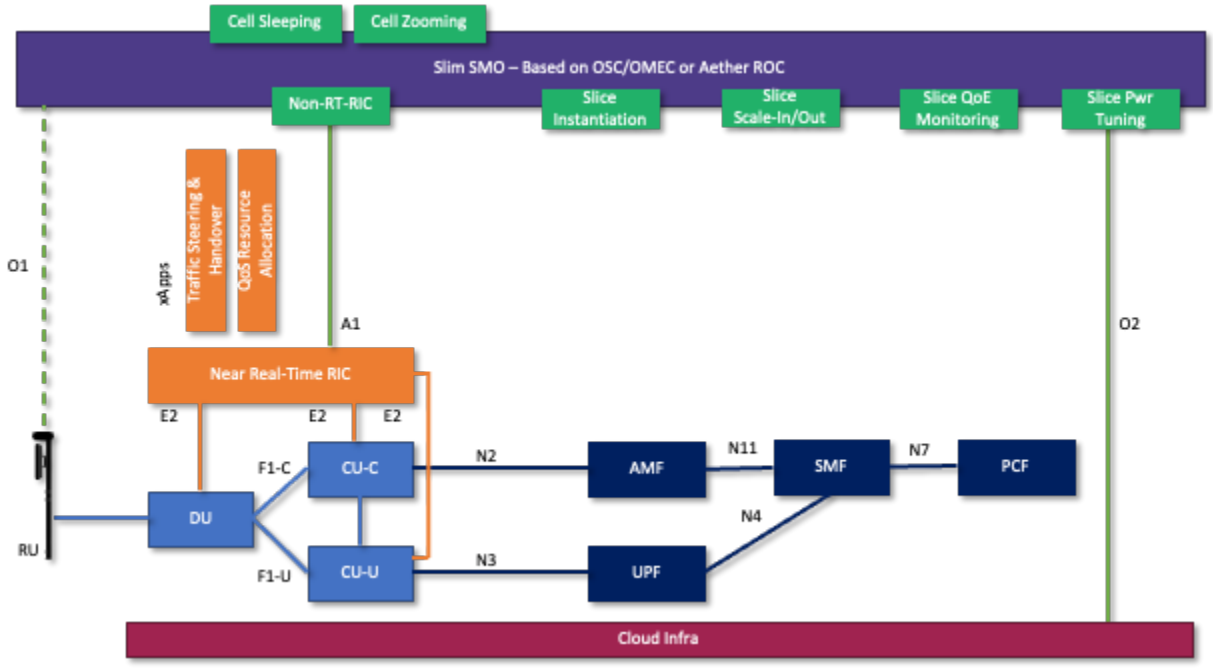


Power Tuning Optimization Strategies

We believe there are two broad strategies to explore for controlling power consumption: **optimizing CPU utilization** throughout the RAN and Core infrastructure and **optimizing power consumed by RAN Radios**. It is expected that the two forms of optimization will interact with each other, and that it is not practical to independently optimize each domain in real time fashion (because, in effect, a feedback loop is likely to resonate throughout the end-to-end system). For this reason, this project will explore optimizing both domains in a holistic way on an end-to-end network, working towards delivering just-in-time capacity to deliver a target QoE for users with optimal power efficiency.

CPU Power Optimization

As RAN and Mobile Core workloads are virtualized and containerized to become CPU workloads (e.g. DUs, CUs, RIC and x/rApps, and as cloud-native workloads for the mobile core), there is promising opportunity to dynamically adjust the compute and power footprint to better match the realtime and projected needs of the mobile network. But this is a complicated system, and we anticipate that adjustments in one component will ripple through the system. For this reason, we are proposing a holistic approach for this research, using feedback loops to guide dynamic tuning of a complete end-to-end system.



All elements of the RAN and Mobile Core (other than the RU and L1 functions in the DU) are predominantly software based, and thus promising targets for CPU optimization.

CPU optimization strategies that will be pursued include:

X86 P-State	Adjust voltage and clocks in response to workloads and demand. This adjusts the entire CPU, and a closed loop system can dynamically adjust the various CPUs throughout the mobile infrastructure to seek to balance capacity with demand.
X86 C-State	Reduce or turn off selected functions of the CPU. The states available are CPU dependent, so we will carefully select target systems in collaboration with vendor partners, and then experiment with the impact of various adjustments for various types of mobile

	workloads (RAN and Core are expected to behave differently to selected changes).
Cloud workload scale-out	The number of processor cores being applied to a task can be adjusted by scaling out a cloud-native workload (adding parallel instances of a service). Well designed implementations of mobile core and mobile RAN should support dynamic scale-out (and scale-in), and this project scope includes optimizing implementations to support scale-out and then researching how dynamic scale-out can be applied to adjust the system to support dynamic traffic patterns (thus optimizing power utilization). Both reactive and predictive scale-out/in will be explored, ultimately using AI/ML to automate the predictive scenarios. RAN pooling/hubbing[6] may also be employed to enable the scale-out of the DUs in the RAN.

RAN Power Optimization

RAN optimization strategies can play another significant role in optimizing power consumption[7]. We proposed to experiment with the dynamic use of RAN components to adjust capacity in various ways, including:

Sleeping strategies	Switching off network components in low traffic conditions. May involve sleeping particular components within RUs (e.g. power amplifiers or cooling equipment) or switching off RUs entirely. Sleeping can also occur at various depths and over various time periods. For example, an approach referred to as discontinuous transmission (DTX) powers-down RAN components during idle periods in the millisecond range. With O-RAN's decoupling of control and data functions, there is the possibility of sleeping discrete DU and/or CU functions, or the RIC and xApp functions as well. But the initial hypothesis is that these are largely software-based elements, and optimizing CPU utilization on these components is the more effective way to optimize power consumption.

[6] https://about.att.com/innovationblog/2022/cloudifying-5g-with-elastic-ran.html

[7] Partially sourced from: https://reader.elsevier.com/reader/sd/pii/S1364032121012958?token=1A82B825B3720DF77791545CF4F8985FE359FA54D298F034CAD2BB51535396D2012ACF253E4718D55813A7A5C41D6984&originRegion=us-east-1&originCreation=20220818215535

Cell zooming[8]	Cell zooming involves adjusting RUs dynamically to alter coverage area and RU transmit power based on the location and Quality of Service (QoS) requirements of users. Cell zooming is often proposed in combination with sleeping strategies in order to fill coverage holes caused by switching off an RU, or to balance traffic across the network in order to maximize the scope for RU sleeping. We propose coupling this with active Traffic Steering and Handover. For example, when there is a decision to switch off a carrier or decrease/increase power, users shall first be proactively moved to different cells in a coordinated manner so that user QoE is maintained.
High Frequency Overlay Optimization	In multi-carrier scenarios (e.g. an RU serving multiple component carriers with, for example, one cell at 800 MHz, another at 2.6 GHz and yet another at 3.5 GHz), we can power down the higher bands when traffic demands are moderate, leaving the lower frequencies powered up to provide coverage.
Density and Overlays	Dense Heterogeneous Networks (HetNets) involve different sizes of RUs serving the same geographical area (usually urban hotspots), most typically a layer of macro RUs with a number of smaller RUs (e.g. microcells) within the footprint of the macro RUs. In combination with sleeping strategies, this strategy saves energy by serving traffic hotspots more efficiently through lower power micro-RUs and shorter transmission distances.
Interference Control	Control of interference is key to energy efficiency. A given data throughput implies a required ratio of signal to interference level at the receiver. If interference is reduced, the transmitted power level can also be reduced while maintaining performance.[9]

[8] Also referred to as “Energy Saving Management” by 3GPP in the 4G era

[9] http://www.mobilevce.com/green-radio

Project Components: Tracks of Activity

We plan to approach this research as an agile software project. Additionally, we will pursue nine tracks of parallel activity that together create a set of building blocks that will allow us to create a holistic end-to-end mobile testbed ultimately instrumented with ML-driven control.

Part of the system will be focused on an SDN-enabled infrastructure testbed where power consumption can be controlled and the resulting network performance (and user QoE) can be measured.

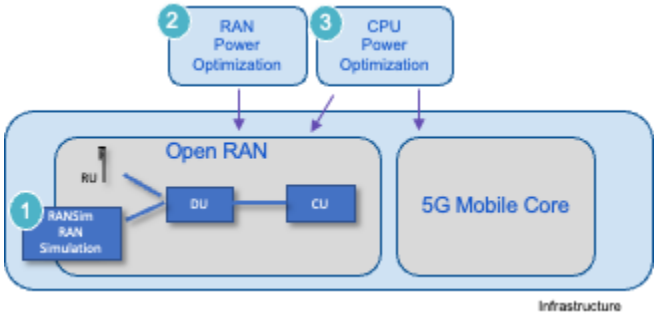
On top of the platform a control loop will be created whereby the power levels across the network can be tuned to meet a target QoE for an established traffic load.

In parallel, an ML engine will be developed capable of predicting network traffic loads. The ML engine will be used to create an outer control loop, ultimately driving the inner control loop with predictive targets towards which the network should be optimized.

We plan to pursue the project building up from the platform, adding the inner control loop, and then wrapping it with the outer control loop. The following describes the components and the corresponding nine tracks of activity.

Platform

This work will leverage the open source Aether end-to-end programmable 5G platform and its component projects SD-RAN, SD-Fabric and SD-Core. This provides an SDN programmable platform well suited for the needs of the project.



The Aether platform will be augmented to provide all the enabling capabilities to pursue this work:

1. **RAN Simulation:** We believe (at least initially) that we can pursue much of this work using RAN simulators. ONF's RANSim will be enhanced to simulate fluctuating traffic conditions across multiple gNBs. This should simulate the decrease of traffic to trigger power reduction, as well as increases as load resumes after quieter periods. From there, modeling environmental attributes such as weather, holidays, public events,

remote work, day of week, time of day, and seasonality can be used to augment the model in more sophisticated ways.

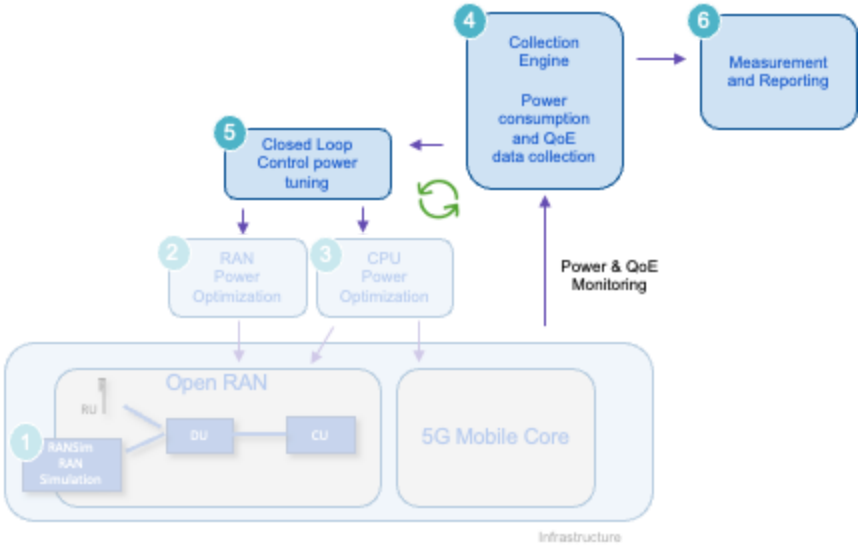
2. **RAN Power Optimization Module (SDK/Algorithms):** Logic to implement the various RAN power optimization strategies, including sleeping strategies, cell zooming, and quiescing high frequency overlays.
3. **CPU Optimization Module (SDK/Algorithms):** The end-to-end system will be dynamically controllable to shrink (and grow) the number of instances of functional components where possible (cloud-native scale-in/out), and also to control the CPU frequencies (P-State and C-States) throughout the end-to-end system.

Avoiding Dependency on Full O-RAN Compatible Platform

We are strategically structuring the project for rapid-time-to-impact, which we define as a measure of how quickly operators can begin deploying power optimized solutions procured from their vendors. We consider it strategic for the capabilities being developed in this project to rapidly work their way into commercial offerings from incumbents. so we don't want to limit the impact of this work to networks that have deployed the full O-RAN architecture. For this reason, incumbent RAN solutions are considered in-scope for this effort. We intend to work with the ecosystem to develop APIs that can work across RAN implementations, and we expect to work with both closed and open source components in parallel so that we can help speed the commercial availability of production-quality offerings.

Inner Control Loop

On top of the platform, an inner control loop will be developed to drive the power optimization modules, and measure current power and user QoE and control power levels throughout the mobile infrastructure (both RAN and Core). This will make it possible to continuously adjust the power to meet QoE and traffic capacity targets.



Control loop components that will need to be developed include:

4. **Network Data Collection Engine:** The power consumption of all mobile infrastructure components will be instrumented, collecting data through enhanced O-RAN and platform interfaces (Kubernetes, CPU and otherwise). This will allow us to correlate power consumption with traffic conditions and user QoE.

5. **Closed Loop Power Tuning:** A closed loop system will adjust the power profile of RAN and Core configurations to align with forecast traffic loads to the targets determined by its northbound counterpart.

   This control loop will provide the ability to experiment with different prioritizations for trading off power optimization and QoE (e.g. for certain network slices (workloads) it might be preferable to deliver lesser performance in exchange for power savings even if QoE suffers). This will allow us to experiment to determine how great a reduction of power utilization is possible in the various system components while still maintaining a target QoE for a network slice (workload).

6. **Measurement and Reporting:** Methods for measuring and recording both power savings and QoE for each network slice will be instrumented. Results will be recorded, logged and analyzed, with the goal of documenting various scenarios that balance power consumption vs. QoE for different goals.

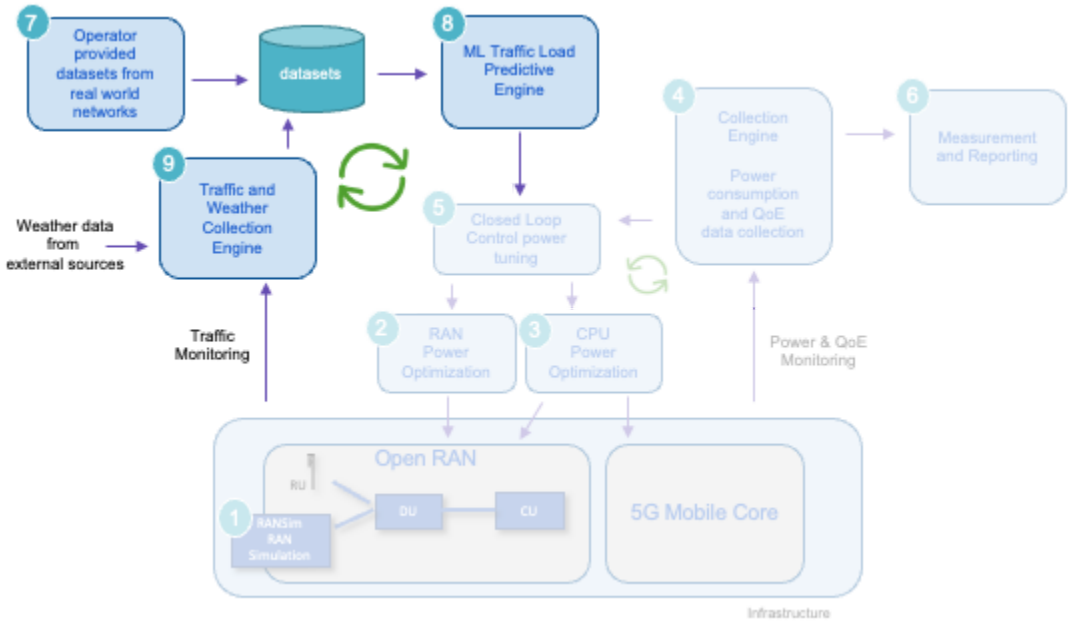
   Spin-up and reaction speed will be tracked and tested, providing data on how much buffer/headroom should be maintained to deliver desired QoE, and how tight a loop should be implemented to optimize power savings and QoE. For this work, we anticipate a need to measure:

   - Power saved vs. baseline
   - Time to spin down capacity, tracking QoE throughout

- Time to spin up capacity, tracking QoE and any negative impact of slow spin-up on QoE.

Outer Control Loop

The inner control loop will be supplemented with an outer control loop driven by an ML-engine predicting traffic loads. We will use real operator data to build and test the learning model, but ultimately we will create an outer control loop that continuously learns from the network itself to repeatedly update the ML model - feeding predictions into the inner control loop.



Components needed for the outer control loop include:

7. **Real World Datasets from Operators:** We propose collaborating with the operators backing this project to collect (and sanitize/anonymize?) real world datasets from production mobile networks in order to train, test and hone initial ML models. Ultimately, this training data will not be used in operating networks, as real data from the network itself will be used to generate predictions. This also helps identify which data points are most indicative of pending workload increases or decreases. Sample data will allow the ML work to progress in parallel with other efforts.

8. **AL/ML-Based Traffic Load Predictive Modeling:** The AI/ML predictive engine will be developed to learn from historic traffic datasets, and then predict traffic demand. We will train this engine both against the collected samples from operators' real networks, and ultimately from data collected from the operating platform once all the components are assembled. We ultimately envision for this model to be able to correlate traffic patterns with external factors such as weather and special events to make more accurate predictions.

- Intel Smart Edge is interested in this collaboration, and is prepared to provide the AI/ML platform to support this work
- Intel Labs has a potential contribution running on Smart Edge that today predicts mobile RAN traffic patterns based on historical data. This work would be leveraged and expanded upon as part of this project.

9. **Building New Datasets From Operational Prototypes and Trials:** In the end state, we plan to collect data from the running network and use it to continuously train/update the ML model. We will collect traffic patterns and external data that are likely to impact traffic (e.g. weather). This will create new datasets that can be fed into the ML-engine, and in turn enable dynamic prediction of traffic demand on an operating network. This will complete the outer closed loop that allows a network to learn over time about its own behavior in relation to external factors, while the inner closed loop continues to optimize the network power load to achieve the desired QoE.

Expected Outputs

Exemplar Implementation in Open Source

We plan to build on the SD-RAN, SD-Core, Aether and OAI open source platforms, and to release a version of the work to open source for the benefit of the ecosystem.

Commercial Prototypes

In parallel with the open source, we want to see commercial vendors implement the APIs on their platforms and demonstrate the control loops running on their commercial RAN and/or core offerings.

Extending O-RAN Service Models

We anticipate developing and implementing prototype service models to expose what is necessary to make well informed decisions, and to expose the control interfaces needed to adjust the mobile infrastructure. All Service Models will be open sourced and contributed back to O-RAN for standardization. Service Models will likely be needed to:

- Measure and track power consumption
- Control power states of the various RAN components

APIs for Control and Cross-Domain Coordination

This work will leverage existing APIs where they exist, and will define new APIs as necessary to enable the ability to control the layers (RAN and core) and coordination of actions that that span multiple layers (e.g., RAN-and-core, or RAN-and-cloud, core-and-cloud, etc.). These APIs will be made available through open source, which may lead to standardization via 3GPP/O-RAN standards.

Accelerating Market Impact via Vendor Coordination

The goal of the work is to help vendors bring optimizations for power efficiency to market as quickly as possible. The work will be done in coordination with the vendor ecosystem, commercial vendor components will be used wherever possible, and outcomes will be published in such a way as to ensure that vendor/commercial implementations can reproduce the results of this work.

Next Steps

We are in the formation stages of this project, building on the ONF's operating model to launch and drive this effort. The ONF's model anyways anchors projects with operator support and backing, and then builds a consortium of aligned parties to pursue the work.

We have assembled an initial consortium of operators who have come together to first define the direction, verify the use cases, and commit to trialing the results of the work. Operators will support the project both financially through directed grants to ONF, with staff to help guide the project to ensure we are collectively delivering an end result that is of value to the operators, and lab/trial support to verify and validate the work.

The consortium is currently recruiting a **select number of vendor partners** to engage as peers in the project. Vendors are expected to contribute resources (engineering and financial support) and commercial components to be assembled into the solution. Contributed products can be closed-source (as long as open APIs are supplied) or open sourced into the project.

ONF will be dedicating a team of engineers to this work funded by the project, pairing this additional talent with the existing ONF team to create the core team to drive the work. As with all ONF projects, this core team will serve as a catalyst engaging with engineers from all the participating companies working together in an agile manner, with operators serving as the product owners to ensure the work continues to deliver against the requirements of the ultimate end-users of the work.

Summary

We believe this project, as an open source exemplar implementation, will significantly advance and accelerate the move towards more power efficient RAN and mobile infrastructure.

Advancing sustainability is a top priority for mobile operators as pressure mounts both to optimize their OPEX and worldwide concerns around global warming continue to mount. We expect to learn as we proceed to demonstrate what is possible in the realm of power optimization. We will seed the industry with open source that vendors can leverage to more rapidly and cost efficiently advance their own product offerings, thus invigorating a pipeline of innovation that the operators will be able leverage to their benefit. And we will create an open platform for future research, allowing other researchers to build on the work.

There is a clear and urgent need to make mobile networks more energy efficient, and ONF's track record tells us that the proposed approach will deliver both a rapid route to impact as well as an important driver towards making open interoperable implementation commercially available. We encourage you and your organization to consider joining us as a founding member of this important initiative.