IEEE

IEEE INGR

International Network Generations Roadmap

2022 Edition

Energy Efficiency

IEEE Future NETWORKS

An IEEE 5G and Beyond Technology Roadmap

futurenetworks.ieee.org/roadmap

Table of Contents

Tables

Figures

ABSTRACT

This 2021 Edition of the IEEE International Network Generations Roadmap (INGR) contains a new Chapter dedicated to Energy Efficiency, which builds upon the initial white paper released in April 2020 [1]. For this purpose, the Energy Efficiency Working Group developed an analysis of the energy efficiency constraints across the whole ecosystem of the Fifth Generation "5G" and following network infrastructure, which can be leveraged by all stakeholders to prioritize resources allocation and technology development to ensure that both technical and economic forecasts can be met. The complexity of the ecosystem and the traditionally siloed approach within the Industry has often prevented the adoption of a holistic approach to addressing the fundamental problem of energy, which is the ultimate constraint to any complex deployment. The proposed framework facilitates an assessment of bottlenecks and their implication on the network: it may be used by both academic and industry stakeholders to develop solutions that address the real issues and enable a healthy ecosystem.

After a comprehensive survey of the ecosystem and its challenges, the following key areas were selected for a more in-depth analysis:

- Network Efficiency
- Small Cell Migration
- Base Station Power
- Economic Factors
- Grid/Utility

This Chapter also identifies the need for a comprehensive "Systems-of-Systems" (SoS) analysis to address the complex inter-relations among the multiple layers, which the infrastructure leverages. An initial proposal describes how a model can be built to enable a comprehensive assessment of energy requirements across such a diverse ecosystem. A future step in the process will consolidate a proposal for standardization of this model, which can be utilized by all stakeholders for both analysis and forecasting of capabilities and return on investment.

Key words:

Energy Efficiency, 5G Energy Gap (5GEG), Power Value Chain (PVC), Power Cost Factor (PCF), Systems of Systems (SoS), Energy Harvesting (EH), Sustainable Power, Embodied Energy, 5G Economic Gap (5GEcG), 5G Equality Gap (5GEqG), 5G Derate Factor (5GDF), Assessment Framework

CONTRIBUTORS

Brian Zahnstecher, PowerRox (Co-chair)

Earl McCune, Eridan Communications (***IN MEMORIAM, RIP***)

Doug Kirkpatrick, Eridan Communications

Rick Booth, Eridan Communications

Kirk Bresniker, Hewlett Packard Enterprise

Lin Nease, Hewlett Packard Enterprise

Anirban Bandyopadhyay, Global Foundries

Bruce Nordman, Lawrence Berkeley National Laboratory

Francesco Carobolante, IoTissimo LLC (Co-chair)

Magnus Olsson, Huawei

Emil Björnson, KTH Royal Institute of Technology

Laurence McGarry, pSemi/Murata

Steve Allen, pSemi/Murata

Paul Draxler, Stonecrest Consulting

Frederica Darema, formerly of NSF and AFOSR

Mohamed-Slim Alouini, KAUST

INGR ROADMAP

1. INTRODUCTION

NOTE: This working group roadmap does not endorse any one solution, company, or research effort.

The technologies being assembled for 5G and Beyond network infrastructure (referred to from here on as "5G&B") offer the potential for significantly higher communication performance than prior generations; in fact, many represent a disruption both in potential and the underlying physics that it utilizes. To this end, they also demand new techniques for design, analysis, and operation: dynamic rather than static, 3D and time-dependent behavior rather than 2D and constant. New tools and techniques, such as the Systems-of-Systems (SoS) approach, are required to predict, measure and control the operational performance of what could be the most complex, human-built infrastructure to date. Along with this technical complexity, the variety of economic and regulatory models under which it will operate adds new demands in being able to provide multi-stakeholder understanding of costs and benefits. Whether operated for profit in a competitive market or exclusively for the public good, from design through buildout, operation and eventual decommissioning and replacement, the dynamic return on investment needs to be calculable with sufficient confidence to warrant the risks. Finally, 5G&B represents an expansion in the expected behavior of the network from a system of data transmission to a more complex system of data transmission and transformation. We expect the 5G infrastructure to either incorporate distributed information computation and storage elements or to admit them as co-located distributed endpoints, and therefore we must account for their footprint in the planning and operation of the network.

While the scale and complexity of such an endeavor has opened the way for a new vision of the communication infrastructure, it has also exposed some dramatic shortcomings in the current approach to system design, whereby performance parameters are often addressed without proper attention to the energy implications of engineering decisions. From data centers to base stations, from smartphones to IoT devices at the Edge, both technical and financial viability of the ecosystem is fundamentally limited by its energy requirements - at each node in the network. Both CAPEX and OPEX are directly affected by energy consumption, be it for the cost of thermal mitigation or the actual energy and servicing costs of running the hardware.

Besides the obvious limitations provided by available technology capabilities and cost trade-offs, unexpected bottlenecks in the system are created by multiple unrelated causes, for example:

- physical constraints, as may be the case of Massive MIMO antennas and the need for the electronics to *fit the envelope*
- deployment constraints, as the ability to find suitable locations for cell and Edge computing hardware with appropriate access to reliable energy and fiber communication
- regulatory constraints, which may limit the power of Radio Frequency (RF) transmission due to safety or local requirements on noise levels

The ability to create a description of the complete system, which can provide insights on the energy-induced bottlenecks, and thus inform the stakeholders on realistic expectations for its performance and cost, can only be achieved by defining a model that enables all subsystems and diverse participants to

share a *common language*, based on the energy requirements for delivering the expected functions and performance.

This approach requires that we first identify and analyze the components that are most sensitive to energy availability and cost of deployment, and then formulate a methodology to model the relationships among all subsystems, which can highlight energy-related bottlenecks, trade-offs and expected return on investment. In this Chapter, we provide an overview of such a methodology, which will constitute the foundation for a standardized methodology and analysis, thus enabling all stakeholders to understand how they can contribute real value to the overall solution.

While there are several disciplines and methodologies that will prove invaluable to gaining the critical insight to shape the successful build out and operation of these breakthrough networks, we believe energy, the ability to do work, is a particularly useful metric to shape decisions. In particular, energy efficiency, the interplay of energy and its costs, provides a robust mechanism for stakeholders to evaluate the difference from theoretical potential costs and benefits to what can actually be realized.

1.1. 2022 Edition Update

Key updates from the 2021 (inaugural) edition as follows –

- Section 1.1: newly added to summarize changes for 2022 Edition
- Section 2.1: content added to expand upon boundary/framework methodology language
- Section 2.2.1: newly added sub section on “Cross-Collaborative Activities Across INGR WGs”
- Section 2.2.1.1: newly added sub section on “2021 5G World Forum”
- Section 3, Figure 3: Figure 3 updated with more legible color scheme
- Section 3: newly added sub section numbering assigned to content headers
    - NOTE: addition of new Section 3.1 caused previous Sections 3.1 & 3.2 (respectively) to increment to the *new* Sections 3.2 & 3.3.
- Section 3.1: content updated with clarifying language
- Section 3.3.6: content updated to expand upon the use of sounding signals
- Section 3.3.6.1: newly added sub section on “Renewable Energy-Enabled Cellular Networks”
- Section 3.3.8: content updated with clarifying language
- Section 5.2.2.2: content updated with paragraph discussing EE and user scheduler
- Section 5.2.2.6: content updated with appendage relating EH to WPT
- Section 5.3.2.1: content updated and appended with clarifying language
- Section 5.5.1.1: newly added sub section on “Challenges in the Optimization of Energy Use and Environmental/Financial Impact of 5G&B Network Infrastructure”
- Section 5.5.2.1: newly added sub section on “Solutions for the Optimization of Energy Use and Environmental/Financial Impact of 5G&B Network Infrastructure”
- Section 6.1.1: newly added sub section on “Collaborative Opportunities” with SustainableICT standards

- Section 9: new references appended to references table

2. WORKING GROUP VISION

Charter

The Energy Efficiency (EE) Working Group (WG) is dedicated to ensuring awareness, resources, and proper linkages are captured and disseminated in a meaningful way to enable the most pragmatic (and therefore minimal) utilization of energy and associated carbon footprint for global communications networks (including mobile telephony and fixed IP networks).

Vision Statement

The ultimate success of any new technology development is intrinsically tied to its energy requirements—be it a battery-operated device or a data center, energy is the currency that determines its business viability. Energy infrastructure is often a significant capital expenditure (CAPEX) driver and energy consumption is often the primary operating expenditure (OPEX) driver, increasingly displacing service and maintenance costs from edge to cloud to core. It is also fundamental to sustainability, and without sustainability systems are inherently inequitable. Its optimization will lead to a digital future rich in content and functionality for all to benefit. This vision is accomplished via inclusion in the IEEE Future Networks (FN) International Network Generations Roadmap (INGR) and the critical interactions with the many cross-functional stakeholder areas that are all inexorably dependent on the intricacies of energy architecture, distribution, and utilization.

5G&B creates both new opportunities for innovative Science & Technology capabilities, but also poses challenges of how to manage the 5G&B infrastructure itself. We will discuss methodologies and examples of tools that would be needed to both support the design, operation, interoperability, evolution and adoption of new infrastructure technologies as they evolve, with particular consideration to energy management of the resources in the 5G&B infrastructure itself, and energy/power delivery of end-user applications that will be supported by the 5G&B advances. This leads to the need for developing such methods for the 5G&B infrastructure and the applications it supports, in the context of 5G&B full situational awareness and how systems are affected and interact with related factors, such as their environments, and which systems they interoperate with, or rely on.

2.1. Scope of Working Group Effort

The EE WG is committed to education on energy-related issues/concerns/opportunities across all industry stakeholders and associated, extended ecosystems. Ideally, all industry stakeholders will come to realize the importance of an obsessive focus on optimizing energy efficiency/utilization at every level (i.e., from component to system to network) as a critical area, as early in the development/deployment/standardization processes as possible to maximize positive results when deployed at all scales (i.e., from edge or small cell to the full network and utility levels).

There are currently utilized metrics that have brought important visibility to energy consumption, such as Power Usage Effectiveness (PUE) [2], but they are over-simplified and do not enable the level of granularity that is necessary to fully understand the trade-offs at the system level and optimize efficiency across the ecosystem, all the way to the utility-scale. Whether the interest comes from technical,

business, and/or sustainability motivations, new metrics such as the concepts and associated critical dependencies of the Power Value Chain (PVC), Power Cost Factor (PCF), and the 5G Energy Gap (5GEG) [3] must be internalized and applied appropriately. We must also provide a mechanism to seamlessly move between technical, economic, and socioeconomic analyses that will ultimately make or break the success of 5G deployments, which is why the additional concepts of the 5G Economic Gap (5GEcG), the 5G Equality Gap (5GEqG), and the 5G Derate Factor (5GDF) are introduced.

Deceivingly, we often disregard small amounts of energy that are consumed "at the edge of the network" without realizing that the farther a device or system is from the power plant and the closer it is to the edge, the higher the multiplication factor of its energy requirement.

Conversely, when it comes to data center energy consumption, there is often misinformation and/or lack of understanding in how to interpret piecemeal efficiencies versus other constraints and how it aggregates to global consumption. Today this total data center contribution is around 1% of the global energy consumption [4] [5].

As many diverse stakeholders have an impact on the overall energy efficiency of the network, from energy generation all the way to its distributed consumption across the ecosystem, it is important to define the boundaries of each subsystem involved and how the interactions across boundaries impact the overall optimization of network EE. Individual subsystems/ecosystems will have their own unique boundary parameters intrinsic to their area of focus. While these parameters are caused by the different, siloed entities involved, part of the scope of the EE WG is to help make sense of this overwhelming field of view with the help of organized commonality yielding a more structured analysis via the introduction of a systematic modeling framework. In order to develop recommendations for the optimization of the system performance, it is essential that we develop a quantitative analysis that is global in nature and overcomes the current siloed approach, which publicizes the achievement of local minima, sometimes at the expense of overall performance. This objective requires the creation of metrics, Figure of Merits (FoMs) and Key Performance Indicators (KPIs) that bring commonality in energy characterization to a very diverse landscape and overcome the fragmentation of ownership, which leads to a lack of global optimization.

A roadmap format is an ideal way to accomplish the vision as it provides awareness, guidance, and tiered approaches for near- (~3 years), mid- (~5 years), and long-term (~10+ years) action.

2.2. Linkages and Stakeholders

There is no stakeholder in the industry without a direct linkage and dependency on this EE WG. A critical focus on energy and power requirements, architecture, distribution, and utilization are essential to the success of any player in the ecosystem no matter how big or small. Of course, this encompasses all the INGR WGs, so it seems unnecessary to list them all out here for that purpose, but the figure below is a tool referred to as the "IEEE INGR Cross-Cut Matrix" used to help facilitate the most salient and direct linkages between WGs.

Specific linkages to content in other WGs and reflection upon the relationships and interdependencies could easily fill another 100 pgs here, but we did want to make clear the work and focus of this WG is most closely aligned to that of the Hardware, Deployment, System Optimization, and Applications & Services WG. Follow-on work to this inaugural publication of the EE WG roadmap Chapter will seek to explore these linkages in more depth.

	INGR Working Group Name	ACCESS				NETWORKS			SYS. & STANDARDS			ENABLERS & USERS				
		1. Massive MIMO	2. mmWave & Signal Processing	3. Hardware	4. Energy Efficiency	5. Edge Automation Platform	6. Satellite	7. Optics	8. Standardization Building Blocks	9. Testbed	10. System Optimization	11. Deployment	12. Applications & Services	13. Security	14. Connecting the Unconnected	15. Artificial Intelligence/ Machine Learning
	ACCESS	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	Massive MIMO		X	X					X	X						
2	mmWave and Signal Processing	X					X		X	X			X			
3	Hardware	X	X		X	X	X		X	X			X	X		
4	Energy Efficiency	X	X	X		X			X		X	X	X		X	X
	NETWORKS															
5	Edge Automation Platform	X								X			X	X	X	
6	Satellite	X	X	X	X	X		X	X				X	X		X
7	Optics*															
	SYSTEM AND STANDARDS															
8	Standardization Building Blocks	X	X	X	X	X	X	X		X	X	X	X	X	X	X
9	Testbed	X	X	X		X	X		X				X	X		
10	System Optimization*															
	ENABLERS AND USERS															
11	Deployment	X	X							X			X	X		
12	Applications and Services	X	X	X	X	X	X	X	X	X	X	X		X	X	X
13	Security	X	X			X			X	X	X		X		X	X
14	Connecting the Unconnected				X	X	X		X	X	X	?	X	X		
15	Artificial Intelligence/Machine Learning*															

Figure 1. The IEEE INGR Cross-Cut Matrix for the EE WG

It has been amply articulated how delivery of 5G, and more so the "beyond" services, will involve both a complex infrastructure for information transmission ranging from fiber and ground-based transmission towers to satellites and many other means in between, mobile devices (ground and aerial platforms, including drones, etc), and also at varying transmission capacities and powers; e.g., Millimeter Wave (mmWave) technology to increase the system bandwidth and highly directional antennas on stationary and mobile platforms, and also unprecedented power requirements for some of the delivery of service requirements. Thus, these infrastructures entail multitudes of multilevel heterogeneous resources, which are of varying capacities and availabilities and under varying demands from the slew of application classes they will support; therefore, a comprehensive support infrastructure is needed with methods and tools that will allow coordination (orchestration) of such resources optimized for delivery of service to meet diverse and varying end-users and end-user applications requirements, and at the same time with optimized management of the power requirements of the 5G infrastructure. Moreover, robustness of 5G&B infrastructures will require them to be cognizant of real-time environmental factors, not only the users'/applications' requirements, but also for example the ability of energy grids to deliver the required energy/power (and variability of said service deliveries), and withstand disruptions in the communications infrastructure per se, due to failures of the infrastructure components, and other perturbations and disruptions which may be caused by space weather, adverse atmospheric weather, floods, earthquakes, tsunamis, localized/buildings or ambient fires, etc.).

The scope of this investigation, and the definition of the tools to address such complexity, certainly exceeds our ability to deliver a comprehensive and detailed review within this Chapter. We will therefore limit ourselves to identifying possible approaches to developing a SoS analysis and simulation capability, but will also initiate the development of a standardization infrastructure, which will enable the diverse contributors from the different parts of the ecosystem to harmonize their inputs and models, so that they can interact with each other and provide the basis for an end-to-end optimization, both from a system design perspective and a dynamic operational management.

2.2.1. Cross-Collaborative Activities Across INGR WGs

Several INGR WGs held multiday workshops since the 2021 release [6] to deep dive on their respective subject matter as well as invite external folks to participate in comprehensive and multidisciplinary panel sessions. In particular, the Massive MIMO WG hosted the IEEE Future Networks 1st Massive MIMO Workshop on 8-10 November 2021, which included a keynote from EE WG member Emil Björnson and a "IEEE Future Networks Panel" featuring EE WG member Paul Draxler [7]. The Massive MIMO & System Optimization WGs captured summaries of the EE perspective in their respective WG chapters in which they summarized their overall efforts [8] [9].

2.2.1.1. 2021 5G World Forum

Additionally, the 2021 5G World Forum [10] hosted a slew of extensive content across the broader INGR/Future Networks ecosystem, which included more of such cross-collaborative panel discussions and captured as part of the event's proceedings. We expect to continue these fruitful discussions and collaborations as the 5G World Forum morphs into the IEEE Future Networks World Forum beginning in 2022 (https://fnwf.ieee.org/). The EE WG participated in the following portions of the program (with associated EE WG members noted) –

- IP-4: INGR Panel: Deployment, Energy Efficiency, and Applications and Services Comprehensive Plans (Francesco Carobolante, Frederica Darema, Brian Zahnstecher)
- IP-5: INGR Panel: Combining Energy Efficiency and Systems Optimization for Network Sustainability (Francesco Carobolante, Brian Zahnstecher)
- TV-9: IoT applications in Energy Sectors (Brian Zahnstecher)
- TUT-14: 5G & Beyond: Addressing the Energy Challenge (Francesco Carobolante)

3. TODAY'S LANDSCAPE

The complexity and diversity of the stakeholders is one of the key challenges to the implementation of a coordinated approach to system-wide energy optimization. It is therefore important that we understand their interactions and motivations, so that we can find a way to gain a shared perspective and achieve synergy of intents. This challenging and multifaceted nature of the stakeholders responsible for various aspects of network components (from a black-box perspective) leads to siloing that inhibits collaborative efforts to bridge gaps.

The system-wide perspective must include a life-cycle assessment of the energy consumption. It is not only the network operation that contributes to it, but also the production and deployment of the hardware infrastructure. Even if new hardware solutions are developed to vastly reduce energy consumption during network operation, stakeholders might keep older equipment during its original intended life span. Hence, there can be long delays between when energy-optimizing technology is developed and when it has an appreciable impact on the overall energy consumption.

Figure 2 exemplifies the elements and interactions occurring in the system, and it intentionally includes Hardware (HW), Software (SW) and operators, as all elements create a meaningful interaction. If this looks like a completely discombobulated mess of contributing elements loosely connected with some type of common purpose not clearly articulated, then that is because that is exactly what this represents.

Each node in the ecosystem can be analyzed by assessing its optimization objectives, based on the inputs and outputs, as well as its constraints, which can be CAPEX, OPEX, environmental regulations, public policies, etc.

By viewing all interactions as the effects that outside forces exercise on the physical infrastructure, we can better identify constraints and opportunities. Figure 2 provides a high-level view of the approach that can be adopted to develop a methodology to address them.

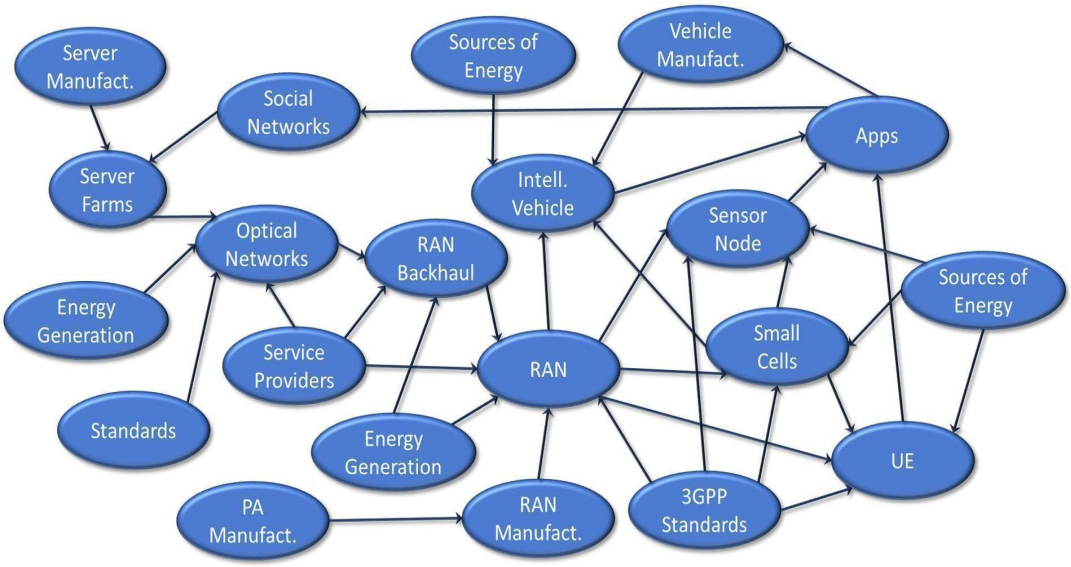


Figure 2. Outside forces affecting the infrastructure

Image courtesy of IoTissimo

As the viability of an application is dependent on its total cost, we identify energy as a measure of such viability, as it affects both the cost of HW (CAPEX), due to power supply and heat mitigation requirement, as well as OPEX. Monetary costs that are not directly energy-related can also be transferred into an equivalent energy consumption, by dividing by the energy price, or ranges of energy prices, if a sensitivity analysis is required.

The complete, global view of this SoS block diagram can be seen in Figure 3 (below). While the flows outlined in the key are missing from this version, it is meant to show how we view the entire 5G ecosystem and how these constituents fall into major categories of stakeholder focus. Now we have taken the disorganized mess of Figure 2 (above) and turned it into the organized mess before you in Figure 3 (below).

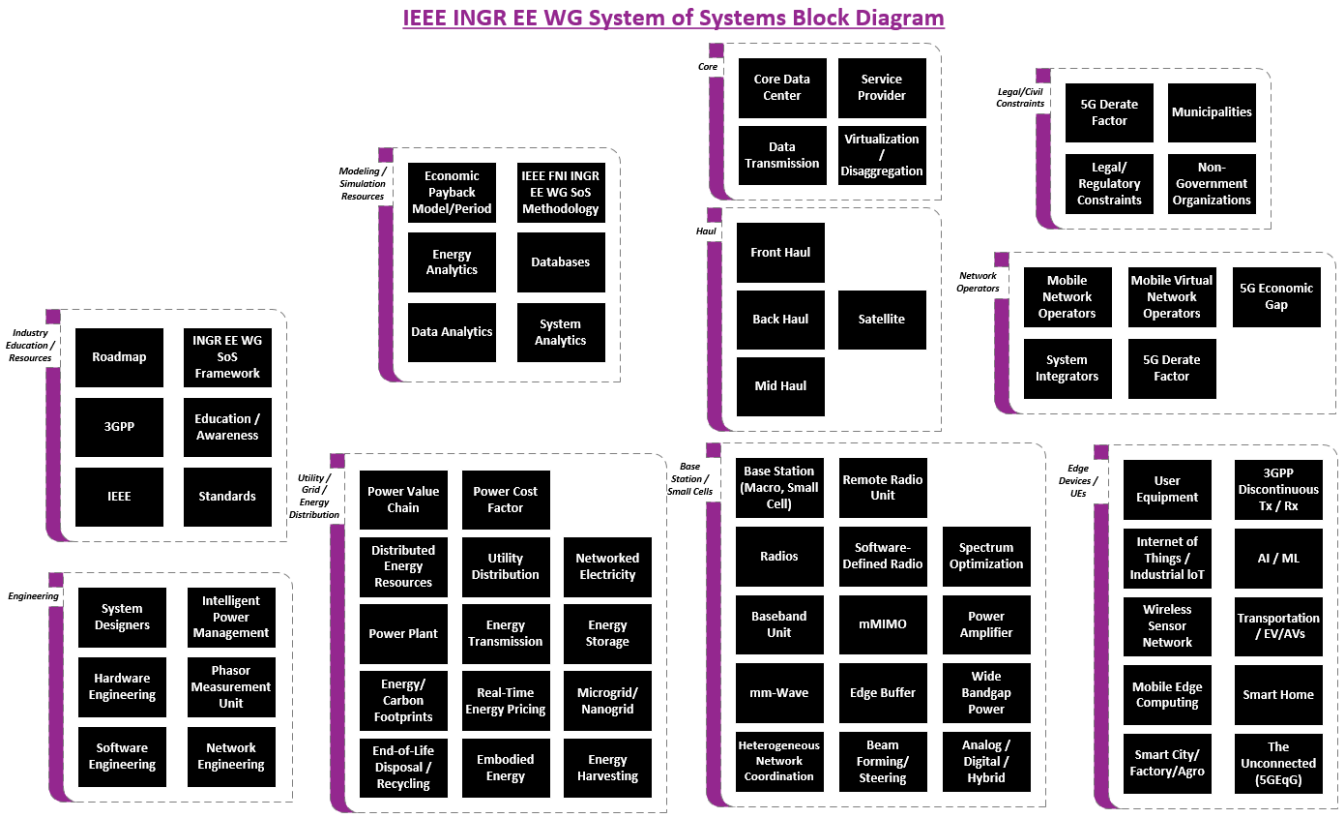


Figure 3. The 5G System-of-Systems (SoS) Block Diagram

Image courtesy of PowerRox

3.1. Global Mobile Telecommunications Energy Footprint

In order to truly appreciate the motivation for EE analysis and optimization in mobile telecommunication networks, one must understand the breakdown of the global telecommunications energy footprint and relate these pieces of the energy consumption pie to the network constituents that consume them. This entire energy footprint represents around 1-3 % of global, electrical energy consumption (including everything from core data centers to smartphones) [11]. The vast majority (60-80 %) is consumed by base stations alone [12]! Within a typical macro base station, 60-80 % is consumed by the Linear Power Amplifier(s) (LPA), which are constantly active, transmit at orders-of-magnitude higher power than the User equipment (UEs), and have relatively low efficiency. It should also be noted that even though the global telecommunications energy footprint represents 1-3 % of the global energy consumption, it has direct control/impact on closer to 20 % of the world's loads [13]. Even relatively small levers can have a highly disproportionate impact on much larger ones when considering second-order effects (reduction of OPEX and/or CAPEX in upstream infrastructure) and have a direct path to the influence of other important environmental metrics such as carbon footprint. Many of these second-order effects are explained extensively in the next section.

This simple analysis implies the majority of the global telecommunications energy footprint could be greatly reduced if we focused on nothing else than improving the LPA. Luckily, the largest consumer of energy in the global network also provides the most opportunity for EE optimization.

3.1.1. EE System Design Best Practices

When evaluating the power subsystems of electrical systems, especially with the purpose of optimizing for energy efficiency, it is particularly important to understand the overall system architecture and how all the energy sources relate to all of the loads that will utilize them. It is quite common to see engineers far more focused on increasing available power (i.e., increasing the source) than decreasing the system power budget (i.e., decreasing the load), which is somewhat counterintuitive. From big power systems down to tiny power ones, there is almost always more opportunity to reduce a system budget than there is to get a bigger (or more efficient power source). This challenge is particularly compounded in battery-powered devices, where Moore's Law and packaging innovations have significantly increased the function-to-power ratio at a rate much faster than the increase of energy storage density, which is completely at the whim of innovations in chemistry and physics.

For mobile devices, it's not Moore's Law alone that increased function/power because Dennard scaling ended in 2005. Since then, transistors stopped using less power when they shrunk, CAPEX decreased, OPEX per transistor stayed flat, and OPEX per die would have increased, but three things happened:

1. The System-on-Chip (SoC)
    a. You can still pull-out cost and power by putting everything on one die, so you do not pay to pump up energy to go through packages/Printed Circuit Boards (PCBs).
2. Heterogeneous Compute [14]
    a. Not just another core, lots of application specific accelerators and bit/little cores, precision design.
3. Active energy management (aka - Intelligent Power Management, or IPM)
    a. Modern silicon would melt if you turned on all the transistors at once. You need to actively ration charge at electrical potentials and keep stuff dark that is not needed immediately. The more stuff you need on, the slower it all runs, hence derating performance and overclocking and all the rest.

This is the lesson we should apply at a larger scale, but with the additional wrinkle that we need real-time, data-driven Artificial Intelligence (AI) making the decisions instead of the static power heuristics in an SoC.

From a philosophical perspective, one should note the same concepts and best practices are applicable to any "system" or "network" or PVC in a SoS, whether it be from one end of a global 5G network to the other or between power supply and microprocessor on a system board.

From the source perspective, there are a handful of simple concepts that can greatly impact the assessment of energy utilization (and therefore optimization). The first of these is the fact that all power sources have an efficiency curve that changes with respect to the load. In other words, if one assesses the power commutation efficiency of any energy source as a single value, then they are likely doing themselves a great disservice and result in measured data with large errors deviating from calculated expectations. As alluded to before, how system budgets are architected from Day 1 plays a big role in one's ability to ideally characterize and optimize system energy performance.

From the load perspective, there are many methodologies to ascertain a system power budget to determine needs as well as opportunities for optimization. The prevailing method seems to be the sum of maxima in which an engineer looks up the data sheet (can also vary wildly in accuracy and not account for manufacturing variations) for each major load in the system, identifies the absolute maximum power

for that component, then uses that along with all the others to sum into what is referred to as the system power budget. Unfortunately, this is a poor approach because it results in a worst-case summation that is well-beyond any realistic scenario, thus adding unnecessary margin into the design. While the detail is a bit outside the scope of this roadmap document, it should be noted this outdated, shortcut approach to power budgets is not the only way to ensure reliability, Quality of Experience (QoE), or whatever one's KPI is for uptime. For instance, if one has a system in which a sensor captures data, transmits it to a processor, has some analysis performed, and stores some data to local storage, then these processes are mostly serialized and therefore do not justify a power source that needs to support the absolute maximum power of all 3+ peripherals concurrently. The takeaway message is to invest the time and resources to properly characterize loads and understand their standalone behavior and in an application to get a realistic, maximum, worst-case power budget. Some safety margin can be added to this (and/or use some localized energy storage to mitigate risk with peak shaving techniques), but it is still likely to pale in comparison to the sum of maxima.

Even worse, there are unforeseen, rippling effects of adding margin on top of margin, which can have negative ramifications on both CAPEX and OPEX all the way up the PVC to the power plant. It may not seem like a big deal to take that sum of maxima and add 10% to be on the safe side, but that 10% will end up translating to 20-30+ % on the power supply nameplate rating, which determines the electrical infrastructure legally required (according to local code), which necessitates larger breakers/back-up power/cabling/etc., which will vary by jurisdiction. This can ultimately restrict the number of systems that can be supported by a single circuit (also translates to room-, building-, and transformer-level limitations) and how much power is allocated by the utility. Hopefully, it is not a far stretch to see how a little bit of well-intentioned safety margin at the system (or even subsystem) level can quickly get out of hand. The reason this typically goes overlooked is that there are too many different stakeholders operating too many different budgets to consider the holistic, end-to-end (e.g., PVC) impacts.

There is no system in the world that will dissipate less power than one that is turned off. The next most efficient scenario (again, as measured by minimal dissipated power) is a load operating at the peak point of the source's efficiency curve. Whether an IP block on an IC (i.e., dark silicon), a peripheral/component in the system, a server in a data center, or a base station in a Radio access network (RAN), a primary goal should always be to turn-off as much as possible when not utilizing, with the next goal being to put in standby (i.e., lowest power state). In application, optimization occurs when loads are in the peak area(s) of the efficiency curve so even considering consolidation of loads to take advantage of higher power supply efficiency at the effective, single load point can be a worthy approach. Also in the RAN, this can be done at the highest resolution with the use of Discontinuous Transmission (DTx) features, which allow even individual Orthogonal frequency-division multiplexing (OFDM) symbols to be dropped, effectively turning circuitry off (or down) even in ~50 us increments, which adds up to a lot of energy (increasing with frequency) over time.

3.1.2. Global Metrics & Hierarchy for 5G&B

Before directly jumping into the discussion on the strategy for determining the needs, challenges, and potential solutions metrics can bring to the table, it behooves all to first consider a hierarchical approach. It is important to have this discussion, and for stakeholders to be in alignment with their approach to metrics, before undertaking very large analyses like that of a cellular network PVC. Metrics can be just as hurtful as they are helpful if either are not carefully crafted in a way that takes into account not just units, but the appropriate application space and desired goal of usage of the metric or they are intentionally skewed to highlight particular data and/or deemphasize other data. As an arbitrary

example, one could argue that a town with 100 reported robberies is not as safe as a town with 50, but when looking at robberies per capita, one might find the town with more robberies has 10x the population of the one with fewer robberies (or the greater number of robberies occurred over a much longer period of time than the smaller number), thus giving a far more amenable metric when considering in this context. The real takeaway message is metrics are useless without the context so always be sure to take this into consideration before getting to the assessment of the actual numbers and measured data.

It is easy to get lost in a sea of metrics that individual stakeholders create to focus on their specific area/needs and can be biased by selfish priorities (i.e., highlight performance vs. energy consumption). Given the selfish priority of this WG is EE, this hierarchy should start with the most fundamental relation of load to application, which is energy per data or Joule/bit. This inverse of this ratio, data per energy or bit/Joule, is another common and equivalent metric. In High-Performance Computing (HPC) data centers, for instance, current advanced development in Dense wavelength division multiplexing (DWDM) Silicon Photonics is estimating 3.5 pJ/bit (or 0.28 Tbit/J) represents a state-of-the-art target for the end-to-end journey of that bit from creation to consumption [15].

As we go down the hierarchy to sublevels, this fundamental metric of joule/bit can be expanded as appropriate for the sublevel. For instance, one focusing on optimizing the RF transmission energy between a base station and particular user equipment (UE) may focus on a metric like minimizing pJ/bit-km or maximizing Tbit/J/km to expand the fundamental relation into the physical domain. A slight modification to this might be from the perspective of one deploying base stations and focused on coverage area, thus minimizing a metric like pJ/bit-km2 or maximizing a metric like Tb/J/km2, as that is more pragmatic for optimizing the data rate amongst users in a particular area of coverage for that particular base station. Going even further to expand into a metric for optimizing the spectral efficiency of the (very costly and precious resource of) bandwidth, it makes more sense to utilize something like Tbit/Joule/Hz or a hybrid with the coverage area metric to yield something like Tbit/J/Hz/km2.

It is no coincidence that this hierarchical relationship of metrics also tends to track the hierarchy of the network. If one follows a bit from a core data center through the haul distribution to the base station, then through RAN and wireless transmission to UE, then it becomes clear why there is a need to keep things as simple and universal as possible in the beginning (e.g., bit/J) and have a far more complicated metric like Tbit/J/Hz/km2 that only considers one piece of the network. At the end of the day, the available bandwidth and the required coverage area can be viewed as inputs, thus J/bit or bit/J will represent the overall performance of the network.

3.1.2.1. Defining a Universal Architecture & The Universal Currency

When assessing energy utilization and attempting to optimize for maximum energy efficiency, it is critical to internalize the relationships between all sources and loads. The ultimate and most comprehensive oversight of this relationship can only be accomplished by finding the absolute extrema of the flow of energy from generation to heat dissipation in the end load. It should be noted that this same concept (and ensuing assessment framework/methodology) applies from micro power-levels to macro power-levels (i.e., across an IC or across a full network). To visualize the concept, a Power Value Chain is proposed and defined as follows (and displayed visually in Figure 4, below):

- "**Power Value Chain (PVC)** is a systematic representation, which describes the energy flow across all the distribution/conversion steps between source and load, that ties together the siloed stakeholders."

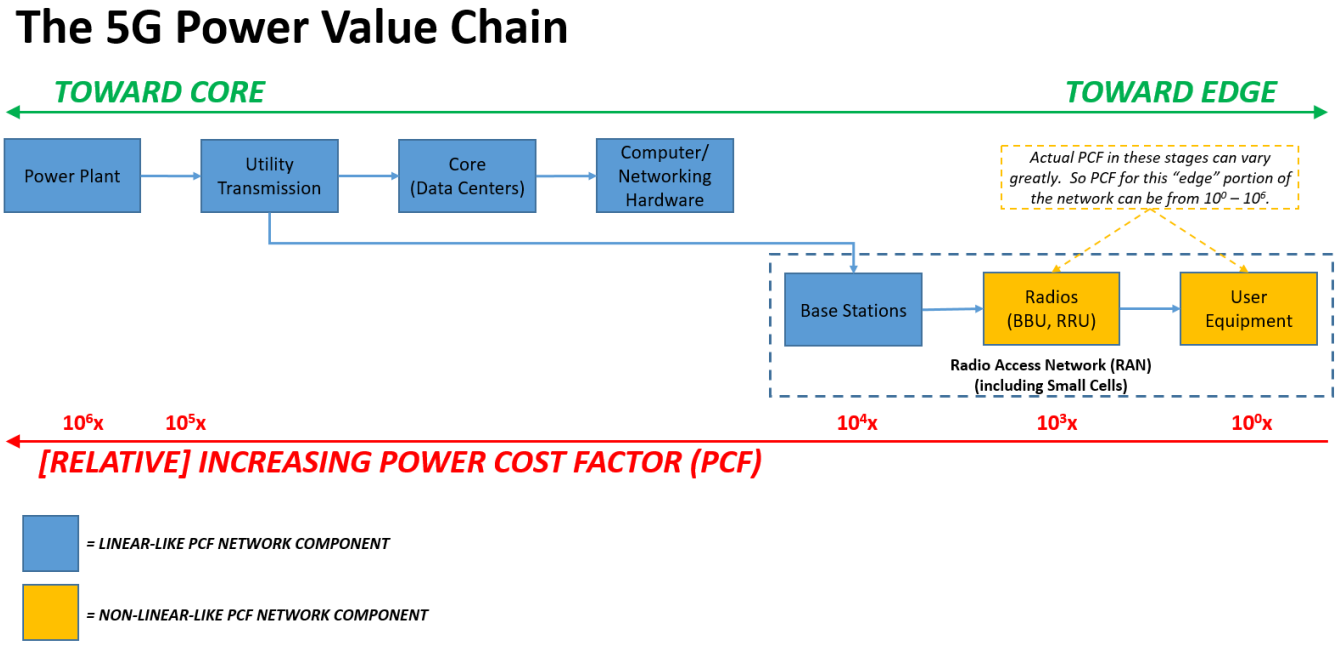


Figure 4. The Power Value Chain (PVC) from Network Edge to Power Plant

Image courtesy of PowerRox

Now that we have defined the PVC and have a simple, visual representation of the absolute power flow from source to load, there must be a metric to help provide a way to differentiate power utilization based on where it is consumed within the PVC. Such a metric is at the heart of a comprehensive EE analysis and optimization because it concurrently brings a technical and economic weighting to the "cost" of those watt-hours that is in the context of the device consuming them as well as their physical location within the PVC. The metric must lend itself to supporting a "universal currency" and common language, which for us is energy. For instance, 1 kWh consumed right outside the utility power plant requires very little extra energy to be generated because there are nearly no distribution losses and therefore has an overhead factor of around 1x. On the other hand, if the same 1 kWh is consumed in a system that has gone through many transmission steps, voltage converters, etc., then it could have an overhead factor that can be orders of magnitude higher.

3.1.2.2. Defining Universal Metrics for 5G&B

For the EE context of this roadmap, the first metric to introduce is the Power Cost Factor, which is defined as follows (and displayed visually along with its relation to the PVC in Figure 4, above):

- "**Power Cost Factor (PCF)** is a unitless number that represents the multiplication factor required to quantitatively assess the overall cost of energy utilization at any given point within the PVC."

This simple, yet very telling metric can be an eye-opening perspective that helps to expand energy consumption in a more nuanced manner. The most revealing aspect comes when one removes the constraints of the very siloed nature of the black box constituents of a 5G PVC and understands just how much energy overhead is associated with having the desired amount of energy at the point of consumption. This is further exemplified by a simple analysis in the following text, which goes through a quick derivation of the PCF associated with 1 mW of received power at the edge.

A good example of how applying the PCF metric assessment to the PVC model can completely augment the technical and business justification for a technology is related to the application of energy harvesting (EH). EH refers to the capture of free, ambient energy from the surrounding environment and can be from every form of energy in the world of physics. Common examples of EH are things like photovoltaic (PV) solar panels, wind generators, and hydrokinetic turbines. Additional examples are thermoelectric generators and far-field RF rectifiers (a.k.a. rectennas) though it should be noted things like wireless power transfer (WPT) from a directed source are not true, scavenged energy sources as they are the same as offline power delivery in which one of the "wires" happens to be an RF resonant connection. While EH technologies cover a wide swath of solutions and can range from µW to MW applications, there is a general trend of them being more economically feasible the more power they can produce. Conversely, the technical perspective provides for increasing feasibility to apply EH technologies to lower power loads. In other words, the lower the system power budget, the greater the opportunity to enable a completely self-powered solution.

There is an undeniable trend of increasing the number of "things" on a network the closer one gets to the edge so by the time we get to the lowest level, it is realistic to imagine 1,000s or even 10,000s of connected devices to a single base station or local network with the application of ubiquitous sensing, such as what one might find in any number of widely-touted 5G use cases such as Wireless Sensor Networks (WSN), Industrial Automation, Asset Tracking, Smart Cities/Buildings (including Smart Grids), Smart Agriculture, or Autonomous Vehicles [16]. The great majority of these network touch points will be low-bandwidth, high-latency kind of devices communicating over low power communication protocols (i.e., Bluetooth, LoRa, SigFox, etc.) and/or utilize local networks (i.e., Wi-Fi, mesh network, etc.) for aggregation and haul purposes. This implies their system power budgets are relatively small (most << 1 W) and therefore great candidates for the use of EH. Supplementing a few mW of a WSN device's power budget may not only be hugely beneficial in reducing the amount of power a plant must generate if far upstream in the PVC, but may theoretically be a huge part of a solution to a major risk currently faced by all deploying and/or depending on the use of 5G&B technologies. This assessment and risk (known as The 5G Energy Gap) is defined and analyzed below. So that a few mW of EH-generated power may not seem like a big deal, but if doing so mitigated the risk of destabilizing or even bringing down the local utility grid with the highly scaled deployment of tiny power devices and things, then it may suddenly become a huge deal. Further considerations for the application and justification of EH technologies is reviewed in Section 5.2.2.6.

One can quickly gain a perspective of how important it is to assess upfront the energy budget by doing a cursory examination of the energy required to power an edge device for industrial applications if we had to supply its energy by WPT, i.e., beaming RF power to it.

A simple link budget calculation reveals that at 25GHz, using a TX antenna with 1,024 elements and a RX antenna with 16 elements placed at 10m distance, we receive less than 0.1% of the energy consumed by the transmitter, and that is in the best-case scenario of line of sight and antenna planes that are perpendicular to the transmission direction. Since one antenna needs to service multiple devices in different directions, we quickly reach efficiencies that are less than 0.01%. Similar outcomes are found

if we use 2.4 GHz frequency with 64 TX elements and 4 RX elements, but in this case the size of the antennas is becoming quite large (50x50 and 13x13 cm respectively), which makes both transmitters and receivers difficult to conveniently place.

Therefore 1 billion devices that utilize only 1 mW each would require 10 billion watts, which is actually more than the average consumption in the whole county of Los Angeles (including 10 million inhabitants and all the commercial activities) [17]. It is clear that this is not sustainable: the ecological impact would be worse than using batteries!

Although such solutions may be useful when other options are difficult to implement, we cannot envision this as the general solution to implementing the vision of ubiquitous sensing and actuation.

While the PCF metric facilitates assessment of any PVC network constituent, the ability to conduct a practical investigation into the consideration and application of energy storage technologies at any (and all) points of the PVC should be noted. The trend of disaggregation has been applied in the networking data center (specifically discussed in greater detail in Section 3.2.2) for many years now and is finding its way downstream to the base stations and rest of the RAN. Disaggregation can apply to just about anything including energy storage so the PCF metric and associated analyses outlined in this document can be used to modularize energy storage for operational purposes, whether it be critical energy for back-up/hold-up, peak shaving for reduction of underutilized infrastructure, or in more economical applications such as for storing/utilizing energy based on the availability of solar energy or the dynamic energy market.

3.1.2.2.1. KEY ANALYSIS: What is the true cost of 1 mW?

In order to have an appreciation for the risk potential of the 5GEG, it helps to do a simple analysis to understand the true cost of power using the point of consumption within the PVC as a frame of reference. As we develop such analysis, we need to distinguish the two different ways in which each device impacts the overall energy footprint of the system.

The first part of the energy calculation relates to the source that provides the energy, which enables the device to operate. For example, let us consider a communication processor connecting the fiber input to a local server: its operating energy will be delivered by a local power supply, which is fed by a secondary ac/dc or dc/dc converter, which derives its energy from a HV to a low-voltage converter that is connected to the grid; finally, such energy is produced by a power station that has its own efficiency and losses in the distribution network. Such "end-to-end" efficiency becomes further reduced when we also consider that the energy associated with its fan cooling system, environmental temperature regulation, etc. needs to be multiplied by the inefficiencies of the supply chain all the way to the electrical power generation.

The second part of the energy footprint calculation relates to the computing and communication infrastructure required to support the activity of such devices. An edge/IoT device may be battery powered (with its own energy consumption footprint), but the activity generated by such a device in the network creates the need for additional communication, computing and data storage/buffering, all of which requires energy consumption. That is, each added device in the system generates additional energy consumption both statically (i.e., equipment sizing to provision such functionality) as well as dynamically (i.e., energy consumption required by the system to support the instantaneous or time-delayed activity generated by the device and its associated applications). As the complexity of the system increases, provisioning such an enormous number of devices and applications further generates

the need for powerful controllers, which can optimize the system performance and dynamically allocate resources.

A third component of the energy footprint is the energy "embedded" in the device, i.e., the total energy required "cradle to grave," which includes the energy spent to produce the device and to dispose of it at its end of life. This analysis, though very important to achieve a complete picture of energy consumption, is beyond the scope of this Chapter, at least for this edition of the INGR.

A simple analysis of the true cost of power consumed at the edge, following it all the way back up the PVC reveals astounding figures:

- What is the TRUE cost of 1 mW received at the edge?
  (All ranges represent the best to worst case.)
- 1-2 W = transmitted by antenna of base station/access point
- 17-50 W = input of base station/access point
- 8-15 % = lost in transmission from power plant
- SUMMARY:
  1 mW EQUIVALENT OF RECEIVED POWER AT THE EDGE REQUIRES 18,000 - 60,000 TIMES THAT POWER GENERATED (or 18 - 60 W) AT THE POWER PLANT.

This would not be such a big challenge if the increase in the number of receiver devices had no significant influence on the total power consumption, but that is no longer the case with 5G.

Until the 4th Generation Wireless Network (4G), we lived in the era that Marconi created a century ago: that of a "broadcasting" technology, where the cell transmits with a fixed radiation pattern into the cell area. Adding more user devices does have an effect: as the number grows, the requirements on signal-to-noise ratio (SNR) of the tower electronics grows, and so does the power associated with mixers, analog-to-digital (A/D) converters, etc. Yet, the growth of energy consumption with the number of receivers is relatively contained.

In the 5G era, we increase the performance of the network along three different trajectories: growth of bandwidth, reduction of latency, and ability to spatially multiplex multiple devices and layers per device, which are orders of magnitude larger than in the past. Thus, many issues arise, which determine a non-linear growth of the energy consumption.

Increasing the bandwidth in the RF section has the following consequences: (i) the need to operate in the mmWave to access large transmission bands causes an exponential increase of signal loss in the air (also dependent on atmospheric conditions); and (ii) the need to adopt beamforming to deliver enough power at a useful distance with a reasonable power consumption in the cell leads to Massive MIMO implementations, which require a large number of elements as well as multiple antennas. These issues require radically new hardware architectures, even if they should not lead to an exponential compounding effect on power dissipation. In mmWave bands, the required densification of the nodes (cells), the frequency, complexity, and performance requirement of the electronics are challenging. For this reason, 5G has so far focused on Massive MIMO in conventional bands, such as 3 GHz, where it is possible to spatially multiplex many users without getting into the issues related to mmWave spectrum and transceiver hardware.

The quest for low latency to enable mission critical services leads to data management and processing at a significantly higher rate than normally acceptable, thus stressing every aspect of the system: from protocol execution to speed of the computing electronics. Additionally, the "mission critical" aspect of these applications forces both redundancy and guaranteed Quality of Service (QoS), which increase

complexity, power and cost. Furthermore, achieving such low latency requires very large computing power to be available close to the Edge rather than concentrated in the Cloud, thus requiring an enormous growth of an infrastructure that is "electrically distant" (and therefore less efficient) from the power generation plant. QoS and safety requirements for the new applications force additional demand on the power infrastructure, such as energy storage and/or local back-up generation, which further contributes a large overhead to the energy footprint. Particularly challenging is the example of connected vehicles (e.g., V2I, I2V, V2V), where such infrastructure would have to exist everywhere before such vehicles could operate in a consistent way.

Enabling billions (or trillions!) of low-power edge devices means lowering their transmission power, so that they can be operated by battery or more ideally by energy harvesting, due to ecological considerations. New standards allow such lower power protocols, but that also implies a densification of the cells, in order to reliably communicate.

Additionally, the control mechanisms required to operate such a complex system imply optimization at every level: from cell transmission (bands, coverage, etc.) to RAN (cell cooperation, shedding, etc.) to computing and storage infrastructure (edge allocation vs. centralized cloud operations). Dynamically supporting such activity requires powerful AI engines that are capable of performing such optimization in real time, thus imposing additional energy provisioning both at the Edge and in the infrastructure.

3.1.3. Articulating the Energy Risks of 5G&B

5G infrastructure represents a new class of complex systems, which are the confluence of three supply chains that must be kept dynamically in balance. Two are familiar: infrastructure and energy. We must build out the physical infrastructure and contend with the physics of its deployment, where the feature size of interest will be measured in tens of meters while the area of interest will be metropolitan to continental. In parallel, we must build out the energy delivery and waste heat extraction infrastructure. Again, here the new dimensions and scales mean that centralized generation and delivery, either continuous over wire or discrete via batteries, may no longer be viable, and harvested or intermittent dynamic power may need to be accommodated. The final supply chain to be balanced is the data supply chain, and here it is the bearer plane as well as the control plane data, both of which will contribute to the continuous training and inference phases of machine learning and AI models. In order to operate these disruptive technologies at the required levels of optimization, not only dynamic rather than static analysis is needed: the objectives demand active and predictive control systems utilizing the latest and anticipated advances in Machine Learning (ML) and AI. These techniques themselves voraciously consume data both during the now continuous process of training as well as inference, thus significantly adding to the energy footprint.

When these three supply chains are in balance, the potential societal and economic benefits of operating the network are realized. When they are not, we not only suffer the economic penalty of an under-utilized resource but also the opportunity cost of not applying those resources to other challenges. Similarly to the derating process, we apply to the potential performance of an individual transistor in a novel semiconductor processor captured on a curve tracer compared to the actual performance of billions of transistors in a modern Central Processing Unit (CPU), under varying load, subject to process and temperature variation and with power delivered through the dynamic AC response of die, package, board, point-of-load regulation and bulk supply, we may need to derate our anticipation of the realizable performance of individual 5G networks and potentially of the class of 5G networks overall.

As we strive to provide understanding to all of the potential stakeholders in a particular 5G network, it is useful to use the Systems-of-Systems (SoS) dynamic, analytic approach to examine the sensitivity to variations each has on the infrastructure, energy and data supply chain, understanding that they have potentially nonlinear inter-relationships. In particular, we want to identify when gaps between the potential and actual performance of a network will be especially hard to close.

Applying the PCF metric to the PVC concept described above yields a hypothesis of the risk such a dramatically increased number of relatively tiny power devices can cause due to their disproportionate PCF at the edge. This hypothetical risk content is referred to as the 5G Energy Gap and is defined as follows:

- "**The 5G Energy Gap (5GEG)** is a hypothetical representation of the disparity between available energy (sources) and demand (loads) of the devices representing the network, fixed co-located resources, and both mobile and fixed endpoints that constitute the majority of "things" in the highly-scalable edge space of the network, based on the dynamic workloads of the proposed 5G use cases."

The gap arises when either the extant or anticipated power infrastructure is unable to satisfy the dynamic response required for a use case. It should be noted that this gap can arise even when there is sufficient bulk centralized energy but the inefficiencies, losses or dynamic response of delivering to the often micro-powered endpoints cannot be overcome, especially when low latency 5G services are introduced. In the same way that the gigahertz frequency current response of a transistor in a modern CPU design may be limited by the in-package capacitor dynamic response regardless of the size of the bulk-power supply, the ability to satisfy a low-latency 5G use case may be dominated by the ability to deliver energy to the last hops in the network microsecond to microsecond.

While we might therefore view the dynamic energy delivery infrastructure to the network, core to edge, as a degree of freedom, given the economic, environmental and regulatory burden of adding variable on demand distributed power capacity, which is also sustainable and at least climate neutral, it is instructive if not also pragmatic to consider it as a constraint. The US Energy Information Agency IEO2020 [18] projection anticipates only an 11% increase in global energy production from 2020 to 2030 and 66% of that modest increase will come from renewables which may be challenging to adapt to power an always on, always dynamic network. If we constrain the static and dynamic energy generation and delivery characteristics, then for a particular network, we can determine a 5G Derating Factor (5GDF), which identifies how much of the theoretical performance and benefits of 5G networks can be realized and is defined/described in detail later in this document.

3.1.3.1. Relating Energy Risks of 5G&B to Financial Risks

The 5GEG has brought visibility to technical risk of grid stability at the utility-level, but it does not adequately enable one to assess the impact energy efficiency has at a system-level and particularly how that impact directly influences "the promise of 5G" one has expectations set to. A 5G Economic Gap is proposed and defined as follows:

- "**The 5G Economic Gap (5GEcG)** is a hypothetical representation of the disparity between available power a system can deliver, and the increasing load demands on its outputs, which means a power-limited system and/or network component will not be able to utilize all its designed potential and therefore be inhibited from delivering on the calculated economics of the payback period."

At first glance, the differences between the 5GEG and the 5GEcG may not be so obvious as they both seem to relate to negative network ramifications resulting from there being more load demanded than sources can provide. The fundamental difference is with one (5GEG) being an energy-focused limitation, which puts more focus on the impact to the grid and utility side of things resulting from too many base stations demanding too much energy at any given time, which can cause grid stability issues. The 5GEcG is a power-focused limitation, which puts focus on the impact of the network hardware (typically base station radios) that must scale-back edge functionality as a result of hitting power/thermal limitations within the system design envelope. For instance, edge devices may have their bandwidth limited if too many are demanding too much at the same time from the same base station, thus resulting in power-gated radio performance, likely due to hitting the maximum thermal dissipation the system packaging is designed to accommodate.

Either of these gaps can be technically limiting as just described, but also both imply economic limitations. From the need to derate network performance due to these energy limitations (see 5G Derate Factor, described below) to change in the calculus of economic payback models, the financial impacts of these gaps are major and far-reaching. It requires very significant investment from both industry and government/municipalities to deploy and instantiate a 5G network and these investments are critically tied to assumptions about applications, user behaviors, network performance, frequency/bandwidth utilization, and ultimately, cost. This WG recognized a void of information in this space, which enables any industry stakeholders to make a more accurate assessment based on realistic derate factors. Power availability may have not been considered for projected deployments and equipment may need to be redesigned to "fit" into the power- and/or energy-limited footprint.

3.1.3.2. Quantifying the Risks of 5G&B Deployments

With the definition and understanding of the technical and economic limitations of the gaps articulated here, one must now attempt to translate these concepts and risks into explicit metrics to enable assessment of these complicated PVCs and various gap analyses. Even converging on the type of metric to use can be quite challenging.

When originally thinking about this metric gap, it seemed important to have something that was comprehensive enough to translate the many different, complicated needs of all the economic and technical stakeholders in the 5G PVCs (and complete SoS, by association), yet still be simple enough to be internalized by numerous folks of many different backgrounds (some technical, others not so much). These requirements were reminiscent of a well-known metric from the data center industry known as PUE. PUE simply measures the data center IT equipment power as a ratio of total building power, which quickly quantifies how much energy goes directly to the IT load and how much goes to overhead and highlighted areas to focus on for energy optimization. This oversimplified metric can easily be understood and communicated across the industry. It quickly became obvious this was a good starting point for a discussion, but does not get us to a complete solution and therefore, required additional, lower-level metrics, particularly ones focused on utilization (versus homogenous, static analysis). Regardless, the shallow learning curve for the PUE metric brought key stakeholders to the table and resulted in much better, more detailed metrics down the line.

In the context of wanting to assess complex PVCs that combine to bring the 5G network to fruition, a new metric is proposed that combines the simplicity of PUE with the energy-optimizing potential of the SoS. This metric must be able to understand and characterize a network at its maximum potential as well as derate the effective, achievable network resulting from the assessment of energy-related

bottlenecks in the PVC. Furthermore, this metric must be willing to do all this, while translating both inputs and outputs into the "universal currency" of energy. A 5G Derate Factor is proposed and defined as follows -

- "**5G Derate Factor (5GDF)** is a unitless coefficient (<1) representing a scaling factor for the application of technical and economic risk factors to the ideal 5G network deployment model that will reduce the optimal, maximum designed capabilities of a network due to energy-limited (5GEG) and/or economically-limited (5GEcG) and/or socioeconomically-limited (5GEqG) factors."

An ideal 5GDF of 1.0 represents a system that lacks any energy-related (i.e., including power and/or thermal) bottlenecks that would prevent the solution from fully delivering on "the promise of 5G" as indicated by its specifications. This is really more of an aspiration than realistic goal in network design because there will always be some weak link in the PVC that limits all constituents from being fully utilized, thus limiting the PVC to some 5GDF < 1.0. More importantly, 5GDF can serve as the measure by which all other components of the SoS can work together to optimize their own EE as well as that of the overall PVC (i.e., the common good or full network stack). Like PUE, it should help to more quickly and easily identify the system contributors most likely to inhibit overall network performance when that performance is based on assumptions of the availability of necessary energy/power resources, which simply may not be realistic in practice. Does your payback model assume a 5GDF of 1.0 to enable success? If so, then your modeling assumptions may be flawed for these reasons as well as inadvertently drive the digital divide.

3.1.4. Strict Payback Period Targets Driving Socioeconomic Disparity

All of the concepts, metrics, and best practices described so far have focused on the technical and economic viability of the 5G&B network. While still tied to these factors, there is another consideration that is important to observe, which adds an ethical dimension (and ideally obligation) to the more socioeconomic impacts of 5G deployment. While not trying to make any political statements, we see equitable network access and an effort to connect the unconnected (or underconnected) to be just as business savvy as it is ethical. Similar to 5GDF, economic payback calculations may make assumptions about the equity of access when considering their subscription models and deployment scenarios. Stakeholders and network owners must recognize the game-changing socioeconomic impact the "promise of 5G" can bring to underprivileged communities. This representation of the socioeconomic impact as related to network energy infrastructure is referred to as The 5G Equality Gap and is defined as follows -

- "The **5G Equality Gap (5GEqG)** is a hypothetical representation of the socioeconomic disparity between those that will be able to adapt their infrastructure and end use cases to unanticipated underperformance due to energy-limited (5GEG) and/or economically-limited (5GEcG) factors, and those that will not have the resources to be flexible enough to do so."

To truly internalize the 5GEqG, one must explore and understand the differences between the inconvenience of a poor (or non-existent) network connection versus the absolute life changing opportunities a cellular connection can bring to a family, particularly those in non-OECD (Organisation for Economic Co-operation and Development [19]) countries. In places like India and sub-Saharan Africa, regions with high poverty and lack of basic resources like electricity and clean, running water to homes, can still get cell service. Even if a generation or two back (i.e., 4G-LTE, 3G, even 2G), individuals can live off their phones for basics like finance/currency exchange, socializing, and even an

untapped awareness of what is going on in the rest of the world they would otherwise be isolated from in poverty [20]. A flip phone charged by a single solar panel and operating on a 2G network can have generational improvement opportunities in quality of life for people in these communities.

It should also be noted being "unconnected" does not purely apply to rural and/or (non-OECD regions) as being "underconnected" can apply to someone in a densified environment that does not equitably service all parts of the community the same and therefore drive socioeconomic disparity, while also not meeting payback goals. Just as energy can define limitations of achievable 5GDF, derated functionality can have disproportionate impact to some users over others, thus also having disproportionate impacts to the expectations of the economic benefits and modeling for economic optimization. Along this vein, one should consider these second-order factors in many different assessments dictating the viability of a next-generation network deployment as it is architected, financed, tested/qualified, and finally deployed to communities of all socioeconomic situations. These considerations should be present in everything from base station optimization to regulatory changes and utilization of new spectrum.

3.2. Current State of Technology and Research

5G initiated a revolutionary reshaping of the network where not only data rate increases are expected, but also flexibility in the support of a diverse set of users with different performance needs in terms of latency, QoS and reliability. Addressing all of these challenges requires an overhaul of the infrastructure, including additional bands as well as new protocols to enable proper management of such diverse communication links. All of these new requirements affect the energy consumption at different levels in the system, from RAN performance to core transport through fiber, switches and routers, all the way to data processing in the application servers.

The description of the complexity in baseband processing, computational and routing requirements generated by the new protocols is beyond the scope of this Chapter: suffice to say that they cause an increase of energy consumption, both on the UE side and the cell/backhaul side. We will therefore focus more broadly on the infrastructure requirements, and in particular the new developments imposed on the RAN as a consequence of the expected service level needs.

To appreciate the challenges posed by the higher bandwidths sought by 5G&B, it is necessary to briefly take a short digression on RF signal propagation and interference management.

3.2.1. The Physics of RF Transmission

The use of multiple frequencies will require factoring in the propagation behavior at each frequency through air, with a range of temperature and humidity, walls, with a variety of construction materials, and other terrestrial propagation effects. One major input to the terrestrial link budget is the average attenuation in air, including oxygen and water absorption (See Figure 5 below, where curve A is in warm, humid conditions, curve B is in cold conditions with low humidity).

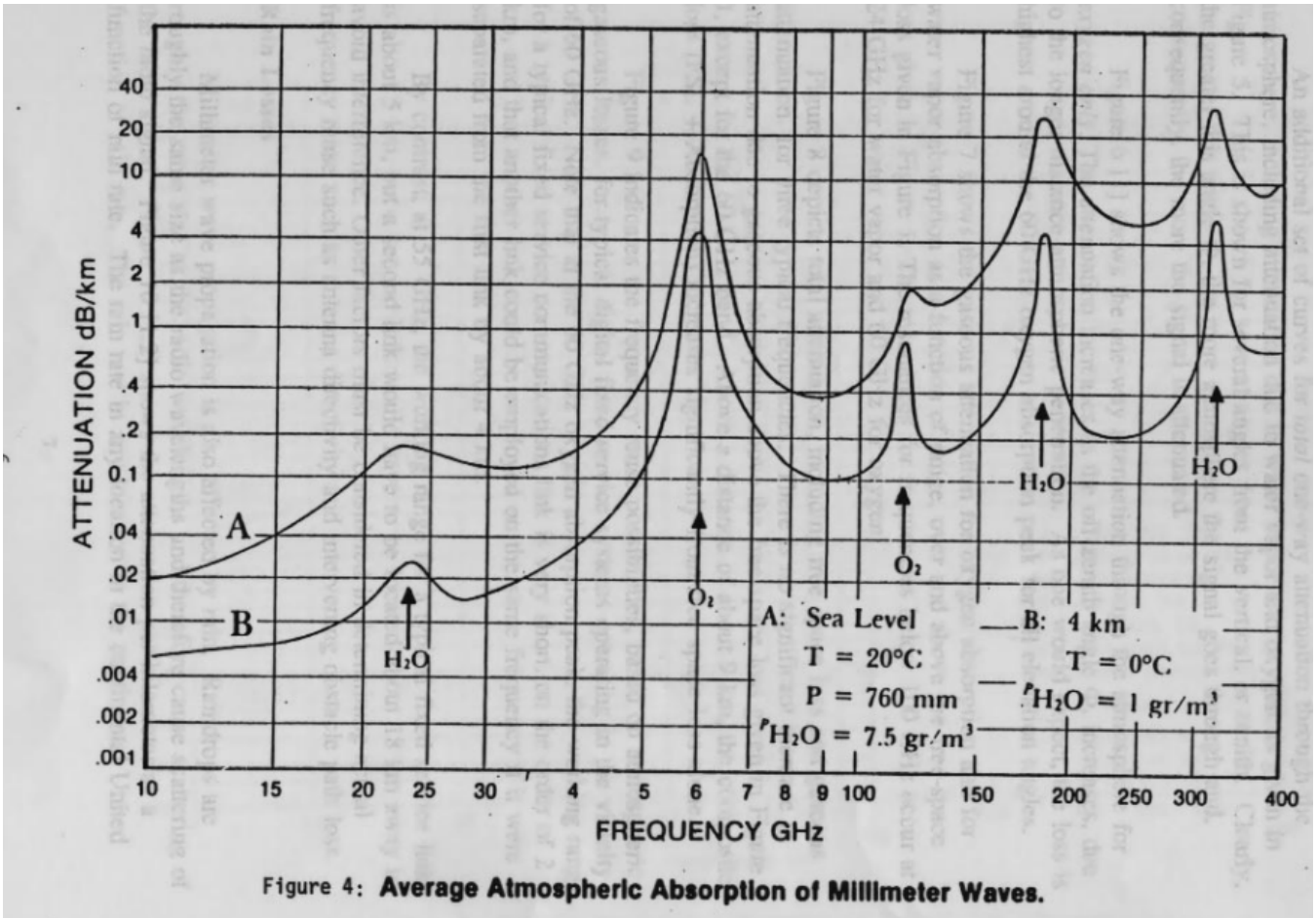


Figure 5. Attenuation of mmWave Transmission over the air [21]

3.2.2. Requirements on Unwanted Emissions Must Be Satisfied

Specific requirements for existing 5G infrastructure systems may be found in Third Generation Partnership Project (3GPP) standards documents such as European Telecommunications Standards Institute (ETSI) document TS 138 104 - V15.2.0 or 3GPP 38.104 [22]. Note that the 3GPP community aggregates these requirements as "unwanted emissions" which consist of "spurious emissions" and "out of band" emissions. By this definition, "out of band emissions are unwanted emissions immediately outside the base station channel bandwidth resulting from the modulation process and non-linearity in the transmitter but excluding spurious emissions." Similarly, "Spurious emissions are emissions which are caused by unwanted transmitter effects such as harmonics emission, parasitic emission, intermodulation products and frequency conversion products, but exclude out of band emissions."

Practically, out of band emissions are characterized in a relative sense by something like the Adjacent Channel Leakage Ratio (ACLR) that scales the allowable leakage in the adjacent channels relative to the power output of the base station. An ACLR requirement of 45dB is common. Spurious emissions, on the other hand, are characterized by an absolute limit on the maximum emissions allowed and are dependent only on the power class of the base station (local, medium area, wide area). The basic limit for the spurious emissions is approximately -90dBm measured in a 100 kHz band and this is intended to protect the base station receiver from unwanted interference from other nearby transmitters.

These unwanted emissions requirements place a very high premium on the cleanliness of the transmitter signal as well as a strong physics incentive for the densification of the base station network and thereby reducing the individual absolute base station power (because the performance of the subsequent diplexer filters is measured relative to the base station power but yet must meet an absolute cleanliness requirement).

3.2.3. The Need to Address the Whole Ecosystem

The constraints highlighted above, due to physics and interference, are informing the current infrastructure deployment and driving the choices for its evolution. The current infrastructure shows significant limitations in network efficiency, which creates a challenge to the proposed trend of migration to Small Cells. Such densification of the RAN infrastructure, without proper redesign, leads to an untenable situation, due to both the high energy consumption in idle mode for existing networks and the excessive Base Station power, which is escalating because of the reduced efficiency of RF amplification in the higher frequency bands.

Concerns exist today about equity and accessibility to the services - both for rural areas and for underserved communities - and 5G is certainly not a panacea for such challenges. Both technology and legislation will have to work hard at alleviating the current limitations.

Network densification poses further strain on the Grid/Utility, not just because of the need to deliver more power to an increasingly complex network, but also to guarantee its reliability, which is required by the type of applications being touted by 5G.

Control of such a complex system is reaching the limit of traditional methods, thus begging for new capabilities, which are mostly sought in the realm of Artificial Intelligence, to be able to establish reliable management and coordination of each subsystem.

Furthermore, the lack of a comprehensive analysis of energy requirements and bottlenecks to achieve the expected network performance makes the Return on Investment (ROI) and Payback Period analysis quite difficult, leading to challenges in broad deployment, which would ultimately help address the 5G Equality Gap.

This is a small list of the challenges we are facing today and that will be addressed in Section 4 and will inform the roadmap in Section 5.

3.3. Drivers and Technology Targets

3.3.1. 5G&B Applications Driving EE System Design Needs

When one goes from the individual system to the network-level, needs and opportunities for EE change and scale with the network. An overview of network architecture and environmental factors helps to bring these opportunities to light. Particularly when it comes to EE, the poorest performers today are often rife with opportunities for drastic improvements. Given base stations are not only the most energy inefficient constituents of the network, but also the far and away majority consumer of global mobile telecommunications energy, it makes the most sense to focus in this area and apply the most resources toward EE optimization in base station consumption.

To start with, networking HW is subject to some of the most rigorous operating environments (i.e., extrema of external temperature/humidity, requirement to use only passive cooling, etc.) and must meet

stringent quality and reliability standards. In general, the need to design for wider environmental operating scenarios is in conflict with increased reliability and compounded by the need for passive cooling (mostly in the case of smaller cells). From an economic standpoint, the extremely competitive markets compound this by dictating razor-thin margins for carriers as they must make many assumptions and predictions about their payback model to achieve financial viability. So even taken at a high level, one can quickly gain an appreciation for the delicate art of determining a payback algorithm and how dependent it is on network performance, which is at the mercy of HW reliability and inexorably tied to energy consumption.

As seen from earlier discussions around the 5GEG, 5GEcG, and 5GDF, a PVC is only as strong as its weakest link. In this case, the maximal network performance is limited by the power/thermal bottlenecks in the PVC that limit 5GDF to its maximum value. Now we have further compounded the aforementioned challenges by also adding energy reliability (e.g., grid stability) to the mix.

The seismic shift 5G&B brings to the physical network architecture is an increased number of small cells at the edge due to the introduction of mmW spectrum. Going from the traditional, macro-base-station model to offloading on networks of small cells, a.k.a. Heterogeneous Networks (HetNet), carries a whole host of EE opportunities. At a first glance, adding a new layer of cells should lead to increased network energy consumption, but if they are deployed precisely where they are needed (traffic hotspots, city centers, etc.) there are studies showing that the offload from macro cells compensate the extra layer power consumption, especially if on/off functionality is implemented in these capacity-boosting small cells. The speed with which a smaller cell can be turned on and off differs by orders of magnitude, thus hypothetically enabling the single most EE operating mode in the world for any piece of equipment, when it is off. This translates to an ability to optimize EE based on real-time network traffic instead of prediction models, making the network more "always available" than "always on" [23].

Furthermore, the idea of applying the "always available" model applies at many different time bases. Reacting to actual traffic patterns can be adjustments in increments of minutes, seconds, and even milliseconds. Counterintuitively, the smaller the slice of time EE optimization can be applied to, the greater the yield of benefit as a whole when assessed in terms of global network energy consumption. The finest increment for savings occurs in the 10s of microseconds realm in which the 3GPP standard provisions for individual OFDM symbols to be dropped, thus mitigating wireless transmitting energy, which carries the highest PCF (and therefore overall highest cost due to lowest EE) at the utmost edge of the network.

Lastly for base stations, the incredible EE optimization opportunities at the edge translate to additional, network-wide improvements because of the shift in the point of data assessment and consumption. By now, one has likely heard many times about the momentous shift of data (and therefore energy consumption) from core toward network edge. Concepts such as Mobile Edge Computing (MEC), Edge Buffering, TinyML, mesh networking, etc. all revolve around the consideration of minimizing the movement of data as much as possible since this movement is the single biggest consumer of energy in the entire digital world (telecommunications or otherwise). EE is able to yield many of the same benefits data latency enjoys for all these same justifications. Specifics of this are analyzed in much greater detail in Section 5.3.

Bear in mind that all these challenges apply to deployments currently under way (Q1 2021 from perspective of this document release). This does not even touch the surface of the thermal/energy challenges associated with the use of mmWave, or the absolute worst-case scenario (perhaps representing the largest disconnect between the "promise of 5G" and reality), which is applying Massive

MIMO to mmWave frequency bands. Different stakeholders are motivated by different (often conflicting) priorities so it should be noted there is a need to consider energy scarcity as opposed to spectrum scarcity. These are often in conflict and attention will typically be given to spectrum optimization as that is seen closer to delivering on QoS/QoE targets. This means global EE optimization may not be apparent as operators are likely to favor spectrum in more densified environments unless they improve their localized, offloading abilities. This is discussed further in Section 5.4, where the forward-looking challenges (primarily related to thermals and packaging) in Massive MIMO and mmWave are analyzed. Alternative, potential solutions such as real time dynamic spectrum sharing as an alternative to and/or augmentation of Massive MIMO implementations are considered.

3.3.2. The Impacts of a Virtualized World

From the allocation of networking HW resources to various SW solutions at every layer of the network stack, the concept of virtually pooling resources to allocate specifically to a particular application is another critical enabler to 5G&B. The idea of using physical disaggregation to modularize resources for specific workloads has been employed in data centers for nearly 20 years now with the use of virtualization tools. Fundamentally, the idea is to pool all the (traditionally) HW resources constrained by physical systems and dole them out only as needed to serve a specific application. For instance, this can result in a physical microprocessor being virtually carved into 100s or even 1000s of pieces to be combined with other resource fragments to form a "virtual machine" serving a dedicated purpose. This methodology is what led to what is now known as Software-Defined Networking (SDN) and Network Function Virtualization (NFV).

The impacts of SDN and NFV on the data center were stark in terms of performance, EE, and even to the overall business models of this equipment was built and deployed. Without going into too much detail here (out of context), virtualization was a major reason development shifted from traditional OEMs to the many contract manufacturers (CM) that form what we refer to as "whitebox developments" today. Once the physical constraints of the box were lifted and utilization could be modularly optimized to the load, an incredible opportunity for consolidation (and therefore EE) presented itself. Similar to today's macro base stations, the servers and computer microprocessors of the early 2000s would consume a large percentage of their power budgets even when the workload was very light. With virtualization, many underutilized systems could be combined into much fewer, highly utilized systems, which resulted in achieving the two top goals of EE: 1.) Nothing saves more energy than a piece of equipment that is off. 2.) A system is most EE when running at the peak point on its efficiency vs. load curve (typically optimized for full load in systems like this).

In 5G&B and other communication networks, virtualization is making its way toward the edge and into the RAN. Virtualization of radio resources (a.k.a. - Software-Defined Radio or SDR) enabled the fast reconfiguration and/or more efficient utilization of frequency spectrum, which many in the industry consider the most precious commodity of all. Just like the creation of virtual machines in the data center, we are seeing the virtualization of network resources dedicated to specific traffic loads in what is known as Network Slicing. And just like the data center, we hope these virtual network slices and SDR Smart Radios can be utilized to optimize EE along with spectral efficiency.

3.3.3. 5G&B Business Drivers

There are many business drivers in the 5G&B ecosystem, so we have attempted to capture the most salient ones in this work as well as address some of the secondary considerations and even touch on more ancillary factors. It is no secret that the deployment of major, generational cellular infrastructure requires a massive amount of capital investment and therefore has a lot of critical payback calculation associated with determining the ROI. This CAPEX is so great that even very large, multinational players are facilitated by government subsidies and/or investment, though this can obviously vary from country to country and assorted levels/structure of local government and policy. Given the INGR has a far more technical than business focus, this EE WG strives to put a lot of this payback perspective through the filter of EE and how careful attention to this area plays a very large role in the realization of payback calculations and realized ROI. This is more obvious in some ways, such as direct OPEX associated with delivery and utilization of energy, but less obvious opportunities also exist and are not necessarily of equal weighting with the current attention given. This work and its contributors hope to really expand this perspective for all stakeholders in the 5G&B ecosystem.

QoE/QoS are typically of the utmost importance to any service provider and/or network deployment stakeholder. If the user experience suffers, then the network will not be attractive to end users and therefore fail to drive the necessary number of subscriptions (or however the network solutions are monetized) and thus make the entire effort all for naught. Analogous to EE descriptions of the PVC and use of PCF metric, most successes are dependent on the edge and work their way back from there so if the revenue is not there, then the entire ecosystem suffers as a result and is not attractive for either economic or technical resource investment.

As highlighted in the last section with the introduction of concepts and metrics such as the 5GEG, 5GEcG, PCF, and 5GDF, the ability to recognize and appreciate the need for EE goes far beyond the RAN and even has the potential to be disruptive at the grid/utility level. Whether one is an end user of the 5G&B network or has never even heard of it, pretty much any stakeholder imaginable will be very vocally upset (and publicly, especially in terms of lobbying for policy) if their power goes out and/or becomes less reliable (i.e., rolling blackouts, brownouts, etc.).

As we expand well beyond the first order in consideration of business drivers, there should be attention paid to other, dynamic game changers in the economy. The fast and meteoric rise of the Gig Economy [24] is an excellent example of this point. From a technical standpoint, a whole lot of traditional network traffic and application is moving from static management by a person sitting at a desk to highly mobile and dynamic (in terms of both economic and network demand) support model. Even the fundamentals of learning and collaboration have evolved into real time application across the world on projects and problem solving, the knowledge economy, telemedicine, logistics, maintenance and repair, etc. These changes bring forth a whole slew of challenges and pivots from the traditional ways of architecting and deploying networks and it seems this trend is bound to continue for the foreseeable future.

Last but not least, there are trends within key constituents and contributors to the network that play integral roles in the success (and therefore ROI) of the network. One example referenced up above is the concept of virtualization, which started in the data center and is now finding value in more expansive aspects of the network. The idea of being able to virtually pool a bunch of tangible, physical resources for optimal utilization yields major benefit in operational and electrical efficiencies by utilizing HW to its full potential, while turning off underutilized resources. At the RAN node level, where base stations (and even highly capable end points such as autonomous vehicles) themselves are required to essentially become mini data centers to enable the increased demand and intelligence of modern networks,

virtualization along with its network analog of network slicing are critical to delivering on something like the "promise" of 5G.

3.3.4. Data Center Efficiencies to the Edge and Corresponding Data Processing Architecture

5G networks will require both (i) more base stations per coverage area, and (as early empirical observation indicates) (ii) more energy consumption per base station than preceding technology generations. Not all energy is "created equally"; the true total cost of power is much lower when conveniently consumed near a power source, versus when consumed in a remote location with no existing power infrastructure. Furthermore, the practicality of generating power in a multi-acre, centralized location is much greater than generating power in a confined urban closet or rooftop. A confined location cannot support a towering wind turbine, diesel generator, solar farm, or hydroelectric power plant. In fact, a confined location probably cannot even consume power from a high-voltage transmission line!

Thus, the expected proliferation of 5G base stations and microcells poses a poignant business dilemma for network operators – how to get enough power to these numerous, distributed locations. Furthermore, given the power loss through the PVC to reach these remote locations, there is an environmental concern with the profound increase in total energy consumed by these emerging 5G networks. Clearly, reducing the power each base station requires is of significant value for both concerns.

Compounding the problem is the increased deployment of software-defined functions at the edge (i.e., in these base stations) to support advanced 5G-related services for content distribution and low-latency event processing. These services cast a very different shadow on the energy consumption compared to traditional communications-only services. In particular, the ratio of service value to energy consumption is far more non-linear; for example, using a neural network to recognize and count cars in a city intersection might require wildly variable amounts of compute power, depending on the traffic patterns of arriving or departing vehicles, as well as background objects that might command scaled processing attention as they enter the visual field. When implemented on conventional processing elements, AI inference algorithms can be immensely power-hungry. And that compute power demand is likely to occur at the same time as peak communications demand.

Thus, there are two concerns for energy efficiency of base stations and cells:

1. Reducing peak data/compute power consumption at each base station / cell – critical to the deployment practicality and cost of 5G services.
2. Reducing total data/compute service power consumption – important for the overall energy footprint of 5G services, and thus environmental and total cost concerns.

If we (for good reason) regard a base station as a miniature data center, then we can look to data center energy efficiency methods for help. Although building base stations next to hydroelectric dams or in the arctic tundra are not options (as they are for data centers), many of the modern energy efficiency advancements in data centers should be adopted in the 5G roadmap. However, given the constrained nature of base stations / cells, the 5G roadmap needs to go farther than conventional data center efficiency technologies. Some of the advancements are listed below:

- Active thermal management − using mechanical / fluid apparatus to reduce the energy spent cooling electronics − used in modern data centers today
- Compute and storage resource virtualization and orchestration
    - Support for minimal "working set" of resources – i.e., the ability to liberally "sleep" resources to the maximum possible extent when not needed for demand
    - Containerized applications that can be activated in real time ("real people time") with known energy requirement footprint, on any discrete computing resource within a base station working set of systems
    - Energy-sensitive, service-sensitive orchestration and workload migration – the ability to move workloads among resources to minimize the working set of powered resources while still meeting service levels
    - Specifications and Software Development Kits (SDKs) for building applications that can support the above orchestration and workload migration capability.
    - Storage tiering for energy optimization (rather than purely cost) – the ability to keep data at hand without consuming energy when not being accessed; involves energy-sensitive and service-sensitive caching algorithms
- Purpose-specific compute and storage technology enablement
    - Provisions in the roadmap to support re-usable function-dedicated resources, for example Field Programmable Gate Arrays (FPGAs), as energy optimized alternatives to general-purpose resources.
- Federation of workload resources among nearby base stations / microcells
    - The ability to provide shared, specialized services from one base station to other base-stations nearby
        - Presumes some form of high-performance backhaul meshing among proximate base stations
        - Enables some base stations to be extreme low-powered "clients" of edge-based services provided by other base stations nearby
        - Reduces the minimum energy footprint required for a microcell that exists primarily to address communications coverage problems

These capabilities and more will be necessary to achieve the coverage associated with the 5G business predictions.

Regarding the base station as an analog of the data center, then it may be useful to consider the dynamic aspects of operations of some of the most demanding data centers, the High-Performance Compute data center, especially at the "capability tier" of a national laboratory exascale (1018 floating-point operations per second) supercomputer. The capital (infrastructure) and operational costs (energy) of these systems is dominated by data movement. Exaflops of compute demanding Petabytes per second of bandwidth even if only requiring picoJoules of energy dissipates MegaWatts of power. The design goals of the exascale pursuit teams is to achieve 10X performance while capping power increase at 2X.

At the same time, these computational factories which consume data and energy and produce insight, are themselves extremely complex distributed networked systems with upwards of 100,000 compute nodes with power and cooling designs in excess of 25MW within the confines of a single data center. Failures in the power and cooling infrastructure can lead to extended downtime due to irreparable thermal damage to components as well as the loss of mission data. Straightforward instrumentation, imperative management (exhaustive if-then-else planning) and thresholding/dashboarding can lead to the situation of systems which increase in complexity while suffering reductions in reliability and availability. Work at the US National Renewable Energy Lab (NREL) on AI Operations of data center compute and infrastructure [25] applies AI-based anomaly detection to find faults early and reduce false positives, correlation engines leverage domain expertise to link metrics with causal behavior, and neural networks to identify metrics.

3.3.5. Network Energy Architecture

Power and thermal management, both in terms of telemetry as well as control, has been an active area of concern and industry development for data center equipment for some time. For more than a decade, the operational costs of an industry standard rack server have been a significant multiple of the capital costs of the equipment. The desire to gain the most demonstrable benefit from every single Watt dissipated in a modern data center has driven such diverse innovations as virtualization and containerization in software, Graphic Processing Units (GPUs) and a new class of application specific accelerators in hardware and novel power delivery and thermal management technologies such as high voltage DC distribution to the rack and ambient liquid cooling. The desire for interoperability and competitive choice in all of equipment choices of servers, storage, and networking have given rise to industry standards of manageability across equipment categories and vendors which could be considered for adoption into the network energy architecture for 5G and beyond. In particular, the DMTF actively publishes the Redfish [26] standards for management of equipment targeting the Software Defined Data Center and is currently drafting enhancements targeting power and thermal management inclusive of novel power delivery and cooling mechanisms [27].

3.3.6. RF Base Stations Today

A traditional base station consists of three main components: a baseband unit (BBU) that takes care of digital signal processing, a radio unit that creates the analog radiofrequency (RF) signal, and a passive antenna that emits the RF signals with a constant radiation pattern. Due to the size and weight constraints of masts and towers, the radio and BBU were traditionally deployed underneath, and a long RF feeder cable was needed between the antenna and radio, resulting in substantial power losses. This is illustrated as "Step 1" in the figure below. A single BBU can support multiple radios that are deployed on the same site, which might cover different frequency bands or cell sectors (this is not illustrated in the figure).

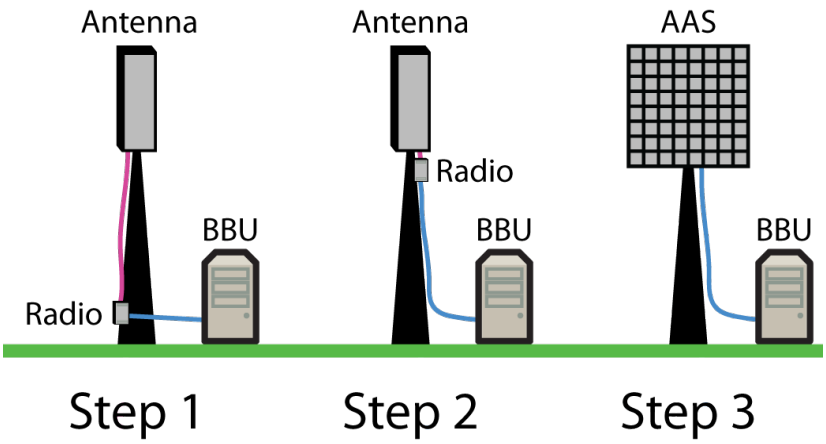


Figure 6. [28]

When the radio hardware is reduced in size, it becomes common to use remote radio units that are deployed in the tower, close to the antenna instead of close to the BBU. This is denoted as "Step 2" in the figure above and became common in the 4G era. Only a short RF feeder cable is then needed, while an optical fiber can be drawn from the BBU to the radio. The next step in the development is active antennas that integrate the antenna and radio into a single unit. There are many types of active antennas, from single-antenna units with constant radiation patterns, via phased arrays units with a single input that is fed to many antenna elements via phase-shifters that enable a steerable radiation pattern, to fully digital "Massive MIMO" with many individually radios that are connected to one or a few antenna elements each. The latter two options will be utilized in 5G and beyond, and the term advanced antenna system (AAS) is being used in the industry to describe this category. This setup is denoted as "Step 3" in the figure. To limit the capacity of the optical fiber between the AAS and BBU, an AAS might perform a limited set of baseband processing to compress/decompress the signals. The BBU doesn't have to be underneath the tower, as illustrated in the figure, but can be moved to a nearby edge cloud that is shared by many base stations.

There are three classes of AAS that are important to discuss from an energy efficiency perspective:

- Phase-shifter-based Phased Array: A single radio unit is connected to an array of antenna elements. The elements are phase synchronized at the carrier frequency and tunable phase shifter modules are placed between the single source and each antenna element which allow the array to perform beam steering; that is, controlling the angular direction of the transmitted downlink beam (and the corresponding angular sensitivity in the uplink) without any external feedback. This type of array is well suited for deployments where a single beam is to be transmitted at a time and the propagation channel is dominated by a single angular direction (typically a line-of-sight path). One example is rural areas with relatively low traffic and where a steerable beam is preferred over a fixed beam since the coverage area is wide in the azimuth domain and a high gain steerable beam can put the power on where the current customer(s) reside and not on empty regions. The first 5G millimeter-wave products for outdoor line-of-sight deployments are making use of this technology.
- Massive MIMO: A large number of radio units are connected to a correspondingly large number of antenna elements. The radios might share a common oscillator but are not necessarily phase synchronized to the point that this can be utilized to steer beams in known angular directions.

Instead, the phase differences are treated as part of the propagation channel and estimated in the uplink or downlink from pilot reference signals, once per coherence time interval. Having multiple radios give two key benefits: Multiple streams of data can be spatially multiplexed (transmitted at the same time and frequency) and different beams can be used in different parts of the band. The first feature is desirable in deployment scenarios where there are many users, while the second feature is desirable in propagation scenarios with substantial multipath propagation (leading to frequency-selective channels). These features are particularly useful in urban deployments, but also some suburban cases. The first 5G deployments in the 3 GHz band are utilizing this technology. The technology can be utilized in both Frequency Division Duplex (FDD) or Time Division Duplex (TDD) spectrum, but the TDD case is preferred since the number of "sounding" signals is then proportional to the number of users, rather than the number of radios at the base station. The reason is that one can then send sounding signals only in the uplink and let an arbitrary number of radios receive the signals simultaneously. The uplink channel estimates are used both for uplink data reception and downlink data transmission, enabled by the fact that TDD systems use the same frequency band in both directions. The vast majority of deployments are in TDD bands. Since the multipath propagation characteristics sensed by the pilot signals are highly frequency dependent, the sounding signals must be spread over both time and frequency.

- Digital at the Edge MIMO Array: This is an emerging technology capability. Recent advances in direct digital synthesis of RF waveforms and ultra-low phase noise RF power modulators have demonstrated the ability to deliver the required phase coherence for phased-array-like performance from an array of direct digital synthesis RF modules connected to a single clock. This eliminates the need for costly and lossy phase shifter modules because this function is now performed digitally within the signal streams passed to the digital RF generators placed at each antenna element. A key capability of this digital technique is the ability to steer different subcarriers of an OFDM waveform to different azimuth-elevation angles from the array, thus it combines the best aspects of phased arrays and the first Massive MIMO products [24]. If the location of the customers is known a priori, this type of array does not require pilot signals for beam configuration but only to enable coherent demodulation of data. This feature is particularly useful in rural environments where customer connections may be grouped together in localized enclaves, which are constant for long periods but may move as a function of time of day, and where the interstitial spaces between the enclaves require minimal service.

It is not a matter of choosing one technology over the other, but an energy efficient deployment relies on deploying the right hardware and the right place. If we compare two systems having the same total radiated power, aperture, and radiation pattern, then the distinguishing factor from an energy efficiency perspective is the power dissipation in the RF and BBU. It will clearly be higher in a Massive MIMO system if the same components are being utilized, but by tailoring the components to the special properties of Massive MIMO (e.g., low-power radios and simple signal processing), there is hope that the differences will vanish with time [29] [30]. Each type of AAS has its own strengths. With both phased arrays, Massive MIMO, and their evolutions, a key feature is the ability to put the power on the customer and not on the empty cornfields in between groups of customers. Moreover, the key feature of Massive MIMO is the spatial multiplexing: One can achieve higher capacity in situations where this is needed.

Finally, it may be the case that in 10 years the above categorizations evolve into a distinction without a difference. The rapid evolution of digital at the edge RF modules may make all of the hardware

configurations become similar where the only difference between systems is the back-end software configuration and future systems could, in theory, shift on the fly between MIMO and multi-beam phased array operations.

3.3.6.1. Renewable Energy-Enabled Cellular Networks

Renewable energy (RE)-powered base stations (BSs) have been considered as an attractive way to support the transition to a low carbon footprint for cellular networks. These types of RE-powered BSs are also very useful for the regions where reliable power grids are insufficient and infeasible to deploy such as far-flung, low-density population, and/or hard-to-reach areas. A recent survey paper [31], provided an up-to-date overview of RE-enabled cellular networks, detailing their analysis, classification, and related works. In particular, this paper introduced first some of the important components of RE-powered BSs along with their frequently adopted models. The paper offered then a variety of strategies and design issues for RE-powered BSs that can be incorporated into cellular networks and categorize them into several groups. Finally, the paper introduced feasibility studies on RE-powered before suggesting future research directions on RE-enabled cellular networks from the perspective of the new characteristics of the emerging Internet of Everything paradigm.

3.3.7. The Role of AI Deep Learning

Energy consumption is set to increase dramatically for 5G just on the basis of the required hardware deployment. To that, we need to add the impact of AI applications required to manage and optimize the complex communication system infrastructure.

ML processes are both a "user" of the Communication Networks and a tool to improve their efficiency. As such, we need to understand the different impact that utilization of ML has under the two different scenarios.

3.3.8. Applications Deployment Optimization

As currently developed in many market areas, the use of ML can provide substantial improvements to the effectiveness of applications deployed both locally and/or globally. The expected gains are largest at layers where the models that have been used to develop traditional algorithms are insufficient, or where good algorithms exist in theory but are too complex in reality. Since such AI processes are extremely computationally intensive and require a large amount of energy to implement, the first step required in the deployment of such processes across the network is the optimization of their execution to balance energy utilization across the different nodes.

Such efforts are accomplished by distributing the AI and other computational processes across the different layers of the system: device-level, edge computing, and cloud computing (also see the above example of autonomous driving). Not only the planning of the data architecture and computational execution needs to be optimized, but also the dynamic performance and availability of hardware components, which can be throttled/augmented as the demand fluctuates, both unpredictably and predictably (i.e., time of day, day of week, etc.). These are the type of time-varying and non-intuitive characteristics that traditional algorithms often fail to model and utilize properly. Studies are required to assess where the benefits of AI algorithms can overcome the additional energy required to run them.

Performance requirements and economic trade-offs in each use case may determine how to distribute functions across the network: big data aggregation (especially for ML training purposes) with longer access latency would reside in the Cloud, while customized data with low access latency would be hosted at the Edge, and vision/sensor aggregation would occur in the device or UE. Inference and use case optimization may reside between Edge and Devices.

This way, an application computing requirement can take advantage of large-scale computing as well as fast response time, while accommodating for limited energy resources available in battery-operated devices.

3.3.9. AI Use for Network Optimization

The second area of system optimization that can take advantage of ML involves dynamically reshaping the infrastructure to respond to the dynamic demand generated by users and applications. This requires intelligent routing of users and communication processes, especially within heterogeneous RAN environments [32] [33], in:

- Spectrum optimization
- Traffic management
- Self-healing networks
- Network Security
- Network services

While current network optimization is fairly static and typically human-driven, the complexity requirement of dynamic heterogeneous-networks optimization will only be possible through autonomous systems, which are capable of adapting communication links, bands orchestration and RAN configuration to efficiently address user requirements.

The speed of light multiplied with the latency requirement determines how far away from the base station that a processing task can be carried out. While beamforming, scheduling, and spectrum management has to occur within the cell hardware or at a nearby edge cloud computer, to achieve a latency in the sub-ms range, other optimization processes can handle larger latencies, thus enabling more complex control structures that can be distributed across the RAN, or even the whole network.

More sophisticated control is required, for example, when scheduling base station shedding and network re-configuration. One example is to put base stations into sleep mode when the traffic is low (e.g., at the late nights), while still enabling the occasional users to access the network. Since the traffic pattern in a given cell is hard to model a priori, a data-driven approach where an AI predicts the usage patterns and makes decisions on what hardware to shut down and what the duty cycle should be. Network densification creates the opportunity for a more nuanced approach, where load balancing allows trading off energy efficiency for user level performance and one can turn off a subset of the radios in a Massive MIMO array. If two base stations operating in different bands are deployed on the same site, the one operating in the highest band can be turned off without reducing the network's coverage area.

The third level of optimization is further out in the future, and entails the ability to operate at what we call the "Systems-of-Systems" level. At this level, all subsystems and infrastructures (RAN, distributed computing, energy delivery, security, etc.) can be dynamically orchestrated to deliver optimal performance at the lowest (or affordable) cost, thus creating system-level savings at a scale that cannot

be obtained when having only visibility at the subsystem level. This approach can be utilized to optimize different metrics, including but not limited to energy efficiency. The implementation of this vision is explained in the following sections.

4. FUTURE STATE (2032)

With such a vast umbrella of coverage in technology, stakeholders, application needs, and market predictions, the further the outlook, the greater the opportunity for variability, error, and understatement or overstatement of forward-looking needs. These points are conflated that much more in the context of 5G&B due to the extensive timescales required to even fully deploy a new network generation.

This work delivers good-faith efforts to culminate extensive past experience and marry it with the state-of-the-art to provide a detailed articulation of the extensive opportunities and challenges for EE optimization in telecommunications in order to paint a pragmatic picture of the future. This roadmap has the objective of looking out a decade in time, which is essentially just the next "G" as viewed from the lens of this highly, multidisciplinary team of dedicated technologists so bear that in mind when digesting our vision of the needs and path to 6G (from today's perspective).

4.1. Vision of Future Technology

Any system designer, solution architect, or network operator will typically have some kind of desire and/or strategy for improving energy efficiency, but dramatic, global change will only be seen with the embracement of the many philosophies and optimization techniques summarized by this work. Such a melting pot of stakeholders (in any industry, not just specific to telecommunications) must work together in a common spirit with a unified goal of driving EE into deployed solutions, which is not done today. Given the many moving pieces of a large network and even vastly different development timelines across the spectrum (i.e., 1-2 yrs for a server or radio, decades for a utility grid), planning for the future must be part of today's priorities as well as tomorrow's.

If following the advice and SoS framework laid bare by this document, then there should be momentum within this melting pot to start harmonizing languages/concepts initially (a primary objective of this work) and ideally move forward with standardization and execution of metrics like PCF and KPIs like 5GDF. 5GDF assessment within the SoS framework is meant to be a very powerful tool that enables people to assess their ideal solutions (i.e., 5GDF = 1.0), closer in alignment with current thinking, and apply more real-world filters to those idealities and end up with a far more pragmatic prediction for performance and economic payback even before actual deployments occur, and application data can be analyzed. As we look to 5G&B, it becomes increasingly important to automate this kind of intelligence and further enable with HW and SW hooks at descending levels of granularity. Today we do much of this kind of EE optimization in the data center and are starting to apply to the RAN, but do very little of this at the utility grid level.

Looking to the future, we need to not only apply EE optimization at all levels, but need to do so with increasing cooperation. The ultimate goal should be individual and collaborative EE optimization from chip to grid. Furthermore, this goal should be inclusive of completely bidirectional feedback enabled by communication links and intelligent power management between every network and grid constituent. Power/energy management solutions and/or subsystems are the most obvious place for such communication links to exist. The kind of end-to-end optimization described here can only happen if

such solutions are enabled with interfaces and protocols allowing the solutions to be aware of each other as well as inform each other of their intentions. Intentions in this context may be a server's desire to consolidate virtual loads to another machine and turn itself off, or it could be a base station (or group of regional base stations) sensing a large change in traffic loading and requesting more efficient spectrum to operate in.

If we truly shoot for the moon, then we can imagine a future in which all major loads and sources talk to each other in the spirit of perpetual and global EE optimization. Imagine a networked world in which performance optimization merges seamlessly with economic/resource optimization. Envision a truly Smart Grid that communicates the real-time price of energy on a dynamic market that may change in intervals down to as little as five minutes, which trickles down to the edge and instantiates a network-level EE optimization of all constituents. This may sound highly optimistic, but whatever perspective one wants to focus on in this material from energy to thermal limitations to socioeconomic disparities, we are simply not on a sustainable path to the future and without this kind of optimism becoming highly contagious, we are not correcting the course either. The more promising revelation is that the many, many bits and pieces required to turn the optimism into reality are being investigated, many of which are outlined in Section 5 of this document.

4.1.1. Cell-free Architectures

The cellular divide-and-conquer approach to organizing the network infrastructure is gradually coming towards an end. In the future, the network will be cell-free and user-centric, which means that every user device is served by a variety of neighboring base stations. Instead of assigning each user to a single base station, the network is organically identifying the best combination of distributed antennas and wireless resources to serve each user. This is the end to a journey where the networks will first become increasingly densified in terms of deploying small cells, addressing demand and begin to reduce the propagation losses. Additionally, some of these systems will be equipped with Massive MIMO functionality to target radiated power to the specific users further reducing the ambient radiated interference. Finally, an extra software level that enables cell-free operation can be utilized to mitigate the interference that otherwise limits the efficiency of dense HetNet small-cell networks. This flexible coverage architecture will require Intelligent Power Management throughout the various layers, coordinating download assets, turning on/off access points and supplemental downlink channels as needed.

4.1.2. Ubiquitous HetNets of Small Cells

Ubiquitous HetNets demonstrate their benefits when service providers blend their networks with the WiFi and shared spectrum, making energy and service the primary objective for the end users. These transformations will enable service providers to provide the high reliability 5G communications with low latency by providing improved fundamental infrastructure in their captive spectral assets, augmented with the shared spectrum assets available.

Many of the current service provider cell deployments have amplifiers with excessive distortion, limiting the allowed throughput and increasing the radiated power. These base station deployments are to address the current peak capacity needs, but the transmit power and energy utilization is excessive (primarily due to the fact they are built with historical architectures). Most of these networks have been structured with layers, but have not been updated with systems that maximize spectral utilization. As

systems are deployed with high fidelity outputs, the signal power can be reduced through beam forming and directing signals to specific users, allowing Tx bursts at higher data rates, getting packets of data to be delivered in a shorter period of time. Trends over the next 10 years will utilize smaller cells, coordinated signal layering, high fidelity transmission and directed signaling so that the wireless signals will be operated at lower power levels and converted to the conducted system as quickly as possible.

Layering of cells will allow the spectral separation of coordinating messaging from data packets. The coordinating messaging should be transmitted at a higher power on channels at frequencies that have wider coverage and penetrate into buildings more effectively since they are more critical and much shorter. The data packets can be delivered to the user terminal through supplemental downlink channels as the user terminal moves into areas of high-fidelity coverage (which can be tracked with historical behavior and AI). In addition, this data deployment can be achieved in bursts. As a result of this layering, the total radiated power will be reduced, lowering the self interference realized, reducing the radiated power footprint, and increasing the overall capacity of the system. Lower power supplemental downlink cell sites with Intelligent Power Management (IPM) will reduce the overhead by converting AC units to fans and passive cooling and intermittent operation based on demand.

A future dynamic HetNet system will have the above features, plus the ability to deactivate resources when not needed and determine handoffs between cells dynamically, based on which cells are available and currently awake. For the system to have this level of dynamic operation, there needs to be sufficient communication between layers, with IPM control and reconfigurability.

Full HetNet, will be disruptive to service providers in that there will be a seamless handoff and cooperation between corporate entities is needed. Part of the benefit of this cooperation is the sharing of cost sharing the OPEX associated with the power consumption of the WiFi access points with the end user. As part of this cooperation there needs to be the HetNet IPM management of that spectral resource.

Migration to small cells has started and will continue to reduce the ambient radiated interference or OTA noise limited operation as we move into the future with the rollout and coordination of these HetNet technologies:

- Captured vs. Shared Spectrum, using WiFi or Long-Term Evolution Unlicensed Spectrum (LTE-U)
- mmWave vs. sub-6 Ghz
- Impact of Mesh networks in the future.

We are still missing key pieces that will be presented in section 5.3.2. Specifically, the "power/energy" model of the different layers and the relationship to spectrum, Tx fidelity, link budgets (over frequency) and data throughput. This is not just needed for HetNet systems, but the overall multi-layered wireless deployment.

4.1.3. Enabling/Deploying Energy-optimal Control Feedback Loop(s)

Ideally, we shall one day live in a world when all power supplies and powered devices talk to each other with the utopian goal of universal EE optimization. Put more definitively, we hope to see expanded control planes and feedback paths that continue to grow outward from the lowest level of granularity (e.g., individual black boxes in SoS block diagram, see Fig. 7 in Section 4.2.1 below) all the way up to the complete network/grid levels. This work explores a handful of proposed concepts, methodologies, and existing efforts to achieve all of this.

In the very general sense, some key initiatives driving EE optimization feedback opportunities (to be utilized at all levels) in this area are captured as follows:

- Turn off what is not needed (i.e., provide such design hooks in HW/SW)
- Throttle what is under utilized (i.e., optimize low-power modes)
- Consolidate what is partially used (i.e., provide dynamic reallocation of tasks)
- Create a real time analysis and forecast of utilization at all levels (i.e., need the ability to communicate parameters up and down the chain) so that different control planes can leverage the information

It is important to consider black box perspectives from a global viewpoint to identify existing hooks/gaps at black box-level and propose how to connect the boxes, then focus on how to apply feedback (in each direction) and optimize all black boxes for energy consumption. Some black boxes have existing HW/SW infrastructure for detailed energy management (i.e., server power capping), where others may have more need for enhancement to enable this functionality. Imagine giving a direction on power capping to HW, then have it make optimal adjustments as well as provide feedback (forward and backwards) for consideration in all black boxes.

The SoS framework introduced below and further defined in the "Energy-Efficient Architectural Framework" content (Section 4.2) combines some historical/existing approaches and digests them in a way that has led to the more novel approach and specific methodologies of the SoS. Section 5.6 (NEED 5 - Grid/Utility) delves deeper into how the SoS can be applied specifically to the black box constituents involved in the grid/utility level, particularly in the roadmap structure and perceived timing outlooks.

4.1.4. Model Complexity

The 5G&B increased communications capabilities (increased bandwidth, ubiquitous connectivity transfer of data at unprecedented rates among users, edge devices and powerful computing platforms) will enhance abilities in applications and create new applications in all sectors. Both the 5G&B infrastructures as well as the increasingly complex applications systems, already at play and more so in the future, are not and will not be operating in an isolated fashion, but in concert with other systems and under dynamic conditions. Overall, they interact and interoperate with other systems as "systems-of-systems". This is the case whether they are natural/physical systems, engineered or other systems, such as civil infrastructures, business-processes, social-systems, and defense infrastructures, to name some, together with the 5G&B infrastructures themselves.

Such capabilities cannot be enabled through ad-hoc and/or static methods. We need more systematic methods, whereby comprehensive, principle-based models of systems and their components are used not only for the design phase of such systems, but also in their operational cycle. This allows cognizant, real-time decision support, as well as other aspects – systems interoperability, evolution, maintenance, test & evaluation; whereby such capabilities are based on more comprehensive methods rather than simple decision tables or data-fitting methods such as ML. We should also emphasize here that "data analytics" (such as that provided through ML-methods per se) are not sufficient. We need "systems analytics," which reflect the dynamically changing conditions and the complex interactions across constituent components of a system (internal factors) as well as external factors that are manifested when a system interoperates or is impacted by the behaviors of other systems (for example adverse weather impact on power-grids; or impact of time-varying renewable power-grid resources; or time-

varying demand of multiple consumers with multiple levels of priorities). ML algorithms alone are not sufficient to characterize dynamic systems and their behaviors. In addition, ML algorithms (or ML-models) alone suffer from issues of "transparency" – that is it may not be fully understood how does the MLmodel change itself based on new training data. The ML algorithms also suffer an "interpretability" issue in which it may not be fully understood how well the changed algorithm/model still represents the system - accurately or adequately, and whether the inferences for decision-making are for example still valid. Thus, the ML algorithm or model can go rogue.

The methods we consider for addressing the needs of such complex environments involve systems-of-systems modeling defined so as to allow multiple-levels of detail and abstraction for predictive behavior of individual components and the system in which these components belong to, as well as when that system needs to interoperate with other systems. Here we use the term modeling to encompass simulation, emulation, as well as numerical, statistical, graph-based modeling.

In particular, we are considering methods that apply the Dynamic Data Driven Applications Systems (DDDAS)/Infosymbiotics paradigm [34] [35], whereby executing application models can incorporate additional data (referred to as "dynamic data" inputs, and collected in real-time or archival) into targeted parts of the model phase-space in order to make the model more accurate or replace parts of the computation to make the model faster. Also, in reverse, the model controls the instrumentation (search in the data space or adaptively coordinate multiple and heterogeneous data measurements and their resources – sensors and actuators of controllers).

Furthermore, methods and capabilities that have been developed under the DDDAS rubric can use off-line and on-line full-scale solutions integrated with real-time sensor and actuator data to extrapolate or interpolate through the manifold of the solution space, and derive real-time assessments with the accuracy of full-scale models. DDDAS-based modeling methods can be considered as methods where the model learns from data. However, distinct from ML methods that can go rogue, in DDDAS methods the model is cognizant of the system and the physics of the system. Because of this cognizance, it can create safeguards, so the DDDAS-based modeling does not go rogue. ML and the related methods can be a useful tool together with the more comprehensive modeling methods (such as DDDAS) to provide cognizant decision-support.

Examples of use of DDDAS cover many areas [36] from aerospace applications (structural health monitoring and mission planning [37]), and electrical power-grids (for optimized resource scheduling under variable power resource availability or sudden disruptions, and multiple consumers and multiple levels of priorities per consumer [38] [39] [40]). In addition, examples of other capabilities include smart cities applications (smart transportation [41]), multi-platform coordination for situational awareness [42] [43] [44], energy-aware optimizations [45], and cybersecurity [46] [47]. Such methods and capabilities have counterparts in adaptive and optimized management of 5G&B infrastructures.

See Example 1 (later shown in Section 4.2.2.1).

4.1.5. Model Validation

Validation [42] is the process of determining the accuracy by which a model (or the collection of models in the case of multi-modal, multi-level, multi-scale modeling) represents the actual system, the accuracy by which it can characterize and describe the actual system, and/or the accuracy by which the model can predict behaviors of a system. Verification [48] checks whether a product, service, or system meets a set of design specifications. In the case of modeling, for example, it checks the accuracy by which a

computational representation (computer program) of the mathematical model implements the mathematical model. Validation asks: are the right equations used and solved mathematically? Verification asks: does the computer program solve these equations correctly?

For the complex systems-of-systems we are concerned with here, including electrical power grids and networking infrastructures, such as those emerging with 5G&Beyond, end-to-end delivery of service guarantees, anomalies or the onset of such potential disruptions, and resilience under such conditions, are difficult to model with traditional methods. DDDAS-based methods [49] [50] [51] bring the power of system-cognizant models and adaptive model-based real-time monitoring, to effectively utilize real-time and archival data for real-time or near-real time response to ensure optimized and continuing operation of the systems at hand.

DDDAS-based models on one hand need to adhere to the traditional validation and verification requirements, and their convergence properties under perturbations from the dynamic-data inputs need to be assessed, and error bounds need to be determined. On the other hand, DDDAS approaches inherently support continuous test and evaluation of a model, and thus continuous validation and verification of the modeling used, as the system they represent evolves or needs to interoperate with other systems. That is, since DDDAS-based models continually interact (integrate) with instrumentation, they can be continually validated (as the system they represent evolves, or as it interacts with other systems, etc.); and the results (computed data) they produce can be compared with the instrumentation data. Additionally, DDDAS-based methods can use the dynamic data to invoke models and data of the appropriate levels of fidelity to represent the conditions at hand, thus optimally satisfying validation and verification requirements.

4.2. Energy-Efficient Architectural Framework

4.2.1. Overview of Systems of Systems (SoS)

As we shared in the previous section, the complexity and heterogeneous nature of the stakeholders in the 5G ecosystem is a major barrier to addressing energy efficiency in a comprehensive way, thus often leading to local optimization, which helps push to the market a particular hardware or software, but with the global effect of actually increasing energy consumption. This ineffectiveness at driving system efficiency by focusing on just one aspect of the ecosystem has led the contributors in this Chapter to seek a more global solution, which can only be addressed by the development of a model that encompasses all subsystems: by definition, these are “Systems of Systems” (SoS), and the ability to model their interactions both statically and in real time is the objective of our analysis.

Additionally, the time frame of intervention of the multiple control planes spans across many orders of magnitude in the time domain (which helps in the nesting of control planes), as beamforming, scheduling and fast spectrum management can occur in the 10s μs to few ms; while handovers, QoS and load balancing decisions occur in the 10s to 100s ms; and finally, optimization via analytics and automated management can be accomplished in a much longer time frame.

Solving such complexity and heterogeneity requires establishing a system description and modeling approach in which all these black boxes can “talk” to each other in our “global currency” of energy. Once all black box outputs are translated into a global currency, both static and dynamic analyses can be performed to provide an extremely useful tool that allows any one network constituent to assess the state of any network configuration as well as assess the impact of any black box (or boxes) on another. Ultimately, the best situation is to have such a tool that can also be used to optimize a network

configuration for energy efficiency. At the very least, it should be able to recognize network bottlenecks and provide reporting and/or recommendations on how to alleviate these bottlenecks and/or optimize specific blocks for maximal energy efficiency.

We are making an attempt to address the near-impossible feat of providing a mechanism that can enable a technical and/or business analysis across the entire network so it requires some kind of standardized process any stakeholder can follow to apply their expertise to a particular black box' content. One can argue every single black box in the SoS has its own set of unique inputs though a compromise must be found to not only enable translation into common outputs in the universal currency, but also incorporate network or performance-related inputs (i.e., QoS/QoE, payback periods, real-time energy pricing, utility caps, etc.).

Any set of boxes can be chained together in a system-wide or network-wide analysis that evaluates energy from source to load (sometimes bidirectionally) in the Power Value Chain (PVC), which was previously described. The PVC is a systematic representation, which describes the energy flow across all the distribution/conversion steps between source and load, which tie together the siloed stakeholders. In order to provide comparable quantification, we also previously introduced the metric of Power Cost Factor (PCF). PCF is a unitless number that represents the multiplication factor required to quantitatively assess the overall cost of energy utilization at any given point within the PVC. Ultimately, a system is assessed for bottlenecks and used to determine an overall metric we call the 5G Derate Factor (5GDF). The 5GDF is a unitless coefficient (<1) representing a scaling factor for the application of technical and economic risk factors to the ideal 5G network deployment model that will reduce the optimal, maximum designed capabilities of a network due to energy-limited and/or economically-limited and/or socioeconomically-limited factors.

The proposed methodology is to break the analytical approach into three, independent methods of analysis:

1. **THE BLOCK ANALYSIS** = this is where each black box captures the unique inputs that are required for that particular box to be able to output the common, energy-related and utilization values. The stakeholders and subject matter experts (SME) must determine the most important factors to impact the box's energy footprint and help to craft the formula(e)/model(s) that will yield these outputs. The template model for the block analysis is shown in Figure 7 (below).
    a. NOTE: this is the key analysis for ANY stakeholder in the 5G network to participate in, and enable the full SoS tool.
    b. Mathematically, if X is the input vector and Y is the output vector, then for each of the output variables the block needs to have a definition of yi=fi (x1, x2, …xn), where i=1 to n and n is the number of output variables.

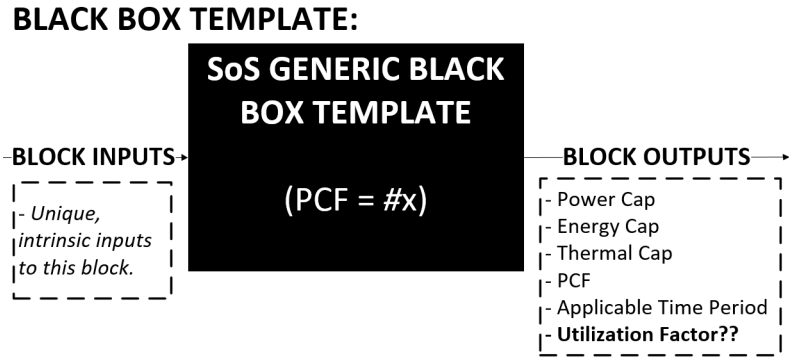


Figure 7. The Systems-of-Systems (SoS) Block Analysis Template Model

Image courtesy of PowerRox

2. The inputs represent the "performance demand" placed on that block: this can be a static set of values or a dynamic one, depending on the analysis we want to conduct.THE PVC CHAIN ANALYSIS = this is a static analysis in which all black box outputs are captured and assessed to determine the power/energy/thermal bottlenecks for the given chain. These bottlenecks can be used to estimate a preliminary 5GDF and the gaps from an ideal value of 1.0. An assessment of relative PCF is performed in this step to enable recommendations for optimization. The template model for the block analysis is shown in Figure 8 (below).

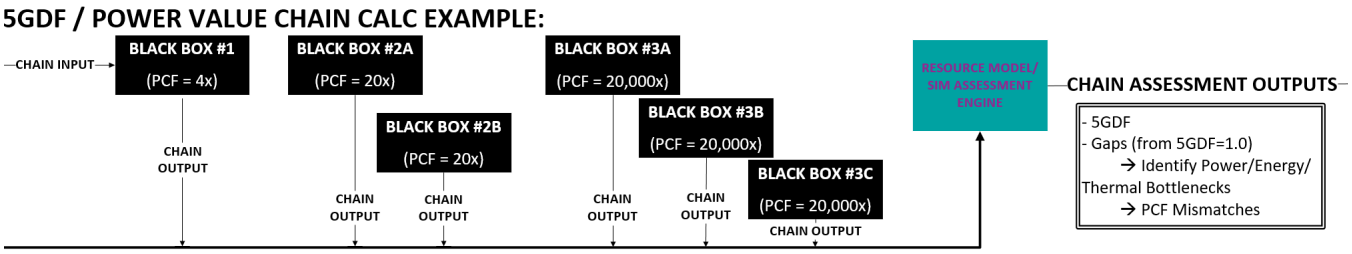


Figure 8. The Systems-of-Systems (SoS) Power Value Chain (PVC) Chain Analysis (Static) Example Flow

Image courtesy of PowerRox

A first analysis can be done by assuming that the system is operating at full capacity (i.e., maximum utilization of each block). This will provide the "energy footprint" of the system, under the assumption that each block has been designed to carry exactly the required workload.

A second analysis can be done by applying a "static" workload to the system, which is based on the performance expectations for which the system was designed. In this case, we will note that each block has a different "utilization factor," i.e., some blocks' capabilities cannot be fully utilized, even when the system is fully loaded (e.g., a computing device is I/O limited and all cores cannot be fully engaged; or an RF Front End is incapable of servicing enough customers to "fill" the data pipeline, etc.), while other blocks reach the limits on their output variables without being able to deliver the required performance (e.g., an RF Power Amplifier reaches its thermal limit and cannot deliver the communication to a distant user; or a communication node cannot manage all the requested data traffic). This analysis will therefore pinpoint which blocks

constitute a bottleneck for the system to achieve the promised performance and will also indicate the cause of the limitation, thus providing a derating factor for the system performance (5GDF) and the ability to assess the upgrades required to deliver the required performance.

This analysis will also likely point to geographical or temporal unbalances in the services provided and facilitate a discussion on the investments required to deliver the expected performance.

Since many different "loading conditions" can be applied as input vectors to the system, the ability to deliver on specific applications (and the CAPEX and OPEX required to support them) can be assessed (e.g., what are the system-level requirements to support autonomous driving?).

This proposed static analysis provides a "worst-case scenario" to the required investments; in the next section, we will address how a dynamic analysis (i.e., the ability to change over time the input vectors for the blocks) may provide both a more realistic scenario as well as the ability to optimize the system efficiency.

3. **THE NETWORK/SYSTEM OPTIMIZATION CHAIN ANALYSIS** = this is a dynamic analysis in which the static analysis is now combined with the performance requirements/inputs to perform more economically-focused and what-if type of analyses based on performance and/or economic targets. This is the method that will instantiate engines for assessing payback period and optimizing a given chain for energy efficiency, thus outputting true limitations to achievable 5GDF and predictions for consumption of resources and cost estimates. The template model for the block analysis is shown in Figure 9 (below).

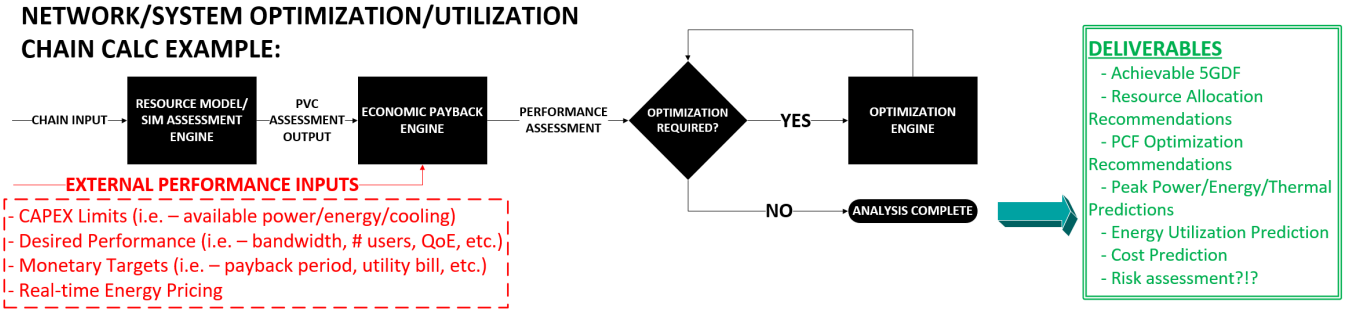


Figure 9. The Systems-of-Systems (SoS) Network/System Optimization Chain Analysis (Dynamic) Example Flow

Image courtesy of PowerRox

Having the data available on energy utilization as well as bottlenecks, we can see study various optimization options (e.g., re-routing users on a different frequency band, cell tower or network; or moving computing tasks over different resources – in the cloud or at the edge). Besides the already described block-level functions, we can also consider additional inputs to the sub-blocks (Figure 10): for example, where the energy is coming from (the grid, renewable sources or local energy storage) and the cost of energy (which varies both as a function of the source as well as the time of use).

This is obviously a very ambitious goal, which seems difficult to achieve when little has been assembled, but the ability to distribute a "black-box" model enables us to unify and coalesce the inputs from a very diverse and siloed ecosystem into a model that can continue to grow step by step.

Additionally, simply changing the "specs" of the blocks with our assumptions for future available performances will allow us to both estimate cost and performance three, five and ten years from now, as well as to understand where new bottlenecks will emerge and therefore where more innovation is required to achieve a fruitful and cost-effective system upgrade.

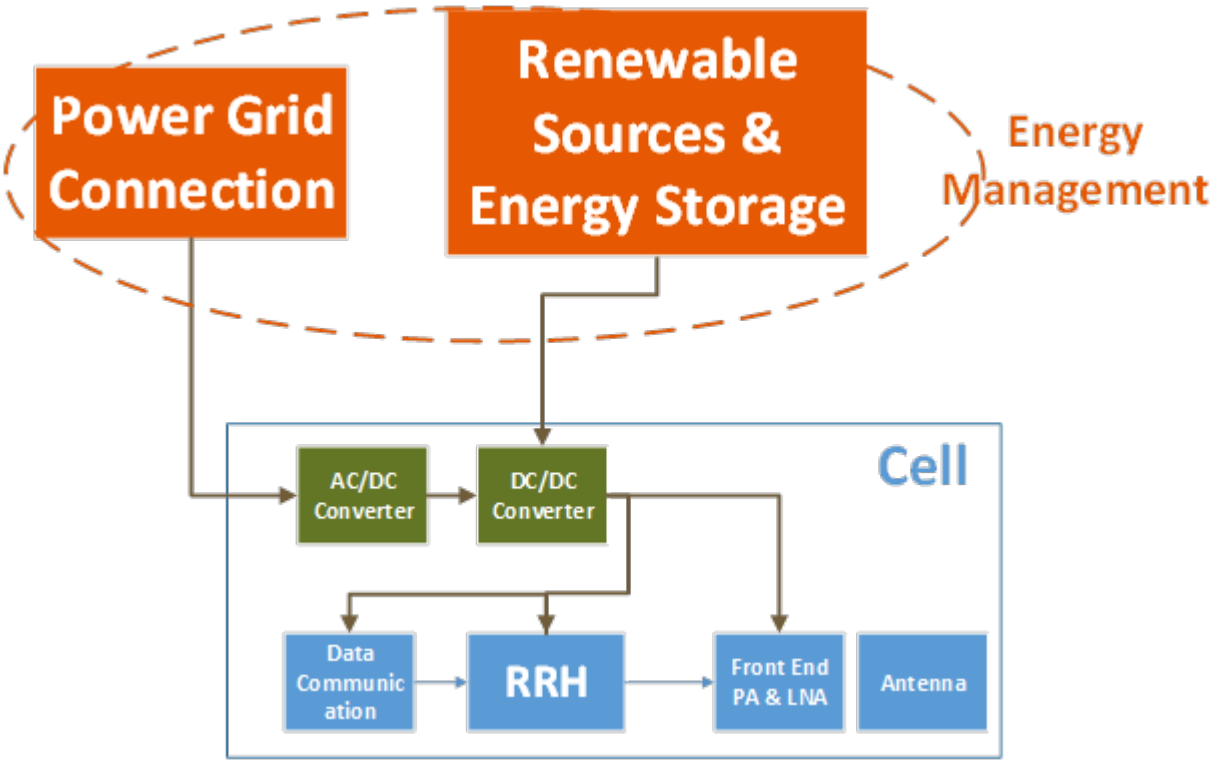


Figure 10. Management of Energy Sources as an additional input to sub-block level description

Image courtesy of IoTissimo

The framework described above is the initial nucleus by which we can construct a systems-of-systems representation of the 5G ecosystem. By adding a data plane and a multi-level control plane, we can now both statically assess the system and its bottlenecks, as well as dynamically (and in real time) exercise the model to identify its behavior and develop optimization strategies. Having all subsystems connected, the simulation can leverage the different layers of control to take into account multiple constraints, including local energy availability and dynamic pricing, network status, hardware capabilities, thermal constraints and weather impact on subsystem performance. A graphic example of a portion of the system, which includes a RAN subsystem, is shown in Figure 11.

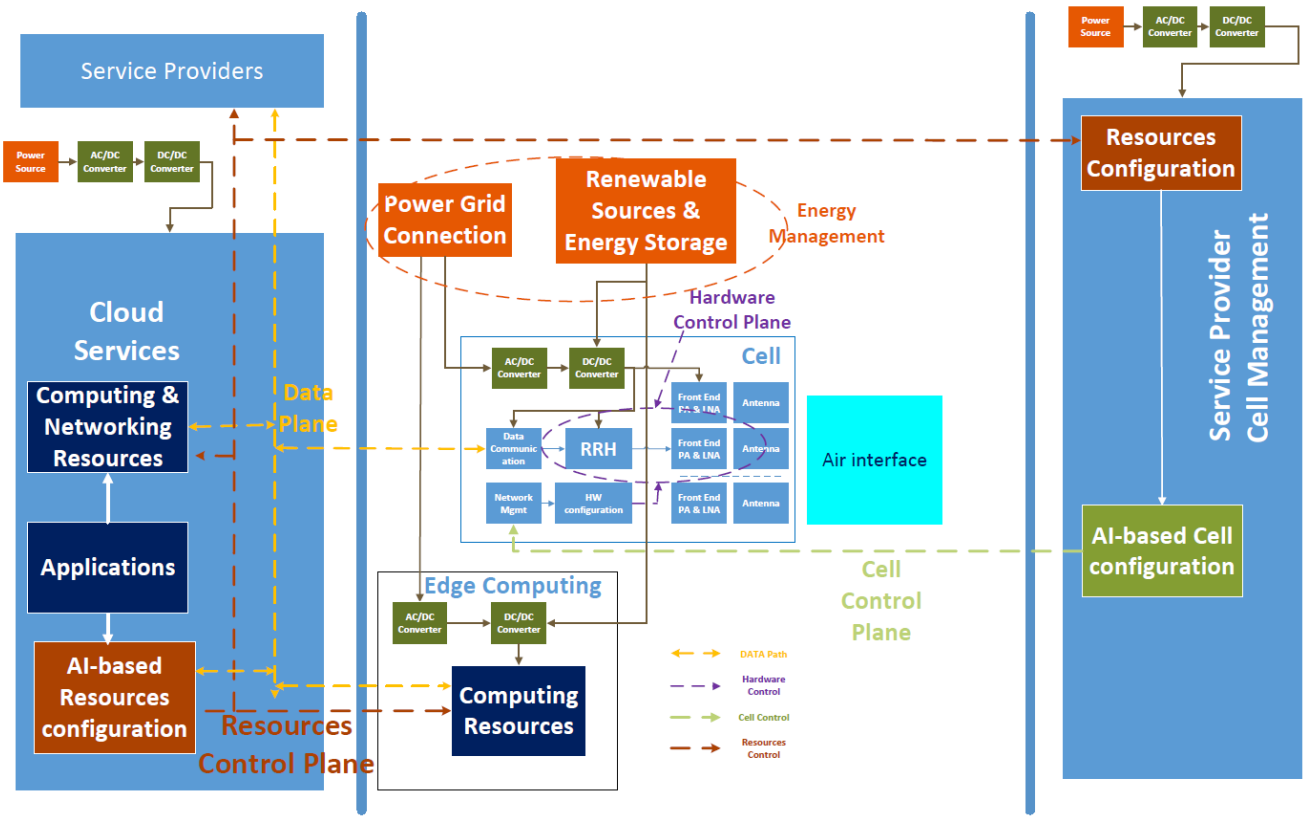


Figure 11. An example of Systems of Systems incorporating a RAN subsystem

Image courtesy of IoTissimo

As the same black-box approach can be used at any level of the system - from subcircuits in a RF Power Amplifier all the way to the global communication and IT infrastructure - one can tailor the analysis to the portion of the system that can be addressed as the model continuously grows with updates from all stakeholders.

4.2.2. SoS Case Studies

The following two case studies exemplify the use of advanced modeling for Energy Efficiencies: 1) real-time decision support and optimization (as also referenced above), and in particular work by [52] and references there-in; and 2) the case of estimation of data transmission costs in a fiber-optic connection.

4.2.2.1. Example 1: Cognizant, Real-time Power-grid Systems Management Under Variable Energy Resources Availability

Works [49] [50] [51] has developed capabilities for real-time electrical power grids systems management under variable resources availability and optimized delivery of service to multiple consumers (& classes of consumers) and of variable per consumer priorities levels. The energy source

variability includes variability of energy resources such as that manifested by renewables (solar, wind, etc) as well as variability in energy (electricity supply) due to sudden disruptions – equipment failure, natural effects (adverse weather, rain/windstorms – tornadoes, hurricanes), or intentional (attack).

Using DDDAS-based methods this work has developed statistical and agent-based algorithms/ models and instrumentation for adaptive, multiscale methods for multi-objective optimization in the presence of large sets of data from multiple heterogeneous sources (resulting from monitoring the power grid – such as Phasor Measurement Units (PMUs), environmental factors and power generation), large numbers of variables and non-linearities within the system considered. These modeling methods are used also for coordination and actuation of distributed sensor networks and collecting information of intensive and time-critical data. They can be used for system-status estimation and optimized delivery of services (electric power) under constraints such as fluctuating demand profiles, power-generation (both conventional and renewable), transmission capacity, differences in planning technologies (that is location, size, short commissioning of power generation units), costs and availabilities of energy sources.

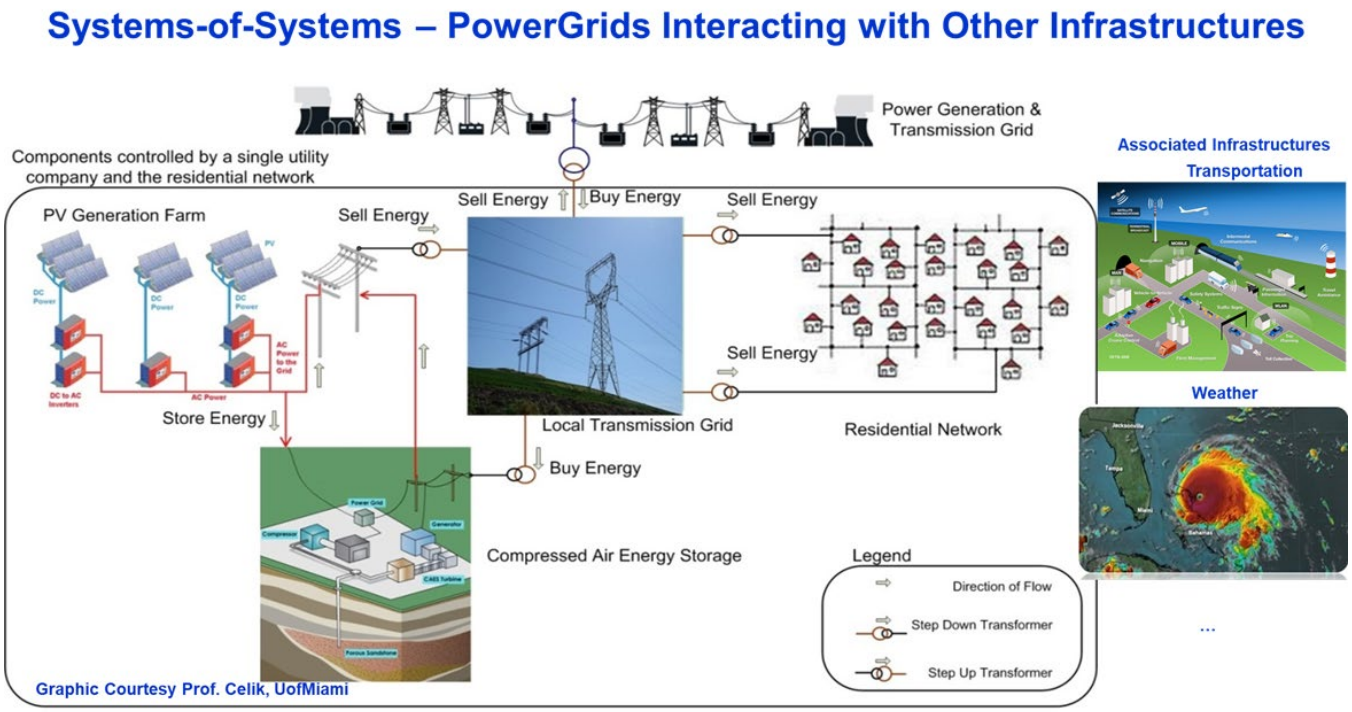


Figure 12. The Power Grids form complex Systems of Systems

4.2.2.2. Example 2: Estimation of Data Transmission Costs in a Fiber-Optic Connection

geographically located markets; however, this example also lends itself to the SoS energy consumption analysis.

The conventional public network takes historical routes and links between two locations via a mix of fixed fiber and wireless links. Light in a vacuum travels 186 miles in a millisecond. From Chicago, IL to New York City, NY this distance would be roughly 800 miles with additional delays due to public network switches. With a dedicated fibre optic network, the distance would be 710 miles and the network switch delay can be minimized on the private network. The public route is longer than the shortest point to point route (~0.5 msec), including over or around mountains, whereas the shortest route would go through the mountain. The switch delay is best evaluated with a "ping" to a machine IP address at a distant location, and can vary from 5-150 microseconds per switch. From an energy, operational cost perspective, fiber networks have an ongoing low fixed energy cost / mile to regenerate the signal with minimal latency. The wireless links have a higher energy cost to transmit the signal over the air, overcoming atmospheric losses and high-power transmission over large distances, even with directional, high gain antennas.

4.2.3. Systems-of-Systems (SoS) Tool Roadmap

The referenced work on the SoS tool are demonstrations based on ongoing research. The frontiers of research and development for this tool include many more features, capabilities and interfaces. This section explores features that would be desirable moving forward in a toolset that is universally applicable and available.

Overall, there are certain features, represented as "blocks" in a systems-of-systems simulation, that are globally defined so as to allow multiple-levels of detail in the simulation for predictive behavior of individual components. This is referred to as polymorphic representation. Most of the blocks within a simulation modeling have inputs and outputs that are specifically related to the neighboring blocks, reflecting the related nature of the system's deployments. Each block can send a query to the connected blocks (upstream and downstream) to gather context information.

For example, in fiber optic network connection, pulling location information from the Geographic Information System (GIS) of the connected blocks to estimate fixed and overhead costs of the data transmission. These costs need to have an associated tag identifying who is responsible for these costs. Some outputs (such as thermal footprint) will have a tag "environment" to identify that the local environment will absorb these costs.

4.2.4. SoS Tool architecture

Features that are in the SoS tool already include static and dynamic analysis, including variations with time (day/night) (workday/weekend/holiday), Monte Carlo engines that are able to examine and compensate for disruption events (modified flight operations due to airframe damage.

As the SoS tool develops, a clean graphical User Interface (UI) is beneficial that allows these features:

- Top level inputs (TLI): time, temperature, GIS, traffic flows, population activity, driving consumer demand
- Low level block inputs: TLI, network connection, power grid connection
- Outputs:

    - Specific outputs to adjacent blocks (Mbps to wireless consumers, Gbps to cloud compute farm)
- Internal black box structure, calculations
    - Fixed cost, ongoing costs.
    - Local effects (for example solar panel or other local power generation)
    - Power dissipation and overall thermal footprint.
    - Sub-blocks that use the same structure as the higher-level blocks.
- Features of interconnect lines
    - Also blocks, generally very simple with a matching input / output pair.
    - GIS information of input and output used to calculate fixed cost and ongoing costs and information delay between input and output.
- Computational engines available for each block
    - Linear tools
    - Predictive tools
    - Local memory history
    - Frequency domain tools

The feedback provided by the SoS tool would be configurable to provide specific insights to the user beyond energy use, specific to the objectives of the user. Some examples are:

- What are the ongoing costs of a generic base-station deployment?
- What is the cost different between local computation vs. cloud-based computing for user equipment based on the current status of the battery, network and other local resources?

Many organizations have custom point tools that need to be accessed. The SoS tool needs to be able to interface to these point tools through a variety of mechanisms: files, sockets, linked or compiled. Some examples of these point tools include

- Ray tracing tools with integrated GIS databases for network planning, including carpeting and frequency planning
- Circuit simulation tools
- Weather forecasting tools
- Power plant simulation tools

In order to develop the SoS tool roadmap, it is essential to explore a wide variety of target examples that will define requirements and drive development insights. A few specific examples, beyond those in the references and those examined in the previous sections include:

- A deep dive into the Texas power grid in winter: typical vs. cold weather events (2011 and 2021).
- Hydroelectric plant providing power to Las Vegas and the southwest

- Solar thermal plant (at Kramer Junction) providing power to LA & Las Vegas
- A specific example cell site

These are questions and challenges that need to be addressed for viability as we develop future generations of electrical systems from power grids to communication infrastructure and beyond.

5. NEEDS, CHALLENGES, AND ENABLERS AND POTENTIAL SOLUTIONS

5.1. Summary

This section is where the true value of all the content comes together in the culmination of all the needs, challenges, and enablers that define this roadmap in the EE context. The following subsections break the content into the five major categories as outlined and described in Table 1 (below).

We start by focusing on inhibitors to overall, network EE and those that can be addressed at the largest points of energy consumption. From there, we move toward the edge and explore how we need to reimagine today's network in terms of a major migration to many more HetNets of small cells to build and enable the network of tomorrow. Once there is a good understanding of the solution EE needs and how it will change the architectural landscape of the network, we put a major focus in the single greatest consumer of network energy, the base station. Given how critical base stations (and essentially anything with radios) are to an economically viable scenario for the 5G&B networks as well as the planet that must survive them, there is extra emphasis and detail on exposing the inefficiencies of modern networks and articulating these shortcomings as gaps requiring the most attention and therefore, the highest ROI for both economic payback and sustainable growth.

The section concludes by taking all the EE best practices in design and implementation of a network and helping to translate into solutions and actionable, quantitative assessment in a way that marries pragmatic, economic goals with constrained, yet qualitative modeling. This is done via the articulation of novel concepts and metrics to simplify these very grand and multifaceted challenges with attainable concepts and design philosophies speaking to stakeholders across the full spectrum. Finally, all these EE optimizations and recommendations are applied to the highest level of a global network and applied at the utility level of the overall, network hierarchy so EE optimizations, financial payback calculations, and ultimately more sustainable 5G&B networks connect to the ecosystem around them in every facet of operation, predictability, real-time adaptability, and market impact.

Table 1. Overall, Major Need Categories Identified by the INGR EE WG

Needs	*Description*
#1: Network Efficiency	Edge Optimization, EE System Design Philosophies, Micro-to-Macro Assessment, 3GPP DTx, Data Centers
#2: Small Cell Migration	Macro-to-Micro Control Plane, Real-time Power Optimization, mmWave Impacts, Cell-free Architectures
#3: Base Station Power	Massive MIMO Impacts, Multi-band Support, Telemetry/Analytic Needs, Energy-centric Feedback Loops
#4: Economic Factors	Technical/Economic Analysis Enablement, Industry Metrics, Socioeconomic Impact, Energy-centric Network Simulation Models
#5: Grid/Utility	Utility-level Impacts/Risks, Networking Electricity, Real-time Energy Market Impacts

5.2. Network Efficiency - Need #1

5.2.1. Challenges

A first step towards identifying the challenges and solutions related to network efficiency is to determine how it can be practically measured and quantified, to enable comparison of current technology with different potential future solutions. Efficiency metrics that focus on particular components of the network can be utilized to locally optimize those components. This is often a suitable way to approach the problem from an engineering perspective. However, there is the risk of missing the bigger pictures since the different components can play widely different roles when considering the network efficiency as a whole. To put it differently, components that are easier to optimize or that play a bigger part in the overall picture should be prioritized when designing the networks of the future. In this section, we will describe challenges at the different layers of the network. Then different potential solutions will be described.

5.2.1.1. Inhibitors to EE System Design

One of the greatest challenges lies in the sheer number of stakeholders involved in the design, deployment, and maintenance of a large-scale network. With a large number of stakeholders comes an even larger number of priorities and motivations (technical and business alike) to drive attention toward the challenges highlighted here and generate momentum for the proposed solutions. There are even likely scenarios in which stakeholders are specifically NOT motivated to work with others in the common spirit of optimizing EE because it can contend with an individual's bottom line. In other words, the concepts associated with EE and "greening" only seems to be embraced when it can be proven to generate a different kind of green (that goes into a bank account). We are not so naive as to expect altruism to come at the cost of profits, but there is logic in trying to raise awareness of the many EE-centric initiatives covered in this document and educate stakeholders so they may come to realize being green is profitable as well once all the true, cost-analysis factors are taken into account. This is explored in more detail in Section 5.5.

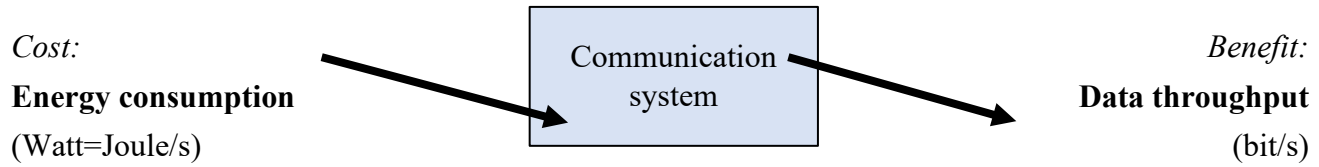
Given this massive amount of global stakeholders all required to "play nice together in the sandbox" to maximize the EE benefits and best practices identified in this report, a whole lot of standardization is necessary to not only motivate stakeholders to come to the table in the spirit of working together, but also build confidence there is ROI on such a large resource investment of people and time. From a high-level framework, such as the SoS proposed by this work, to more established consortiums and governing bodies such as for utilities, 3GPP, or Open-RAN, the pieces can be standardized along with the method in which they complement and/or work directly with each other. This implies all these standards bodies and documentation are at various levels of development from feasibility (essentially non-existent) to very mature, which tends to align with the industry they were born out of. Navigating the necessary, enabling web of standards can be a major inhibiting factor and challenge to forward progress. This is explored in more detail in Section 6.

Time, itself, presents a whole host of challenges. Time must be considered and managed across many orders of magnitude in a 5G&B network when considering optimizing EE from data packet-to-packet (i.e., microseconds) to adapting solutions to the real-time price of energy (i.e., minutes to days). From an architectural standpoint, a glacially slow time interval may even have to be measured in many years or decades for a full, generational deployment. This work attempts to enhance understanding of implications across all the aforementioned time domains and assess how many, asynchronous timelines dictate optimal EE (and therefore 5GDF) as well as how they must work together to determine overall

bottlenecks for focusing assessment efforts. From the phase jitter of a timing signal to the delay in shifting virtual loads around, this is an important consideration for those looking to perform network-level analysis across multiple time, control, and energy domains.

5.2.1.2. Key Challenges at the Physical Layer

When considering the physical layer, there are two modes of operation to consider: active and idle mode. The active mode refers to the case when the network is actively transmitting payload data in the uplink and/or downlink. We can then make a benefit-cost analysis to identify the appropriate efficiency metric. The benefit is the data throughput (DT), measured in bit per second. The cost is the energy consumption at the physical layer, measured in Watt or Joule per second. We can view it as follows:



The benefit-cost ratio is then obtained by taking the ratio of the two terms, resulting in a physical-layer energy efficiency metric measured in bit per Joule:

Energy efficiency [bit/Joule]=(Data throughput [bit/s])/(Energy consumption [Joule/s]) .

In wireless systems, the data throughput can vary greatly, depending on the network traffic load (e.g., number of simultaneously active users and their respective data traffic) and communication channels (e.g., SNR, hardware capabilities). Previous network generations have been rather poor at managing these load variations, in the sense that the traffic load only has a minor impact on the energy consumption [54]. For example, the energy consumption could be at 80% of its maximum value when the load is close to zero and then grow linearly with the load to reach the maximum value at maximum traffic load. Based on the formula above, this leads to much lower energy efficiency when the data throughput is low than when the data throughput is high. This is not a desirable design, but we need future physical-layer technologies that can keep the energy efficiency equally high irrespective of the traffic load. To give a concrete example, [55] from 2015 reported that the network traffic in Sweden had increased by 13 times over a six-year period, while the energy consumption only increased by 40%. This might sound like a great achievement, but it is rather the opposite: it shows that the existing cellular technology was very energy inefficient since it consumed a large amount of energy even when the traffic was low.

From the data center perspective, pJ/bit is used in HPC data centers because when data throughput is constant, the data rate is known, therefore the desired metric energy consumption (EC) is going to be:

EC (J/s) = DT (b/s) * EE (J/b)

The EE is the knob focused on for turning and the EC determines what is required to build to support that DT. The DT is determined by the operations/s rate of the compute times a ratio of bits of data needed per operation:

DT (b/s) = Compute (operations/s) * (bits/operation)

An old rule of thumb was a byte of bandwidth per operation, figuring that arithmetic operations need arguments and produce results that have to go somewhere, but there are also non-logic instructions and there are caches, etc. With caches, local high bandwidth memories, etc., the rate is now more like a bit per operation, but when you're talking about an exaflop, that is real money.

The general implications are illustrated in the following figure, where the left graph is schematically illustrating the energy consumption of a base station as a function of the traffic load. The desirable behavior is that energy consumption is a linear function of the traffic load, which is not the case in existing networks, such as 4G. As illustrated in the right half of Figure 13, this would result in an energy efficiency that is equally high irrespective of the traffic load, which also not the case in existing networks.

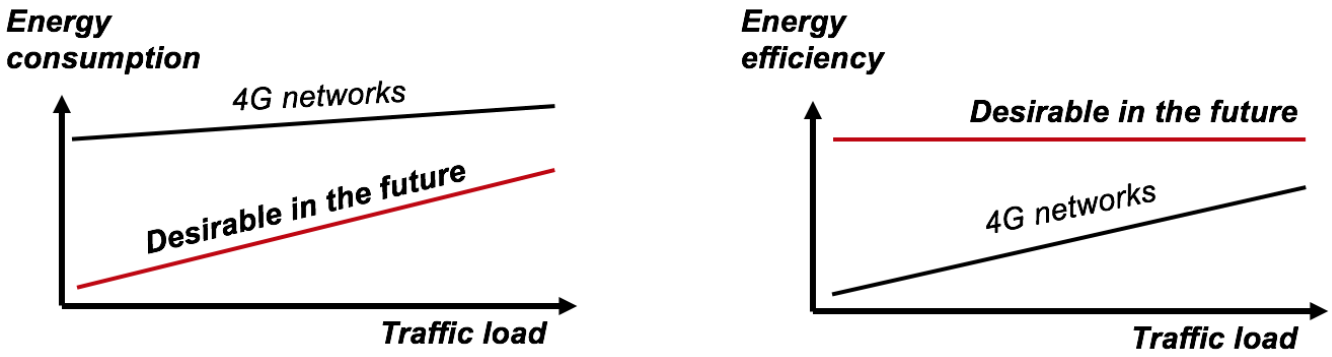


Figure 13. The energy consumption and energy efficiency of a base station depend on the traffic load. It is desirable to the energy consumption proportional to the load, to get a constantly high energy efficiency.

The physical-layer energy consumption in active mode is caused by baseband processing, RF processing, and power amplifiers. Each of these parts will have a traffic-independent and a traffic-dependent part, but it is necessary that the latter part dominates to achieve the aforementioned behavior, and this is a major challenge to achieve. The OFDM modulation/demodulation, up/down-conversion, and filtering are examples of operations that essentially are traffic-independent; as long as there is something to transmit, the signal must be generated at the transmitter and processed at the receiver. Other operations, such as encoding/decoding of the data and RF signal power are naturally traffic dependent. However, even if they provide the desired behavior, it is still important and challenging to reduce their impact on the energy consumption.

The first 5G deployments are focused at sub-6 GHz frequencies, primarily, considering the sub-6 GHz band. Massive MIMO capable RF equipment is dominating the deployments, and this will naturally increase the energy consumption per site. Hence, a near-term challenge will be to make efficient use of the new capabilities to increase the spectral efficiency and thereby the energy efficiency.

In the mid-term and long-term, a major network densification is expected. This is particularly covered in Need #2 (Section 5.3), but could have positive energy efficiency gains, but the challenge is to make efficient use of the reduced propagation losses and also put the cells into sleep mode when not utilized to achieve the behavior outlined above.

New frequency bands in the mmWave ranges are supported by the 5G standard and will eventually be added to the networks, particularly in small cells. A major challenge will be to achieve energy efficient operation also in these bands. The hardware will likely not be as refined, thus the losses in the RF domain will likely be much larger than the sub-6 GHz bands. A major mid-term challenge will be to improve the energy efficiency of the hardware, as well as identifying deployment scenarios and multiplexing techniques that can increase the overall efficiency, even if the energy consumption is increased. In the long-run, even higher bands in the sub-terahertz and visible range can potentially be used, but although the bandwidth can be very large, the range and hardware efficiency are typically reduced.

5.2.1.3. Large-scale Deployment of IoT Devices

It is envisioned that a large number of IoT devices will be deployed during this decade. Some of these can have dedicated power supply, but the ideal situation is that they are battery-driven or even battery-less so they can harvest energy from the environment to carry out their tasks and transmit/receive information signals wirelessly. A mid-term challenge is to minimize the energy consumption during sleep mode, so energy is only consumed in active mode. A long-term challenge is to build networks that can enable this, both at the network side and the device side.

Table 2. Challenges Associated with "NEED #1 - Network Efficiency"

Near-term Challenges: 2022-2025	*Description*
Defining the 5G Energy Gap (5GEG)	**LITERAL DEFINITION** = a hypothetical representation of the disparity between available energy (i.e., sources) and demand (i.e., loads) of the [mostly] "micro-power" devices representing the majority of "things" in the highly scalable edge space of the network, based on proposed 5G use cases.
Defining the 5G Economic Gap (5GEcG)	**LITERAL DEFINITION** = a hypothetical representation of the disparity between available power a system can deliver, and the increasing load demands on its outputs, which means a power-limited system and/or network component will not be able to utilize all its designed potential and therefore be inhibited from delivering on the calculated economics of the payback period.
Spectral efficiency improvement	Increase the spectral efficiency of macro cells, which are necessary for coverage and consume a large amount of power.
High energy consumption in idle mode in new networks	As more base stations are deployed, there will be more idle mode when there is no user in the cell and therefore, we need to reduce the energy consumption in those cases.
Mid-term Challenges: 2026-2027	*Description*
Densification	Reduce the required transmit power per cell by densification, while avoiding that the energy consumption grows when the load is low due the larger number of base station sites.
Energy efficient deployment use of Massive MIMO and mmWave	A Massive MIMO radio requires many more hardware components than a traditional base station, which could lead to high energy consumption. The hardware for mmWave bands is not as refined and energy efficient. The Massive MIMO technology might only be capable of serving one user at a time, so no spatial multiplexing gains are achieved.
High energy consumption in idle mode in old networks	Even if new more efficient networks are deployed, the idle mode energy consumption will remain to be high in the existing 4G networks.
Battery drained in idle mode for devices	For IoT devices with very low duty cycles, the main energy consumption originates from idle mode.
Long-term Challenges: 2028-2032	*Description*
Battery-less equipment	The energy consumption of IoT devices is a concern, not primarily from an energy efficiency perspective, but to alleviate the situation where the lifespan of a device is defined by the battery life.
New frequency ranges with more bandwidth	The data throughput is proportional to the available bandwidth and also operates at the most energy efficient level when the SNR is low and bandwidth is large.

5.2.2. Potential Solutions

Optimize System Design for Power & Energy Utilization

A clear example of how the complexity of new applications requires a total system perspective on the design of HW and SW to achieve a viable product can be found in the development of autonomous vehicles. The "data management" side of autonomous driving represents a drastic departure from prior "dashboard" applications (entertainment, GPS, etc.), as it represents an order of magnitude increase in computing complexity. Some basic sensors and functions have been already deployed in existing vehicles (radar and proximity detectors previously enabled automatic braking and even self-parking of a vehicle), but the requirements to manage a vehicle at any speed and in a dense urban environment led to a plethora of sensors subsystems [56].

All of these inputs need to be processed in real time to ensure the safe operation of the vehicle in all conditions.

A functional technology decomposition of autonomous and connected vehicles leads to the analytic results shown in Figure 14 and Figure 15. The energy required to move data in a subsystem is often larger than that required for processing it, and the energy spent for over-the-air-transmission is even more taxing. For an Electric Vehicle (EV) the energy required for enabling a self-driving experience has a significant impact on the travel range afforded by a battery charge, thus forcing an extreme optimization of the system.

Camera, Radar, Lidar and Sonar data streams need to be processed to extract relevant perception information [57].

Outputs are then combined in a "Fusion" engine that correlates data from various sensors and extracts "validated" information. This output is then combined by a processor (which utilizes AI's "Deep Learning" technologies) with position/location information coming from other sensors (GPS, accelerometers, gyros, etc.) and from other available data (maps and detailed road conditions provided by the cloud) to generate a Prediction of future events and enforce a Driving Policy. Such algorithms, based on desired destination, generate a Motion, Maneuver and Trajectory Plan, which finally informs the Actuation of all subsystems to achieve the desired outcomes.

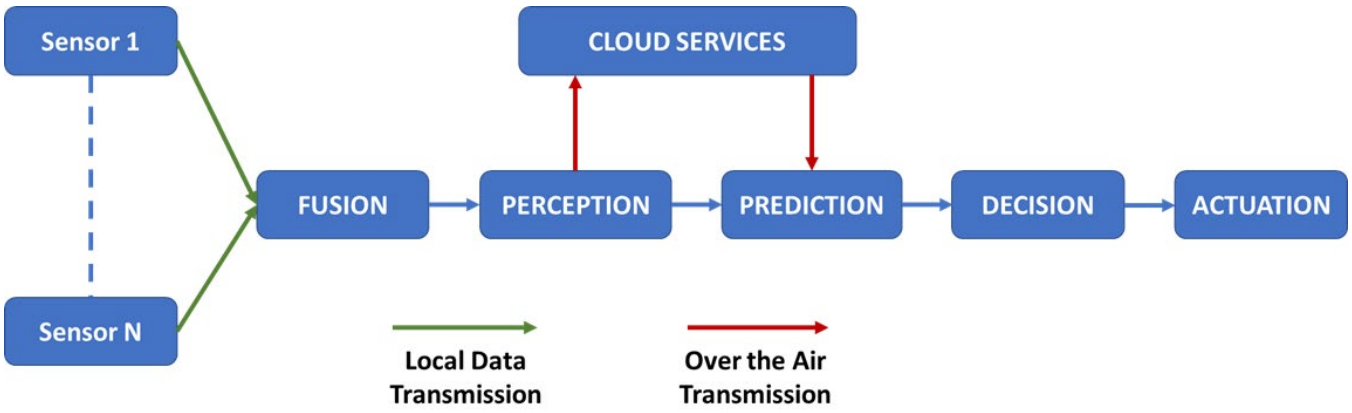


Figure 14. Sequential processing of data in a self-driving vehicle

Image courtesy of IoTissimo

A key application of 5G technology revolves around self-driving cars, where Artificial Intelligence (AI) plays a major role in supporting the interaction of the vehicles with the environment, by combining the locally processed data with information from the cloud and other vehicles.

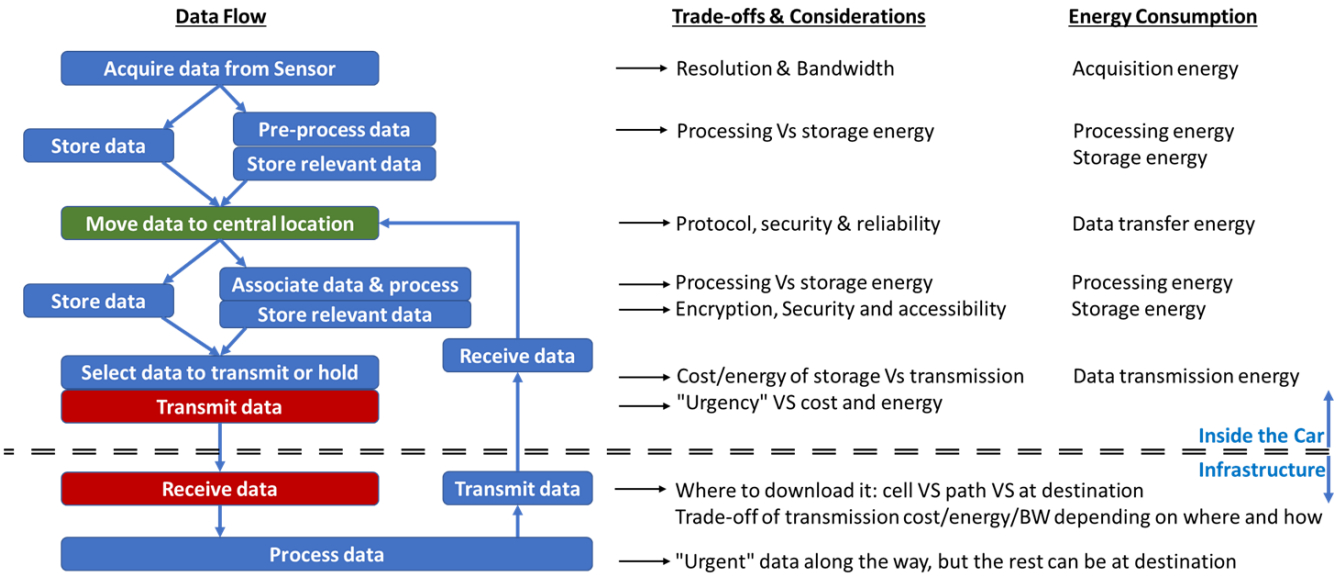


Figure 15. Flow of data in a self-driven vehicle and trade-offs in energy optimization

Image courtesy of IoTissimo

The amount of data that needs to be processed for each autonomous vehicle is staggering: "Just one autonomous car will use 4,000 GB of data/day," and according to Intel, self-driving cars will soon create significantly more data than people — that is, 3 billion people's worth of data [58].

Elon Musk estimated that the energy required for self driving today is around 25% of the energy required to move the vehicle, thus significantly impacting the economy of the application. Since data transmission generally requires more energy per byte than data processing and data storage, the job at hand is therefore that of reducing as much as possible the datasets and as early in the pipeline as possible, so that the problem is manageable according to both timing and energy (and cost!) budgets. We therefore accept significant inaccuracies, and "hybridize" the system (localized vs. centralized) to reach the best possible accuracy given the constraints: pre-processing at the sensor level, then at the local computing engine, then at the edge and then in the cloud, progressively reducing "data" and sharing "information."

The need to consider trade-offs at different levels in the hierarchy of the system leads to developing the ability to use multi-dimensional analysis and simulation tools, which allow us to implement changes in any sub-system and assess their impact on the whole energy footprint.

Such capability can be achieved by modeling the whole activity as a SoS, which accounts for hierarchical datapath and control layers as well as energy sources and their dynamic availability. This

effort will be discussed in the following sections and can provide the basis for a realistic implementation of an AI-optimized system.

5.2.2.1. Mitigating the Inhibitors to EE System Design

There are far too many inhibitors and overall challenges and concepts to comprehensively cover here in terms of ideal, global optimization for EE, but some of the more high-level ones (and perhaps some more on the fringes) are tabled here. Some of these concepts are more philosophical in nature in that they tend to apply across the full spectrum of electrical applications from the microwatt to the megawatt levels. Some are more "big picture" concepts that bring attention to dependencies and ROI factors that may not be the first to come to mind. Then, of course, there are the more direct concepts that are more obvious in their application to design, qualification, and standardization. Much of this content is not meant to be the core focus of this section and is therefore tabled as simplified, bulleted lists as follows:

- **Philosophical Challenges**
    - Source vs. Load: putting more focus into optimizing loads than increasing sources is somewhat counterintuitive to where a lot of design effort is placed. In a grand majority of use cases, it is far easier to reduce the power demand than it is to provide an unlimited energy source.
- **"Big Picture" Challenges**
    - Global Perspective (e.g., the SoS): simply articulating complicated issues and interdependencies in a way that is easily understood and digestible by a wide breadth of stakeholders is key. Once we speak the same language, we can start working on common solutions.
    - Embodied Energy: in order to truly assess the lifetime energy footprint of resources and applications/products/services "from cradle to grave" stakeholders need to start considering the full cost of energy from the harvesting of raw materials to manufacturing to use to even post-deployment considerations such as recycling and management of hazardous waste.
    - The Mitigation of Primary Batteries: disposable batteries need to become a thing of the past whether due to landfill space, major contribution to hazardous waste, stress of supply chains of limited/precious resources, or simply the cost of replacement.
        - This is mostly supported by education/awareness of advancements in energy storage technologies and techniques (i.e., energy harvesting) employing such solutions to meet the goals outlined here.
    - Common Metrics: not only should we agree on the concepts to articulate the major issues, but also on the [initial] metrics proposed to assess the issues at hand and devise a methodical, analytical approach to very complicated analyses.
    - Socioeconomic Factors (i.e., 5GEqG): in our focus to optimize financial returns and energy utilization, we must never forget the need to identify and address socioeconomic factors.
- **Design Challenges**

- Design For Energy (DFe): with all the various forms of Design For X (DFx) checklists, rarely is there a provision specifically for energy minimization. This is certainly addressed in bits and pieces based on various design stakeholder perspectives and motivations, but it is atypical to see EE treated as a checklist item in the same way many other DFx analyses are performed.
    - A great example of this is the classic consideration of Data Processing vs. Data Transfer. Put more fundamentally, just because you can do something does not mean you should. In other words, simply because we have the ability to collect and transfer massive amounts of data does not mean it is the most EE (and perhaps economical) approach. As demonstrated by the PCF metric, the more data processed at the source in the hopes of mitigating transmission, the greater the opportunity for massive EE optimization. More energy is spent transferring data than any other aspect of computation or analytics, whether it be between UE and core network or processor and memory.
- Proposed Strategy Standardization: Open vs. Proprietary (or both?)
    - Perhaps start with an established, related ecosystem to test the waters. The Open Compute Project [59] is a good example of this, where you have all the proper stakeholders/constituents in a bustling, open community quickly growing solutions in both the HW and SW domains as well as partnering with other, analogous groups (e.g., Open Networking Forum (ONF) [60]; Open RAN, O-RAN Alliance [61]; and Open RAN Policy Coalition [62]).
    - Modeling & Simulation Tool(s)
        - Make it painfully obvious (though not too complicated) that it is in EVERYONE'S best interest to adopt these practices, whether your motivation be financial, environmental, and/or socioeconomic.
    - Test & Measurement Protocols
        - From automated test equipment (ATE) to pragmatic metrics that paint EE in a way that speaks to different stakeholders from different backgrounds (i.e., PCF, 5GDF, joule/bit, bit/Hz/W/km2, etc.), standardization of testing and qualification approaches are just as important as design-related ones.

5.2.2.2. Efficient Physical Layer Operation in Active Mode Using Spatial Multiplexing

Since the energy efficiency in active mode is the ratio of data throughput and energy consumption, and improved energy efficiency can be achieved in different ways, as illustrated in Figure 16.

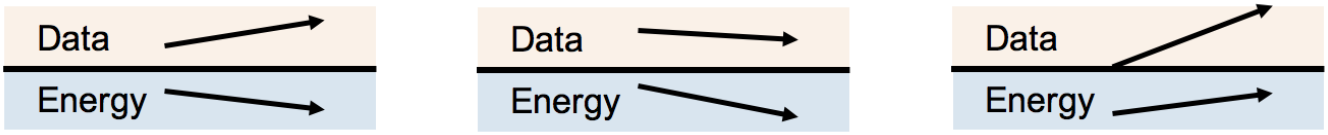


Figure 16. Since the energy efficiency in active mode is the ratio between data throughput and energy consumption, there are three different ways that it can be improved.

The first option (left) is that the data throughput is increased while the energy consumption is constant or decreasing. The second option (middle) is that the energy consumption decreases, while the data throughput is constant or slightly decreasing. The third option (right) is that both the data throughput and energy consumption are increasing, but the throughput increases faster. Different future developments of the network technology led to different types of improvements but given that the data throughput has been dramatically increasing for decades it is likely the first or third option that will materialize.

Recall that the energy consumption consists of a traffic-dependent and a traffic-independent part. Even if the traffic-independent part is constant, its relative impact can be reduced by making use of spatial multiplexing [63]. Instead of transmitting one signal per carrier frequency, several signal layers can be multiplexed using Massive MIMO [64]. These signals can either be directed to a single user (so-called single-user MIMO) or multiple users (so-called multi-user MIMO), or a combination of them. In this way, the energy consumption of the traffic-independent parts per signal can be reduced inversely proportional to the number of spatially multiplexed signal layers [63]. This corresponds to the third option above, where the hardware is evolved in a way that leads to higher energy consumption, but also the ability to spatially multiplex several signals (to one or multiple users) so that the energy consumption per user can be drastically reduced [63].

Since the data throughput per multiplexed signal is a logarithmic function of the Signal-to-interference-plus-noise Ratio (SINR), it is more energy efficient to multiplex many signals, each having a relatively low SINR, than a few signals with high SINRs. This is another explanation for why Massive MIMO can be an inherent energy-efficient RF solution, in some deployments. However, in macro cells where many users are located at the cell edge, spatial multiplexing gains might not be available since the SNR is too low to obtain the benefits. In those cases, cell densification is preferable over the use of Massive MIMO and spatial multiplexing [63].

There is an inherent tradeoff between energy efficiency and latency in the user scheduler. To minimize latency, the scheduler should let the base station transmit packets immediately when they arrive to the packet queue. In low-traffic hours, the base station will then have to switch between active and sleep mode at the same rate as the packets arrive, and never get the chance to multiplex multiple signals. An energy-efficient scheduler will accept slightly larger latencies to gather multiple packets and transmit them with spatial multiplexing. The latter will allow for longer sleep cycles and more energy-efficient data transmission, but at the expense of larger delays. Since scheduling algorithms are not standardized, it is up to each vendor to decide what metric to use in their design.

In the near-term period, the use of Massive MIMO technology will dominate the early 5G deployments, in the 2.5 and 3.5 GHz bands. This trend can already be observed in the USA and China. While the energy consumption of such base stations is larger than for traditional base stations, the spatial multiplexing capability will enable a leap in energy efficiency at the physical layer. This improvement is particularly evident when the traffic in the network increases, so that the spatial multiplexing capability

is utilized more regularly, and can be further improved by software updates that refine the spatial multiplexing capability (e.g., making use of reciprocity-based beamforming).

In the mid-term period, the hardware of Massive MIMO base stations are evolving and becoming more streamlined for its specific properties. The first generation of the technology is heavily overdesigned, relying on traditional components for macro base stations with a high traffic-independent energy consumption. However, with a hardware design that is tailored for the Massive MIMO operation, with many low-power radios instead of few high-power radios, the traffic-independent energy consumption can be vastly reduced [65].

5.2.2.3. RF Hardware Evolution

Looking closer at the hardware evolution, the efficiency of the RF hardware can be improved in a number of different ways. A key distinguishing factor between Massive MIMO and conventional sector antennas is the number of radios; instead of having one radio connected to a large set of antenna elements, there are many radios that are connected to one or a few antenna elements each. Since the total transmit power of a base station is determined by regulations, each radio in a Massive MIMO base station is generating a signal with a lower transmit power than conventionally. Hence, it will be easier to design efficient power amplifiers, for example, in terms power added efficiency (PAE). One way of viewing this is that the base station of the future will consist of many handset-grade low-power components instead of a few high-power components [29]. By redesigning the base station architecture to make full use of this new implementation paradigm, the power losses in the RF hardware can be greatly reduced. Since the hardware evolution is gradual, the improvements are expected to be largest in the mid-term period.

If one compares a conventional 4-antenna sector antenna for LTE (Long Term Evolution) with a future 100-antenna Massive MIMO base station, one might get the impression that the Massive MIMO base station will consume 25 times more power, while only being able to serve 5-10 times more users, thereby decreasing the energy efficiency. However, this is not the case, as illustrated by Figure 1 which is reproduced from [30]. This figure compares the conventional base station with three different types of Massive MIMO deployments that are specified in the legend. While the energy consumption of the analog frontends increases, it is not a 25-fold increase if the frontends are properly redesigned for Massive MIMO operation. Moreover, all the other components of the energy consumption (e.g., power amplifiers, baseband processing) can be greatly reduced and, yet the throughput is the same or increased. For example, Option 2 gives a 7-fold improvement in energy efficiency, while the Option 3 and Option 4 give roughly 30-fold improvements. Interestingly, the energy consumption mix is entirely changed, from one being dominated by RF power and baseband processing to one being dominated by the analog frontends. This development has not occurred yet, since the first Massive MIMO base stations in 5G makes use of legacy components, but is expected to occur gradually over the near-term and medium-term time interval.

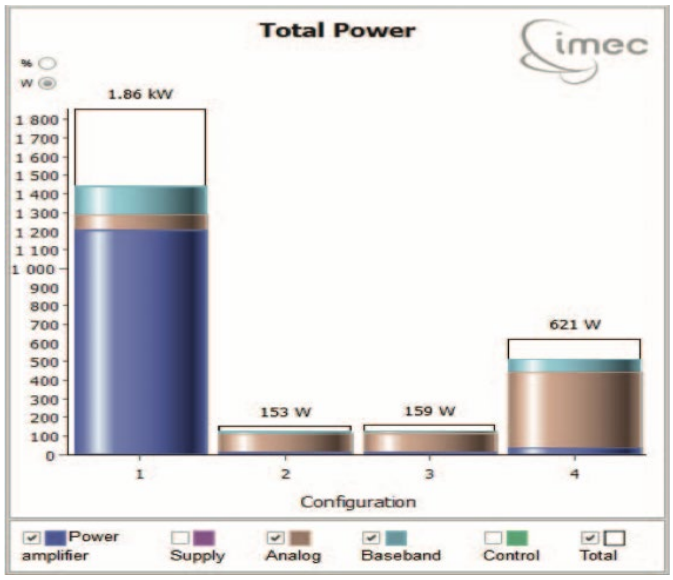


Figure 17. Power consumption for a conventional 4x4 MIMO base station (Reproduced from [30])

Option 1, with three Massive MIMO scenarios. Option 2: 100 antennas serving 10 users using simple signal processing; Option 3: 100 antennas serving 25 users using advanced signal processing; Option 4: 400 antennas serving 100 users using advanced signal processing.

The hardware that is utilized for mmWave communications is not as mature as that for conventional bands, thus the losses are relatively large but there is also a better chance of a swift improvement. Fully digital solutions are plausible, and the energy consumption can be on par with phased arrays, if one utilizes the fact that ADC resolution per radio can be reduced and the insertion losses from phase-shifters are alleviated [29]. These improvements will likely not materialize in products until in the long-term perspective.

5.2.2.4. Efficient Physical Layer Operation in Idle Mode

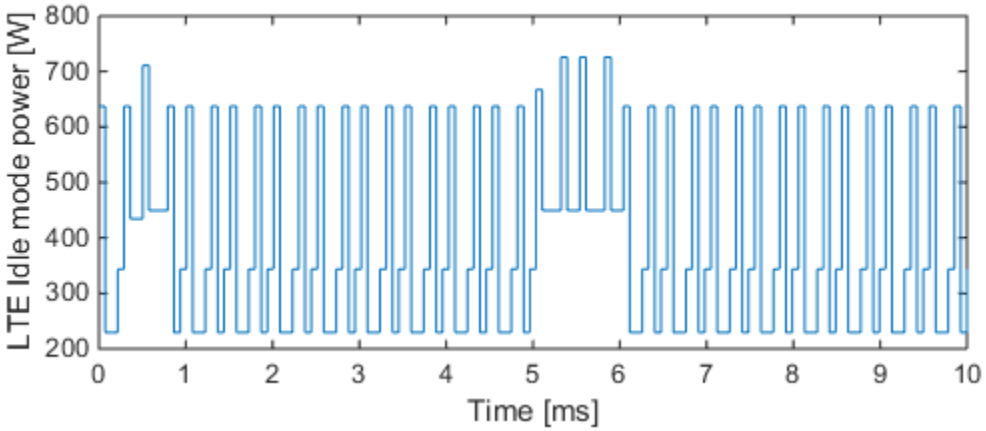
The physical-layer energy efficiency must also be measured in idle mode, which is frequently happening on the network level, both due to the daily temporal variation in traffic demand (low-traffic hours vs busy hours), the spatial distribution of users in the network, and the burstiness of packet data [54] [66]. In fact, when moving towards smaller cells, it will be even more common that base stations are operating in idle mode due to increased randomness in which small cells that the active users are currently residing in.

A different metric must be utilized to measure the efficiency in idle mode since the data throughput is zero in this case, but the necessary control and synchronization signals must still be transmitted to enable access and paging. Typically, the aim is to reduce the load-independent energy consumption, which is done by applying sleep modes. The type of sleep mode depends on whether the carrier or base station have a capacity or coverage role in the network. Capacity carriers or nodes can be de-activated completely during low traffic hours; however, this requires that coverage is provided by another carrier or base station. For coverage-providing carriers or base stations, there is a tradeoff between the ability of

a base station to enter sleep mode when idle, and the increased delay that is caused if a user device attempts to access the network when the base station is sleeping. The 4G standard required the network to transmit a large number of signals even in idle mode, which prevented sleep-based energy savings and is one part of the reason for the aforementioned small difference in energy consumption between low traffic load (when the network is often in idle mode) and high traffic load (when the network is seldom in idle mode). In fact, a 4G carrier transmits the cell-specific reference signals (CRS) approximately every 0.2 ms, which means that only very short (2-3 OFDM symbols) micro-sleeps are possible. 5G has addressed this issue by drastically reducing the signal load in idle mode, to allowing for prolonged base station sleep modes. The most necessary control signals for access are still broadcasted but many other signals are only transmitted on-demand from the accessing user devices. Furthermore, the periodicity of the necessary control signals can be configured between 5 ms and 160 ms. For a coverage-providing carrier the default periodicity is 20 ms. This means that in idle mode a 5G carrier can have approximately 100 times longer silent periods than a 4G carrier (20 ms vs. 0.2 ms), during which sleep modes can be applied.

Longer silent periods not only mean that sleep mode can be applied for a longer time. It also means that more hardware components can be de-activated and then re-activated, keeping in mind that it also takes some time for hardware to be turned on and off, to avoid transient behaviors. Consequently, a possibility is to design sleep modes with even lower energy consumption [67]. This same work has analyzed base station hardware components with respect to their de-activation and activation times, and based on that designed sleep modes tailored to different sleep lengths. The longer the sleep length, the more hardware components can be de-activated, and the lower the resulting sleep mode power consumption.

In Figure 18 below, the potential sleep modes from energy consumption [67] have been applied to comparable 4G (LTE) and 5G New Radio (NR) base station configurations operating in idle mode. The figure shows the instantaneous energy consumption of the base stations, and it can be estimated that the average idle mode power consumption of the 4G (LTE) base station is approximately 400 W, while the average idle mode power consumption of the 5G (NR) base station is approximately 45 W. This illustrates the energy saving technology potential of the 5G air interface in idle mode compared to the 4G air interface.



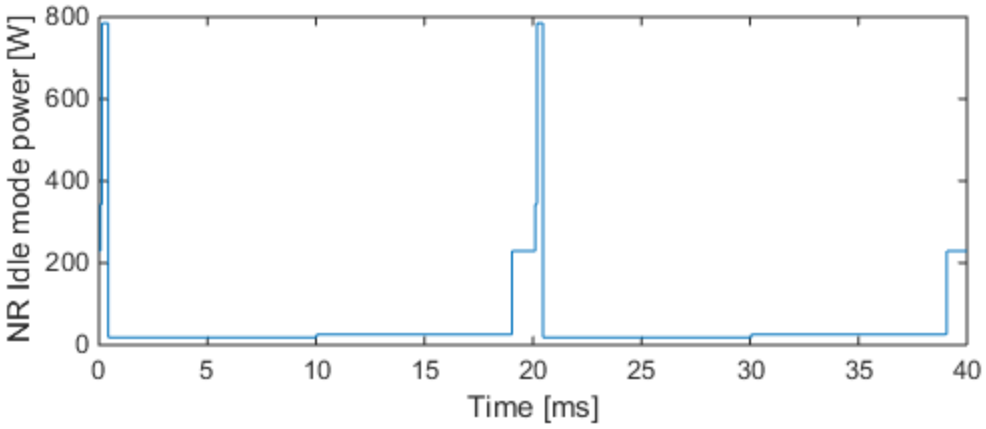


Figure 18. The idle mode power consumption in comparable LTE and NR configurations, showing that the average energy consumption is substantially smaller in NR thanks to the new sleep mode features

(Reproduced from [68])

So, the 5G air interface has been designed to facilitate use of aggressive sleep modes in idle mode. But, as seen in this discussion the resulting energy savings are highly dependent on the sleep mode capabilities of the RF hardware. Even though cellular base stations [67] outlined the technology potential, today's RF hardware is not there yet. It is therefore of outmost importance that current and future RF hardware research and development not only focus on efficiency improvements in active mode, even more important are the sleep mode capabilities in idle mode.

In the near-term, these sleep mode technologies can be utilized to reduce the energy consumption in the new 5G bands, as compared to the case of using 4G technology in those bands. In the mid-term, when the support for 5G has increased among user devices, it is desirable to transfer using the NR radio interface in bands that were previously using the LTE radio interface, to enable a larger fraction of the base stations to make use of these sleep modes.

5.2.2.5. Wake-up Radio (WuR) for User Devices in Idle Mode

The ability to enter sleep mode is important also at the device side, to reduce the energy consumption in idle mode. This is particularly important for battery powered IoT devices which can have very low duty cycles of the applications, but should still be reachable for paging so that the network can request it to send data or carry out certain tasks. One solution is to equip the device with a special wake-up radio receiver, which is used in addition to the conventional receiver unit. The device can then turn off the conventional unit when in idle mode, while the wake-up radio is actively listening for paging signals. These signals are designed specifically to enable a simple energy-efficient detection and determination of whether the device is being paged or if the packet is directed to another device. This feature might be utilized in IoT devices in the mid-term perspective.

5.2.2.6. Energy Harvesting (EH)

A complementing way to improve EE, especially where it has the highest value (in terms of PCF) at the edge, is to mitigate the power losses that are not directly connected to the desired communication link,

particularly the losses over the electricity grid. If an edge device sourced power from the utility grid, then it can have a PCF of 106+, where if the same device sourced the same power locally (e.g., EH), then it can have a PCF of 1 (See Section 3, "What is the true cost of 1 mW?" topic header for more detail on PCF calculation/analysis). This can be achieved by EH, which comes in different forms and enables the capture of ambient energy in every form afforded by physics. The value proposition of applying EH technologies to 5G&B use cases may have even greater value for those unconnected/underconnected to robust electric and network service infrastructures, thus addressing what this group had identified as 5GEqG (see Section 3).

The amount of ambient energy that can be converted into usable energy for systems can vary greatly depending on the form of EH modality (i.e., solar, thermoelectric, kinetic/electrodynamic/piezoelectric, triboelectric, etc.) and the operating environment. It is important to note EH is not an all-or-nothing opportunity so just because it may not be able to source an application's full power budget needs, subsidizing a portion can a have a lot of value from extending battery life between recharges to decreasing overall system size by the elimination of larger, standby power rails in a main power supply. For these reasons, sizing appropriate energy storage to EH application opportunities is also a critical enabler. In general, the smaller the system power budget, the greater the opportunity for supplementing it with EH-sourced power, but there are EH application opportunities spanning from microwatts to gigawatts.

At the device level for example, solar energy from light and/or heat, thermal differentials between device and external environment, kinetic energy from user movement, and RF energy leakage from other transmissions can be utilized to charge the battery and/or supplement energy for subsystems (i.e., charge a super cap enough to enable WuR). These are numerous ways of making practical use of energy that otherwise would be wasted. At the base station level [69], outdoor base stations can make use of solar panels, wind generators [70], and even fuel cells [12] [71]. This is particularly suitable for small-cell base stations that don't require active cooling, since they will tend to have smaller power budgetary needs than their macro counterparts, and therefore might have a total energy consumption that can be sustained by the local EH. Data centers could have direct access to a variety of different EH, including a local set of wind turbines and even the eventual repurposing of the massive, thermal energy expenditures of the end loads.

There are several important aspects to EH. On the one hand, it can bypass the losses on the grid since the energy is extracted locally. On the other hand, the overall harvesting efficiency might be substantially lower than in a conventional large-scale power plant, even if one accounts for the losses on the grid in the latter case, though this point may be mitigated if energy is harvested from "free" sources that would otherwise add no value regardless of the conversion efficiency. Of course, this ROI assessment must be done reasonably to incorporate the added cost of the EH enhancement along with the payback period in application. Such a payback calculation must not only consider the direct cost-adder of the EH, but the savings seen in OPEX and CAPEX throughout the entire PVC in question. These secondary factors are commonly overlooked in a typical flaw of the appropriate analysis. For instance, consider some IoT self-powered WSN that runs off a secondary coin cell battery. While the cost of supplanting the battery with EH may exhibit even 1-2 orders of magnitude in cost increase in terms of system Bill of Material (BOM), e.g., CAPEX, the cost of labor (e.g., OPEX) to replace that battery (even once) over the product life will greatly exceed the cost of applying the EH front end. This cost discrepancy becomes painfully obvious in applications where the WSN exists in a harsh environment (i.e., factory high temp telemetry data) or even impossible-to-access scenarios such as ubiquitous sensors embedded in structures.

The environmental advantages can be quite high even if the financial justification is a little more challenging in the near term. Not all energy is equal in terms of the footprint. If one compares harvesting from renewable energy sources with power plants making use of fossil fuel, it might be desirable to choose the renewable sources even if they are less efficient in the generation of usable energy. The mitigation of primary batteries alone is a great motivator for sustainable energy when considering the space and hazardous materials associated with the grand majority of primary cells.

The cost analysis is even more definitive if we compare EH to WPT: the transmitter required for a WPT implementation will certainly cost more than $100, thus greatly exceeding the additional cost for EH enhancements in the receivers, and then likely have an operating cost between $50 and $200/year in energy consumption.

5.2.2.7. Backscattering Communication

More energy-efficient communication schemes should be adopted for sensor communications given the extensive penetration of IoT networks in smart homes, smart cities, smart factories, etc. Here we can rely on energy-harvesting based communication techniques or, more radically, on backscattering-based communication [72], which essentially enables sensors to run in a battery-free fashion by retro-reflecting ambient RF signals using, for example, on/off modulations type-scheme without the need of any active RF transmission. Backscatter communications is particularly interesting since we can bypass power-hungry components such as oscillators, mixers, and amplifiers and just rely on low cost and complexity devices that need a small amount of power typically harvested from the surrounding environment.

We have three categories of backscatter communication systems: (i) ambient, (ii) mono-static, and (iii) bi-static, depending on the adopted architecture. First, in the ambient case, the so-called reader uses the surrounding ambient transmissions for receiving the backscattered signals from the tags. Second, in the mono-static backscatter scenario, a backscatter transmitter in the tag reflects back while modulating a signal that is generated by a reader. Finally, to overcome the fact that the round-trip path-loss can be a limiting factor in the mono-static case since both the RF source and the backscatter receiver are co-located in the same device, in the bi-static case, a signal generated by an RF source is modulated by the backscatter transmitter/tag, and it is reflected to an independent backscatter receiver in the reader. A large-scale utilization of backscattering technology is envisioned to appear far into the future.

5.2.2.8. Wider Bandwidths and Visible Light Communication

If there would be no restriction on the bandwidth and transmit power, the highest energy efficiency is achieved when using a very large bandwidth and operating at relatively low SNR, where the capacity is a linear rather than logarithmic function of the SNR [73]. This is a motivation for going further up in the frequency domain, beyond mmWave spectrum, to gain access to even wider bandwidth. There is certainly a tradeoff when it comes to how practically achievable such gains are, particularly if the hardware efficiency reduces, but it is a direction to consider in the long-term perspective. A dynamic scheduling of users between different frequency bands will be necessary to make the most efficient use of low bands that can feature good coverage but limited spectrum resources and extremely high bands that can feature abundant spectrum resources but very limited coverage.

One exciting and relevant energy-efficient technology is visible light communication (VLC), known also as Li-Fi [74], which promises to provide optical-fiber-like performance by relying on the visible spectrum which is a portion of the electromagnetic spectrum that is entirely untapped, free, safe, and

provides a high potential bandwidth to piggyback on energy-efficient LED technology used for illumination in indoor or even short-range outdoor environments to also transmit information a high speeds. Today, available VLC indoor technology is able to deliver 100 Mb/s over a range of about 5 m and we should be soon on target to provide the multiple gigabits-per-second needed for beyond 5G networks. As such, this technology offers a great potential especially in view of the expected scarcity of RF spectrum. Due to its inherent short-range characteristics, it is a complement to other wireless technologies.

Table 3. Potential Solutions to Address "NEED #1 - Network Efficiency"

Near-term Challenges: 2022-2025	*Potential Solutions to Near-Term Challenges*
Addressing the 5G Energy Gap (5GEG)	Considering the high risk to network continuity and even electrical grid reliability identified by the 5GEG, a number of solutions have been identified to soften or even mitigate this risk. The high-level solutions identified and detailed in this document to address the 5GEG are as follows - ● Optimizing System Design for Power & Energy Utilization ● Mitigating data transmission via implementation of localized analysis and consumption (Edge Buffering techniques, localized AI/ML processing, Mobile Edge Computing (MEC), and data consolidation). ● Consolidate/Turn off resources when not being used in all timebases (i.e., sleep modes, dark silicon, virtualization, network slicing. ● Energy Harvesting (device-level) ● Migration of Data Center Efficiencies from HPC/ Exascale to Enterprise Applications
Addressing the 5G Economic Gap (5GEcG)	With the 5GEcG defined and articulated in application, it has been shown how it can be addressed by characterizing with metrics such as PCF and 5GDF and applying analyses from the system level to the network level via a framework like the SoS to mitigate the gap. This approach addresses the challenges presented by the 5GEcG by enhancing awareness of energy limitations in system design that limit functional ability and provide clarity for the true bottlenecks in a PVC that inhibit EE best practices.
Spectral efficiency improvement	Deployment of Massive MIMO base stations capable of spatial multiplexing, which leads to slightly higher energy consumption but a vastly higher spectral efficiency, thus increasing the energy efficiency at the physical layer.
High energy consumption in idle mode in new networks	The NR radio interface in 5G allows for longer periods of sleep mode when the cell is in idle mode, which reduces the energy consumption in cells that make use of 5G technology.
Mid-term Challenges: 2026-2027	*Potential Solutions to Mid-term Challenges*
Energy Harvesting (Base Station Level)	Enable a higher-power-level class of devices (i.e., 10s of watts) to supplement (or ideally fully meet) system energy needs with scavenged energy from the immediate environment, thus continuing to decrease the multiplicative demand on the local utility grid.
Energy efficient deployment use of Massive MIMO and mmWave	The hardware implementation will become more energy efficient, by utilizing components that are tailored for the specific new use cases. Better PAE in power amplifiers and less overdesign will contribute to this, leading to a situation where the energy consumption per base station is reduced even if Massive MIMO is used.
High energy consumption in idle mode in old networks	The ability to upgrade radio features can be accomplished through SW to enable legacy, or brownfield, deployments of 4G base stations to have some 5G capability, but require HW (i.e., greenfield) upgrade to take full advantage of the latest sleep mode capabilities.
Battery drained in idle mode for devices	Wake-up radio receivers can be utilized in IoT devices to limit the energy consumption in idle mode, so that the battery is mainly used for active mode operation.
Long-term Challenges: 2028-2032	*Potential Solutions to Long-term Challenges*
Energy Harvesting (Network Level)	Ubiquitous implementation of EH solutions for most of the HW (i.e., 100+s of watts) beyond the core via an SoS (or similar) framework that identifies candidates of highest PCF and optimizes consumption for meeting QoE expectations at the bare minimum of system power. This could extend to data center HW as well via implementation of waste heat recovery and conversion to usable, electrical energy from chip to rack levels.

Battery-less equipment	EH and backscattering are solutions to enable battery-less equipment that can perform basic tasks and communicate the results without the need for a wired power supply or a limited life-span determined by a battery.
New frequency ranges with more bandwidth	The energy efficiency is increasing when using more bandwidth, at least if reductions in hardware efficiency are neglected. This motivates the use of new frequency bands including visible light communications.

5.3. Small Cell Migration - Need #2

5.3.1. Challenges

If one can increase the spectral efficiency without increasing the energy consumption, or if the increase is small compared to the gains, then the energy efficiency is improved. The area spectral efficiency is traditionally increased by densifying the network infrastructure. The traditional approach has been to deploy more base stations, thereby decreasing the cell size which leads to better signal-to-noise ratio (SNR) and more frequent reuse of the spectrum over space. The gain in area spectral efficiency comes from deploying more cells, while the spectral efficiency within each cell might be roughly the same as in the past. In 4G and 5G, the MIMO and Massive MIMO technology enables spatial multiplexing within each cell. This is a way of increasing the spectral efficiency within a cell.

Hence, a denser network deployment achieved by smaller cells has the potential of improving the energy efficiency at the physical layer. Recall that the energy efficiency definition for active mode, described in Need #1, is the ratio of the data throughput and energy consumption. The energy consumption can be divided into two categories: the radiated power (including the heat dissipation in the power amplifiers) and the circuit power (for analog RF processing and digital baseband processing). Which part dominates depends on the hardware characteristics and the transmission range (average transmit power required to deliver the service in the coverage area). In large macro cells, the transmit power is dominating over the circuit power [54]. Hence, the first step towards a more energy-efficiency architecture is to reduce the coverage area of each cell so that the average transmit power can be reduced. This leads to migration towards small cells.

When the cells become sufficiently small, the circuit power will become the dominant factor in the energy consumption, thus one needs to determine which path to take when it comes to hardware architecture. One can either minimize the energy consumption by using passive antennas and other simple components, which only enable one user to be served at a time in the coverage area and gives limited ability to suppress interfering signals from other cells. Alternatively, one can make use of active antennas, such as Massive MIMO, to achieve spatial multiplexing of multiple users (or layers per user). This will increase the total energy consumption in the cell but can nevertheless lead to higher energy efficiency since the circuit power is shared between the users/layers. Importantly, these efficiency gains only occur in cells where the spatial multiplexing capability can be regularly used. If there is only one user device with a single antenna within a cell coverage, the data throughput gain from Massive MIMO is small and will likely not compensate for the increased energy consumption. Hence, there is not one but multiple types of small cells that must be considered and utilized when building efficient networks.

The migration towards a small-cell based network infrastructure is inevitable, and it is also associated with new challenges. One limiting factor when it comes to any type of densification is interference. When the cells become smaller, the distance-dependent path-loss exponent is typically reducing. This has a positive effect on the intra-cell SNR, but is problematic since the interference grows as well and becomes increasingly complicated; there are more neighboring cells that cause interference, which makes managing them more complicated, and the total inter-cell interference becomes larger in relative terms. Another issue is mobility; with a denser network the number of handovers between cells

increases. A handover is typically associated with signaling between the UE and the cells involved in the handover, which applies to both active and inactive UEs. An increased number of handovers consequently increase signaling, which in turn causes extra energy consumption both in the UE and on the network side. Control signaling is another challenge when there is much interference between cells and a large number of cells that contribute to interference, making it hard to coordinate.

Table 4. Challenges Associated with "NEED #2 - Small Cell Migration"

Near-term Challenges: 2022-2025	*Description*
Complicated control plane	The control plane signals must be protected from interference, which is cumbersome if there are many potentially interfering cells.
Increased intra-cell and inter-cell interference	As the data traffic per area unit grows, the interference between concurrent transmission will increase. This will naturally occur between cells, but the spatial multiplexing feature of Massive MIMO will lead to decreased interference within each cell.
Improved coverage in mmWave bands	mmWave signals are low power and easily blocked so the coverage in these bands will be spotty.
Mid-term Challenges: 2026-2027	*Description*
Coordinated multipoint	There is a limit to how much interference that can be dealt with locally at each access point. Coordinated multipoint methods are needed to co-optimize the scheduling and physical layer transmissions between cells.
Improved coverage depending on traffic variations	Since the traffic is varying with time, it will be very costly to deploy small cells that are capable of handling the peak hour traffic variation on their own.
Long-term Challenges: 2028-2032	*Description*
Cell-free architecture to alleviate inter-cell interference	When the network deployment becomes very dense, inter-cell interference will be the main performance limiting factor, which calls for new cell-free architectures that can deal with the problem in a scalable manner.
Improved signal rank and coverage for specific users	The majority of the energy consumption in active mode is due to the service requests from a few specific users that have low SNR or request very high rates. It is desirable to have the ability to adapt the propagation conditions to handle such users without deploying a huge number of active antennas.

5.3.2. Potential Solutions

5.3.2.1. Characterizing Energy-Centric Coverage as "Carpeting"

The propagation of cellular signals has a complex behavior. There is a first order behavior that is a function of frequency, the height of the antenna above the terrain, and the nature of the terrain itself [75]. Additionally, foliage, structures, geographic obstructions to the line-of-sight and even the curvature of the earth itself all contribute to this complexity. In Figure 19 below, we present a series of data from the Canadian Research Council scaled for the case of a 1 W Small Cell and a 200 W Macro Cell, where the cited power is the average RF power emanating from the base station antenna. The height of the base station antenna is the same for both cases and is only dependent on the selection of frequency and morphology (Urban/Suburban). Superimposed on the RF propagation curves are the estimated minimum necessary signal level at a mobile receiver for correct decoding of signals with different modulation orders, from 16 QAM (Quadrature amplitude modulation) to 1024 QAM. The figure considers a single cell in isolation, while inter-cell interference will reduce the range of successful decoding when there are interfering base stations at similar distances as the serving one.

Beyond this first-order consideration, many service providers have proprietary, derating function models that account for considerations such as building penetration rates and multiband coverage parity to

achieve specific performance and QoS metrics. These objectives then influence the detailed scheduler decisions that allocate spectrum for individual users on an ongoing basis.

Looking at the bottom of Figure 20, it is immediately clear that the range of propagation for the 200 W signal corresponds to a coverage area that is significantly smaller than 200x the area of the coverage of the 1 W signal. Indeed, as the subtitle of the figure calls out, a 200x increase in power only results in a 11-41x increase in area coverage depending on the conditions. In other words, the 200x increase in power results in a 5-20x decrease in efficiency.

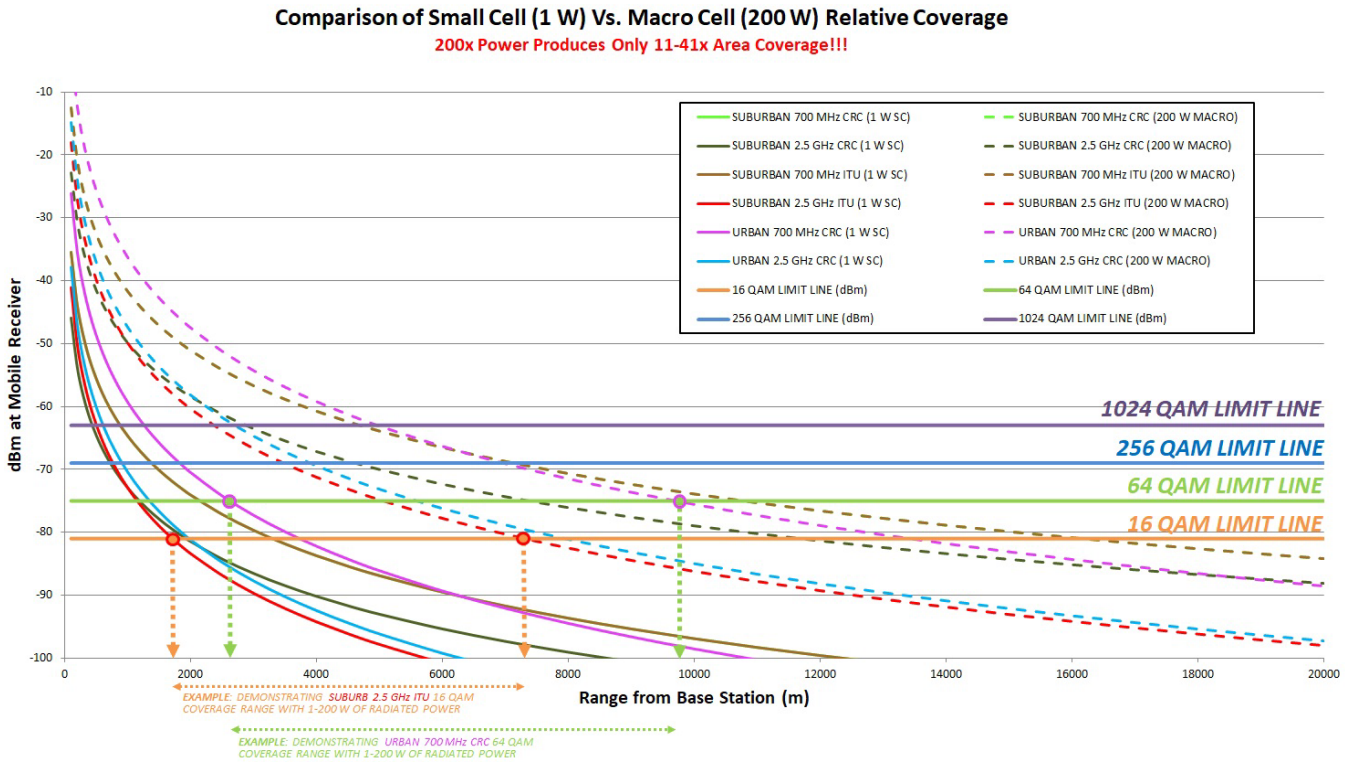


Figure 19. Comparison of Base Station "Carpeting" at Various Power Levels and Frequencie

Image courtesy of Eridan Communications

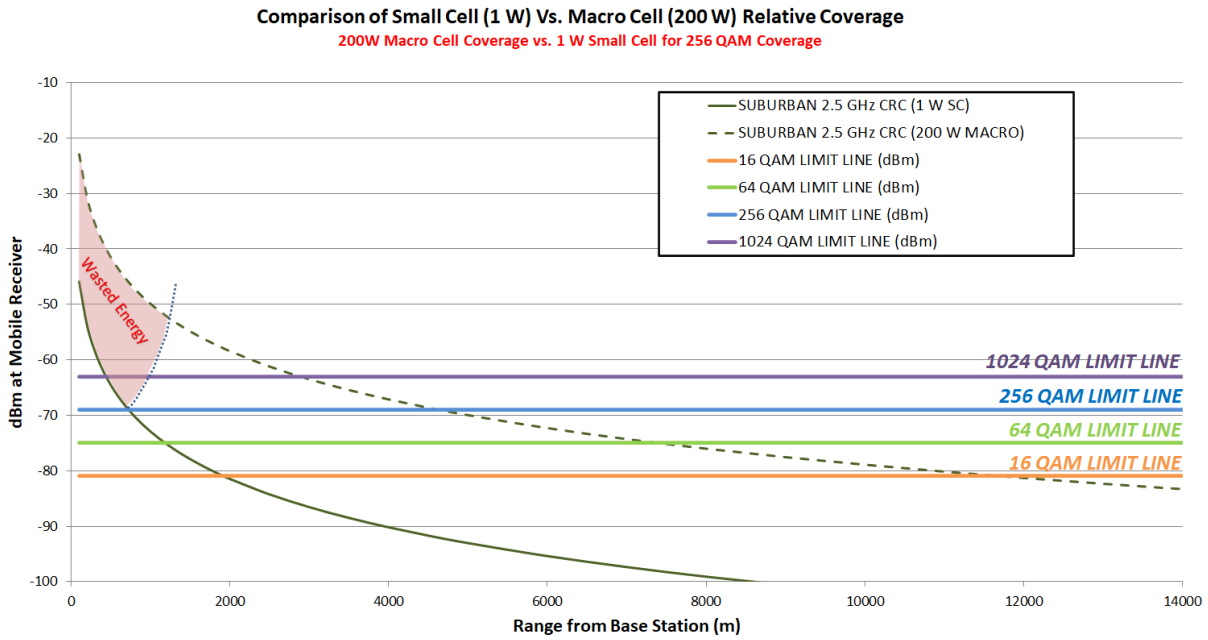


Figure 20. The propagation comparison for a specific case of 2.5GHz using the CRC model data for a Suburban configuration. While a 200W Macro Cell provides extensive coverage, the 1W Small Cell enjoys a large efficiency advantage because RF propagation is significantly more lossy than 1/r2.

Image courtesy of Eridan Communications.

The origin of this decrease is illustrated for a particular case in the companion figure, Figure 20. The "wasted power" of the 200 W Macro Cell has been filled in red shading, bounded by the RF signal from the 200 W system on the top end, the 1 W system on the near side, and a complementary image of a 1 W system on the farther side for the 2500 MHz, 256 QAM, and suburban CRC data called out in the figure. Note that the axis is logarithmic: each vertical tick represents a 10x increase in RF signal intensity.

This fundamental characteristic of the physics of propagation of RF is a key force driving Mobile Network Operators towards network densification and an energy-centric network topology that has been coined as "carpeting." As described earlier, the energy efficiency gains of densification are at their largest when the RF power is the dominant part of a base station's total energy consumption.

For discussion's sake, let's say that we are performing coverage planning in a suburban setting bisected by a commuter freeway. Our carpeting plan would start with a messaging and control chanel, transmitted by a 200 W, 700 MHz macrocell basestation in an urban setting that has a 13 km range at the 16QAM limit (even further with a QPSK constellation complexity) which penetrates walls and building better than higher frequency channels. Small cells along the freeway with high gain antennas, would overlap along the freeway to cover the demands during rush hour and sleep when not needed. Similarly, to augment the macrocell coverage and address the suburban demand, HetNet small cells (a mix of supplemental downlink basestations, 1 W basestations and WiFi access points (AP)) are distributed throughout the neighborhoods every 0.5-2 km, which will be activated as needed to ensure the macrocell has capacity to coordinate the overall messaging and variable upload/download capacity. Some WiFi access points will be household centric and handshake with the macrocell as needed. These household

centric APs must be IPM enabled to conserve energy when not needed for local WiFi interactions and coordinated HetNet downloads.

By performing a rigorous tactical carpeting plan, regions with a mix of high density and low-density utilization will be able to deploy a sub 1 GHz omnidirectional primary and control channel with a geographically rich set of higher frequency, directed secondary and supplemental channels focused on the time varying high-density load being serviced moving rapidly on the highway. These secondary and supplemental channels will breathe (via a time varying range of coverage in distance and angle) and be activated/deactivated as demand changes with time of day and traffic patterns, resulting in reduced interference on other receivers within range. A dynamic metric to evaluate the cost and performance of an advanced future, complex tactical carpeting plan will include an ensemble of base stations, access points, gNodeB and eNodeB from a heterogenous network (HetNet) wireless technology deployment being dynamically activated and deactivated for improved power utilization that are being utilized to satisfy a dynamic user demand. Queueing theory augmented with machine learning, utilizing historic user behavior patterns, will be used to load balance the download packet throughput (TPUT) of the system. The planning and use of a variety of frequencies will also rely on the theoretical (propagation) and historical performance (coverage) to optimize the TPUT scheduling on the fly. This carpeting plan will include beam steering components as one of the elements in a palette of coverage technologies to address usage demands and ensure the macrocell has sufficient capacity for messaging and coordination of assets.

5.3.2.2. Interference Management Within a Cell

The Massive MIMO technology gives the ability to process the antenna signals at a base station to suppress interference in the spatial domain. Despite the word “massive”, it is possible to utilize multiple antennas and radios also in small cells, both at conventional frequencies and in mmWave bands. The interference that can be mitigated at a base station can originate from multiple users that are spatially multiplexed within the cell, but in a small-cell context with a limited number of users per cell, it might be even more important to reject interference that leaks from other cells. Transmit precoding and receive combining schemes that take such interference into account, such as so-called multi-cell MMSE processing [76], can be utilized to limit the interference in the spatial domain. This solution will start to be utilized in the near-term.

5.3.2.3. Efficient Control Plane Transmission

Small cells will exist under the umbrella of macro cells. By decoupling the control and data plane transmissions, the data can be transmitted from a serving small cell while the control plane can be emitted from other points, including a macro cell operating in an entirely different band [77]. In this way, the reliability of this signaling can be increased and interference can be alleviated without requiring cooperation between multiple base stations at the physical layer. To reduce the energy consumption and overhead for control signaling, some signals sent from the base stations can be sent on-demand rather than being broadcasted periodically. This was described earlier in Section 5.2 when discussing idle mode. Moreover, an intermediate step between idle mode and active mode now utilized in 5G for devices that are inactive but likely to soon switch to active mode. This feature can reduce the control signaling that is required to handle devices that are switching frequently between active and idle mode, as is commonly the case with bursty traffic. This feature will address control plane issues in the near-term.

5.3.2.4. Interference Management Between Cells

While intra-cell interference is relatively straightforward to manage by Massive MIMO, if the base station is purposely scheduling users that it can separate in the spatial domain, the inter-cell interference is more complicated to deal with. Firstly, because it can change rapidly and, secondly, because there can be a much larger number of interferers than the base station can handle on its own.

Evolved coordinated multipoint techniques can then be utilized to let multiple base stations cooperate in serving the users and mitigating interference. Compared to deploying larger arrays in each cell to handle the interference better, the multi-cell coordination can be more efficient since it requires less transceiver hardware. The interference suppression is implemented on a software interface that connects the base stations. If the neighboring cells are sharing the same edge-cloud baseband processors, the implementation might not even require any further infrastructure deployment [78]. In the mid-term, these methods are predicted to play an important role in limiting the interference. The algorithms implemented in the edge-cloud can utilize machine learning to detect traffic patterns, learn the local propagation environment, and in other ways tailor their operation to better deal with interference and increase the energy efficiency. These features will enable a situational awareness, where neighboring cells know which users are around, which other base stations are around, etc. These base stations might not use the same spectrum but anyway enable seamless handoff. Base stations with overlapping coverage can cooperate to enable longer sleep modes and other types of energy-saving features [79].

5.3.2.5. Cell-free Architecture

The ultimate long-term solution to the interference issue is to remove the cell boundaries all together, leading to what is called a user-centric cell-free architecture [80] [81]. The locations of the base station antennas don't have to change, but they are no longer creating independently operating cells. Instead, for each user, the network divides the base stations into two categories: those that can affect the user with its actions and those that cannot (or whose potential impact is negligible). The concept of base stations, as distinctly different transmission units, become obsolete and we can instead talk about Transmission and Reception Points (TRPs). With this in mind, entirely new deployment scenarios can be enabled where the antennas are located close to the users instead of being gathered at distinct points far from the users. This transition is illustrated in Figure 21.

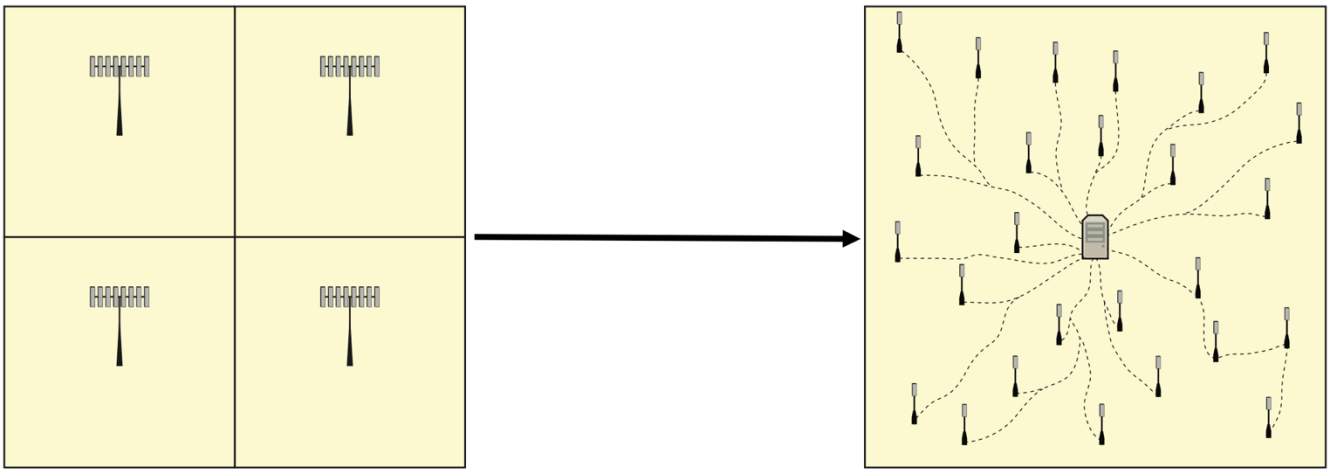


Figure 21. The networks will gradually transition from the cellular architecture to the left to the cell-free architecture to the right.

Geographically speaking, the cell-free architecture implies that each user is associated with all the surrounding the TRPs. These TRPs are cooperating in the decoding of uplink signals and are jointly transmitting the downlink signals. In this way, the interference that the TRPs would otherwise cause to users in "other cells" can be controlled and limited, although not removed entirely. It is called a "cell-free" architecture since the distinction between cell-center users (those with high SNR and little interference) and cell-edge users (those with low SNR and high interference) is removed. Every user will experience a situation where they are in the middle of a cell. There will still be SNR variations, but they are reduced and particularly the interference is reduced. Another advantage of this, is that the signaling related to handovers can be reduced as the network seamlessly can handle this without having to inform the user when moving through the network.

The price to pay for creating a cell-free network is that signals must be co-processed by multiple TRPs. Since current networks are already moving towards a situation where neighboring TRPs are sharing the same baseband processor in an edge cloud, the implementation cost (monetary and in terms of energy) can potentially be limited. If a data packet can be delivered faster, thanks to less interference and thereby a higher spectral efficiency, the cost for the interference cancellation will potentially outweigh the increased computational complexity. Or to put it differently: cell-free functionality is not a feature that should always be utilized, but the network should intelligently identify which users need this additional layer of interference mitigation and only apply it when it can increase the energy efficiency.

A first step towards a cell-free architecture could be so-called hyper-cell abstraction [82]. A hyper-cell is formed by a number of transmission and reception points (TRPs) that jointly serves the users in the coverage area to provide a consistent user experience across the network. A TRP can be a macro base station, a small cell base station, a Remote radio unit (RRU) or any other kind of light node. A more long-term approach to implement a true cell-free architecture is to make use of so-called radio stripes [64], where multiple TRPs are deployed along the same cable to limit the hardware footprint. The cable could either be an ethernet or optical fiber, which delivers fronthaul and power supply to the small-sized TRPs that are deployed along the cable.

In summary, the convergence towards a cell-free architecture is a long-term vision. The networks should first be densified to reduce the propagation losses, then MIMO techniques should be utilized to make

more efficient use of each TRPs, and finally the cell-free architecture can be utilized to deal with interference between cells. While the local processing at each TRP can be developed using conventional theory [83], the joint operation of many TRPs, and particularly the resource management, will be complicated and likely make use of machine learning methods to achieve an efficient operation that exploits the local propagation conditions, traffic patterns, typical user movements, etc.

5.3.2.6. Coverage Improvements with Intelligent Reflecting Surfaces

An alternative to deploying additional base stations is to improve the coverage area of existing base stations. In its elementary form, passive repeaters can be deployed to circumvent shadowing by creating additional signal paths. This case is illustrated in Figure 22, where the user device is located in a valley that is shadowed from the base station, but a passive billboard is deployed to create an additional signal path [84]. The strength of this path is limited by how large an area it should provide extra coverage to. A larger surface can gather more energy, but will also focus it on a smaller region. The benefit of deploying a passive surface, instead of another base station, is that a higher data rate can be achieved without the additional energy consumption and cost of having an extra base station. There is a renewed interest in deploying passive surfaces in 5G since the range of mmWave signals is short due to additional blockage and penetration losses. In the near-term, this technology can be utilized to improve the network coverage and thereby the energy efficiency.

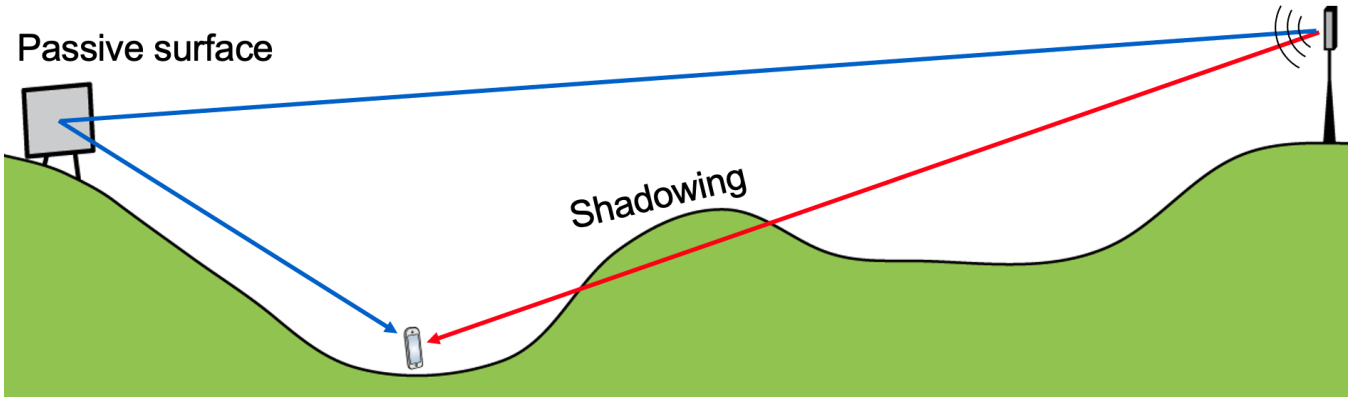


Figure 22. A passive surface can be deployed to reflect signals from a base station towards shadowed areas.

The concept of artificial radio space [85] takes this technology one step further when it comes to addressing the increasing energy consumption of current and emerging networks and the desire to go towards more green networks. With this promising concept, the wireless propagation environment is turned into an intelligent reconfigurable space that plays an active role in transferring radio signals from the transmitter to the receiver. One way of viewing it is to turn the passive surface in Figure 22 (above) into a reconfigurable surface as shown in Figure 23 (below). Depending on which users that is served, the reflected beam can be focused differently and larger surfaces with a stronger directivity can be utilized since the coverage area is not static but adaptable.

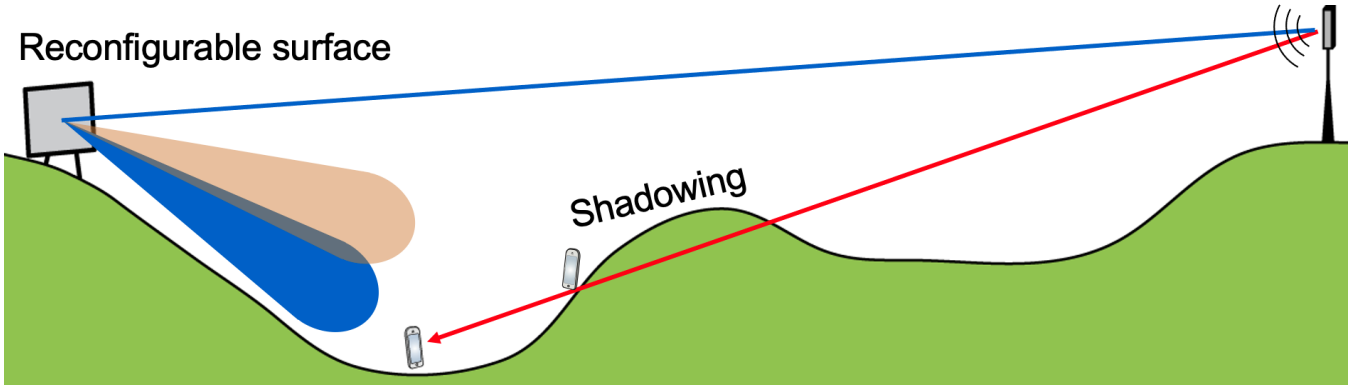


Figure 23. A reconfigurable surface can be utilized to direct signals from base stations to shadowed locations in an adaptive manner.

This concept can be enabled by the use of Intelligent Reflecting Surfaces (IRSs) in the environment [86]; these IRSs can be realized by low-cost, low power consuming, essentially passive elements of metamaterial that scatter the electromagnetic waves to form a directive beam pattern towards the desired target locations in a passive manner (i.e., without generating new radio signals and basically without incurring any additional energy consumption) to alleviate “not-spots”. The surface can be viewed as a steerable mirror and can even be more effective than mirrors since it not only changes the direction of the signal but also the shape of the waveform [87].

In the paper [88], tools from stochastic geometry are utilized to study the effect of a large-scale deployment of IRS on the performance of cellular networks. It was shown that the deployment improves the coverage regions of the base stations without the need for extra transmit power. Furthermore, it was shown that to ensure that the ratio of blind spots to the total area is below 10-5, the required density of IRS increases from just 6 IRS/km2 when the density of the blockages is 300 blockage/km2 to 490 IRS/km2 in a denser environment when the density of the blockages is 700 blockage/km2.

A key challenge in implementing IRS technology is not to reconfigure the surface, but to learn and adapt the reconfigurability over time [87]. In the mid-term, IRS technology that can adapt slowly to long-term traffic variations, which does not require instantaneous channel state information, can be utilized. In the long-term IRS technology, which can also adapt their behavior to specific users, will be developed. In this way, additional signal paths can be created to deal with interference and the number of signals that can be spatially multiplexed [85].

Table 5. Potential Solutions to Address "NEED #2 - Small Cell Migration"

Near-term Challenges: 2022-2025	*Potential Solutions to Near-Term Challenges*
Complicated control plane	Decoupling of control and data plane enables a more flexible operation of the control plane, potentially in a different band. An intermediate mode between idle and active mode is currently utilized in (i.e - 5G-NR “inactive state”) to reduce the signaling for devices that are frequently switching between being active and idle.
Increased intra-cell and inter-cell interference	The use of multiple antennas at the access points can be utilized to suppress inter-cell interference in the spatial domain. The intra-cell interference, caused by spatial multiplexing, is locally controllable and will only be caused if the local algorithms determine that the throughput gains are positive.

Improved coverage in mmWave bands	A combination of additional small cells and passive repeaters can be utilized to increase the worst-case SNR in the coverage area.
Mid-term Challenges: 2026-2027	*Potential Solutions to Mid-term Challenges*
Coordinated multipoint	Clusters of neighboring cells, connected to the same edge-cloud processor, can coordinate their resource allocation as well as physical-layer transmissions to limit interference and increase energy efficiency. Machine learning methods can be utilized to adapt the algorithms to the local, time-varying conditions.
Improved coverage depending on traffic variations	Slowly reconfigurable IRS technology can be deployed to reflect signals from base stations to locations with high traffic.
Long-term Challenges: 2028-2032	*Potential Solutions to Long-term Challenges*
Cell-free architecture to alleviate inter-cell interference	A cell-free network architecture is the long-term goal for a small-cell deployment, where all transmission points surrounding a user are collaborating in the physical transmission as well as resource allocation. To achieve a scalable implementation, machine learning methods will be key to avoid the need for a central control entity.
Improved signal rank and coverage for specific users	Real-time adaptable IRS technology can be utilized to direct signals towards users and follow them as they move around, to increase the SNR and create additional signal paths.

5.4. Base Station Power – Need #3

5.4.1. Challenges

The main challenges to be overcome for future base station requirements center around delivering high power signals at ever increasing frequency bands, while minimizing unwanted emissions in the same or other bands.

5.4.1.1. Challenges with Unwanted Emissions

Fundamentally, wireless systems are interference limited and that interference can be self-generated or inflicted by other transmitters. For greatest utilization of spectrum, it will be required to get the signals onto wired connections as quickly as possible, at highest spectral efficiency through higher constellation complexities, transmitted at lower power levels. To accomplish these objectives, the output signals will need to be transmitted with high fidelity (very low EVM or Error Vector Magnitude) and low spectral regrowth, in beams that are directed toward the user terminal / access point. Future technologies will utilize:

1. Higher fidelity transmitters to generate higher constellation complexities with lower EVM and lower spectral regrowth.
2. Antenna arrays to further reduce the transmitter power level by directing the radio waves more effectively towards their intended target, in addition to reducing the interference impacting other receivers within range.

Future high-density wireless networks will need to benefit from the cost and technology trends of innovation by having every transmitter turn to the lowest possible power, generate signals efficiently with the highest fidelity, leverage cell densification to service closer receivers and direct more focused energy to the receiver; this way, interference will be reduced, and greater capacity will be realized for a given band of spectrum.

5.4.1.2. RF Semiconductor Process Technologies Evolution and Application Fit

The efficiency of the transmitter is highly dependent on the radio architecture, waveform Peak-to-Average Power Ratio (PAPR), frequency, and the semiconductor technology being used. Fundamentally, the ability to precisely follow the signal requirements and faithfully transmit the constellation of symbols requires a circuit design that is pushing the limit of transistor technology.

Analog applications typically require semiconductor process technologies offering transistors with an transition frequency (Ft) approximately an order of magnitude higher than the frequency of the signals being processed (at least when dealing with large signals - e.g. Ppeak > 1 W). These constraint, being driven by either distortion requirements or simply slew rate of the signal edges, limit the number of technologies that can be effectively utilized, whereby for example Gallium Nitride (GaN) has gained acceptance in sub-6GHz RF application, thus yielding the ability to process relatively large voltages, but cannot yet support mmWave RF power amplifiers, which can be satisfied by CMOS Ft's ranging between 200 and 300 GHz, albeit limited by a breakdown voltage of only a few volts. As processes continue to improve in performance and reliability, III-V semiconductor materials will continue to extend their reach into higher frequencies and power levels; yet MIMO arrays at lower power may still favor silicon solutions, due to cost and integration capabilities.

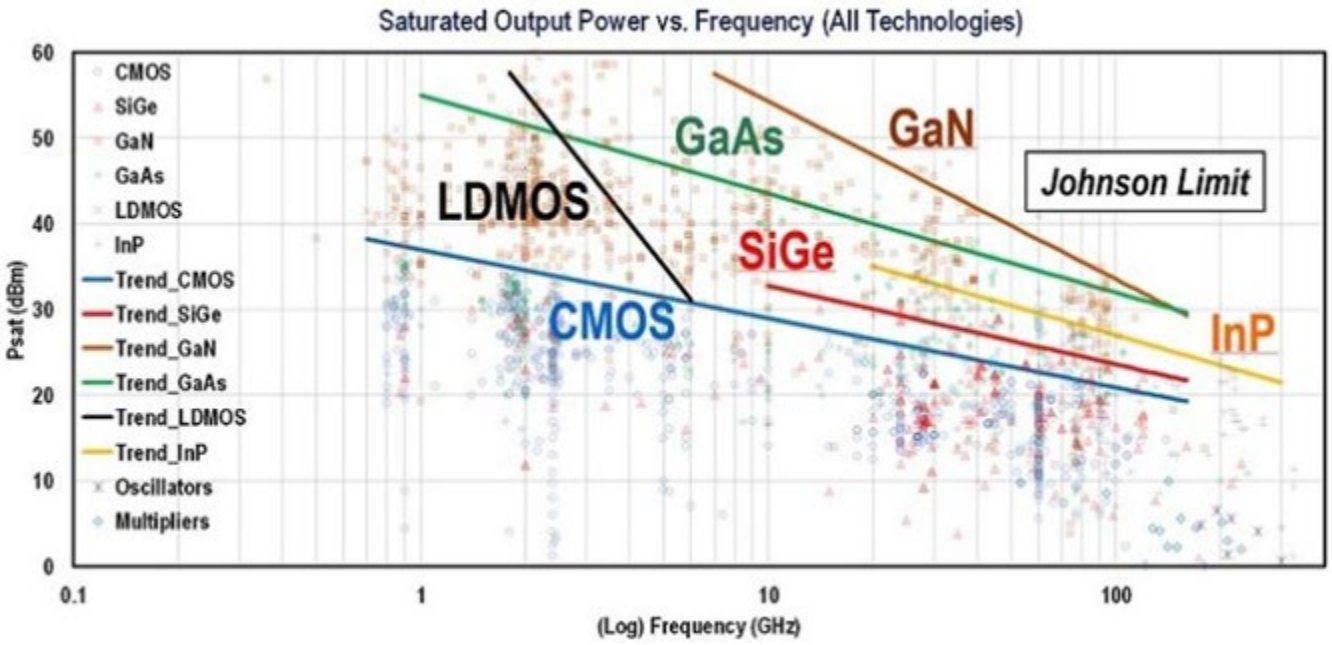


Figure 24. Power Amplifiers Performance Survey 2000-Present [88].

The Georgia Tech database [89] identifies the peak power level trends for a variety of technologies, which factors in the technology ft, Ron and operating conditions. The waveform statistics, combined with the efficiency in backoff, can be used to perform a quick estimation of the final stage efficiency. In many cases this may include an assumed, efficient digital signal processing (DSP) or predistortion (DPD) to achieve a transmitted signal with the highest possible fidelity. Engagement with the standards committees will be needed within future standards efforts to ensure that lower PAPR waveforms are available for the primary and control signals. Between the reduced output power capability at higher

frequencies and the increased attenuation associated with the circuitry and propagation, the ever-increasing PAPR of waveforms coming from the standards committees is having a negative impact on efficiency and the amount of power required per bit transferred.

Significant innovations are needed to address the future energy demands for wireless systems, and historically there have been barriers to entry for product development and integration that have hindered startup companies in contributing innovation. For product development, the challenges for startups include establishing robust and accurate test platforms. There is momentum in the GNURadio community [90] to generate radio resources that enable rapid prototype development. Some in the industry are addressing this by utilizing open-source hardware, the barrier to entry for innovative startups may be significantly reduced by ever increasing modulation complexity, enabling increase innovation of components with greater efficiency, thus supporting complex modems and making RF hardware components more accessible. In addition, these energy saving components have a clearer path to market and product integration via the OpenRAN initiatives, defining open interfaces within the RAN components and subcomponents, enabling the selection of best-in-class performance. It should be noted that discussions regarding OpenRAN, and other open-source initiatives are a sensitive topic of the debate in the industry that has yet to gain wide acceptance.

5.4.1.3. RF Semiconductor Challenges & Limitations for Massive MIMO and/or mmWave

The directive radiation pattern of antenna arrays helps lowering the transmit power per power amplifier (PA) needed to meet the specified Effective Isotropic Radiated Power (EIRP) and, hence, enable silicon technologies to address both sub-6 GHz and mmWave 5G radios. However, EE to generate the required power remains one of the biggest challenges both for user equipment (Handsets, CPE or Customer Premises Equipment) and base stations. The problem is much more enhanced in mmWave applications compared to sub-6 GHz as the power amplifier efficiency is low (below 25%) at frequencies above 28 GHz. III-V compound semiconductor technologies like GaAs, GaN can generate higher power, but cost, supply issues and challenges of integration have in the past constrained these technologies to only final stage PA for sub-6 GHz base stations. New developments in processes and integration capabilities are addressing these limitations.

For RFSOI (Radio Frequency Silicon on Insulator) & Silicon Germanium (SiGe) technologies, the coverage (max uplink Tx power limit) and long-term reliability and ruggedness for mmWave applications are additional challenges that are being addressed.

A mmWave handset today consumes 10x the power in the radio when operating in the mmWave frequency domain versus sub-6 GHz. For a compact handset this can bleed the battery completely after a couple of hours of continuous mmWave communication. Taking a closer examination of the handset architecture regarding 5G-NR transceivers, we see that a common modem is used for the baseband processing and initial signal generation. In fact, both FR1 (sub-6 GHz) and FR2 (mmWave) transceivers are configured with multiple transmitters and receivers to avoid hand blockage and enable MIMO configurations. As a result of the better signal control, efficiency and propagation characteristics, the sub-6 GHz (FR1) signals have options to send signals at higher constellation complexity, that require higher signal fidelity. The final stages are 30-50 % efficient for FR1 PAs and typically GaAs HBTs with higher voltage breakdown. This technology has been used for >20 years and proper reliable operational conditions have been established. In comparison, the mmWave (5G-NR FR2) handset requires multiple (4 or more) transceiver modules per device to avoid hand blockage and

obtain the best link. Each module is reconfigurable, with multiple sub-arrays available to transmit or receive signals from a variety of general directions, but has limited range, and is inefficient (<10 %). The FR2 front end amplifiers are built in low voltage silicon processes, so to get higher power levels, higher voltages are being pushed by stacking devices which tend to stress the long-term reliability and ruggedness of silicon devices.

One of the key attributes that limits the transmit power level for handsets is the amount of radio energy that is absorbed by tissue. There are national safety boards throughout the world that establish safety limits. There are significant differences between the way the FR1 and FR2 radio energy interact with tissue resulting in very different requirements. These requirements are changing and will have an impact on the future evolution of user devices. In a base station, one can reach an EIRP of >60 dBm but it needs a large array size (1024 array elements) for CMOS ICs, while the use of SiGe or RFSOI based beamforming devices can reduce the number of array elements to 1/4th for the same EIRP, due to their higher power handling capability. One can reach the max EIRP limit of 71 dBm using either SiGe or RFSOI using ~500 array elements, although human and/or animal safety limits for EIRP may constrain the ultimate available power delivered.

Table 6. Challenges Associated with "NEED #3 - Base Station Power"

Near-term Challenges: 2022-2025	*Description*
Requirements on Out-of-band Distortion Must Be Satisfied	Higher fidelity transmitters and phased arrays capable of operating at relatively lower power per element in a denser cell environment are required to minimize unwanted emission while increasing EIRP
Determining Macro vs. Small Cell Size/Needs vs. Freq.	Three different physics drivers interact in this set of tradeoffs. Excess power - the power radiated from the cell that is greater than that required by the UE and which depends on the range to the UE - is a source of inefficiency for macro cells relative to small cells. Working in the opposite direction, the power consumption of the protocol stack at the Radio Unit (RU) and Digital Umit (DU) is independent of range and represents a baseline level of power consumption that scales with the number of cells and therefore advantages macro deployments. Finally, riding on top of both of these considerations, the propagation characteristics vary as a function of frequency and local topography which suggests that a heterogeneous network of macro cells at lower frequencies and small cells at medium and higher frequencies is called for. Key challenges include systems of systems that are inherently suited to dynamic heterogeneous operations, minimizing the baseline power consumption of the protocol stack at the DU/RU, and a functional systems-of-systems model that allows for both overall optimization and the identification of system efficiency roadblocks.
Support for Many RF Bands	The increase in the number of spectrum bands allocated for 5G and future standards requires the implementation of heterogeneous transmitters with different RF technologies to provide optimized performance in different bands
Heatsink/Package Size Becomes Impractical for PAE	A modular transmitter approach that fits within the λ/2 spacing (5mm at 30GHz, 50mm at 3GHz) between each antenna element requires high levels of integration. In the 30GHz case this is required to achieve the spacing in the first place; in the 3GHz case this is required to achieve an array structure that can be "transparent" for the purposes of wind loading. Such high concentration of transmitters poses a significant challenge in heat extraction if the transmitters have efficiency meaningfully lower than 50% (e.g. 10%), thus adding cost and energy dissipation if active cooling is required.
Mid-term Challenges: 2026-2027	*Description*
Power/Energy Telemetry Data Acquisition	The need to optimize EE across the RAN deployment and ensure that each node is optimally utilized without requiring performance throttling will require monitoring and real time measurement of energy utilization, thus enabling dynamic reconfiguration and utilization of the infrastructure. Commensurate instrumentation will be needed to provide measurement data on time-dependent energy demand and availability, as well as cognizant management of resources and demand.
Power/Energy Data Analytics	Evaluation of dynamic energy data to predict performance requirements over time of day, day of week, weather condition, and other statistically significant parameters, is needed to enable a proactive approach to energy management and a predictive capability to assess peak energy demands.

Defining Energy-optimal Control Feedback Loop(s)	Dynamic HW-level optimization is required to minimize energy expenditure at the Cell level, based on real-time requirements, to prevent throughput reduction due to exceeding thermal limits.
Long-term Challenges: 2028-2032	*Description*
Enabling/Deploying Energy-optimal Control Feedback Loop(s)	A multi-layered control approach, based on predictive models, is required to optimize real-time Cell energy consumption based on real time requirements, which leverages the nesting of control loops, starting from HW-level optimization through RAN nodes coordination and dynamic deployment of computing resources across Edge and Cloud.

5.4.2. Potential Solutions

5.4.2.1. RF Semiconductor Path to 6G

Continued research on Si technologies including RFSOI & SiGe and use of GaN-on-Silicon technologies can take us to higher frequencies (95 GHz and beyond). A lot of innovation in mmWave packaging is needed to reduce loss, parasitics and form factor [91].

5.4.2.2. Power Electronics in 5G&B

The requirements for power management in the 5G ecosystem span many orders of magnitude in the power scale: from energy harvesting applications where 80 % conversion efficiencies have been achieved down to the µW level all the way to Grid-connected converters for server farms, which handle MW power levels at >98 % efficiency.

Covering all power management applications is not within the scope of this chapter, but we will highlight here some of the critical technology constraints and some significant innovations that are on their way to overcome key bottlenecks affecting efficiency.

The continued migration to lower silicon geometries for improved performance of SoC's and computing cores leads to the need to deliver very low voltages to high current loads. All of these loads (Core Processors, BB Processors, MCU's, DSP's, FPGA's, HW Accelerators and Memory) must become more efficient, and system integration is the key to achieving such gains. This is accomplished by co-packaging heterogeneous technologies to enable a tight interaction of Power Management with the subsystem blocks, thus dynamically optimizing system performance and power delivery.

It has been acknowledged for many years [92] how physical separation of such components progressively leads to increased inefficiencies with each new generation of semiconductor process; yet, Power Management has traditionally provided incremental rather than disruptive solutions, due to a siloed industry and the fact that Moore's Law has historically provided a predictable way to improve the "energy/bit" FoM. As this free ride has lost momentum, now SoC developers are finally addressing head-on the challenge of improving system efficiency, which is no longer just a lack of "greenness" but an effective brick-wall toward higher performance. See Section 5.5 for more detail on this.

As now the majority of "digital" supply voltages are < 1 V and often down to 0.5 V for battery operated devices, power supply regulation requirements are becoming exceedingly difficult to meet. Since the power dissipation of a CMOS core increases with an exponent greater than 2 for the supply voltage, even 10's of millivolts of supply margin become significant multipliers to power dissipation. Current solutions, where the Point of Load (POL) regulation is on a separate package from the Load and connected through copper traces on standard organic substrates, are forced to adopt supply margins well in excess of 100 mV, due to the need to ensure operation in the presence of very fast load transients. This causes very large efficiency losses that are growing at every new generation of silicon lithography

node. Maintaining an accurate regulation when the frequency content of the energy required by the load extends all the way to more than 1 GHz requires advanced Power conversion architectures and a design of the Power Delivery Network (PDN) that is capable of maintaining a low impedance all the way from DC to 1 GHz: co-location of the Power Management with the load through technologies like Power Supply in Package and Power Supply on Chip (PSiP and PwrSoC) as well as "Granular Power Supply" are now becoming fundamental to achieving acceptable system efficiencies. Co-packaging power supply and SoC can provide as much efficiency advantage as moving to the next node in silicon process technology used to, but no longer does. To achieve such integration, passive components need to be miniaturized, which requires converters to operate at very high switching frequencies (tens to hundreds of MHz compared to few MHz, as typically used.

Not only the physical distance of the Power Supply to the load is important, but also that of the "data Path." As more energy is lost "transferring bytes" than "computing bytes," interconnect technology has often become the limiting factor for intensive computing applications and communication processors, even at the board level: chip to chip interconnect for high bandwidth systems is moving to solutions that avoid the signal path going through organic substrates and rather uses silicon bridges or other 2.5D and 3D packaging solutions to eliminate the need for driving low impedance I/O's. For longer paths, optical communication is becoming standard and electrical to optical interface technologies (die to fiber) continue to progress, both in performance and cost. Extensive information and presentations on heterogeneous integration technologies can be found in the Heterogeneous Integration Roadmap webpage of the IEEE Electronics Packaging Society [14].

Besides fast transistor technologies, high Q passive components are required for power conversion and distribution: new magnetic materials with lower hysteretic losses are facilitating the migration to higher converter frequencies and hence higher power densities; and silicon trench capacitors are providing very low impedance for low-voltage high-performance cores, thus improving the PDN performance.

To transport the higher power demands of 5G, continued investment in a higher voltage distribution infrastructure is required. Higher voltage distribution is not new to the power management industry with +/-48 V commonly used in telecom infrastructure systems. Higher voltages of 48 V, and increasingly popular 54 V, are favored as they reduce the current losses squared (I2R) associated with copper board losses and interconnects. Both 48 V/54 V voltage levels operate below the Safety Extra Low Voltage (SELV) of 60 V and require limited or no safety isolation compliance depending on the type of equipment being deployed.

A higher voltage 48 V distribution can have other second order benefits. In general, ACDC front-end, power supplies with 48 V output will see an efficiency improvement of typically 2 % compared to a lower 12 V voltage output system commonly used today. Additionally, there is considerable investment and innovation in the step-down process, from 48 V down to the low voltage Point-of-Load silicon level. This includes high efficiency resonant conversion and 'lossless', open loop capacitor divider techniques. The latter can operate with efficiencies as high as 98 %. The higher voltage transportation of power does open opportunities for wide bandgap processes such as Gallium Nitride (GaN) and Silicon Carbide devices (SiC). In general, both technologies tend to provide greater efficiency benefit in higher voltage applications: in the ACDC Power Factor Correction (PFC) system, the DCDC switching architecture and the 48 V step-down to the point-of-load.

5.4.2.3. Power Packaging

As “More than Moore” technologies have been advancing, the ones that are having the greatest impact to system performance are related to packaging. Increased frequency of operation and reduction of application size (and thus physical distances) lead to reduction of parasitic elements and losses, but since energy density grows inversely to the volume, efficiency needs to be drastically improved to maintain cost-effective thermal management solutions.

Innovative packaging technology has become widespread in the power management industry and has helped to reduce the footprint of power regulators through advanced FET technology combined with passive device integration. However, the ability to shrink packaging further can only be achieved if the efficiency and losses are maintained or improved, through further component improvements, and/or the thermal impedances of the package are further reduced.

A SoC presents multiple functions on the same chip, usually a mix of analog and digital with the advantage of fast, low parasitic interconnections. Presently, this remains popular in many applications but may be limited in the high density 5G arena. Generally, the mix of analog and digital technologies results in complex design with slow time to market and high development costs – sometimes exacerbated by noise and interference issues due to the complex integration. Often, the resulting IC device lacks the flexibility to be re-configured to another customer or adjacent market. This may work well for some well-defined, custom markets but unlikely to serve a broader strategy. Additionally, the development of analog and digital silicon technologies moves at different speeds. Therefore, the combined process is usually a compromise of cost and performance between the two technologies – never an optimum.

For this reason, many companies have turned to Power System in a Package (PSiP) or Multi-Chip-Modules (MCM) to improve power density and save valuable board space by integrating silicon and passives on a substrate material. The multi-chip module idea enables silicon from optimized analog and digital processes to be used as separate dice with low parasitic, high-speed interconnect between the ICs. The integrated module size can be reduced by using fine pitch components and packaging the devices more tightly. An injected resin molding maintains the insulation between components while most importantly reducing the thermal impedance and facilitating heat spread across the module. There are further examples where companies have integrated separate power stages – combining FET, driver and inductor – that can be driven by a low voltage ‘brain’ controller in a more digital friendly process. This strategy uses the best processes for the power analog integration and for the digital control portion. Additionally, modules can be specified carefully as ‘black boxes,’ thus a multi-sourcing specification can be derived to enhance supply chain security and competitive pricing.

The PSiP or MCM modules use a variety of sophisticated techniques to increase power density: from ICs mounted on a regular PCB to high density Ball Grid Array (BGA) or flip chip on HDI (High Definition Interconnect) substrates. Size reduction is further achieved through embedded processes where the die (or dice) and passive devices are ‘embedded’ in the cavities of the PCB or substrate material. In addition to the size and weight advantages associated with embedded package technologies, parasitics between components are dramatically reduced, leading to efficiency improvements, which may allow higher switching frequency and potentially further size reduction.

Sophisticated 2.5 and 3D packaging solutions now allow bringing together heterogeneous technologies, thus optimizing performance at the functional level without the burden of inefficient interconnect outside the chip. Transition from solder bumps to copper pillars, the ability to embed devices in the substrate, Wafer on Wafer bonding (WoW) and Through Silicon Vias (TSV) are providing options to

the system designer that were inconceivable just a few years ago. Liquid cooling is no longer relegated to supercomputer applications, but is now penetrating much broader markets.

Faster semiconductor processes, improved passive components and advanced packaging and integration are enabling new and more efficient architectures: from Envelope Tracking RF Power Amplifiers to resonant capacitive converters that efficiently supply low voltage to the processors from the 48 V bus rails. Of particular interest is the application of these technologies to the high level of integration required to implement Massive MIMO antennas at mmWave frequencies, where each transmit module must fit within a 5x5mm footprint.

As Massive MIMO architectures and cell densification lead to the need for a large number of RF Power channels at relatively low power, the advances in commercial solutions for Mobile devices can be leveraged to develop solutions to tackle Base Station power needs with high levels of integration, programmability and reliability.

5.4.2.4. Thermal Mitigation

In addition to the advances in efficient power transmission and packaging technology, 5G expansion requires the implementation of advanced cooling structures. Ironically, forced air cooling can expend large amounts of power while attempting to cool the power systems. Liquid cooling is relatively well understood at a datacenter room or chamber level, but it may become likely that we will need more localized, innovative ways that can extract heat from localized power sources often buried within the computer system. Cooling technologies do look promising and offer the promise of being able to access local system hotspots and extract the heat to a cooler system node. This does lead to the intriguing challenge of recycling or reusing the excess generated heat. It would certainly be interesting to consider Edge based systems to be strategically located in building and office structures where the excess heat could be reused.

The 5G power challenge can only be solved with a combination of electrical power efficiency improvements at semiconductor level combined with advanced packaging technologies to miniaturize and increase power density and advanced cooling systems [93].

Table 7. Potential Solutions to Address "NEED #3 - Base Station Power"

Near-term Challenges: 2022-2025	*Potential Solutions to Near-Term Challenges*
Requirements on Out-of-band Emission Must Be Satisfied	Development of Massive MIMO radios with many low-gain antennas, utilizing handset-grade hardware instead of few high-gain antennas. Dedicated circuit designs with reduced distortion. Higher fidelity transmitters to generate higher constellation complexities with lower EVM and lower spectral regrowth. Circuit design capable of shifting operations to a variety of bands while maintaining fidelity and efficiency. Continuous improvement in Wide Bandgap (WBG) FET Driver Technology.
Determining Macro vs. Small Cell Size/Needs vs. Freq.	As shown in Section 5.3.2, a densification of cells leads to an optimal distribution of coverage, where band selection depends on density of users and performance requirements. Massive MIMO can be leveraged in a denser environment, where not only lower power is utilized and better directed to users, but also more aggressive power management (cell power-down) can be more effectively implemented as the load varies.
Support for Dynamic Spectrum Access	High levels of integration in the baseband processing and flexible cell architectures supporting heterogeneous integration and intelligent provisioning. Hybrid RISC/FPGA/ASIC baseband implementations for balance of flexibility, speed, and power consumption. Small cell RF transmitter technologies that are full octave to multi-octave tunable at the full required transmit power. Multi-octave efficient antenna designs.

Heatsink/Package Size Becomes Impractical for PAE	Utilization of Massive MIMO and high cell density to reduce individual RFPA Power and facilitate use of highly integrated solutions developed for Mobile Devices. Hybrid integration of semiconductor technologies on advanced packaging for enhanced signal integrity, power delivery and heat management. For high performance computing and low-voltage domains: chiplet approach with silicon or advanced substrate interconnect and embedded power management with integrated passive components, utilizing WoW and 2.5/3D package technologies. For high and medium voltage conversion: optimization of intermediate voltage bus and hybrid integration of III-V semiconductor technologies in multi-die packages with enhanced thermal characteristics.
Mid-term Challenges: 2026-2027	*Potential Solutions to Mid-term Challenges*
Power/Energy Telemetry Data Acquisition	Real time measurement of energy utilization. Efforts already under way (see Section 4.1.4 – Subsection Model Complexity), through dynamically integrated executing models with real-time (as well as archival) data resulting from measurements will address the challenges for optimized operation of the system under variable over time conditions (resource availability and demand), and will cognizantly (model-driven) collect needed operational data (resources and demand).
Power/Energy Data Analytics	Evaluation of dynamic energy data to predict performance. Again, the referenced efforts (see Section 4.1.4 – Subsection Model Complexity) will provide the analytics - "system analytics", not only "data analytics").
Defining Energy-optimal Control Feedback Loop(s)	HW-level optimization to minimize energy expenditure, based on real time requirements, including band selection and Massive MIMO configuration
Long-term Challenges: 2028-2032	*Potential Solutions to Long-term Challenges*
Enabling/Deploying Energy-optimal Control Feedback Loop(s)	Optimization of Real-time Cell Energy Consumption, based on predictive models and coordination of system-level resources, including band selection, dynamic cell loading/balancing, and computing resources allocation/redistribution across cloud and edge to support best possible compromise among data-transfer/data-compute power and availability/latency/level of service

5.5. Economic Factors – Need #4

To have a true appreciation for the challenges associated with optimizing a 5G&B network for EE, one must have a very good understanding of what drives costs. From a first-order perspective, this relates primarily to CAPEX and all that feeds into the architecture and deployment of the network. Moving out to the second order there is the OPEX to run the network. By the third order, the focus goes to the applications and supply chains touched by the network and the functionality it provides.

Cellular Base Station Costs

There are a number of costs in the installation, operation, and maintenance of a cellular telephone base station. The costs depend on the cost metric:

Metric 1: bits/joule. This metric takes the total number of information bits, both sent and received, and divides it by the total amount of energy needed to deliver these bits. This is an energy centric metric.

Metric 2: bits/$. This metric takes the total number of information bits, both sent and received, and divides by the total number of dollars needed to deliver these bits. This is an economic metric.

These two metrics are somewhat related.

bits/$ = 1/(joules/bit * $/joule + $(fixed))

This allows for accounting for a lot of fixed costs that are needed in order to establish a base station.

Using Metric 1, the contributing drains on the energy supply are as follows:

All of these are running in some electronics somewhere in the base station and are consuming power.

- Antenna(s), if active, such as a MIMO array or phased array with active elements

- Transmitter(s), if not part of the antenna, such as a standard cell tower configuration
- Receiver(s)
- Clock reference including GPS
- Baseband modem, both transmit and receive, including A/Ds.
- Software/firmware managing the radio link(s)
- Software/firmware managing the health and well being of the cell site
- External back haul interface electronics such as point to point microwave, fiber, cable, or telephone
- Software/firmware to manage back haul
- Air conditioning/cooling (if needed)
- Power module to create/distribute the power needed to run all of the electronics
- Power management software/firmware to manage power source. Includes mains, solar, external motor-generator, etc.

Using Metric 2, the additional dollar costs are as follows:

- Site purchase/rent/lease including permits, taxes, and site survey to determine suitability and connection to mains, if needed
- Continued taxes and site maintenance
- Tower and other facilities installation including cables
- Hardware cost for radio electronics
- Hardware costs for cooling
- Hardware costs for back haul
- Hardware costs for power management including back up sources such as solar and/or motor-generator
- Continued service for maintenance and failure
- Probably a bunch of other items.

5.5.1. Challenges

If we now are promoting small cells as a “holy grail” for EE, and also bringing in a lot of business and strategy aspects in addition to the pure technology aspects, we must address the most important piece, site acquisition. HetNets and small cells have been on the table for almost 15 years now, but still have not been adopted on a larger scale. One reason is that the macro networks still (and for a foreseeable future) are able to deliver what is needed, but another is site acquisition. This is an enormous inhibitor and cost for operators (in terms of CAPEX or OPEX in the case of site leasing), and a big reason why they prefer to continue building on their existing, macro network sites. In addition to costs, regulatory

obstacles can be an impediment to site acquisition. Without that obstacle addressed, ubiquitous small cells face a very challenging path to adoption.

Table 8. Challenges Associated with "NEED #4 - Economic Factors"

Near-term Challenges: 2022-2025	*Description*
Defining the 5G Economic Gap (5GEcG)	**LITERAL DEFINITION** = a hypothetical representation of the disparity between available power a system can deliver and the increasing load demands on its outputs, which means a power-limited system and/or network component will not be able to utilize all its designed potential and therefore be inhibited from delivering on the calculated economics of the payback period. - In practice, this is about identifying true, functional limitations for each network contributor as put in the context of power/thermal limitations. - This can be fairly straightforward in HW. For instance, a base station radio may be thermally limited by the maximum amount of dissipated power its packaging was designed to handle and therefore caps energy consumption at this threshold to prevent damage to the equipment, but this results in a forced limitation in functional output so it is easier to characterize this network constituent due to the direct relation of radio functionality to QoS/QoE. - In SW or more abstract constituents (like Standards Bodies or Regulatory Agencies), assessing the 5GEcG may not be so obvious. A government or municipality may pass regulation limiting RF energy and/or acoustic noise levels within proximity to others that force reduced functionality of a deployed network solution, but a bit more effort may be required to fully assess this impact in terms of the universal currency.
Defining the 5G Derate Factor (5GDF)	**LITERAL DEFINITION** = a unitless coefficient (<1) representing a scaling factor for the application of technical and economic risk factors to the ideal 5G network deployment model that will reduce the optimal, maximum designed capabilities of a network due to energy-limited (5GEG) and/or economically-limited (5GEcG) and/or socioeconomically-limited (5GEqG) factors. In practice, this is a catch-all to assimilate the culminated result of chaining many black boxes together, after having translated into a universal currency, and optimizing in both static and dynamic scenarios. - While 5GDF has the potential to be a very powerful metric for the universal assessment of network EE performance, there are numerous challenges to overcome to turn this from a concept to a resource. Essentially, one must mix a lot of objective and subjective and/or empirical and/or anecdotal data points to ultimately determine a single figure. Of course, this can be done (as any good Curve Fitter will tell you), but doing so in a qualitative manner that provides universal value in a common language is far easier said than done.
Strict Payback Period Targets Driving Socioeconomic Disparity	**5G Equality Gap (5GEqG)** = a hypothetical representation of the socioeconomic disparity between those that will be able to adapt their infrastructure and end use cases to unanticipated underperformance due to energy-limited (5GEG) and/or economically-limited (5GEcG) factors, and those that will not have the resources to be flexible enough to do so. Payback calculations may consider certain usage models to determine economic viability of an investment/deployment, but there may still be gaps in the assumption of socioeconomic factors that must also fall into place to manage the risk these factors can have on subscription models, utilization, and ultimately financial justification of existence in the first place.
Characterizing Individual Components (specific yet compatible)	In order to fully develop and analyze the kinds of PVCs necessary for a 5G&B network and enable the use of the SoS framework proposed in this document, one must fully characterize the constituents in terms of the "universal currency" of energy. This is far easier done with some network components than others. For instance, it is much quicker and clearer to characterize the energy consumption of a single radio than it is to characterize the EE performance of network-slicing SW.
Mid-term Challenges: 2026-2027	*Description*
Cooperation Between Network Stakeholders	Succeeding as a group (as is necessary for network-wide EE optimization) requires an unprecedented level of cooperation by all the stakeholders involved. Further compounding this challenge is an industry that is historically isolated from itself so that the users are accustomed to keeping their best practices and data to themselves, perhaps even contributing to their "secret sauce" and differentiating them in their slice of the market. The only way to incentivize stakeholders to break down these barriers is to either show them the risks are too great to ignore or the economic advantages too great to justify the old way of conducting business.

Model Complexity	Modeling a network is a highly complex and massive undertaking, which is why such work is traditionally sporadic and not comprehensive. Once through the inherent inhibitor of stakeholder cooperation (described in previous line), a lot of assessment, characterization, concept/project definition, semantics, and even basic approach to a modeling framework needs to be considered and mostly agreed upon to make forward progress.
Model Validation	Even assuming one passes through all gates to define and generate a model that a sufficient number of stakeholders can agree upon and discuss in a common language, it must still be validated. Validating a network-level model can range from challenging to essentially impossible for a full-scale, real-time implementation. The framework for defining the model should take validation into consideration from Day 1 and capture the CHALLENGES and other HW/SW hooks necessary to implement partial validation in a way that can be confidently scaled across the network and rolled up into a larger PVC.
Long-term Challenges: 2028-2032	*Description*
Cooperation Between Network Stakeholders	With a framework such as the SoS identified and deployed, the longer-term challenge comes with getting users to utilize the system in a way that perpetuates its growth and supports a perpetual roadmap via standardization, data repository, analytic tools, and expanding application to additional use cases.
Model Complexity	The longer-term outlook for a comprehensive model is to continue to add to the component models to combine with others for increasingly complex PVCs, continue to drive models for lower and lower levels of detail thus achieving finer levels of network granularity for analysis. The complexity of how the model is applied can expand laterally to applications beyond the original intention. Assuming the initial model is proven successful and useful over time, it can lend itself to other applications and EE challenges to extend value and learning to other industries/markets.
Model Validation	Model validation is a perpetual quest in the pursuit of higher model confidence by comparing to an increasing number of data sets with a decreasing amount of error between the simulated and measured values. Considering PVCs as large as those contained within the 5G&B SoS, one can always continue to drive this confidence with measured data on smaller PVCs, then combine with ever-increasing PVCs to validate at increasing scales until one feels confident in the results of a PVC analyzed from the switching cycle of a single power converter up through impact to global economics and even policy.
Socioeconomic Considerations	It is hard to imagine a world in which the 5GEqG is taken as seriously as things like the 5GEG/5GEcG or even 5GDF so the ultimate challenge is really getting economic stakeholders to put an equal weighting to socioeconomic factors as they do purely economic ones.

5.5.1.1. Challenges in the Optimization of Energy Use and Environmental/Financial Impact of 5G&B Network Infrastructure

The optimization efforts conducted thus far [94] address the problem from the perspective within each cited aspect, but not, for example, of the entire end-to-end network infrastructure, such as minimizing BS energy consumption, while at the same time ensuring or improving QoS. Moreover, the methods used in the past (such as for 4G) for optimization within each one of these aspects (even dynamic switching methods for BS sleep on & off) use oversimplified models and assumptions. For example, they assume uniform traffic distribution and arrival pattern in all cells, as well as other related simplifications and parametrizations constructed “by trial and error," but not by other factors such as traffic load. In fact, the activity levels of both base stations and associated mobile users are limited to only two, i.e., active or sleeping (inactive) in most works. A smart phone today transmitting at different data rates would consume different amounts of energy (not to mention much more sophisticated base stations). Other examples of inadequacies in the present methods include those involved in the deployment and management of hierarchical BS configurations. On one hand, one can reduce energy consumption in each of these layers by utilizing higher frequencies to increase data rates, but on the other hand, additional factors such as overheads of transmission and additional radio interference may decrease performance and QoS.

5.5.2. Potential Solutions

Pragmatic, proven solutions to the multifaceted and complicated nature of payback calculations in the massive CAPEX deployments and incredible OPEX budgets has been one of the most challenging and sought-after aspects of cellular deployments from 1G through 5G&B. While there is nothing new to that story, many of the solutions presented in this work take a fresh and highly underappreciated approach of trying to frame completely in the context of EE and energy utilization. Even more importantly are the novel solutions provided by the SoS framework and the harmonization of traditionally heterogeneous, siloed economic calculation methodologies. Furthermore, these new approaches are presented in a way to generate mutually beneficial momentum for all stakeholders in the very extensive 5G&B ecosystem.

Table 9. Potential Solutions to Address "NEED #4 - Economic Factors"

Near-term Challenges: 2022-2025	*Potential Solutions to Near-Term Challenges*
Addressing the 5G Economic Gap (5GEcG)	Applying the 5G Economic Gap Analysis (5GEcG) - While maximal power/thermal ratings are considered at the system level in HW design, the impact on these physical, system-level limitations are not typically taken into context when implemented as a network constituent. Adding this new layer of analysis has the potential of being very enabling to larger PVC analyses.
Adopting & Utilizing the 5G Derate Factor (5GDF)	Applying the 5G Derate Factor (5GDF) - If successfully implemented as hypothesized, then 5GDF has absolutely tremendous potential to streamline previously near-impossible economic payback analyses in a way that clearly communicates gaps/opportunities to payback calculations with commonality that can be internalized by any stakeholder in the PVC.
Strict Payback Period Targets Driving Socioeconomic Disparity	Considering the 5G Equality Gap (5GEqG) - The gatekeeping stakeholders of the 5G&B ecosystem have a moral and ethical obligation to consider the socioeconomic disparities that exist today and how they can be addressed by the deployment of 5G&B technologies. There may be industrial winners and losers, but a lot of CAPEX will be funded by nation states, which means they should serve all levels of society. - From the growth of the Gig Economy to Telemedicine to the evolutionary, quality of life changes connectivity brings to the unconnected/underconnected, there is tremendous economic incentive to address these socioeconomic factors even for those seemingly uninterested in ethical, altruistic behaviors.
Characterizing Individual Components (specific yet compatible)	Applying the Systems-of-Systems (SoS) Black Box Analysis - The global, SoS block diagram contains a whole lot of black boxes! Luckily, the proposed framework has been designed to allow for incremental additions of individual black boxes that only serve to increase the overall value of the framework (and associated analyses) with each addition. - This WG has put most of its initial focus on the RAN and other network constituents toward the edge of the network since that is where a grand majority of our SME is rooted. The more we can "fill in the blanks" of the global SoS, the more complex the PVCs generated and therefore, the more fruitful and heterogeneous analyses can be performed. - Growing the SME pool is critical to enabling the SoS framework to enhance value and utilization now and sustainably into the future.
Mid-term Challenges: 2026-2027	*Potential Solutions to Mid-term Challenges*
Cooperation Between Network Stakeholders	Disaggregated Network Metric(s) - One could argue there are as many metrics as there are contributors in the 5G&B ecosystem so trying to identify simple ones that are comprehensive enough to be useful, yet simple enough to enable true commonality is very important to addressing historical challenges in this very legacy industry. Mixed History of Stakeholder Collaboration

	- It is unrealistic to immediately expect a lot of industry stakeholders accustomed to working in isolation (whether intentional or by circumstance) for many years to suddenly all decide to fully collaborate for the common, sole goal of driving EE. That is not to say there has not been plenty of industry consortiums, standards, and other collaborative efforts in the past and growing today, but nearly all with different focuses. Speaking of reality, many of the concepts/metrics/proposals in this work have yet to be fully proven in their value to justify all the efforts and resources involved. Assuming some or all of this content is eventually adopted (or at least considered for adoption) by major, industry stakeholders, it will take time to bring them to the table, enable them to capture their concerns/objectives, and negotiate the most pragmatic methodologies to open the doors to actionable change.
Model Complexity	It is no small feat to model portions of a network let alone having all those portions interact in a cohesive model at the macro scale. Even the semantics around what is considered a truly macro-scale model can be complex if it is to consider a regional network, incorporate the full energy generation/distribution aspects, and be enabled to perform assessments all the way up to the absolute, global level. As our models (particularly in the context of the SoS) move up the hierarchy from the lowest levels of granularity to the highest levels of the global energy markets, analyses that previously seemed impossible gradually become a reality over time with the additive contributions of more stakeholders, data notes, and analytic tools.
Model Validation	Demonstration of Energy & Total Cost of Ownership (TCO) Savings - A model has very little value outside hypothetical discussions if it cannot be validated. The increasingly complex models only serve a purpose and provide value if they can be validated. In this context, validation shall come in the form of demonstrated EE and increased ROI in the form of TCO savings. Given the longer timelines associated with a cellular generational deployment, it can take a minimum of several years to collect and assess enough data to comprehensively validate proposed models.
Long-term Challenges: 2028-2032	*Potential Solutions to Long-term Challenges*
Cooperation Between Network Stakeholders	Demonstration of Ability to Optimize Energy Utilization from Micro to Macro Levels - A steady state of EE utilization shall be accomplished when the payback period of any constituent in a 5G&B PVC can be accurately and confidently modeled against the impact of any other. In other words, long-term success shall be achieved when a change in the characteristics of a single black box at a very low level of granularity (i.e., a radio at the edge, daily change in regional energy pricing, etc.) can be taken into consideration for the payback period associated with a black box at the highest levels of granularity (i.e., a power plant, regulations/legislation, global carbon emissions, etc.).
Model Complexity	A full, global-scale realization of PVCs in the SoS is the ultimate measure of success of the proposed framework.
Model Validation	Validated Model Database - Eventually, enough data shall be captured, analyzed by data scientists, and validated to justify repositories to streamline the ability to share data/learnings in the hopes of determining new best practices and paying them forward from one network generation (and its associated stakeholder knowledge pool) to the next.
Socioeconomic Considerations	The true value of this work will be the facilitation of harmonic equilibrium between the [pragmatic] economic payback periods for the CAPEX to achieve ROI profitability and an enhanced quality of life accessible to all members of a given society. Furthermore, a great measure of success will be the justification of investment in under-served regions (i.e., non-OECD countries) traditionally considered “uninvestable.”

5.5.2.1. Solutions for the Optimization of Energy Use and Environmental/Financial Impact of 5G&B Network Infrastructure

All these and other drawbacks call for more systematic methods that allow cognizant decision making under i.) dynamically changing conditions of resources in the network infrastructure; ii.) time-varying needs of the applications and users supported; and iii.) multi-objective optimizations across all these diverse resources and constraints. Works [95], referenced in the Applications and Services WG chapter of the INGR [96], have developed dynamic and multi-objective optimization methods and end-to-end or SoS approaches, encompassing an advanced Digital-Twin implementation (“DDDAS-based Digital-Twin” [97]) for the management of electric powergrid infrastructures. This includes renewable (variable) energy sources and support of multiple users with multiple levels of priorities. The referenced

methods enable support of advanced automated control systems and in this case, self-healing power grids along with the need to harvest and adaptively distribute and deliver energy in a secure, reliable, and low-cost fashion. The referenced work enables self-healing microgrid infrastructure, which leverages traditional (coal, hydroelectric, nuclear) power sources for reliable electricity distribution. As discussed in [98] [99], these methods (which have been applied to microgrids, powergrids, as well as to enterprise resource planning systems, cybersecurity, and other areas) benefit from the advanced capabilities promised by the 5G&B network infrastructure. Additionally, these dynamic, multi-objective optimization methods can be applied [100] to support the networks themselves and address heterogeneous, multi-level networking (inclusive of communication and computing capabilities) with an SoS methodology.

The SoS construct is very powerful for company-wide simulation architecture and coordinating computing resources with large scale simulations. The details of the simulation are going to be very specific and proprietary to the organization. What we are advocating is folding in calculations for energy generation/use and power dissipation (at all levels and an additional high-level energy use summary) to provide energy insights into the overall system simulation.

5.6. Grid/Utility – Need #5

Energy Grid Implications of 5G Deployment

For the last two decades, we have seen repeated exaggeration of the amount of energy use required for data centers, which caused needless alarm, but also obscured the very real and large amounts of energy required. Luckily, a series of studies (e.g. [101]; some supported by DOE) has brought real data and credible estimates to the table, as well as insight about underlying drivers and efficiency opportunities [102]. Data centers do use large amounts of energy, but the designers of their IT hardware and of the facility power and cooling infrastructure have worked hard to make dramatic progress in the efficiencies of both domains. The result has been to shift from large annual increases in the total through about 2007, with only very modest increases since that time.

A similar trajectory unfolded with IP network equipment (not including mobile infrastructure), including studies to document aggregate consumption and savings opportunities [103]. In addition, key technologies were developed to reduce network equipment energy use, such as Energy Efficient Ethernet [104].

Just over 1% of global electricity demand comes from fixed and mobile network infrastructure, according to a recent study [105]. While 1 % may sound small, 1 % of a very large number is still a large number.

For any network, energy is used for the infrastructure devices as well as for the endpoint equipment connected to it. For IP networks, the networked equipment (devices like computers) uses much more energy than switches and routers. For example, an earlier estimate for all electronics found it consuming about 10 % of U.S. electricity use, with data center electronics covering about 10 % of this (for 1 % of the total) and IP network equipment about half of the data center total (none of these figures including power and cooling infrastructure for relevant facilities). While much of the rest of electronic devices was not networked at that time (e.g., almost all TVs), the fraction of energy use of electronics that is networked is ever rising.

For devices that are connected to a network, it is common for this to increase their energy use, for the hardware that implements the network interface, for extra computation required to participate in the

network, and in the common result to stay in a higher power state than otherwise due to the network connectivity (e.g., on rather than asleep, or asleep rather than off).

When devices are powered by batteries or through energy harvesting, there is a huge incentive to optimize both the communications and the rest of the device for efficiency. However, when devices are mains powered—and the vast majority of energy use of networked devices is in this category—the effort on efficiency often drops off dramatically. The combination of the dramatic rise in wirelessly connected devices with a major migration to heterogeneous networks of small cells in 5G has the potential to shift this energy balance towards the edge of the network. Many barriers to efficiency are systemic rather than under the control of a single designer or product manufacturer, requiring collective action to produce them, e.g., through technology standards and public policy. A lack of visibility and understanding of energy demand requirements from the complete, end-to-end network power value chain (from power plant to wireless) abstracts most stakeholders from the realization they play an active role in global EE.

Current Activities

Some efforts are being made to focus on the intersection of energy and 5G communications. Notably, the GreenTouch Project (2010-2015) investigated how much more energy efficient networks must be by 2020 to be viable [106]; there was a workshop in 2018 on "IEEE 5G Energy Efficiency Tutorial" [107]; and there is an Energy Efficiency Working Group component (this work) to the IEEE International Network Generations Roadmap (INGR) [108]. There is likely attention to energy issues within the 5G community, both for infrastructure equipment and for end-point devices, but history has shown that industry by itself is not sufficient to produce clear analysis and reveal all important efficiency opportunities.

5.6.1. Challenges

5.6.1.1. The 5GEG & Overall Risk 5G&B Applications Pose to the Utility Grid (Smart or Otherwise)

As first reviewed in Section 3.0, the 5GEG has outlined the relationship between load demand and resource availability in the 5G&B network and identified a gap between these two as a risk to a grid's ability to deliver requested energy sufficiently and reliably. Given this section covers more of the roadmap content aspect of the document, we shall now explore the risk posed by the 5GEG in greater detail. As there are many risks, they are articulated here in terms of projected impacts in the near, mid, and long terms.

In the near term from today through the next few years (2020-2023), the risk posed by the 5G network currently under deployment is more of an initial litmus test. As legacy base stations are upgraded to 5G radios, their local utility service will be enabled to immediately start assessing impact by the change in individual, base station power draw to see if energy footprints are matching predictions. The aggregated impact may still take more time to become apparent, so this stage is far more about initial turn on and ability to meet deployment (e.g., primarily CAPEX) targets. Once these initial deployments are activated, the assessment of payback period (e.g., primarily OPEX) targets is enabled. Our world is already in the midst of massive deployment of connected devices and smart "things" so the true impact and risk posed by the exponential growth of the IoT and their use in emerging applications predicted to increase energy density (i.e., edge automation, TinyML, etc.) will quickly become apparent.

The major increase of EVs alone (regardless of connectivity) has the potential to completely disrupt today's utility grid economics. The famous "duck curve" was published in 2013 to predict the impact to electricity demand with large-scale deployment of renewables (namely PV) [109]. This curve was intended to highlight the disparity (or energy gap) the State of CA faced due to massive deployment of intermittent, renewable energy sources, which worsens over time. On a daily basis, the worst-case for generation coincides with the greatest increase in demand as people return home from work toward the evening when generation is lowest and demand is greatest. Now just imagine if EVs penetrated a majority of the automotive market instead of the extreme minority they represent today. The duck's head would come clean off its body!

Moving into the mid term period of 2023-2025, it should be very clear if the risk hypothesized by the 5GEG will come to fruition. Regardless of whether or not this occurs within this timeframe, there is a second wave of risk that will align with the production utilization of mmWave transceivers enabling access to bandwidth in the 6 GHz and well-beyond bandwidth. Section 3.2 details the justification for this assumption based on RF physics and advancements in radio electronics. This second wave will further be exacerbated by the production deployment of near- or full-automated vehicles necessitating an extensive increase in networking connectivity and therefore energy to sustain such an increase. While much of the "promise of 5G" hinges on the ability to dynamically-access bandwidth into the mmWave, it is critical to NOTE implementation of mmWave transceiving elements dissipate large amounts of heat and it is widely accepted by the experienced and knowledgeable group of people contributing to this WG that the packaging and thermal solutions needed to make mmWave technologies viable for 5G are neither on the current path to fruition nor is their path correction obvious.

From an economic perspective, the anticipated migration of a major portion of the workforce to the emerging, gig economy is a key item of note. The basic concept of going from an ownership model to a service/lease model will transition traditional business models into completely service-based models, even for those historically dictating a massive CAPEX/infrastructure investment. The change to the economic models around some of the largest items in terms of consumer spend and energy consumption in the transportation sector must be taken into account and given unprecedented emphasis on impact to the use case. Reliability and warranty predictions will need to adjust to usage models once considered outliers.

Looking to the future period of 2025-2030 as well as being on the cusp of the next-generation (presumably 6G) deployment, it is expected to not only have massive antenna arrays support Massive MIMO techniques, but they are also required to support mmWave well into the 10s of GHz in order to meet the "promise of 5G" as articulated today. One would think there is a high degree of risk associated with this need given the power and thermal challenges of these solutions individually, which are absolutely compounded by each other when combined.

It is difficult to have any discussion about risk to the utility grid and not make mention of security. This point is particularly salient here, where we propose the enhanced connectivity and intelligence of a forward-looking Smart Grid. With great connections comes great responsibility so while all this enhanced, bidirectional control and feedback provide an absolute wealth of EE enhancement capability, each and every feature is also a potential risk and security point of failure. Regardless, a deep dive on grid susceptibility and solutions for addressing these security shortcomings are not within the context of this document. It is this WG's goal to provide the vision, risks, and solutions enabling a grid that can self-optimize for EE and other stakeholders' contribution to interpret these things and opine on how to secure such a Smart Grid. It should be noted this INGR effort contains a Security WG [110] that can be referenced accordingly.

5.6.1.2. The Role of the Utility Grid in 5G&B

The utility grid plays a pivotal role in the deployment of any network, which may seem like a rhetorical point, but this takes on a completely new meaning in 5G&B networks. Naturally, nothing in the electronic world works without power so that fundamental function of the grid goes without saying. While more second order, we have discussed how the way energy grids are architected, deployed, and utilized can play a critical role in the economic viability of a network. As we breakdown networks into PVCs and apply all the SoS/5GDF analysis, it quickly starts to become clear how a once benign grid that merely delivered power is now a critical stakeholder as the economic and reliable arbiter of energy.

A grid's number one priority is still to deliver energy from source to load within the specified constraints that define it. Delivering this energy reliably is implied so anything that poses a risk should be a serious consideration for any stakeholder, whether they be closer to the source/distribution or the load/end user side of the spectrum. Furthermore, the far majority of deployed electrical grids (at least in the US) are very old and were originally designed for unidirectional power transmission that predated ubiquitous deployment of commercial and residential renewable generation, thus forcing it to transform into a bidirectional grid. This can cause stability concerns as was observed by Germany's Energiewende initiative in the early 2000s with a major, government-sponsored acceleration of solar PV deployment [111]. Here is a good place to note and remind all those stakeholders beholden to grid performance transcend 5G&B as even someone that has never heard of 5G or even a cell phone likely has utility grid service to their home (particularly in OECD countries). This means there will be very high-profile attention on any grid stability concerns caused by 5G&B and potentially public outcry if citizens are losing their power, whether directly driving 5G&B interests or not. As outlined by the 5GEqG (See Section 3.x), this not only bridges the digital divide for the unconnected/underconnected, but also brings to light the disproportionate impact on general quality of life effects 5G&B networks may hold in their hands.

As we assess where we are now and look to the future, there are few better applications of the "smart" moniker than to the Smart Grid. Like most "smart" applications, this can take on any number of meanings and levels of integration, communication, and control. Given the incredible amount of long-term CAPEX and OPEX involved in deploying a utility grid, we must recognize that upgrades, enhancements, and generational improvements are to be carefully considered and strategized with inputs from a high number of stakeholders. This roadmap documenting process is ideal for such a comprehensive, multifaceted, long-term need since iterative steps can be broken out and explored in more manageable bits across near-, mid-, and long-term timelines.

Adding communications (e.g., data) to power transmission is neither a novel nor emerging consideration. Looking all the way back to the implementation of the first, ubiquitous telephone exchange networks, information and power have gone hand-in-hand and this original symbiosis of data and power still exists today (nominal -48 Vdc remote line power) for those still having analog, landline phone service. As we moved into the digital data era and saw an explosion of system peripheral device support, which led to the Universal Serial Bus (USB), which has also delivered power (increasing from few watts early on to ~100 W today). Power over Ethernet (PoE) is another example of concurrent power/data over long distances. So how can we apply these intelligence, communication, and power management capabilities to create a Smart Grid?

5.6.1.3. Disaggregation of the Utility Grid

Before we can determine what HW/SW hooks should be considered to implement a Smart Grid, we need to have a good understanding of the potential usage scenarios and application enhancements such a thing would bring. In order to align with modern and forward-looking needs, we need to identify the trends in load demand and the energy market. A key trend impacting both load and source sides of transmission is disaggregation. From the disaggregation of the grid to the disaggregation of energy storage, this trend has found its way out of the core data center and up to the global network/utility level [112] [113] [114]. While the justifications for these disaggregating trends can be related to network performance and throughput (e.g., data efficiency), they are often focused on EE (directly or indirectly, even if not considered a priority). Energy storage may be disaggregated and modularized to specific systems for technical (i.e., backup/holdup, peak shaving, etc.) and economic reasons (i.e., store/sell expensive power, utilize cheap power). Whatever the reason, nearly all these techniques should lead to an overall reduction in energy-related CAPEX and OPEX.

This concept of disaggregation of the legacy, centralized grid on the path to the fully bidirectional Smart Grid is ever present when considering Distributed Energy Resources (DER) [115]. DERs are sporadic energy sources spread throughout the grid, often closer to the point of consumption. Common examples of DERs include PV arrays, wind generators, fuel cells, and other kinds of (typically renewable) modular generation such as generators and any number of energy storage systems.

Looking to the future, it is our hope the legacy grid of yesteryear will continue to be upgraded and adapt to become the Smart Grid of tomorrow. Furthermore, we ideally imagine a Smart Grid designed to self optimize for EE, particularly for 5G&B applications. This is much easier said than put into practice, so these desires need to be articulated and proposed in a pragmatic sense aligning with usage/economic factors, backwards compatibility with legacy grids, and within the very long timelines (measured in decades not years) associated with advancement in power engineering and generational grid enhancement. A key, intermediary step will be to apply communication/networking/telemetry capabilities outfitted to existing grid constituents. This is done in bits and pieces today though typically done for selfish reasons such as a data center shifting loads within a locale if power capped and needing to borrow from one rack to power another.

A microgrid is a fully contained grid of DER and load designed to be fully self sustaining. It can be connected to a larger grid or completely isolated (a.k.a. - islanding). As one may imagine, microgrids can be more predictable for managing load demand impact to the grid and accomplishing security since management is focused on a smaller grid than a full, utility-scale deployment. The flip side is they are typically sourced with intermittent and unpredictable energy sources (i.e., sunlight, wind, etc.) and have a higher dependence on localized energy storage than a more traditional, off-grid solution. More importantly, microgrids are smaller than their full-sized counterparts and therefore provide enhanced options for EE optimization in alignment with the trend of smaller system size relating to higher levels of control and automation.

5.6.1.4. Energy Storage

Energy storage is another area of interest that has the potential to have a major, enabling impact on the success of a next-generation Smart Grid. Applications of energy storage, and even just what is implied by energy storage, can vary greatly. All energy storage is not the same. Seems like a silly, obvious thing to state, but it is common for people to barely discern between storage classified as batteries vs. capacitors, let alone taking into account the many different chemistries and physics variations dictating

the application (and likely economics of use as well). At the modular scale, the appearance of a simple, two-terminal, (especially dc) device can be quite deceiving. It is even common for experienced power designers to oversimplify energy storage and lack the appreciation required for appropriate use and economic payback assessment.

Just take batteries for instance, which are actually quite complicated devices when one considers the nuances of operational/environmental factors dictating performance. Such a list may include charge rate, discharge rate, depth of charge, state of charge, cycle life, cycle depth, number of cycles, internal equivalent series resistance or ESR, output impedance matching, an appropriate arrangement of cells in series and/or parallel, cell-to-cell balancing requirements, pack-to-pack balancing requirements, thermal management (internal and external), over-/undervoltage protection or OVP/UVP, overcurrent protection or OCP, quality/reliability impacts, mechanical factors (i.e. – rigid vs. flexible cells), and more. NOW, take into consideration all these factors are different for just about each and every unique chemistry (i.e. – sealed lead acid or SLA, Li-Ion/Li-polymer, nickel-metal hydride or NiMH, nickel-cadmium or NiCad, lithium-iron-phosphate or LiFePO4 or LFP, zinc-manganese dioxide or Alkaline, and many more). NOW, take into consideration the need for a non-rechargeable (a.k.a. – primary) or rechargeable (a.k.a. – secondary) solution. NOW, consider NOT taking all these things into consideration may be as benign as slightly degraded life not meeting the application performance/economic targets or as catastrophic as a localized, negative thermal event.

One should note the development cycles for new energy storage technologies can be quite long and incremental when compared to many of the loads they serve. For instance, there is no Moore's Law to energy storage as it is confined by the bounds of chemistry and physics more so than process, so it will not shrink in size and grow in energy density like silicon gates on an IC. This not only generates a gap in emerging technologies between the rates of advancement for source vs. load, but can create a false perception there is a need for energy storage to exponentially advance to keep up with the load. Many solution providers feel they can create whatever power budget is necessary to meet their needs, then organically (perhaps even magically) have a power source capable of meeting those needs even when this is (often) not the case. In other words, it is far more pragmatic and realistic (and quicker and cheaper) to put the maniacal focus into reducing the system power budget (and/or utilizing more intelligent power management techniques) before putting the human energy into looking for a bigger battery.

Table 10. Challenges Associated with "NEED #5 - Grid/Utility"

Near-term Challenges: 2022-2025	*Description*
Data Distribution	Today's network is a horribly complicated PVC of siloed stakeholders and is reflective of disconnected nature of data collection and distribution as well. The characterization and assessment of EE in a very large system or network or SoS begins with normalization of the way data is captured and shared.
Communications Medium(s)	Not only must data be captured and shared, but it must be formatted in a manner enabling commonality. The medium(s) used to communicate such shared data must also support this commonality.
Mid-term Challenges: 2026-2027	*Description*
Independently owned Infrastructures	It is difficult to get many stakeholders to adopt the common EE enhancements proposed if they are not incentivized with common motivation. If the DER trend continues without commonality for infrastructure design and deployment, this challenge will only exacerbate the 5GEG, 5GEcG, 5GEqG, and other concerns outlined in this document.
5G HW Power Characterization Standard	As with any high-stakes and highly competitive business landscape in very large marketplaces, there is a tendency for stakeholders to want to insert their own bias when it comes to collecting and reporting data that reflects upon the performance of the HW being assessed.

Grid Fault Response Time	Any great PVC is as strong as its weakest link and this point becomes highly obvious and salient the more complex the grid and the more dependent a user is on its reliability. While we can propose all the best monitoring, optimization, and self-healing techniques in the world, it will not matter if one cannot execute these techniques in a meaningful timeframe. The complex architectures of the grids and networks (again, all different variations of a PVC) necessitates a whole variety of time-base needs defined by adequate grid fault response from the millisecond to infinite time resolutions.
Long-term Challenges: 2028-2032	*Description*
Ecosystem Paradigm Shift	When it comes to utility grid ecosystems, the development timelines can be considered glacial at best. This means any improvement proposals must be distributed over many years, very proactively and methodically strategized. In execution the needs must be eased in with difficulty and expense being inversely proportional to deployment. In other words, the bigger and more auspicious the goal, the more expensive it will be justify, the longer it will take to implement, and the more stakeholder buy-in is required to support.
Physical Layers For Digitally Managed Power	Going from grid intelligence to a self-optimizing Smart Grid implies the full networking of power solutions along with all PVC constituents. A PVC contributor needs to be just as good at sharing information (i.e., telemetry, control, or otherwise) as it is at commutating power. There is a need to apply the learning and success of something like the OSI Model [116] directly to the power-control aspects of PVC contributors for bidirectional transfer of information to truly enable intelligent power management (IPM) techniques across the full, 5G network stack.
Adoption by Energy Storage/Generation Products	The trend of DERs and migration to small cells go heavily hand-in-hand to determine the viability of a 5G&B network. The mitigation of risk from the 5GEG and maximization of 5GDF are closely tied to the ability to optimally locate, allocate, and generate a utilization algorithm for energy storage. These challenges are further compounded by an extensive amount of available energy storage technologies and an appropriate learning curve for how to identify the appropriate storage application, determine/design an appropriate storage medium, and utilize that storage within the technical and economic bounds for which it is intended for maximal EE.

5.6.2. Potential Solutions

5.6.2.1. Powering Options for 5G Infrastructure Equipment

For over a century, powering an electrical device required little more than deciding which of the few AC voltages available were suitable and how much peak current would be required. In recent years, we have added the need and value of incorporating renewable energy sources into our electricity systems, and new ways to use storage to deliver reliable power. Telecommunications equipment has always carved its own path for powering, in using DC for easier integration of batteries and due to the all-electronic nature of the load. This combined with the highly diverse nature of 5G equipment and deployment scenarios means that the topic of powering the equipment holds a great deal of opportunity for innovation and efficiency.

5.6.2.2. Powering Opportunities

At present, telecommunications equipment is being powered from traditional mains power, through the long-standing use of 48 V DC power, and through new DC mechanisms. Off-grid base stations are increasingly powered by a combination of PV and batteries, to reduce use of diesel generation, or to eliminate it entirely.

In many deployment contexts, the costs avoided by not installing a traditional AC mains power connection are substantial, particularly for equipment installed in places other than telecommunications service provider facilities. This is particularly true for equipment installed in public places, and for

mobile equipment. Using renewable generation has operating cost advantages, as well as helping companies meet policy goals.

New powering technologies are continuing to emerge from different ways to harvest energy from the environment, new storage technologies, and innovative ways to interconnect generation, storage, and end-use devices. A particularly promising approach is to network the electricity itself, as is proposed with Local Power Distribution [117] [118]. While telecommunications will be an early adopter of many of these technologies, as they should have diverse applications well beyond communications, developing and using them for 5G will help them achieve that wider use. This can have significant cost, environmental, reliability, and other benefits.

With new powering options, and increasingly for traditional grid-connections, the relative availability of power changes over time, over the course of a day, or more dramatically in anomalous circumstances. Equipment can respond to this availability by altering service delivery, charging or discharging local batteries, or changing the operation of associated equipment that provides other services (e.g., lights on a public pole). A consistent mechanism is needed to orchestrate this coordination; the Local Power Distribution technology proposes a 'local price' (with a non-binding forecast) as the means to accomplish this.

Today, many IP network devices are powered with Ethernet (colloquially Power over Ethernet) which now can deliver up to 90 W with IEEE 802.3bt [119]. Even more promising is the emerging technology of Single Pair Ethernet (SPE) technology, which can provide both power and data over just two wires rather than Ethernet's traditional eight wires, to reduce costs, increase efficiency, and offer more options for length, wire gauge, and the like [120]. In many 5G applications, the SPE link can be used for primary, or backup data backhaul. In others, the SPE will be only used for powering, with data backhaul accomplished through other means such as fiber optic or wireless connections.

Standards for interconnecting power infrastructure to these devices are needed to decouple the power equipment from the telecom equipment so that the two technologies can evolve separately.

5.6.2.3. Applications of the Smart Grid

Moving up to the grid level, this becomes more emerging in application as efforts are few and far between, but still exist. Take something as fundamental as monitoring grid health and fault response. One might think no stakeholder is better enabled to perform this monitoring and respond to faults than the grid owner/operator, but this turns out not to be the case. As a matter of fact, it can be quite common for users with major loads on the grid (such as a cable provider like Time Warner or a network operator like AT&T). A good example of this is a collaborative effort between the cable/broadband and utility companies in which a service provider has far more touch points on a grid for measuring line parameters than the utility. Naturally, they require more points since they are closer to the point of service (i.e., more cable than utility junction boxes in a single neighborhood) and monitor with finer resolution, which enables them to be aware of and react to grid health concerns before the utility [121] [122]. Effectively, the service provider has a high-resolution grid sensor network exceeding that of the actual grid operator and power supplies are the key differentiating factor here since many more power supplies are required to deliver the service to homes than HV line transformers for a given area. While the business model of utilizing this high-resolution grid sensor network is outside the scope of this document, the application is noted because it is an excellent case study in how the grid can migrate from legacy to Smart with intermediary steps utilizing existing infrastructure with the appropriate stakeholder coordination.

Looking even further into the future, an ideal Smart Grid will be able to constantly monitor itself (also promoting faster fault response and self-healing), communicate in any direction in a PVC, and participate in bidirectional control feedback loops that prioritize self-optimization of constituents all for the primary purpose of increasing EE. First-order control may involve reporting of telemetry information (from load to source or vice versa), with PVC constituents independently optimizing for EE based on this information. The next level of control occurs when this reporting of telemetry expands to also provide the time-dependent feedback based on a change that has occurred using the original telemetry info and additional info about the ensuing EE optimization. The highest level of control will occur when there is constant, bidirectional control feedback amongst all PVC constituents to self-optimize based on some common, higher-level initiative such as making adjustments to account for the dynamic cost of real-time energy markets and/or meeting dynamic performance goals (i.e., how to increase 5GDF from 0.7 to 0.9 for a given use case OR how to adapt to a sudden emergency situation). Just imagine if the real-time price of energy on the market changes in intervals of minutes instead of days or hours and what kind of HW/SW/logistics will be required to take advantage of such a scenario [123]. Imagine the EE and economic benefits if a 5G&B network IS capable of taking full advantage!

5.6.2.4. Network Power Integration

A system architecture innovation foundational to Internet technology is the separation of technology details of application layer protocols from the physical layer technologies that they transit over – and vice-versa. This allows the two domains to evolve separately and in parallel, and to mix and match the combinations used. It is not unreasonable to assert that without this, we would not today have the Internet we know.

The Network Power Integration (NPI) Model shown in Figure 25 (below) expands upon the very familiar and ubiquitously deployed OSI network architecture. The NPI architecture provides hierarchical layers for connecting the physical layers of power distribution to digitally manage the flows of power over space and time across a network as shown in Figure 26 (below) [124]. The OSI Model is an obvious template for the Network Power Integration Model because it is so well known, understood, standardized, and integrated into today's WW networks. A key value-add here is a model designed and optimized specifically for the optimization of electricity flow separate from technologies for moving data generally. The kind of data exchanged between powered equipment is typically much less demanding than today's high-speed data networks so such a protocol and supporting, physical transport layers can be optimized for lower bandwidth and higher latency, which in and of itself should consume less energy to accomplish the task compared to higher-speed networks.

Layer 1 of the NPI model is managing the flow of electrons across a single power link – analogous to how Ethernet and Wi-Fi manage the flow of bits and packets across their data links. Layer 2 is the mechanism to move from a link context to a network context – analogous to what the Internet Protocol does for IP networks. This is accomplished with price (and quantity) as the fundamental mechanism, analogous to the Internet Protocol. This is the foundation of Local Power Distribution. Managed DC power distribution is present today in a variety of technologies, most widespread of which are USB and Power over Ethernet. The layer 2 capabilities require only a few more messages than what USB already have, as well as bi-directional power links between power infrastructure devices (nanogrid controllers) [125].

In the context of the SoS applied to 5G&B use cases, the standardization of energy awareness communication is hugely beneficial and important when wanting to apply common metrics such as PCF

and 5GDF. With a dedicated protocol for communicating between power solutions, the ability to characterize individual component PCFs and identify the energy-related bottlenecks in a PVC can be streamlined. The result of this finely tuned, energy-related knowledge sharing is an accelerated path to calculating 5GDF and performing the network-level chain analyses that can achieve our higher-level goals of maximizing EE and economic payback concurrently, while also taking into account critical, functional parameters such as QoE/QoS and RF spectrum optimization.

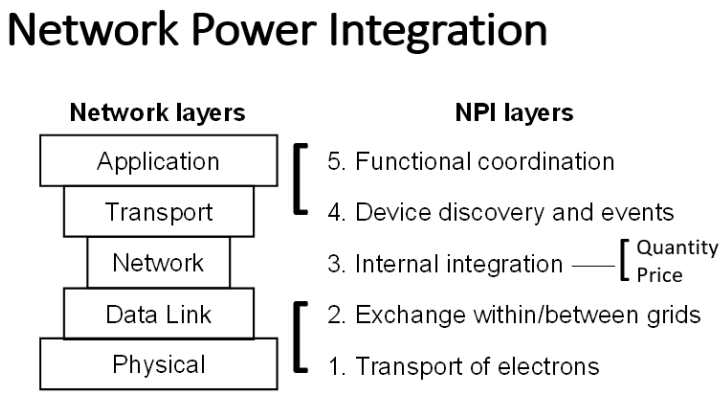


Figure 25. The NPI Model

Image courtesy of Lawrence Berkeley National Laboratory

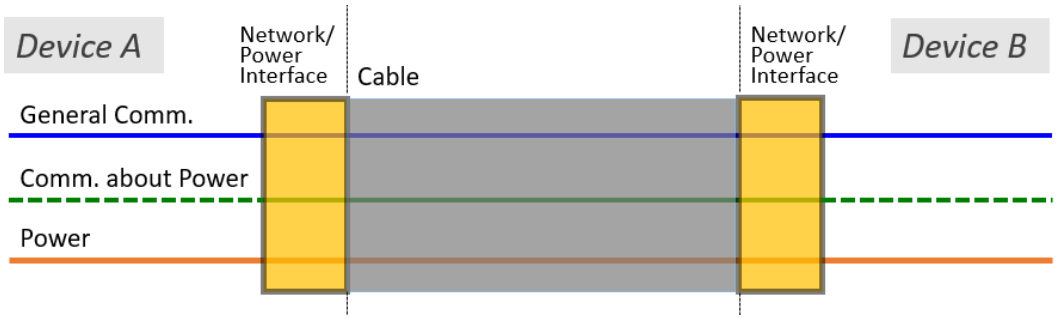


Figure 26. The disaggregation of power, data communications, and communications about power

Image courtesy of Lawrence Berkeley National Laboratory

5.6.2.5. Case Study: A Smarter Grid

An Advanced Research Projects Agency - Energy (ARPA-E) project for Generating Realistic Information for the Development of Distribution and Transmission Algorithms (or "GRID DATA" [126]) is a US DoE-funded effort to provide models and repositories of detailed grid data to enable the kinds of forward-looking analyses and simulations optimistically targeted by this group's efforts. This data is made freely available to those interested in using it for this purpose.

Even working to address systemic issues via developments that evolve on glacial time scales, there are clear signs of progress. Previous group efforts such as the GreenTouch Initiative (mentioned above) and the EARTH project [54] have attempted to address EE concerns by identifying EE improvement requirements projected over a 10-year period and even building a modeling framework for portions of the network based on measured data. More recently, a complete microgrid was built around the DER concept and fully implemented in a small community in the State of CO. The kinds of low-level energy managers, bi-directional grid feedback, and grid-level EE optimization professed by this effort have been demonstrated on a smaller scale. The National Renewable Energy Laboratory (NREL) partnered with ARPA-E to utilize their GRID DATA in a real-world environment in a microgrid smart home demo community, a.k.a. an Autonomous Energy Grid (AEG), in Basalt, CO [127].

One very interesting take away from the AEG experiment is that a successful implementation of AEGs on a larger scale are dependent on a hierarchy of asset control from 10s (i.e., building-level) to 100,000,000s (regional-level). This represents the culmination of many of the needs and challenges summarized in this document in which individual controllers are placed with every network constituent in the PVC and these must be read as well as fed back with control information for its own EE optimization. The thought of accomplishing this with many millions or even hundreds of millions of control points is enough to overwhelm anyone, but the multifaceted values netted from such a system will yield incredible EE benefits for the 5G&B network as well as all the humans and world it touches.

Table 11. Potential Solutions to Address "NEED #5 - Grid/Utility"

Near-Term Challenges: 2022-2025	*Potential Solutions to Near-Term Challenges*
Data Distribution	Leverage Existing Mediums - The lowest hanging fruit for data sharing is to leverage existing mediums. In the near term, this is less of an exercise in data acquisition and more of an exercise in discovery of commonality in data formatting and means of distribution. While the work of this WG has demonstrated how anything being assessed in the SoS (and therefore 5G&B network) can be converted into a common, universal currency, this concept is new and must be adopted by the masses. Bidirectional Feedback Loop - Today's utility grid infrastructure is simply not enabled for feedback loops, let alone bidirectional ones. The information flows in one direction along the PVC and any telemetry making it to other levels is certainly not used to generate feedback and drive behaviors. Feedback Through Aggregate Consumption Metering - While touch points exist today at high levels of granularity in consumption, these touch points are rarely outfitted to collect detailed energy consumption data and certainly not enabled to share it up and down the PVC. In the interim, energy consumption data captured at the point of consumption can still be used to statically assess utilization and generate assumptions to facilitate EE best practices.
Communications Medium(s)	Too many systems with independent monitoring solutions/protocols. Universal communication only comes when there is a universal language and common means of transmission.
Mid-Term Challenges: 2026-2027	*Potential Solutions to Mid-Term Challenges*
Independently owned Infrastructures	Data Sharing - As the emergence of DERs and microgrids become more prevalent, there will be an increase in the numbers of stakeholders and owners of pieces of infrastructure and therefore, the data associated with them. Furthermore, there will be various interests from control to monetization of this data so stakeholders must either be incentivized (i.e., subsidized and/or economic motivation) and/or compelled (i.e., regulation) to contribute the data resulting from their infrastructure to a higher cause.

	Grid Data Analytics/Database - There are already major industry constituents (i.e., broadband companies) and consortiums (i.e., CableLabs) that are collecting, analyzing, and monetizing utility grid data. It should come as no surprise that the companies with the most touchpoints on the grid (particularly as these points increase exponentially the further one gets to the edge) stand to have the highest-resolution data on grid health/utilization and therefore be in the best position to aggregate data. The sky is truly the limit on how such comprehensive databases of grid metrics and analytics can be used, but ideally some of this application will come in the form of feeding data into controllers that can assess and distribute data, optimization guidance, and feedback info for the purpose of maximizing the efficient utilization of energy. - Combining this with high-resolution price changes of a dynamic energy market also provides a great deal of opportunity for driving ROI and OPEX reduction, while concurrently serving the more noble goal of EE utilization.
5G HW Power Characterization Standard	A standardized test plan based on the common concepts and metrics defined in this document can help to normalize data collection as well as ensure everything from raw data collection to analytic assessment is done in a way consistent with the teachings of this document and common across all industry stakeholders.
Grid Fault Response Time	Those that have the most detailed sensor data (i.e., service providers) need to share it with those with the highest degree of grid control (i.e., grid operators). Feeding into and expanding databases of sensor data that can be quickly mapped to physical layouts and PVCs will yield a system that has the potential to greatly enhance the resolution of grid fault detection and enable a faster response (ideally self-healing).
Long-Term Challenges: 2028-2032	*Potential Solutions to Long-Term Challenges*
Ecosystem Paradigm Shift	Major Infrastructure Change - Any real infrastructure change is easily in the 10+ year horizon and never is that more applicable than to energy infrastructure, especially the major distribution grid portions. This is why the true enablement of all the far-reaching, collaborative efforts to maximize EE outlined in this work can only be achieved by the will, investment, and leadership efforts of difference-makers in the ecosystem. In fact, this is so overarching that even defining what that ecosystem entails is an extremely difficult task. A truly smart grid is far easier said than implemented, but the rewards are just as great in terms of economic impact, a migration toward a more sustainable future, and quality of life in terms of socioeconomic impact (i.e., 5GEqG). As this work has outlined, the explosive growth of network utilization and the number of attached devices that all increase exponentially from one cellular generation to the next necessitate these kinds of ecosystem paradigm shifts for survival along with economic growth. Major Device Standard Change - Clearly, if any ecosystem is capable of completely revamping the way energy is generated, architected, distributed, and utilized, then the needs of the many are going to drive the momentum. In this context, the “many” are devices and controllers directly impacting load performance. Given anything in the electrical or electronic world requires power, the power supply (or application equivalent) is the logical place to apply the most effort in terms of upgrading intelligence and communicative ability. - Standardization is the only path for driving compatibility toward common goals amongst vast numbers of devices stretching across vast networks and utility grids. The Network Power Integration model detailed in the text below is an example of such a proposed solution.
Physical Layers For Digitally Managed Power	The Open Systems Interconnection model, or OSI model, that forms the basis for all modern communication systems is an ideal analog for the type of layered model that can be used to harmonize the physical layers of power and managed-power devices and the digital communication links that will enable the intelligence to perform the kind of visionary analytics and optimizations proposed in this work.
Adoption by Energy Storage/Generation Products	- An absolute critical enabler to any Smart Grid must be inclusive of many different types of energy storage at all levels of integration, the same as is required for bringing intelligence to power supplies and powered products. Pragmatic energy storage solutions not only enable the mitigation of energy sourced from fossil fuels, but their strategic placement and utility in ways that are suited to the specific application and environment of focus can be hugely impactful in meeting end power needs, while reducing CAPEX. For instance, applying the concept of peak shaving by using localized storage at the point of utilization (i.e., maximizing PFC in the PVC) increases reliability, while concurrently reducing the need for transmission and generation from non-localized sources.

	- This strategy of energy storage adoption paves the way for scalability of renewable generation and enables improved payback via the justification for driving economies of scale. More importantly, it drives a major departure from the way most stakeholders tend to think of the balance between energy sources and the end loads that consume that energy. These benefits and philosophical approaches to energy storage and utilization can be applied from the macro (i.e., grid-level) to the micro (i.e., end system-level or even subsystem-level as is done in servers and even individual ASICs today).

6. STANDARDIZATION LANDSCAPE AND VISION

6.1. Standardization Opportunities

Given the very large net of topic coverage areas cast by this WG, there are certainly numerous opportunities to standardize many of the concepts, methodologies, and framework(s) described. Many of the principles in this document are emerging proposals in an attempt to generate momentum for commonality and collaboration across all the stakeholders in the network. This means they must first be vetted by a host of industry experts to validate the value in such proposals in order to justify the massive resources required to generate and deliver a new standard (or class of standards). Also, to this point, there is a lot of preliminary, investigatory work that goes into such an effort because of the need to research and culminate the existing/legacy standardization landscape to ensure complementary solutions moving forward (as opposed to redundant, counterproductive, etc.).

The INGR Leadership Team has a Standards Group specifically focused in this area. While expressing interest in turning some of this content into standards proposals, a key distinction was noted between standardizing FRAMEWORK vs. PROTOCOL. This came to light when the EE WG expressed reservations about proceeding with standardization efforts too early in the process of proposing and validating much of this content within the 5G&B community as well as the greater industry. It was clarified that the approaches to standardization of a framework can vary greatly from the effort to standardize a protocol.

In short, the general framework standardization can occur earlier in the process since it is capturing the overall deliverable/tool/output as the identification of need(s), an appropriate ecosystem to serve said need(s), and a methodology to bring these resources together in the spirit of capturing and documenting the content in a usable format for external consumption. Additionally, framework standards can more easily and quickly be modified/updated as it is cultivated and gains traction within the appropriate industry(ies).

Standardizing a protocol, on the other hand, can be a far more intensive, involved effort requiring greater resources and review cycles. So for one, the EE WG wants to further validate and justify the content before determining if this next-level push is appropriate and enabled for success. Initial consideration of standardizing the framework makes a lot more sense as a segue to considering protocol standardization based on the response and adoption rate of this overall content by the 5G&B community.

6.1.1. Collaborative Opportunities

The IEEE SustainableICT Initiative (https://sustainableict.ieee.org/) has been developing a family of nine new standards all focused on various aspects of specifying, implementing, and assessing more sustainable network constituents. Most of these standards address many of the needs and solutions tabled in this document and provide numerous opportunities for collaboration on these as well as

impetus for offshoots and new ones. The applicable SustainableICT standards are listed as follows (grouped by general area of focus –

- Green ICT Emissions
  - **IEEE 1922.1**: Standard for a method for calculating anticipated emissions caused by virtual machine migration and placement
  - **IEEE 1922.2**: Standard for a method to calculate near real-time emissions of information and communication technology infrastructure
- EE Communication HW
  - **IEEE 1923.1**: Standard for computation of energy efficiency upper bound for apparatus processing communication signal waveforms
  - **IEEE 1924.1**: Recommended practice for developing energy efficient power-proportional digital architectures
- EE ICT
  - **IEEE 1925.1**: Standard for Energy Efficient Dynamic Line Rate Transmission System
  - **IEEE 1926.1**: Standard for a Functional Architecture of Distributed Energy Efficient Big Data Processing
  - **IEEE 1927.1**: Standard for Services Provided by the Energy-efficient Orchestration and Management of Virtualized Distributed Data Centers Interconnected by a Virtualized Network
  - **IEEE 1928.1**: Standard for a Mechanism for Energy Efficient Virtual Machine Placement
  - **IEEE 1929.1**: An Architectural Framework for Energy Efficient Content Distribution

7. CONCLUSIONS AND RECOMMENDATIONS

7.1. Summary of Conclusions

One of the most valuable outcomes of this work is its ability to identify many known and potentially unforeseen challenges in the deployment of EE telecommunication solutions, particularly for 5G&B, and the impact these factors have on the stakeholders that bring these solutions to realization as well as the planet as a whole. Seemingly impossible linkages between vast arrays of stakeholders from all corners of our ecosystem have been articulated and introduced through the lens of our universal currency of energy.

From this universal currency, we have proposed the SoS framework to enable a methodology and processes for unifying technical and business drivers in a novel way that we hope will drive traditional and future linkages that place EE optimization at the top of everyone's priority list.

After developing and sharing an understanding of the intricate relationship of Physics, infrastructure, regulations and financial constraints, we have identified some of the most critical efforts required to address the key technical aspects impacting the viability of 5G&B under the lens of EE.

- Network Efficiency stands as the first objective, by leveraging spectral efficiency, low-power operating modes, improved Massive MIMO deployment and Energy Harvesting to supplement/stabilize the Grid infrastructure.

- Migration to Small Cells offers significant advantages in both coverage and efficiency, but future cell-free architectures require major development in control and automated coordination that hopefully does not consume more power than what is saved.
- Base stations are a dominant source of power consumption and will continue to increase their share as their number increases and the need for heterogeneous network operation. Higher fidelity transmitters are required to generate higher constellation complexities with lower EVM and lower spectral regrowth. Massive MIMO radios with many low-gain antennas and highly integrated circuits with sophisticated cooling techniques will require significant advances in Silicon and WB materials as well as packaging technologies. Energy and data analytics will in time open the road to deploying energy-optimal control feedback loops via the coordination of system-level resources.
- The adoption of a common framework of Systems-of-Systems analysis and optimization will help bring together previously siloed stakeholders, enabling them to more effectively understand the impact of different technologies at the ecosystem level and thus achieve a more effective way of predicting investment requirements and payback periods. Such economic analysis will not necessarily lead to a more equitable development in support of the underserved, but may enlighten policymakers when assessing where legislation may be required to correct the development trajectory.
- The need for a significant investment to bring up the Grid infrastructure has been foreseen for a while. The WG analysis points to the areas where coordination between Grid and Telecom infrastructures can lead to benefits affecting both systems in terms of efficiency as well as reliability: two-way communication between the two ecosystems will bring to life the ability to optimize control of the use of energy resources in a collaborative way, thus realizing the potential of analysis and control afforded by the Systems of Systems.

This work is meant to break down the barriers between siloed stakeholders by bringing commonality in the articulation of the existential needs and challenges faced by all. The SoS framework will be a key component of analyzing the benefits of Intelligent Power Management (IPM) and forming multi-party cooperative agreements.

This work has also made extensive efforts to identify risks and interdependencies amongst such a diverse array of stakeholders within the fragile, push-pull of technical, economic, and socioeconomic perspectives and drivers at play in the 5G&B ecosystem. While some of these can be less obvious, perceptible, and/or likely to occur than others, the awareness that is generated and the inspiration to bring key stakeholders to the table to address these risks is a win-win for all.

We recommend this critical education begin with not only reviewing the foundational concepts, terminology, and metrics introduced by the efforts of this WG, but sharing with colleagues, collaborators, and competitors alike. We put this content forward in order to enable the industry to assess the value of the content, apply it as appropriate, CHALLENGE the more nascent concepts to drive qualitative and iterative improvement across the editions, and heed the call to a more EE, sustainable world led by 5G&B deployments and their supporting ecosystems. A key objective of this massive effort of so many stakeholders is to meet these noble challenges with proposals and solutions that are realistic, inspire collaboration, yet also enable stakeholders to interpret this work in a meaningful way to their own bottom line, whether that be technical advancement, economic growth, or mitigating the socioeconomic divide.

Perhaps one of the biggest values a reader can extract from this work is the culmination of such a wide variety of stakeholder perspectives, challenges, and proposed solutions all interpreted through the lens of EE in a common resource. This document, along with the references and external resources tabled within, is meant to build upon some of the more seminal work in this area (particularly that of GreenTouch, EARTH project, and the overall INGR this chapter belongs to) and provide a funnel for that information to be consolidated and further built upon in the perpetual quest for EE optimization.

7.2. Working Group Recommendations

If there is one takeaway from this work above all others, then it is that ALL things contained within this massive, 5G&B ecosystem can (and should) be viewed through the lens of EE. Finding a way to articulate, assess, and implement EE optimization is a universal approach to technical and economic sustainability in the most equitable terms for all touched by these networks. Whether direct or indirect, that impact essentially touches every person on the planet, thus further elevating the importance of the spotlight on this WG area of focus. We urge readers to consider how every mW at the edge is related to every GW of generation and dig deeper into the PVCs that connect them. We present the PCF metric to simplify and facilitate a more realistic approach to dissecting the CAPEX and OPEX of energy at the point of consumption.

This work and its common format of expressing so many technical and economic constituents in the universal currency of energy are presented in a way to not only speak to such a multidisciplinary set of stakeholders, but also to enable impassioned people to approach their own set of (what they consider to be) critical stakeholders. Even some of the boldest actions proposed in this work can be accomplished by starting with the simple action of collecting a group of motivated people in a room (virtual or otherwise) and not only sharing this work, but soliciting their feedback in how it has the potential to impact them through their own filters. The SoS framework is highly enabling for driving such inflection points in the way people collaborate and drive EE optimizations into their own solutions, resource investment justifications, and payback calculations.

It is the hope of the INGR EE WG that this work is seen as seminal reading for a widely encompassing spectrum of stakeholders brought together, perhaps for the first time or in novel approaches of collaboration for making EE optimization the primary business and technical driver of the industry and the many, rich ecosystems surrounding it. To this end, the most important recommendation we can make is for stakeholders to educate themselves on the many concepts, philosophies, challenges, solutions, and capturing of open questions/concerns that will propel us all more efficiently into the future.

As mentioned in Section 2.2, there is a high degree of value in the cross-correlation and cross-referencing of work amongst the WGs of the INGR team. Many workshops, meetings, and discussions produced this vast body of work. This is critical methodology to continue pursuing, not only as WGs hone their content through later editions of the INGR chapters, but also enhancing the quality and depth of cross-cut linkages, particularly with the Hardware, Deployment, System Optimization, and Applications & Services WGs.

Whether these recommendations seem familiar or foreign to the reader, an EE-mindset is required to enable the success of 5G&B networks over the next decade covered in this work and well into the follow-on generations. There is nothing that dissipates less power than something that is off. This is followed by the second least dissipator of power, which is a system operating at the optimal point in its

Load vs. Efficiency curve. The final step to actualization is a solution (or network in this context) that can operate in a way that yields reasonable payback, in a rational amount of time, to encourage economic investment and socioeconomic growth. No matter your metric, our EE optimizations serve your bottom line and serve it well.

We need to identify academics, industry partners and commercial software developers that are interested in developing and deploying SoS toolsets. One way to achieve this is to form a working group and/or standardization effort that will fabricate a long-term roadmap for the SoS framework and associated set of tools.

Naturally, one of the best recommendations this WG can make is to become a more direct stakeholder in these EE optimization efforts by joining us and/or one of the initiatives/organizations put forth in this document.

7.2.1. Future Work

As this is a living document, the future is already upon us today as we look to continually refine, improve, correct, and hone the fundamental messaging and recommendations contained within. Though part of the 2021 Edition INGR, this is the first version of this WG's output. Therefore, the next revision is already on the horizon.

As this first documented collection of topics sets the foundation for this effort, more resources can now be allocated to investigating the linkages across WGs and more importantly, the opportunities to increase the value and accuracy of each others' work to provide more qualitative deliverables for the industries and ecosystems it serves.

The current SoS framework is demonstrative, requiring refinement of ongoing research frontiers and development of the tool. Specific workshops covering these R&D areas as well as detailed examination of additional applications, low and high-fidelity usability studies are recommended moving forward.

As concepts evolve into tools, standards, databases, and new consortiums of collaboration focused on bringing EE to the forefront, there shall always be higher orders of integration/feedback and finer levels of granularity for optimization. A clear indicator of success along these perspectives shall be the enablement and evolution of HW systems connected by increasingly common interfaces and driven by SW that works harder to translate and optimize all performance via our universal currency.

8. CONTRIBUTOR BIOS

Francesco Carobolante (Co-chair), IoTissimo

Francesco Carobolante is Principal at IoTissimo, where he helps global organizations and young companies develop technology and business strategies to compete in the fast-changing high-tech world. Previously, he was Vice President Engineering at Qualcomm and held senior positions in semiconductor firms and start-ups.

Creator of many industry "firsts", including resonant wireless charging, which was Best of Innovation Award Honoree at 2015 Consumer Electronic Show, Carobolante is a renowned innovator and market development leader with extensive track record in establishing technology partnerships. He authored over 70 patents and has been invited keynote speaker at several premier international engineering conferences.

Brian Zahnstecher (Co-chair), PowerRox

Brian Zahnstecher is a Sr. Member of the IEEE, Chair (Emeritus) of the IEEE SFBAC Power Electronics Society (PELS), IEEE PELS North America Regional (R1-3) Chair, sits on the Power Sources Manufacturers Association (PSMA) Board of Directors, is Co-founder & Co-chair of the PSMA Reliability Committee, Co-chair of the PSMA Energy Harvesting Committee, and is the Principal of PowerRox. He Co-chairs the IEEE Future Directions (formerly 5G) Initiative webinar series and is the founding Co-chair of the IEEE 5G Roadmap Energy Efficiency Working Group and has lectured on this topic at major industry conferences. He previously held positions in power electronics with industry leaders Emerson Network Power (now Advanced Energy), Cisco, and Hewlett-Packard. He has been a regular contributor to the industry as an invited keynote speaker, author, workshop participant, session host, roundtable moderator, and volunteer. He has nearly 20 years of industry experience and holds Master and bachelor's degrees from Worcester Polytechnic Institute.

Earl McCune (IN MEMORIUM, RIP), Eridan Communications

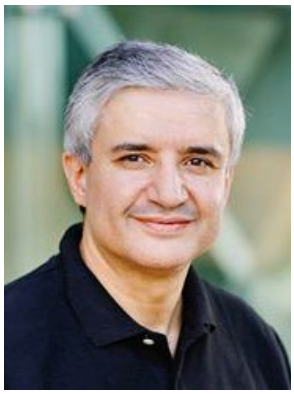
Earl McCune (S'78–M'79–SM'97–F'18) received the B.S.E.E./C.S. degree from the University of California (UC), Berkeley, CA, USA, the M.S.E.E. degree from Stanford University, and the Ph.D. degree from UC Davis, CA, USA. He was a Silicon Valley serial entrepreneur, with 93 issued U.S. patents and the author of two books. His research interests included RF circuits and systems, including modulation design, with an emphasis on the joint optimization of throughput and energy efficiency while also minimizing implementation cost. He was an emeritus MTT Distinguished Microwave Lecturer, a member of multiple IEEE conference and standards committees, and served as the Chair of the Energy Efficient Communications Hardware Standards Working Group. His considerable work experience included stints at NASA, Hewlett-

Packard, Watkins-Johnson, Cushman Electronics, Digital RF Solutions (start-up #1), Proxim, Tropian (start-up #2) and Panasonic, and Eridan (start-up #3) where he was CTO. He was a Professor of Delft University of Technology, where he held the chair of sustainable wireless systems.

Steve Allen, pSemi/Murata

Steve Allen has 40 years of experience in power conversion with approximately 20 years in power modules, and then the remainder in power management ICs, with three startups: Enpirion, Powervation and MIT spinout Arctic Sand – now acquired by Murata/ through their pSemi RF semiconductor division.

He currently leads the power semiconductor business unit developing bucks, boosts and charge pumps based on Arctic Sand technology, using advanced packaging technology. He holds an MBA with distinction from Bournemouth University, and BSEE at Portsmouth.



Mohamed-Slim Alouini, KAUST

Mohamed-Slim Alouini was born in Tunis, Tunisia. He received the Ph.D. degree in Electrical Engineering from the California Institute of Technology (Caltech), Pasadena, CA, USA, in 1998. He served as a faculty member in the University of Minnesota, Minneapolis, MN, USA, then in the Texas A&M University at Qatar, Education City, Doha, Qatar before joining King Abdullah University of Science and Technology (KAUST), Thuwal, Makkah Province, Saudi Arabia as a Professor of Electrical Engineering in 2009. His current research interests include the modeling, design, and performance analysis of wireless communication systems.

Anirban Bandyopadhyay, Global Foundries

Dr. Anirban Bandyopadhyay is the Director, RF Strategic Applications & Business Development within Global Foundries, USA and is located in Hopewell Junction, New York. His work is currently focused on hardware architecture & technology evaluations and market studies for different RF and mmwave applications. Prior to joining Global Foundries, he was with IBM Microelectronics for 8 years where he used to manage design enablement group for wireless applications and also led RF strategic applications and marketing. During 2000-2007, Dr. Bandyopadhyay was with Intel, California where he worked on different areas like Silicon Photonics, signal integrity in RF & Mixed signal SOC's. Dr. Bandyopadhyay did his PhD in Electrical Engineering from Tata Institute of Fundamental Research, India and Post-Doctoral research in Oregon State University, Corvallis. He has more than 15 publications in international journals, wrote a book chapter on Optical Photodetectors and holds 5 US patents. He represents Global Foundries in different industry consortia on RF/mmwave applications.

Emil Björnson, KTH Royal Institute of Technology

Emil Björnson is a Professor of Wireless Communication at the KTH Royal Institute of Technology, Stockholm, Sweden. He is an IEEE Fellow, Digital Futures Fellow, and Wallenberg Academy Fellow. He has a podcast and YouTube channel called Wireless Future. His research focuses on multi-antenna communications and radio resource management, using methods from communication theory, signal processing, and machine learning. He has authored three textbooks and has published a large amount of simulation code.

He has received the 2018 IEEE Marconi Prize Paper Award in Wireless Communications, the 2019 EURASIP Early Career Award, the 2019 IEEE Communications Society Fred W. Ellersick Prize, the 2019 IEEE Signal Processing Magazine Best Column Award, the 2020 Pierre-Simon Laplace Early Career Technical Achievement Award, the 2020 CTTC Early Achievement Award, and the 2021 IEEE ComSoc RCC Early Achievement Award. He also received six Best Paper Awards at the conferences.

Rick Booth, Eridan Communications

Richard Booth is a radio systems design and implementation engineer and has designed receivers, frequency synthesizers, transmitters, slotted wave guide antennas and various phase locked systems. His current interest is the design of polar or envelope elimination and restoration RF transmitters at all power levels and frequencies up to 3 GHz. He has worked mostly at small companies and startups and was most recently retired from Panasonic R&D.

He received his BS from MIT in 1969 and his PhD from USC in 1974, all in Electrical Engineering, and has 25 patents and several publications. In his spare time, he plays with his dogs.

Kirk Bresniker, Hewlett Packard Enterprise

Kirk Bresniker is Chief Architect of Hewlett Packard Labs, a Hewlett Packard Enterprise Fellow and Vice President. He joined Labs in 2014 to drive The Machine Research and Advanced Development program, leading teams across Labs and across HPE business units with the goal of demonstrating and evangelizing the benefits of Memory-Driven Computing. His current focus is accelerating the transfer of technologies from Labs disruptive development portfolio in order to drive differentiating value into existing product categories as well as create new offerings. Prior to joining Labs, Kirk was Vice President and Chief Technologist in the HP Servers Global Business Unit, the capstone to 25 years of innovation leadership in product development.

Joining HP as a PA-RISC system hardware engineer in 1989, Kirk has always been a part of HP and HPE compute team, focusing on high volume, low cost and modular compute platforms. Starting in 1997, Kirk began a decade-long research and development effort to develop novel new modular system architectures which would eventually become a new category of integrated hardware and software offerings known as Blade Servers. This early work was eventually refined and blended with the

contributions of the combined HP-Compaq merger lead to become HP BladeSystem c-Class, the undisputed leader in Blade Server platforms. In 1999, he oversaw a complete re-vamp of the Business Critical Systems product line. Kirk oversaw the transformation of the HP-UX UNIX and fault tolerant NonStop to blades to extend BladeSystem to the mission critical market, culminating in the Superdome X mission critical X86 blade platform, the highest performing HPE Mission Critical ProLiant system created to date. It was also during this period that he led the earliest investigations into what would become The Machine Research program.

Kirk currently holds 30 US and 10 foreign patents in areas of modular platforms and blade systems, integrated circuits, and power and environmental control. He is a Senior Member of the IEEE, a founding member of the IEEE Industrial Advisory Board, a founding member of the Markkula Center for Applied Ethics Tech Ethics Advisory board, and a member of the World Economic Forum Global Futures Councils on The Future of Compute, Quantum Computing and Agile Governance. He graduated in 1989 Cum Laude with a BSEE from Santa Clara University Honors Program.

Frederica Darema, formerly of NSF and AFOSR

Dr. Frederica Darema, a member of the Senior Executive Service, is the (Retired) Director of Air Force Office of Scientific Research, Arlington, Virginia. She guides the management of the entire basic research investment for the Air Force. Dr. Darema leads a staff of 200 scientists, engineers and administrators in Arlington, Virginia, and foreign technology offices in London, Tokyo and Santiago, Chile. Each year, AFOSR selects, sponsors and manages revolutionary basic research that impacts the future Air Force. AFOSR interacts with leading scientists and engineers throughout the world to identify breakthrough opportunities; actively manages a $510 million investment portfolio encompassing the best of these opportunities; and transitions the resulting discoveries to other components of the Air Force Research Laboratory, to defense industries and to other federal agencies. The office's annual investment in basic research is distributed among more than 200 leading academic institutions worldwide, 100 industry-based contracts, and more than 250 internal AFRL research efforts.

Dr. Darema is a graduate of the University of Athens, Greece, and the Illinois Institute of Technology and the University of California at Davis, where she attended as a Fulbright Scholar and a Distinguished Scholar. After Physics Research Associate positions at the University of Pittsburgh and Brookhaven National Laboratory, she received an American Physics Society Industrial Postdoctoral Fellowship and became a Technical Staff Member in the Nuclear Sciences Department at Schlumberger-Doll Research. Subsequently, at the T.J. Watson IBM Research Center she was a Research Staff Member and Research Group Manager. While at IBM, she also served in the IBM Corporate Strategy Group examining and helping to set corporate-wide strategies. From 1996 to 1998, she completed a two-year interagency assignment at the Defense Advanced Research Projects Agency.

Before being appointed to her present position, Dr. Darema was the Director of the Directorate for Information, Mathematics and Life Sciences, at AFOSR. Prior to AFOSR, Dr. Darema was at the National Science Foundation where she held executive positions as Senior Science and Technology Advisor and Senior Science Analyst in the Computer and Information Science and Engineering Directorate. In that capacity she initiated and led multi-directorate and multi-agency initiatives that fostered groundbreaking multidisciplinary research directions in computer sciences and in applications

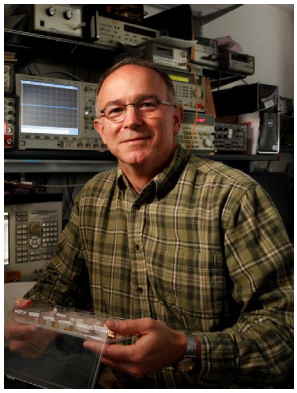
modeling and simulation. Dr. Darema has served on many scientific committees in the United States and internationally. She has an extensive record of publications and has given numerous keynote speeches and other presentations in national and international professional forums. Dr. Darema's scientific and technical accomplishments include seminal contributions in the parallel high-performance computing field, and specifically in: programming models; parallel algorithms; applications modeling and instrumentation systems; and systems performance-engineering methods for the design of applications and software for parallel and distributed systems.

Paul Draxler, Stonecrest Consulting

Paul J. Draxler (S'81–M'84–SM'13) received the B.S.E.E. and M.S.E.E. degrees
(with special focus on electromagnetics, RF and microwave circuits, antennas, and plasma physics) from the University of Wisconsin–Madison, Madison, WI, USA, in 1984 and 1986, respectively, and the Ph.D. degree (with a focus on power amplifiers, behavioral modeling and predistortion) from the University of California at San Diego. La Jolla, CA, USA, in 2013.

Following his graduation from the University of Wisconsin–Madison, he designed hybrid and GaAs monolithic microwave integrated circuit (MMIC) power amplifier circuits with the Hughes Aircraft Company and Avantek. From 1988 to 1995, he held various positions at EEsof and HP-EEsof where he focused on RF and microwave computer-aided engineering: custom design environments, nonlinear modeling, and electromagnetic simulation. In 1995, he joined Qualcomm Technologies Inc., San Diego, CA, USA, to lead a team focused on RF computer-aided engineering and design methodology automation. In this role, he has provided consulting to many design teams on system and circuit simulation, electromagnetic modeling, and board- and chip-level design methodologies. By 2003, he transferred into Qualcomm Research Center, Qualcomm Technologies Inc., as a Principal Engineer involved with research on advanced transmitter and power amplifier technologies. He is currently consulting to a number of organizations through Stonecrest Consulting covering the broad range of his career. He has authored or coauthored over 50 journal and symposium papers on electromagnetic simulation, circuit simulation, system simulation, power amplifier behavioral modeling, and predistortion. He is co-holder of over ten patents with other patents pending.

Dr. Draxler is a Sr. Member of the IEEE and has served on the IEEE Microwave Theory and Techniques Society (IEEE MTT-S) International Microwave Symposium (IMS) Technical Program Review Committee (TPRC) since the late 90s and the Power Amplifier Symposium executive committee for 10 years (general chair, technical program chair, publicity chair, and other roles).



Doug Kirkpatrick, Eridan Communications

Dr. Kirkpatrick is the co-founder and CEO of Eridan Communications, Inc., a Santa Clara based company developing transceiver products for the next generations of wireless communications – 5G and beyond. Dr. Kirkpatrick is also a founding General Partner of InnerProduct Partners (IPP), a San Francisco based early stage VC firm, and the acting CEO of a very early stage startup in rare-earth-free permanent magnets based in Cleveland. Previous to InnerProduct Partners, from 2010-2013 he was a partner at Vantage Point Capital Partners, and from 2002 – 2010 Dr. Kirkpatrick was a Program Manager and Chief Scientist at the Defense Advanced Research Projects Agency (DARPA). In addition to his DARPA role he was simultaneously the Senior Technologist for Technology Productization for the Undersecretary for Acquisition, Technology, and Logistics in the Department of Defense. Prior to his tour at DARPA he was the VP for R&D at Fusion Lighting, a lighting technology startup in the Washington DC area, and before that a VP and Division Manager for Science Applications International Corporation, also in the Washington DC area.

Dr. Kirkpatrick is a Fellow of the American Physics Society, a Member of the IEEE, and a Member of the Materials Research Society. Dr. Kirkpatrick holds a BS degree in Physics and Mathematics (1980) from the College of William and Mary and a Ph.D. in Physics from M.I.T. (1988). He is a named inventor on 16 US patents.

Laurence McGarry, pSemi/Murata

Laurence McGarry has over 30 years of experience in Power Management and Semiconductor industries with a track record in marketing and applications management.

His early career focused on ACDC and isolated DCDC system design for military and consumer power applications before moving into silicon IC definition and management.

He has extensive experience of interfacing with customers and defining differentiated, value-based products. He has a proven track record of developing strategies, growing product portfolios and revenue in Power Management businesses. He is a global professional who has lived and worked in UK, Hong Kong, China and US and an accomplished conference presenter, author of several published papers and magazine articles with two patents. He graduated from Glasgow University in 1988 with a BEng(Hons) in Electronics and Electrical Engineering and holds a MBA from Washington State University.

Lin Nease, Hewlett Packard Enterprise

Lin Nease is an HPE Fellow and chief technologist for HPE Pointnext Services' IoT advisory practice. In this role, he is responsible for setting strategy, building a technology plan, and driving innovation with key enterprise customers and partners of HPE. Lin also consults on IoT strategy with HPE's enterprise customers in manufacturing, transportation, government and smart city initiatives, financial services, hospitality, and other industries. He co-founded HPE's EdgeLine business, drove portfolio enhancements to HPE's GreenLake services, established the company's IoT practice, and led HPE's membership in organizations like the Industrial Internet Consortium.

In his 30-plus years with HPE, he has been a chief technologist and director of strategy for multiple business units, including the company's business-critical servers and networking groups. In addition, Lin has been a chief technologist and general manager for multiple global accounts (including GE and UPS), driven multiple M&A activities and cross-business initiatives, and led launches of numerous successful commercial products, including the industry's first blade solution and HPE's long-lived Superdome platform. He also holds several patents in software-defined networking.

Lin received his BS in computer science at Arizona State University and his MBA from California State University Sacramento. He also served as a computer operator in the U.S. Air Force.

Bruce Nordman, Lawrence Berkeley National Laboratory

Bruce Nordman is a Research Scientist with Lawrence Berkeley National Laboratory, operated by the University of California for the U.S. Department of Energy. His research focuses on the intersection of energy use, electronics, and networks. He works at all network layers from physical layer technology to application layer protocols to user interfaces.

Magnus Olsson, Huawei

Magnus Olsson has been in the wireless industry for more than 20 years. Since 2017, he holds a position as Principal Energy Efficiency expert at Huawei Technologies, based at the Stockholm Research Center in Sweden, where his current responsibilities include RAN Energy Efficiency research and standardization in ETSI and ITU. Before that, he was with Ericsson Research, Stockholm, Sweden, which he joined in 2000. Over the years he has worked on several radio access technologies and areas such as advanced antenna systems, interference rejection techniques, and energy efficiency of radio access networks (RAN). He has authored and co-authored over 30 international journal and conference papers as well as book chapters, and is a recipient of the IEEE Communications Society Fred W. Ellersick Prize (2014). He has held leading positions in both internal and various European collaborative research projects. For example, he was the Technical Manager of the successful €14.8m EU FP7 EARTH project on RAN energy efficiency which received the ceFIMS Future Internet Award (2012). Since 2015, he is on the steering committee of the IEEE Sustainable ICT initiative, a pan IEEE Societies initiative responsible for Sustainable ICT activities across IEEE, and since 2019 member of the Energy Efficiency working group of the IEEE International Network Generations Roadmap (INGR).

9. REFERENCES

[1] IEEE Future Networks Initiative - Energy Efficiency Working Group, "Energy Efficiency - 1st Edition White Paper," International Network Generations Roadmap (INGR), Apr. 2020.

[2] Wikipedia contributors. "Power usage effectiveness." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 4 Mar. 2020. Web.

[3] B. Zahnstecher, "The 5G Energy Gap [Expert View]," in IEEE Power Electronics Magazine, vol. 6, no. 4, pp. 64-67, Dec. 2019.

[4] Masanet, E., Shehabi, A., Lei, N., Smith, S., and Koomey, J. 2020. "Recalibrating global data center energy-use estimates" In Science Magazine. February 28. https://science.sciencemag.org/content/367/6481/984.abstract.

[5] IEA (2019), "Tracking Buildings", IEA, Paris https://www.iea.org/reports/tracking-buildings.

[6] IEEE Future Networks Initiative - Energy Efficiency Working Group, "Energy Efficiency, 2021 Edition" International Network Generations Roadmap (INGR), Apr. 9 2021. [Online]. Available: https://futurenetworks.ieee.org/roadmap.

[7] IEEE Future Networks Initiative, IEEE Future Networks 1st Massive MIMO Workshop, 8-10 November 2021. [Online]. Available: https://futurenetworks.ieee.org/conferences/future-network-massive-mimo-workshop.

[8] IEEE Future Networks Initiative - Massive MIMO Working Group, "Massive MIMO, 2022 Edition" International Network Generations Roadmap (INGR), Mar. 25 2022. [Online]. Available: https://futurenetworks.ieee.org/roadmap.

[9] IEEE Future Networks Initiative - System Optimization Working Group, "System Optimization, 2022 Edition" International Network Generations Roadmap (INGR), Mar. 25 2022. [Online]. Available: https://futurenetworks.ieee.org/roadmap.

[10] IEEE Future Networks Initiative, 2021 IEEE 4th 5G World Forum (5GWF), 2021, doi: 10.1109/5GWF52925.2021.

[11] IEA (2020), Data Centres and Data Transmission Networks, IEA, Paris https://www.iea.org/reports/data-centres-and-data-transmission-networks.

[12] C. I, "5G's Green Journey and More," 1st 5G Energy Efficiency Tutorial (EET), Santa Clara, CA, September 19, 2018.

[13] "Ericsson Energy and Carbon Report," Ericsson, June 2014.

[14] "Heterogeneous Integration Roadmap," IEEE Electronics Packaging Society (EPS), HIR 2020 version, https://eps.ieee.org/hir.

[15] Y. London et al., "Energy Efficiency Analysis of Comb Source Carrier-Injection Ring-Based Silicon Photonic Link," in IEEE Journal of Selected Topics in Quantum Electronics, vol. 26, no. 2, pp. 1-13, March-April 2020, Art no. 3300113, doi: 10.1109/JSTQE.20.

[16] M. Olson, "The Impact of 5G on Utilities Starts Now," T&D World, Sep 09, 2019.

[17] "Electricity Consumption by County," California Energy Commission. [Online]. Available: http://www.ecdms.energy.ca.gov/elecbycounty.aspx.

[18] L. Capuano, "U.S. Energy Information Administration's International Energy Outlook 2020 (IEO2020)," U.S. Energy Information Administration, Center for Strategic and International Studies, October 14, 2020. https://www.eia.gov/outlooks/ieo/ppt/ieo2020.ppt.

[19] Organisation for Economic Co-operation and Development, https://www.oecd.org/.

[20] Linda Capuano, "International Energy Outlook 2018 (IEO2018)," U.S. Energy Information Administration, July 24, 2018.

[21] FCC Office of Engineering and Technology, "Millimeter Wave Propagation: Spectrum Management Implications" FCC Bulletin Number 70, July, 1997. https://transition.fcc.gov/Bureaus/Engineering_Technology/Documents/bulletins/oet70/oet70a.pdf.

[22] “5G; NR; Base Station (BS) radio transmission and reception", 3GPP TS 38.104 version 15.2.0 Release 15.

[23] Y. Jading, "From Always On to Always Available," 1st 5G Energy Efficiency Tutorial (EET), Santa Clara, CA, September 19, 2018.

[24] "Why the gig economy could actually hurt small businesses," Tom Popomaronis, ITProPortal, February 11, 2020. https://www.itproportal.com/features/why-the-gig-economy-could-actually-hurt-small-businesses/.

[25] D. Sickinger, S. Serebryakov, T. Cader, "AIops: Bringing Artificial Intelligence to the Data Center," 19 November 2019. [Online] https://sc19.supercomputing.org/presentation/?id=exforum129&sess=sess370.

[26] DMTF’s Redfish®, https://www.dmtf.org/standards/redfish/.

[27] Redfish Power and Thermal Enhancements. https://www.dmtf.org/dsp/DSP-IS0015.

[28] https://ma-mimo.ellintech.se/2020/09/26/active-and-passive-antennas/.

[29] Beammwave, “Digital Beamforming for Mobile Devices: The power efficient architecture for 5G on mmWave frequencies,” White paper, Sept. 2020. https://www.beammwave.com/whitepapers.

[30] C. Desset, B. Debaillie, “Massive MIMO for Energy-Efficient Communications,” Proceedings of the 46th European Microwave Conference (EuMC), 2016.

[31] Ko, Young-Chai; Jung, Kug-Jin; Park, Ki Hong; Alouini, Mohamed-Slim (2021): Renewable Energy-Enabled Cellular Networks. TechRxiv. Preprint. https://doi.org/10.36227/techrxiv.16550607.v1.

[32] "Artificial Intelligence in the RAN," Ericsson Technology Review, #12, 2020. [ONLINE] https://www.ericsson.com/4ae5c4/assets/local/reports-papers/ericsson-technology-review/docs/2020/artificial-intelligence-in-ran.pdf.

[33] "Enhancing RAN Performance with AI," Ericsson Technology Review, #12, 2019. https://www.ericsson.com/493ce3/assets/local/reports-papers/ericsson-technology-review/docs/2020/enhancing-ran-performance-with-ai.pdf.

[34] Darema F., (2004), Dynamic data driven applications system: a new paradigm for application simulations and measurements, in Proceedings of the 2004 International Conference on Computational Science, pp. 662-669.

[35] F. Darema, "Grid Computing and Beyond: The Context of Dynamic Data Driven Applications Systems," in Proceedings of the IEEE, vol. 93, no. 3, pp. 692-697, March 2005, doi: 10.1109/JPROC.2004.842783.

[36] Orchestrating the Cognitive Internet of Things . In Proceedings of the International Conference on Internet of Things and Big Data - Volume 1: IoTBD, ISBN 978-989-758-183-0, pages 96-101.

[37] D. Allaire, D. Kordonowy, M. Lecerf, L. Mainini, K. Willcox, “Multifidelity DDDAS Methods with Application to a Self-Aware Aerospace Vehicle,” Int’l Conf. on Computational Science, 29: 1182–1192, 2014.

[38] X. Shi, H. Damgacioglu, N. Celik, “A Dynamic Data Driven Approach for Operation Planning of Microgrids,” Procedia Computer Science, Volume 51, 2015, Pages 2543–2552.

[39] Mehrad Bastania, Aristotelis E. Thanosb, Haluk Damgaciogluc, Nurcin Celik, Chun-Hung Chend, “An evolutionary simulation optimization framework for interruptible load management in the smart grid,” Sustainable Systems and Society, 41 (2018) 802-809.

[40] Damgacioglu, Haluk, Mehrad Bastani, and Nurcin Celik. "A Dynamic Data-Driven Optimization Framework for Demand Side Management in Microgrids." In Handbook of Dynamic Data Driven Applications Systems, pp. 489-504. Springer, 2018.

[41] R. M. Fujimoto et al., “Dynamic data driven application systems for smart cities and urban infrastructures,” Proc. - Winter Simul. Conf., vol. 0, no. Oecd 2011, pp. 1143– 1157, 2016.

[42] E. Blasch, “DDDAS advantages from high-dimensional simulation,” Proc. - Winter Simul. Conf., vol. 2018-Decem, pp. 1418–1429, 2019.

[43] A. J. Aved, “Scene Understanding for Real Time Processing of Queries over Big Data Streaming Video,” University of Central Florida, 2013.

[44] E. Blasch, J. Ashdown, C. Varela, F. Kopsaftopoulos, R. Newkirk,” Dynamic Data Driven Analytics for Multi-domain Environments,” Proc. SPIE, Vol. 11006, 2019.

[45] L. Carlone, A. Axelrod, S. Karaman, G. Chowdhary, “Aided Optimal Search: Data-Driven Target Pursuit from On-Demand Delayed Binary Observations. In: E Blasch, et al. (eds.) Handbook of Dynamic Data Driven Applications Systems. Springer, Cham., 2018.

[46] W. Silva, E. W. Frew, and W. Shaw-Cortez, “Implementing Path Planning and Guidance Layers for Dynamic Soaring and Persistence Missions,” Int’l Conf. on Unmanned Aircraft Systems (ICUAS), 2015.

[47] E. Blasch, Y. Al-Nashif, and S. Hariri, “Static versus dynamic data information fusion analysis using DDDAS for cyber security trust,” Procedia Comput. Sci., vol. 29, pp. 1299–1313, 2014.

[48] Darema, F., CyberInfrastructures of Cyber-Applications-Systems, Proceedings of ICCS2010, Procedia Computer Science 1 (2012) 1287–1296.

[49] X. Shi, H. Damgacioglu, N. Celik, “A Dynamic Data Driven Approach for Operation Planning of Microgrids,” Procedia Computer Science, Volume 51, 2015, Pages 2543–2552.

[50] Mehrad Bastania, Aristotelis E. Thanosb, Haluk Damgaciogluc, Nurcin Celikc, , Chun-Hung Chend An evolutionary simulation optimization framework for interruptible load management in the smart grid, Sustainable Systems and Society, 41 (2018) 802-809.

[51] Damgacioglu, Haluk, Mehrad Bastani, and Nurcin Celik. "A Dynamic Data-Driven Optimization Framework for Demand Side Management in Microgrids." In Handbook of Dynamic Data Driven Applications Systems, pp. 489-504. Springer, 2018.

[52] Darema F., (2004), Dynamic data driven applications system: a new paradigm for application simulations and measurements, in Proceedings of the 2004 International Conference on Computational Science, pp. 662-669.

[53] Flash Boys: A Wall Street Revolt, Michael Lewis, W.W. Norton & Co., 2015.

[54] G. Auer and et al., D2.3: Energy efficiency analysis of the reference systems, areas of improvements and target breakdown. INFSO-ICT-247733 EARTH, ver. 2.0, 2012.

https://cordis.europa.eu/docs/projects/cnect/3/247733/080/deliverables/001-EARTHWP2D23v2.pdf.

[55] Ericsson, “Ericsson Mobility Report”, November 2015. https://www.ericsson.com/en/mobility-report/reports.

[56] Terry Costlow, "Fusing Sensors for the Automated Driving Future," SAE Mobilus, February 13, 2019. https://saemobilus.sae.org/automated-connected/feature/2019/02/fusing-sensors-for-the-automateddriving-future.

[57] M. Kaul, “Self-Driving Car Technology and Associated Computational Power Requirements”, CICC 2018, San Diego, California, USA, April 9, 2018.

[58] M. Kaul, “Self-Driving Car Technology and Associated Computational Power Requirements”, CICC 2018, San Diego, California, USA, April 9, 2018.

[59] The Open Compute Project, https://www.opencompute.org/.

[60] Open Networking Forum, https://opennetworking.org/.

[61] Open RAN, O-RAN Alliance, https://www.o-ran.org/.

[62] Open RAN Policy Coalition, https://www.openranpolicy.org/.

[63] E. Björnson, L. Sanguinetti, M. Kountouris, “Deploying Dense Networks for Maximal Energy Efficiency: Small Cells Meet Massive MIMO,” IEEE Journal on Selected Areas in Communications, vol. 34, no. 4, pp. 832-847, April 2016.

[64] G. Interdonato, E. Björnson, H. Q. Ngo, P. Frenger, E. G. Larsson, “Ubiquitous Cell-Free Massive MIMO Communications,” EURASIP Journal on Wireless Communications and Networking, vol. 2019, no. 197, 2019.

[65] C. Desset, B. Debaillie, “Massive MIMO for energy-efficient communications,” European Microwave Conference (EuMC), 2016.

[66] G. Auer, V. Giannini, C. Desset, I. Godor, P. Skillermark, M. Olsson, M. A. Imran, D. Sabella, M. J. Gonzalez, O. Blume, and A. Fehske, “How much energy is needed to run a wireless network?,” IEEE Wireless Communications, vol. 18, pp. 40–49, October 2011.

[67] B. Debaillie., C. Desset, F. Louagie, “A Flexible and Future-Proof Power Model for Cellular Base Stations”, IEEE VTC Spring, 2015.

[68] P. Frenger, R. Tano, “More capacity and less power: How 5G NR can reduce network energy consumption”, IEEE VTC Spring, 2019.

[69] B. Zahnstecher, “5G Opens the Door for Energy Harvesting (EH) in Telco Applications,” Half-Day Tutorial, INTELEC 2017, Gold Coast, Australia, October 22, 2017.

[70] Joshua Israelsohn, "Any way the wind blows....," ECN Magazine, July 21, 2014.

[71] R. Ali, L. Stroyov, S. Patel, "Telstra's Fuel Cell Experience," INTELEC 2017, Gold Coast, Australia, October 22, 2017.

[72] F. Rezei, C. Tellambura, and S. Herath, “Large-scale wireless-powered networks with backscatter communications — A comprehensive survey," IEEE Open Journal of the Communication Society, August 2020.

[73] E. Björnson, E. G. Larsson, “How energy-efficient can a wireless communication system become?,” Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, USA, October 2018.

[74] M. Obeed, A. Salhab, M.-S. Alouini, and S. Zummo, "On optimizing VLC networks for downlink multi-user transmission: A survey," IEEE Communications Surveys and Tutorials, Vol. 21, No. 3, pp. 2947 - 2976, Third Quarter 2019.

[75] ITU-R, "Attenuation by atmospheric gases, P Series, Radiowave propagation," Recommendation ITU-R P.676-9. [Online] https://www.itu.int/dms_pubrec/itu-r/rec/p/R-REC-P.676-9-201202-S!!PDF-E.pdf.

[76] E. Björnson, J. Hoydis, L. Sanguinetti, “Massive MIMO Networks: Spectral, Energy, and Hardware Efficiency,” Foundations and Trends in Signal Processing: vol. 11, no. 3-4, pp. 154–655, 2017.

[77] A. Mohamed, O. Onireti, M. A. Imran, A. Imran and R. Tafazolli, "Control-Data Separation Architecture for Cellular Radio Access Networks: A Survey and Outlook," in IEEE Communications Surveys & Tutorials, vol. 18, no. 1, pp. 446-465, First quarter 2016, d.

[78] E. Björnson, E. Jorswieck, “Optimal Resource Allocation in Coordinated Multi-Cell Systems,” Foundations and Trends in Communications and Information Theory, vol. 9, no. 2-3, pp. 113-381, 2013.

[79] Huawei PowerStar Solution, Network-level AI-based Energy Saving, Mobile World Live, 02 JAN 2019. https://www.mobileworldlive.com/huawei-updates/huawei-powerstartm-solution-network-level-ai-based-energy-saving.

[80] M. Olsson, C. Cavdar, P. Frenger, S. Tombaz, D. Sabella, R. Jäntti. “Towards Green 5G Mobile Networks,” GReen Optimized Wireless Networks (GROWN) workshop in association with IEEE International Conference on Wireless and Mobile Computin.

[81] A. Maaref, J. Ma, M. Salem, H. Baligh, and K. Zarin, “Device-centric radio access virtualization for 5G networks,” in Proc. IEEE Globecom Workshops, Dec. 2014, pp. 887–893.

[82] J. Zhang et al, “PoC of SCMA-Based Uplink Grant-Free Transmission in UCNC for 5G”, IEEE Journal on Selected Areas in Communications, vol.35, no.6, June 2017.

[83] E. Björnson, L. Sanguinetti, “Scalable Cell-Free Massive MIMO Systems,” IEEE Transactions on Communications, vol. 68, no. 7, pp. 4247-4261, July 2020.

[84] A. Ghasemi, A. Abedi, F. Ghasemi, Antennas and Passive Reflectors. In: Propagation Engineering in Radio Links Design. Springer, New York, NY., 2013.

[85] M. Di Renzo, A. Zappone, M. Debbah, M.-S. Alouini, C. Yuen, J. de Rosny, S. Tretyakov, "Smart radio environments empowered by reconfigurable intelligent surfaces: How it works, state of research, and road ahead", IEEE Journal of Selected Areas in Communic.

[86] Q. Wu and R. Zhang, “Towards smart and reconfigurable environment: Intelligent reflecting surface aided wireless network,” IEEE Commun. Mag., vol. 58, no. 1, pp. 106–112, 2020.

[87] E. Björnson, Ö. Özdogan, E. G. Larsson, “Reconfigurable Intelligent Surfaces: Three Myths and Two Critical Questions,” IEEE Communications Magazine, To appear.

[88] ka2020 M. Kishk and M.-S. Alouini, “Exploiting randomly-located blockages for large-scale deployment of intelligent surfaces,” IEEE Journal of Selected Areas in Communications (JSAC)-Special Issue on Massive Access for 5G and beyond, To appear, 2020.

[89] Hua Wang, Tzu-Yuan Huang, Naga Sasikanth Mannem, Jeongseok Lee, Edgar Garay, David Munzer, Edward Liu, and Michael Edward Duffy Smith, "Power Amplifiers Performance Survey 2000-Present," https://gems.ece.gatech.edu/PA_survey.html.

[90] GNU Radio, https://www.gnuradio.org/.

[91] A. Bandyopadhyay, "Silicon Technology Solutions to Address Power and Performance Requirements for Sub 6GHz & mmwave 5G Radio Interface," 1st 5G Energy Efficiency Tutorial (EET), Santa Clara, CA, September 19, 2018.

[92] F. Carobolante, "Power Supply on Chip: from R&D to commercial products," PwrSoC 2014, Boston, MA, October 2014. http://pwrsocevents.com/wp-content/uploads/2014-presentations/ts/S0_3%20Plenry%20Carobolante.pdf.

[93] D. Kirkpatrick, "Mitigating Thermal & Power Limitations to Enable 5G," 1st 5G Energy Efficiency Tutorial (EET), Santa Clara, CA, September 19, 2018.

[94] J. Wu, Y. Zhang, M. Zukerman and E. K. -N. Yung, "Energy-Efficient Base-Stations Sleep-Mode Techniques in Green Cellular Networks: A Survey," in IEEE Communications Surveys & Tutorials, vol. 17, no. 2, pp. 803-826, Secondquarter 2015, doi: 10.1109/COMST.2.

[95] A Dynamic Data-Driven Optimization Framework for Demand Side Management in Microgrids, Haluk Damgacioglu, Mehrad Bastani, and Nurcin Celik; Handbook of Dynamic Data Driven Applications Systems, Volume 1, 2nd Ed., E. Blasch, F. Darema S. Ravela, A. Aved.

[96] IEEE Future Networks Initiative - Applications & Services Working Group, "Applications & Services, 2022 Edition" International Network Generations Roadmap (INGR), Mar. 25 2022. [Online]. Available: https://futurenetworks.ieee.org/roadmap.

[97] The Dynamic Data Driven Applications Systems (DDDAS) Paradigm and Emerging Directions, F. Darema E. Blasch, S. Ravela, A. Aved. Springer 2022.

[98] Dynamic Data Driven Adaptive Simulation Framework for Automated Control in Microgrids, A. E. Thanos, M. Bastani, N. Celik and C. Chen; IEEE Transactions on Smart Grid, vol. 8, no. 1, pp. 209-218, Jan. 2017, doi: 10.1109/TSG.2015.2464709.

[99] DDDAS-based multi-fidelity simulation framework for supply chain systems, N. Celik, S. Lee, K. Vasudevan, & Y. J. Son; IIE transactions, 42(5), 325-341, 2010.

[100] DDDAS @ 5G and Beyond 5G Networks for Resilient Communications Infrastructures and Microgrid Clusters, Abdurrahman Yavuz Temitope Runsewe, Nurcin Celik, Christina Chaccour, Walid Saad, Frederica Darema; Springer (2022).

[101] Shehabi, A., Smith, S.J., Horner, N., Azevedo, I., Brown, R., Koomey, J., Masanet, E., Sartor, D., Herrlin, M., Lintner, W. 2016. United States Data Center Energy Usage Report. Lawrence Berkeley National Laboratory, Berkeley, California. LBNL-1005775.

[102] Koomey, Jonathan, and Samuel Naffziger. 2016. "Energy efficiency of computing: What's next?" In Electronic Design. November 28. http://electronicdesign.com/microprocessors/energy-efficiency-computing-what-s-next.

[103] Steven Lanzisera, Bruce Nordman and Richard E. Brown, “Data network equipment energy use and savings potential in buildings”, Energy Efficiency, Volume 5, Number 2 (2012), 149-162.

[104] Christensen, K., P. Reviriego, B. Nordman, M. Bennett, M. Mostowfi, and J.A. Maestro 2010. “IEEE 802.3az: The Road to Energy Efficient Ethernet.” IEEE Communications, special issue on Green Communications. March, 2010.

[105] Malmodin, Jens, and Dag Lundén, The electricity consumption and operational carbon emissions of ICT network operators 2010-2015, Report from the KTH Centre for Sustainable Communications Stockholm, Sweden 2018.

[106] GreenTouch Final Results from Green Meter Research Study," A GreenTouch White Paper, Version 2.0, August 15, 2015.

[107] IEEE 5G Energy Efficiency Tutorial, September 19, 2018, https://futurenetworks.ieee.org/education/ieee-5g-learning-series/ieee-5g-learning-series-bay-area-energy-efficiency-edition.

[108] IEEE, IEEE International Network Generations Roadmap (INGR), Energy Efficiency Working Group. [Online] https://futurenetworks.ieee.org/roadmap.

[109] Office of Energy Efficiency & Renewable Energy, "Confronting the Duck Curve: How to Address Over-Generation of Solar Energy," October 12, 2017. https://www.energy.gov/eere/articles/confronting-duck-curve-how-address-over-generation-solar-energy.

[110] IEEE, IEEE International Network Generations Roadmap (INGR), Security Working Group. [Online] https://futurenetworks.ieee.org/roadmap/ingr-edition-1-2019/.

[111] V. Smil, "Germany's Energiewende, 20 Years Later," IEEE Spectrum, 25 Nov 2020. [Online] https://spectrum.ieee.org/energy/renewables/germanys-energiewende-20-years-later.

[112] C. Armel, "Energy Disaggregation," Precourt Energy Efficiency Center, Stanford, December 2011. https://web.stanford.edu/group/peec/cgi-bin/docs/events/2011/becc/presentations/3%20Disaggregation%20The%20Holy%20Grail%20-%20Carrie%20Armel.pdf.

[113] U. SIvaram, "Energy use disaggregation is coming. Here are 3 ways utilities can make the most of it," Utility Dive, April 26, 2018. [Online] https://www.utilitydive.com/news/energy-use-disaggregation-is-coming-here-are-3-ways-utilities-can-make-the/52226.

[114] B. L. Berenbroek, "Energy disaggregation to empower utilities and their customers," Smart Energy International Podcast, Jan 11, 2021. https://www.smart-energy.com/industry-sectors/energy-grid-management/energy-disaggregation-to-empire-utilities-and-their.

[115] ARENA, "What are distributed energy resources and how do they work?" ARENAWIRE, 15 March 2018. https://arena.gov.au/blog/what-are-distributed-energy-resources/.

[116] Wikipedia contributors. "OSI model." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 28 Feb. 2021. https://en.wikipedia.org/wiki/OSI_model.

[117] Nordman, Bruce, and Ken Christensen, DC Local Power Distribution with Microgrids and Nanogrids, First International Conference on DC Microgrids, Atlanta, GA, June 2015.

[118] Nordman, Bruce, and Ken Christensen, DC Local Power Distribution: Technology, deployment, and pathways to success IEEE Electrification Magazine, June 2016.

[119] IEEE, IEEE 802.3bt-2018 - IEEE Standard for Ethernet Amendment 2: Physical Layer and Management Parameters for Power over Ethernet over 4 pairs, 2018.

[120] Dwelley, Dave, A Quick Walk Around the Block with PoDL, IEEE 802 Tutorial, Nov 9, 2015.

[121] P. Lee, "The Software-Defined Power Grid Is Here," IEEE Spectrum, 23 Jun 2020. https://spectrum.ieee.org/energy/the-smarter-grid/the-softwaredefined-power-grid-is-here.

[122] T. King Jr., "Sensing & MeasurementGrid Modernization Laboratory Consortium," Oak Ridge National Laboratory Electricity Advisory Committee, February 2020. https://www.energy.gov/sites/prod/files/2020/03/f72/Overview%20of%20GMLC%20Advanced%20Sensors%20and.

[123] S. Jerchich, "Overview of Congestion Revenue RightsIn the New California Energy Market," California ISO MRTU Readiness Conference, Sacramento, March 21, 2006. https://www.caiso.com/Documents/CRROverviewPresentation.pdf.

[124] B. Nordman, "Networked Electricity," 1st 5G Energy Efficiency Tutorial (EET), Santa Clara, CA, September 19, 2018.

[125] DC Local Power Distribution with Microgrids and Nanogrids, First International Conference on DC Microgrids, Atlanta, GA, June 2015.

[126] Advanced Research Projects Agency - Energy (ARPA-E), "Generating Realistic Information for the Development of Distribution and Transmission Algorithms (GRID DATA)," ARPA-E Program Description, Release Date: 01/15/2016. [Online] https://arpa-e.energy.gov/.

[127] B. Kroposki et al., "Autonomous Energy Grids: Controlling the Future Grid With Large Amounts of Distributed Energy Resources," in IEEE Power and Energy Magazine, vol. 18, no. 6, pp. 37-46, Nov.-Dec. 2020, doi: 10.1109/MPE.2020.3014540.

10. ACRONYMS/ABBREVIATIONS

Term	Definition
4G	Fourth Generation (Wireless Network)
3GPP	Third Generation Partnership Project
5G	Fifth Generation
5GDF	The 5G Derate Factor
5GEG	The 5G Energy Gap
5GEcG	The 5G Economic Gap
5GEqG	The 5G Equality Gap
5G&B	5G and Beyond
AAS	Advanced Antenna System
ACLR	Adjacent Channel Leakage Ratio
A/D	Analog-to-digital
AEG	Autonomous Energy Grid
AI	Artificial intelligence
ARPA-E	Advanced Research Projects Agency - Energy
ATE	Automated test equipment
BBU	Baseband Unit
BGA	Ball Grid Array
BOM	Bill of Material
CAPEX	Capital expenditure
CM	Contract Manufacturer
CPE	Customer Premises Equipment
CPU	Central Processing Unit
DDDAS	Dynamic Data Driven Applications Systems
DER	Distributed Energy Resources
DFe	Design for Energy
DFx	Design for X
DOE	Department of Energy
DPD	Digital Pre-Distortion
DSP	Digital Signal Processor
DT	Data throughput
DTx	Discontinuous transmission

DU	Digital Unit
DWDM	Dense wavelength division multiplexing
EC	Energy consumption
EE	Energy Efficiency
EH	Energy Harvesting
EIRP	Effective Isotropic Radiated Power
eMBB	Enhanced mobile broadband
ETSI	European Telecommunications Standards Institute
EV	Electric Vehicles
EVM	Error Vector Magnitude
FDD	Frequency Division Duplex
FN	Future Networks
FoM	Figure-of-Merit
FPGA	Field Programmable gate Array
Ft	Transition frequency
GaN	Gallium Nitride
GIS	Geographic Information System
GPU	Graphic Processing Units
GRID DATA	Generating Realistic Information for the Development of Distribution and Transmission Algorithms
GSMA	GSM (Groupe Speciale Mobile) Association
HDI	High Definition Interconnect
HetNet	Heterogeneous Network
HIR	Heterogeneous Integration Roadmap
HPC	High Performance Computing
HW	Hardware
IEEE	Institute of Electrical and Electronics Engineers
IoT	Internet of Things
INGR	International Network Generations Roadmap
IP	Internet protocol
IPM	Intelligent Power Management
IRS	Intelligent Reflecting Surface
KPI	Key performance indicator
LPA	Linear Power Amplifier
LTE	Long-term evolution

LTE-U	Long-term evolution - Unlicensed spectrum
M2M	Machine to machine
MCM	Multi-Chip-Modules
MCU	Micro Controller Unit
MEC	Mobile Edge Computing
MIMO	Multiple input, multiple output
ML	Machine Learning
mmWave	Millimeter wave
NFV	Network function virtualization
NPI	Network Power Integration
NR	New radio
NREL	National Renewable Energy Lab
OCP	Over Current Protection
OECD	Organisation for Economic Co-operation and Development
OFDM	Orthogonal frequency-division multiplexing
ONF	Open Networking Forum
OPEX	Operational Expenditure
OSI	Open Systems Interconnections
OVP	Over Voltage Protection
PA	Power Amplifier
PAE	Power Amplifier Efficiency
PAPR	Peak-to-Average Power Ratio
PCB	Printed Circuit Board
PCF	Power Cost Factor
PDN	Power Delivery Network
PFC	Power Factor Correction
PHY	Physical layer
PMU	Phasor Measurement Unit
POL	Point of Load
PSiP	Power Supply in Package
PUE	Power Usage Effectiveness
PV	Photovoltaic
PVC	Power Value Chain
PwrSoC	Power Supply on Chip

QAM	Quadrature amplitude modulation
QoE	Quality of Experience
QoS	Quality of Service
RAN	Radio access network
RF	Radio Frequency
RFSOI	Radio Frequency Silicon on Insulator
ROI	Return on Investment
RRU	Remote radio unit
RU	Radio Unit
SDK	Software Development Kit
SDN	Software defined network
SDR	Software defined Radio
SELV	Safety Extra Low Voltage
SiC	Silicon Carbide
SiGe	Silicon Germanium
SINR	Signal-to-interference-plus-noise Ratio
SME	Subject Matter Expert
SNR	Signal-to-noise ratio
SoC	System-on-Chip
SoS	Systems of Systems
SW	Software
TDD	Time Division Duplex
TinyML	Tiny Machine Learning
TLI	Top level inputs
TSV	Through Silicon Vias
TPUT	Throughput
TRP	Transmission and Reception Point
UCNC	User Centric No Cell
UE	User equipment
UI	User Interface
UVP	Under Voltage Protection
WBG	Wide Bandgap
WG	Working group
WPT	Wireless Power Transfer

WoW	Wafer on Wafer bonding
WSN	Wireless Sensor Networks
WuR	Wake-up Radio

ANTITRUST STATEMENT

Generally speaking, most of the world prohibits agreements and certain other activities that unreasonably restrain trade. The IEEE Future Networks Initiative follows the Anti-trust and Competition policy set forth by the IEEE Standards Association (IEEE-SA). That policy can be found at: https://standards.ieee.org/wp-content/uploads/2022/02/antitrust.pdf