CHARTING THE FUTURE OF INNOVATION | VOLUME 107 I 2022

ERICSSON TECHNOLOGY Review

SPOTLIGHT ON

SUSTAINABLE NETWORKS

ERICSSON


-SHOW BUSINESS
-NETWORK
-MUSIC
-CINEMA
-BUSINESS/FINANCE
-WORLD NEWS
ANALYSIS
NEWS
-VIDEO
-MUSIC
-FILMS
-SEARCH
-CONTACTS
-MESSAGES
-EUROPE
-AMERICA
-ASIA
-AFRICA
SEARCH
SCANNING

Ericsson Technology Review brings you insights into some of the key emerging innovations that are shaping the future of ICT. Our aim is to encourage an open discussion about the potential, practicalities, and benefits of a wide range of technical developments, and provide insight into what the future has to offer.

ADDRESS
Ericsson
SE -164 83 Stockholm, Sweden
Phone: +46 8 7190000

PUBLISHING
All material and articles are published on the Ericsson Technology Review website: www.ericsson.com/ericsson-technology-review

PUBLISHER
Erik Ekudden

EDITOR
Tanis Bestland (Nordic Morning)

EDITORIAL BOARD
Hans Bergström, Magnus Buhrgard, Torbjörn Cagenius, Magnus Ewerbring, Dan Fahrman, John Fornehed, Kjell Gustafsson, Jonas Högberg, Sara Kullman, Johan Lundsjö, Cecilia Nyström, Håkan Olofsson, Patrik Roseen and Robert Skog

ART DIRECTOR
Carola Pilarz (Nordic Morning)

PROJECT MANAGER
Susanna O'Grady (Nordic Morning)

LAYOUT
Carola Pilarz (Nordic Morning)

ILLUSTRATIONS
Jenny Andersén (Nordic Morning)

SUBEDITORS
Ian Nicholson (Nordic Morning)
Paul Eade (Nordic Morning)

ISSN: 0014-0171
Volume: 107, 2022

FIGHTING CLIMATE CHANGE
WITH ICT INNOVATION AND ENHANCED ENERGY PERFORMANCE

■ **THE LATEST** Intergovernmental Panel on Climate Change report has once again raised the alarm about the critical need for immediate and drastic climate action, highlighting the indisputable link between climate change and human activity. The science is clear: we must limit global warming to 1.5°C above preindustrial levels to minimize the damage and avoid reaching an irreversible tipping point. To do so, we need to halve global greenhouse gas emissions by 2030 and reach net zero before 2050.

Addressing the climate crisis requires transformational changes that will rely on new and emerging technologies, ceaseless innovation and extensive collaboration across industry sectors. 5G has a fundamental role to play, particularly with respect to the rollout of digital technologies and the digitalization of industrial processes. The ability to support a variety of virtual reality and augmented reality use cases that can dramatically reduce business travel is just one of many examples of the broader impact that networks can have both today and in the future.

Today's networks are already playing an outsized, multiplier role in tackling climate change by helping other sectors cut their emissions. The Exponential Climate Action Roadmap confirms this and estimates that ICT technology has the potential to reduce global carbon emissions by up to 15 percent by 2030. The growing use of advanced 5G-enabled technologies in smart factories and other industrial settings opens up the possibility that our positive impact can be even greater than predicted.

“ENERGY-EFFICIENT AND SUSTAINABLE NETWORKS ARE A NECESSITY IN THE FIGHT AGAINST CLIMATE CHANGE”

At Ericsson, we are keenly aware that energy-efficient and sustainable networks are a necessity in the fight against climate change, and we are committed to using our specialist expertise to enable other industry sectors to transition toward a low-carbon economy. Our extensive experience of network operations and optimization is an invaluable asset to us in our work to continuously identify new opportunities to minimize energy consumption in mobile networks while maintaining a consistently high quality of experience.

The lean design of the New Radio (NR) standard represents a major improvement compared with LTE, enabling unprecedentedly low energy consumption in live 5G networks. Lean design principles are the foundation for improving radio-access network energy performance in the years and decades ahead. It is of the utmost importance that the progress that we made in NR with respect to energy performance can be extended into future standardization and products.

Figuring out how to increase the use of renewable energy to power networks is a critically important aspect of our research. Most recently, we have been greatly encouraged by the results of a trial on Deutsche Telekom's (DT) 5G-enabled network, in which we worked closely with DT to efficiently harness solar and wind energy, while simultaneously optimizing power supply and demand.

Looking ahead, we believe that the main energy performance challenge will be scaling processing with traffic to meet the digital processing needs of high-performing networks with larger antenna arrays and bandwidths, and shorter processing requirements. In the 6G timeframe, we are advocating for a further reduction in fixed idle-mode energy usage and peak power requirements – changes that will enable the use of smaller, lighter products and support novel deployment solutions in 6G networks.

We hope this special issue of our magazine helps you and your organization chart a more sustainable and energy-efficient path forward. Please share it with your colleagues and business partners to help us spread the message as widely as possible. You can find both PDF and HTML versions of all the articles at: www.ericsson.com/ericsson-technology-review

ERIK EKUDDEN
SENIOR VICE PRESIDENT,
CHIEF TECHNOLOGY OFFICER AND
HEAD OF GROUP FUNCTION TECHNOLOGY

Energy-optimized network modernization

Minimizing energy consumption and maximizing the use of sustainable energy sources at ICT sites requires a transition toward a smart energy setup and a holistic approach to energy management that includes close collaboration with the energy utility sector.

ANETTE HÖGLUND, JOHAN PETTERSSON, HELENE HALLBERG, LARS HUMLA, ERIK SANDERS

The evolution toward a smart energy setup at Information and Communications Technology (ICT) sites is a key enabler for the creation of resilient, robust and cost-efficient radio access networks (RANs). At the same time, this approach also opens up new income streams for ICT site owners to support the electricity power system (power grid) with ancillary services.

■ While energy efficiency has always been an important consideration in network development, 5G includes an exceptionally powerful set of tools that communication service providers (CSPs) can use to minimize network energy consumption. As impressive as the 5G energy narrative is, however, it is only one part of the solution that the ICT industry needs to overcome in its energy challenge. Fully addressing the challenge will require broad action across the entire ICT network that includes the evolution of the energy setup to ensure alignment with the needs of the smart grid.

There are multiple interdependencies between ICT networks and the power grid. Alignment and collaboration between ICT site owners and power companies will open up possibilities for new services and income for ICT site owners, while simultaneously contributing to a more resilient and robust power grid.

Growing awareness of climate change [1] is driving a mindset shift in the ICT industry toward a more holistic approach to benchmarking the energy capability needs in a network. With this comes a desire to modernize telecom equipment to increase energy efficiency. Beyond using 100 percent renewable energy, the shift also includes new energy perspectives that focus on optimizing the energy setup at sites. The aim is to ensure high network availability for service assurance and enable intent-based energy management with the ultimate goal of achieving net-zero energy cost.

As the ICT network matures, ICT site owners have an ongoing opportunity to reflect on potential improvements from an energy setup perspective. While a wider range of options are available when building new ICT sites, there are many ways to reduce energy costs at legacy sites as well.

Figure 1 illustrates Ericsson's comprehensive energy perspective on network modernization where the focus evolves during the journey toward net-zero energy cost. The ICT site owner's focus will shift from pure energy consumption reductions in the first phase to enabling smart energy in the second, and ultimately gaining new income in the third.

As highlighted in Figure 1, there are five evolution steps toward net-zero energy cost and beyond:

1. Explore possibilities to reduce energy consumption in each network
2. Build and optimize for a low network energy consumption
3. Ensure autonomous, resilient and holistic energy usage in operations
4. Build power grid and ICT interdependency to ensure high availability
5. Open a new source of income by provisioning ancillary services for the power grid, such as an open energy bid and clearing support.

These steps provide opportunities to reduce energy costs as well as opening up the potential to turn energy generation into a new source of income.

Key definitions

ICT site – a distributed edge compute site with radio communication capabilities. Future sites that are located far out in the telecom network will include the kind of compute capacity that is currently only available in large, centralized data centers.

ICT site owner – ICT sites have traditionally been owned by CSPs, but it is increasingly common in many countries for ISPs to own ICT sites today. The meaning of site ownership varies from case to case, but it often includes responsibility for supporting equipment such as the shelter, energy setup and climate control, as well as application equipment such as radio and data servers.

Transmission system operator – a natural or legal person who is responsible for operating, ensuring the maintenance of and, if necessary, developing the transmission system in a given area.

Distribution system operator – a natural or legal person who is responsible for operating, ensuring the maintenance of and, if necessary, developing the distribution system in a given area.

Energy system setup – a technical solution at an ICT site that provides the applicable type of energy and the interface to the power grid, which can provide energy backup and relies on autonomous central or distributed energy management.

Net-zero energy cost – the reduction of an ICT site owner's energy bill (the opex cost) to zero.

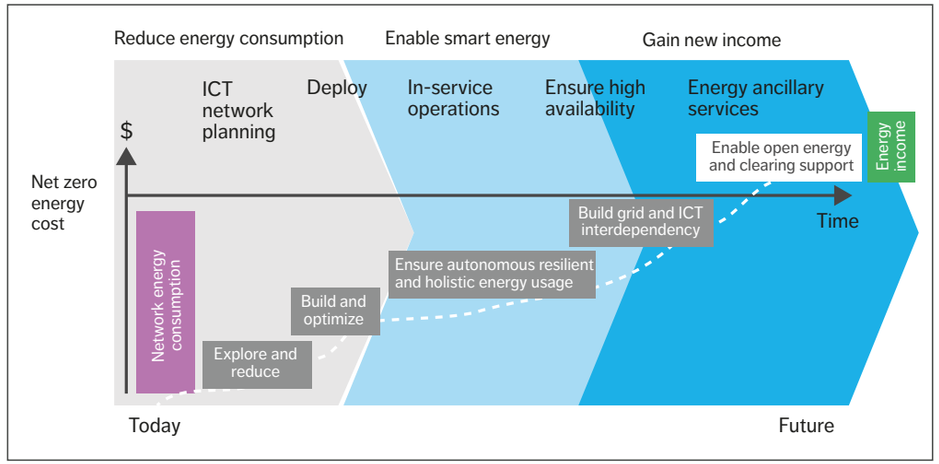

Figure 1 Ericsson's comprehensive energy perspective on network modernization

Approaches to reducing energy consumption

There are four primary aspects that must be considered when reducing network energy consumption:

1. Hardware modernization and new software features
2. Energy-saving software features and artificial intelligence (AI) enabled automation
3. Network energy optimization services
4. The support of digital twins.

Hardware modernization and new software features

The biggest reductions in network energy consumption to date have resulted from the introduction of new hardware and the removal of old systems. Major hardware modernization projects are usually triggered by new releases of spectrum and/or new releases of network technology (such as 5G). Looking ahead, one of the best ways that CSPs and infrastructure service providers (ISPs, also known as tower companies) can speed up their progress toward net-zero energy cost is to consider energy consumption when performing traffic simulations at the network planning stage.

Energy-saving software features and AI-enabled automation

The continuously expanding scope and growing complexity of 5G applications is putting greater demands on networks to deliver high availability, ultra-reliability, low latency and high security. More automation is essential to cope with these demands. To be successful, this automation must be built on a solid understanding of site capability, capacity, performance, user demand, weather conditions and energy availability at any point in time.

The energy-saving software features in 5G enable CSPs and ISPs to minimize energy consumption during low traffic periods, reduce energy waste during peak traffic hours and turn equipment on and off at exactly the right time [2]. Software functionality with AI agents at the site will improve the CSP's/ISP's understanding of the energy perspective, identify the root causes of inefficiencies and take autonomous action to turn assets on and off when needed.

More advanced software functions will make it possible to reduce manual work, find faults early and maximize network availability. The use of sensors at each site to measure temperature, humidity and electrical characteristics will provide valuable information about the local conditions that make it possible to optimize energy utilization and maximize the life cycle of site equipment.

Intelligent agents capable of handling complex processes are needed to optimize trade-offs between long-term benefits of the agent's behavior and short-term benefits from the immediate steps to be taken; for example, how to optimize a network in multiple steps and how to operate a network. These processes must be learned autonomously, without the intervention of a human domain expert.

Network energy optimization services

Network energy optimization services help CSPs/ISPs explore the full potential of reducing energy consumption in each technology network (4G and 5G, for example). CSPs/ISPs gain the confidence to activate RAN energy-efficiency features by learning how optimization cycles, radio KPI monitoring and the tracking of crowd-sourced consumer experience data ensure that the network's Net Promoter Score remains high. A network energy optimization service can also analyze the system-embedded energy metering and network-wide hardware installed base to evaluate the need for hardware modernization.

> DIGITAL TWINS ARE AN IDEAL SOLUTION FOR ICT SITE OWNERS THAT WANT TO AVOID THE EFFECTS OF ERRATIC INITIAL EXPLORATIONS ON LIVE MOBILE NETWORKS

The support of digital twins

Digital twins are an ideal solution for ICT site owners that want to avoid the effects of erratic initial explorations on live mobile networks. A good example of where they can be useful is in the case of network parameters that are configured on a per-cell level but may have a strong impact on the performance of surrounding cells. A change in any of these parameters also affects users served by the surrounding cells.

Finding the optimal configuration for these types of parameters is a complex exercise. Digital twins make it possible to test on an external entity that mimics the behavior of the live network. Once the agent has acquired all the necessary knowledge from the digital twin, the achieved policy can be safely applied to the live network.

Looking ahead, AI will be fundamental in the digital transformation of many sectors as they optimize or introduce new services to use energy more efficiently, to plan and predict maintenance.

Optimizing the energy setup at ICT sites

Energy experts predict more frequent disturbances in the power grid in the years ahead due to extreme conditions such as wildfires, heavy rain and flooding. Increasing reliance on renewable energy sources and the growing popularity of electric vehicles is also expected to have a volatile impact on the power grid. With this in mind, CSPs/ISPs that want to ensure high network reliability without increasing the energy cost should review the capabilities of the

Terms and abbreviations

AI – Artificial Intelligence | **CSP** – Communication Service Provider | **DC** – Direct Current | **ICT** – Information and Communications Technology | **ISP** – Infrastructure Service Provider | **Li-ion** – Lithium-ion | **PV** – Photovoltaic | **RAN** – Radio Access Network | **VDC** – Volt Direct Current

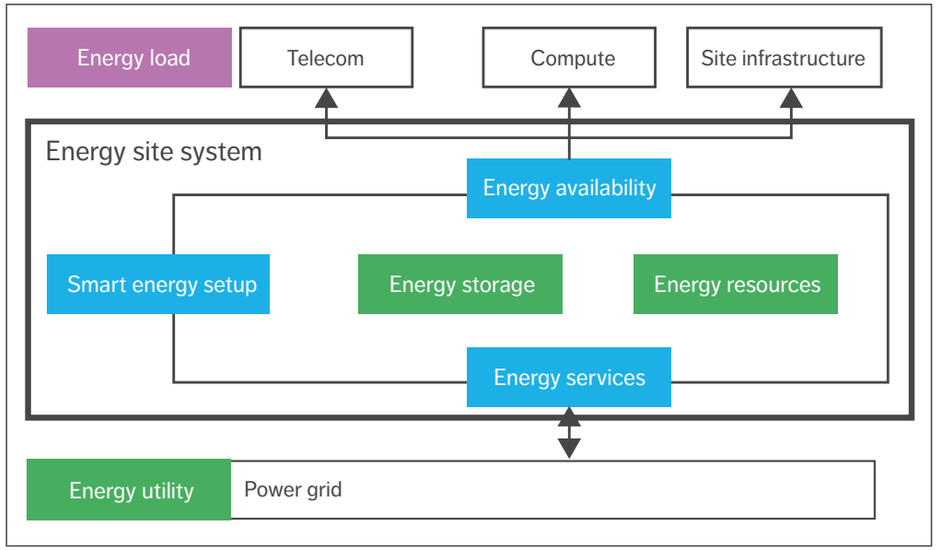

Figure 2 Energy system setup at an ICT site

energy setup at their ICT sites and look for opportunities to improve their resilience.

There are three core actions that ICT site owners can take to ensure the provision of a robust and reliable ICT network, while simultaneously minimizing their impact on the climate and the environment:

1. Introduce local renewable energy generation (solar and/or wind) and add services such as weather forecasting where AI predicts the control of energy storage to optimize the usage of renewables.
2. Improve the overall efficiency of the power system by reducing the number of conversion steps, improving the efficiency/unit and/or increasing system voltage.
3. Establish a smart energy setup capability, where the energy availability at the site is based on the power grid, storage and resource status, and optimize the utilization of site assets.

Figure 2 illustrates the energy system setup that we recommend for ICT sites. The creation of a robust autonomous energy solution makes it possible for ICT site owners to ensure energy availability even at the sites that are furthest out on the edge.

Traditionally the telecom network has only had radio and core equipment at network sites and has relied on data centers for compute capacity. As more advanced use cases are introduced in 5G and beyond, the need for compute at network sites on the far edge will arise, and network sites will turn into ICT sites. The implication of this is that the energy site system needs to evolve in parallel to ensure the allocation of the available energy is done in an intelligent way. A smart energy setup will be key to achieving resilient and sustainable energy usage.

Energy storage and resource technology

The power grid is the most common source in the energy setup at today's ICT sites, with complementary sources utilized when and where required. At off-grid sites and for sites with seasonal variations, diesel-powered combustion engines are still the main complementary source, although ethanol and liquefied petroleum gas are replacing diesel in some countries. In locations with available space, photovoltaic (PV) silicon panels are sometimes used. While lead acid batteries continue to be the most common choice for backup, batteries based on lithium-ion (Li-ion) technology are becoming increasingly popular.

To make the site energy setup more sustainable, multiple evolutions paths are possible, but some of them may have limitations. To reduce the usage of diesel, fuel cells can be used with fuels in gas, liquid or solid form. Combustion engines will also use fuel variants that have lower greenhouse gas emissions.

The need for Li-ion batteries in the coming years will exceed the production capacity and that will open up for new battery chemistries. PV panels will have improved efficiency by tandem cells or other chemistries. Even if the efficiency of low-cost, thin-film PV technologies is lower than traditional technologies, they could gain market share and be suitable for areas with fewer site restrictions.

Future evolution steps in excessive solar locations will hopefully generate hydrogen that can be stored and used as long duration backups. Combustion engines that can run on hydrogen are also expected to emerge. A future wind turbine design that can harvest low wind speeds while still being robust enough for high winds would make wind turbines a potential contributor to the energy mix at ICT sites.

The energy setup for ICT equipment includes power supply units with remote access and has intermittent peak shaving capability. The energy evolution will provide the hybrid control and intelligence to impact the peak load. The system operational voltage may move from 54VDC (volt direct current) to 57VDC, including DC/DC (direct current) backup. The next step is to evolve from 48VDC to nominal 400VDC. The higher system voltage introduces a reduced size of the cable area and decreases the deployment cost of DC distribution at ICT sites.

> THE ICT NETWORK HAS THE POTENTIAL TO PLAY A SIGNIFICANT ROLE IN REDUCING ENERGY CONGESTION AND ENSURING STABLE FREQUENCY

Collaborating for greater resilience and robustness

The transformation of power generation from a centralized model to a decentralized one, characterized by the generation and storage of vast amounts of renewable energy sources, creates new challenges that require reforms to the electricity market. Looking ahead, it will be the responsibility of the transmission system operator [3] not only to ensure resilience and robustness in the power grid, but also to enable interactions in a wider electricity sector/community.

Government strategies will have a big impact on how energy markets can develop flexibility in energy services at international, national and regional levels. Given the opportunity, the ICT network has the potential to play a significant role in reducing energy congestion and ensuring stable frequency.

Collaboration will be required to ensure an optimized and interoperable path that enables adaptive energy consumption by the infrastructure at the ICT sites to contribute to balancing energy services at the national level.

ICT and power grid interdependency

Many of the requirements on the operational performance of ICT networks are similar to those of energy utilities, particularly the new decentralized power grid from the transmission and distributed system operators. Both need to provide capabilities that enable the generation and storage of renewable energy, boost operational efficiency and increase resiliency and reliability.

In terms of renewable energy and storage

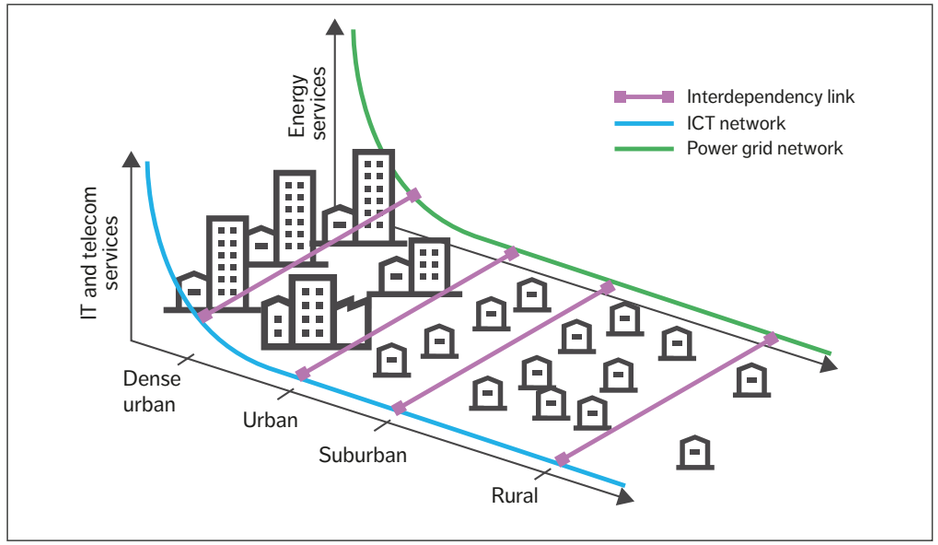

Figure 3 The interdependency between ICT and the power grid

resources, there is a need to adapt and utilize what is available at the edge or local area with small-scale production intermittent availability and reduce/increase energy consumption from time to time. Operational efficiency requires capabilities that optimize assets and utilization, as well as providing high network and service availability. Automation to mitigate service disruptions and power outages and redirect traffic/power flow in case of disasters is key to ensuring resiliency and reliability.

Figure 3 illustrates the interdependency between the ICT network and the power grid. Significant mutual gains can be achieved by enabling communication and service collaboration between the two. Reliability, availability and resilience are essential to ensure high network availability and service assurance, as well as increasing robustness and improving maintainability, which reduce total cost of ownership in the long run.

> SIGNIFICANT MUTUAL GAINS CAN BE ACHIEVED BY ENABLING COMMUNICATION AND SERVICE COLLABORATION BETWEEN THE ICT NETWORK AND THE POWER GRID

The ICT network and the power grid can mutually benefit in their respective service sectors by sharing relevant performance and status information at multiple intersection points, particularly when the situation at sites is challenged by extreme situations like wildfires. This approach enables both networks to maintain service in the local geographical area, ensuring their ability to support local residents and businesses when they are needed most.

Creating ancillary services for the power grid

ICT site owners can start now to prepare for the possibility to support the power grid's needs of different types of ancillary services, thereby opening up a new source of income [4]. This can be done in five steps:

1. Establish an energy profile for the site and optimize site power system capacity.
2. Introduce a segmentation usage capability of energy storage for different purposes such as ICT network backup and ancillary services. Consider adding storage to increase the income potential from the energy ancillary service.
3. Optimize the ICT site's own energy consumption and local energy production while balancing it with energy storage at the site.
4. Manage the energy consumption of multiple grid-connected ICT sites as one entity to maximize the impact on the power grid.
5. Utilize energy arbitrage by producing energy at the ICT sites and sharing it with the power grid when needed.

An efficient deployment includes AI-based automation to best optimize business intents and support the smart energy capabilities for both the ICT network and the power grid [5].

Supporting the digital transformation of the energy sector

Energy utilities currently use a variety of telecommunication technologies including fixed, fiber, microwave and mobile. In all cases, the communication has to be secure and highly available, as energy services must be reliable and in balance, according to the redundancy policy in operation.

Looking ahead, it is clear that the digital transformation of the energy utility sector would benefit from the introduction of LTE/5G technology [6]. These technologies make it possible for ICT sites to provide energy utilities with insights for the geographically distributed low-voltage power grid with characteristics of outages and performance degradation. Further, with the help of AI agents, LTE/5G networks can help predict performance degradations/deviations and their potential impact.

5G networks are particularly well suited to smart-grid management [7]. The 5G network can be either commercial or private depending on preferences on the energy utility. With a 5G network in place, an energy utility can ensure that its critical functions are implemented in a resilient, reliable and secure way. A 5G network is also the most efficient way for an energy utility to establish a cost-effective strategy for connecting grid assets over a wide geographical area. All new devices in the power grid use wireless technology for communication. Long-term, stable technology evolution is guaranteed, in alignment with the long life-cycle considerations of the telecom segment.

Conclusion

Whenever the need or desire to modernize ICT sites arises in a local or regional area, new energy-related opportunities arise with it. A holistic approach is key to maximize the potential for energy consumption reductions in existing and future ICT networks across the whole service life cycle. To be successful, it is essential that ICT site owners explore and identify candidates to reduce energy consumption as early as possible – ideally in the network planning stage. During the deployment stage, it is crucial to build and optimize for low energy consumption. With respect to in-service operations, artificial-intelligence-based automation, together with active network energy optimizations, make it possible to minimize energy consumption without negatively impacting quality of service (QoS) in the networks.

As they plan for the future, ICT site owners also need to consider the fundamental changes that are taking place in power generation and distribution. These changes, along with new energy consumption patterns, have created multiple challenges for the electricity utility sector. Mitigation alternatives will be costly and take a long time to deploy. By investing in a smart energy setup at ICT sites, ICT site owners can gain access to a resilient energy source that reduces their energy costs to net zero, while simultaneously opening up a promising new income stream by connecting with the smart grid.

Further reading

- **Ericsson Technology Review, Building robust critical networks with the 5G system, available at:** *https://www.ericsson.com/en/reports-and-papers/ericsson-technology-review/articles/building-robust-critical-networks-with-the-5g-system*
- **European Association for the Cooperation of Transmission System Operators for Electricity website, available at:** *https://www.entsoe.eu/*
- **Ericsson blog, Digitalization and 5G climate action, available at:** *https://www.ericsson.com/en/blog/2021/1/digitalization-5g-climate-action*
- **IEA report, Energy service companies, available at:** *https://www.iea.org/reports/energy-service-companies-escos-2*
- **Ericsson, Pioneering a sustainable future, available at:** *https://www.ericsson.com/en/about-us/new-world-of-possibilities/pioneering-a-sustainable-future*
- **Ericsson, Artificial intelligence, available at:** *https://www.ericsson.com/en/ai*
- **Ericsson, Ericsson Silicon, available at:** *https://www.ericsson.com/en/ran/ericsson-silicon*
- **UBBA, The importance of private broadband for grid modernisation, available at:** *https://www.ubba.com/ubba-resources/smart-energy-international-the-importance-of-private-broadband-for-grid-modernisation/*
- **UBBA, Rural broadband playbook, available at:** *https://www.ubba.com/ubba-resources/rural-broadband-playbook-2/*

References

1. **GSMA, Mobile Net Zero: State of the Industry on Climate Action 2021, available at:** *https://www.gsma.com/betterfuture/wp-content/uploads/2021/04/Mobile-Net-Zero-State-of-the-Industry-on-Climate-Action.pdf*
2. **Ericsson Technology Review, Ensuring energy-efficient networks with artificial intelligence, April 13, 2021, Vandikas, K; Hallberg, H; Ickin, S; Nyström, C; Sanders, E; Gorbatov, O; Eleftheriadis, L, available at:** *https://www.ericsson.com/en/reports-and-papers/ericsson-technology-review/articles/ensuring-energy-efficient-networks-with-ai*
3. **Directive 2019/944 of the European Parliament and of the Council of 5 June 2019 on common rules for the internal market for electricity and amending directive 2012/27/EU, available at:** *https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32019L0944&from=EN*
4. **Ericsson white paper, Ancillary services to utilities using mobile network power infrastructure, Eleftheriadis, L; Pettersson, J; Hallberg, H; Palma-Serrano, M, available at:** *https://www.ericsson.com/en/reports-and-papers/white-papers/balance-smart-grids-with-5g-backup-for-utilities*
5. **IEA, IEA Energy and Carbon Tracker 2021, available at:** *https://www.iea.org/data-and-statistics/data-product/iea-energy-and-carbon-tracker-2021?utm_campaign=IEA+newsletters&utm_source=SendGrid&utm_medium=Email*
6. **Ericsson, Energy utilities, available at:** *https://www.ericsson.com/en/industries/energy-utilities*
7. **UBBA, Utility Networks: The inherent risks of doing nothing, available at:** *https://www.ubba.com/wp-content/uploads/2021/12/UBBA-Position-Paper-Inherent-Risks-of-Doing-Nothing-Oct-2021COMP.pdf*

THE AUTHORS

Anette Höglund

◆ is the strategy execution driver for energy at the Ericsson CTO office. She has combined the perspectives of new business opportunities, product management and deployments in the field as well as development in different areas since she joined Ericsson in 1995. Höglund aligns Ericsson's technology strategy for energy solutions and evaluates the energy opportunity space with customer project collaborations. She holds an M.Sc. in management from Henley Business School, University of Reading, United Kingdom.

Johan Pettersson

◆ joined Ericsson in 1993 and currently serves as the strategic product manager for remote site management. For the past 10 years, he has mainly been working with power and energy-related products linked to site equipment and the ability to monitor and reduce energy consumption at sites. He also drives Ericsson's

business-wide energy program, whose purpose is to execute the technology strategy solution approved by the Ericsson Technology Board. Pettersson holds a B.Sc. in electrical engineering from Uppsala University, Sweden.

Helene Hallberg

◆ is a senior specialist in energy-efficient radio systems who joined Ericsson in 1988. She is involved in energy-efficiency development and collaboration across the company, as well as in energy-related regulatory discussions and standardization activities. She is also a delegate in the International Electrotechnical Commission Technical Committee standardization of electricity supply systems encompassing transmission and distribution networks with their network interfaces. She has also filed a patent in the energy-efficiency area. Hallberg studied electrical engineering at a technical college in Stockholm, Sweden.

Lars Humla

◆ is a senior specialist in battery and solar systems at Ericsson. After working with battery solutions for mobile and fixed systems for the Swedish Defense Material Administration, Humla joined Ericsson in 1996. He spent a brief period working for Emerson and as an agent for a battery technology company from Switzerland before rejoining Ericsson in 2006. He has been active in the Swedish standardization committee for batteries since 2010. Humla studied electrical engineering at a technical college in Stockholm, Sweden.

Erik Sanders

◆ is a product manager for AI and automation. He joined Ericsson in 2005 as an engineer in the 3G RAN area and continued with hardware development for radio base stations. In his current role, he drives innovation programs for Ericsson in the area of machine learning and reasoning. Sanders holds an M.Sc. in mobile communication from Linköping University, Sweden.

Improving energy performance in 5G networks and beyond

Continued focus on energy performance in 5G and 6G development will be essential to enable new deployment scenarios with smaller and lighter telecom equipment, as well as minimizing the climate impact of mobile networks.

CECILIA ANDERSSON, JONAS BENGTSSON, GREGER BYSTRÖM, PÅL FRENGER, YLVA JADING, MY NORDENSTRÖM

As the 5G rollout continues and we look ahead to 6G, the exponential growth in processing is going to be one of the biggest challenges from an energy performance perspective.

■ The telecom industry has a long history of prioritizing peak performance and high capacity. Until recently low energy consumption was typically perceived as a nice-to-have extra benefit rather than a crucial feature of state-of-the-art mobile network equipment. As awareness about the need to optimize energy performance in telecom networks continues to grow, several challenges must be addressed.

In 5G networks, digital processing in base stations can increase more than 300 times compared with early Long Term Evolution (LTE) products, primarily due to an increasing number of antenna branches, broader bandwidths and shorter transmission time intervals (TTIs). This increase is expected to become even larger in 6G. To handle this increase responsibly from an energy consumption perspective, it is essential that the future 6G standard is as lean as possible. Most importantly, the amount of mandatory and always-on signaling must be kept to a minimum, which will enable underlying components and subsystems to provide sufficiently high levels of load dependence in their energy consumption.

The percentage of energy consumption originating from compute and digital silicon will continue to rise as the New Radio (NR) rollout continues. The trends of increased bandwidth, more antenna ports and shorter TTIs are driving an exponential increase in processing needs on the digital frontend, which propagates into the radio unit, beamforming and layer one processing. Processing needs for layer two, packet processing and control functions are also increasing, but not as quickly.

What is energy performance?

In existing mobile systems, where energy consumption has low dependence on the load and traffic growth is high, an energy-efficiency metric is not particularly useful. Even if no effort is made to reduce energy consumption, traffic growth will cause energy efficiency to improve from one year to the next.

The term energy performance broadens the focus beyond energy per bit to total energy consumption, highlighting the similarities between achieving high system performance and low energy consumption. Optimizing energy performance means minimizing the energy consumption for a set of performance requirements (user throughput, capacity, latency and so on). The concept of energy performance makes the trade-off between energy consumption and performance requirements transparent.

Energy performance is relevant in three separate contexts – economy, ecology and engineering – that have different stakeholders and use different terminology. In the economic context, optimizing

> THE TERM ENERGY PERFORMANCE BROADENS THE FOCUS BEYOND ENERGY PER BIT TO TOTAL ENERGY CONSUMPTION

energy performance results not only in lower operating expenses for mobile network operators, but also in the potential for lower capital expenses due to the use of smaller and lighter equipment. This equipment enables new and simplified deployments as well as less-costly power supply and energy storage solutions. Low energy consumption is therefore of particular importance when discussing the costs for enhanced availability and enhanced national security.

In the ecological context, the value of optimizing energy performance is best demonstrated by the fact that about 94 percent of Ericsson's total carbon emissions impact is associated with the operation of our products [1].

The engineering context is arguably the most underestimated and underutilized aspect of energy performance. A sharp focus on reduced energy consumption in the product engineering phase will yield desirable reductions in product size and weight. Tighter energy performance product requirements drive technical, managerial and organizational innovation in a greater change process that leads to closer collaboration over multiple areas of technology.

Terms and abbreviations

AI – Artificial Intelligence | **C-RAN** – Centralized RAN | **CSI** – Channel-State Information | **D-RAN** Distributed RAN | **KPI** – Key Performance Indicator | **LTE** – Long Term Evolution | **MB** – Megabyte | **Mbps** – Megabits Per Second | **MHz** – Megahertz | **MIMO** – Multiple-Input, Multiple-Output | **ms** – Millisecond | **NR** – New Radio | **PA** – Power Amplifier | **PRB** – Physical Resource Block | **RAN** – Radio Access Network | **RAT** – Radio Access Technology | **RF** – Radio Frequency | **Tbit** – Terabit | **TTI** – Transmission Time Interval | **UE** – User Equipment | **W** – Watt

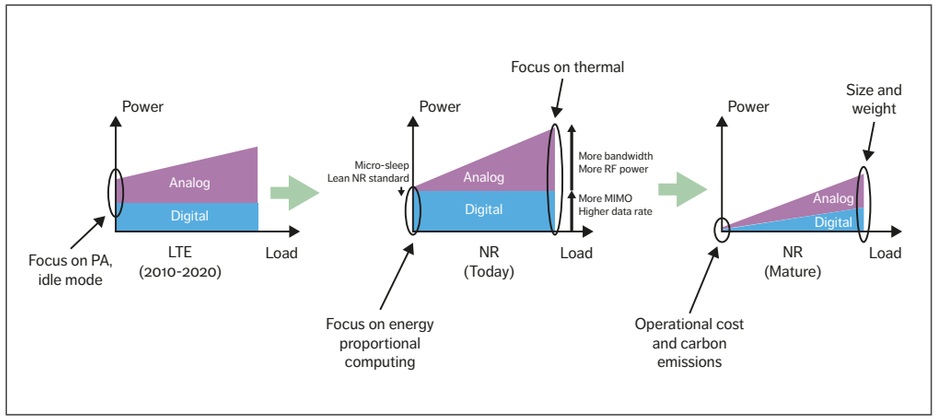

Figure 1 The energy performance journey of mobile networks

Depending on the context, the impact of energy usage can be expressed in terms of money, carbon emissions, thermal design constraints and so on. But regardless of how the benefit is viewed, the solutions required to reduce energy consumption are very similar. It is important, however, to differentiate between targets whose purpose is to reduce average energy use (mostly relevant for operational costs, carbon emissions and the like) and those aimed at reducing peak energy use (mostly relevant for product dimensioning, thermal design, size and weight). Fortunately, though, effective solutions for energy reduction often reduce peak and average energy use simultaneously.

The network energy performance journey

Figure 1 illustrates the energy performance journey of mobile networks. The graph on the left shows the power consumption of a typical LTE base station in the period 2010-2020. The graph in the middle shows the power consumption of a typical New Radio (NR) base station today. The graph on the right shows the projected power consumption of a mature NR base station beyond 2025.

In LTE, the energy consumption of the radio access network (RAN) was dominated by base stations that comprised around 80 percent of the RAN electricity use. Furthermore, within each base station, around 80 percent of the energy consumption was used in the power amplifiers (PAs) [2]. During this time the main energy performance goal was to reduce the idle-mode consumption of the PAs.

Network traffic was much lower in the early days of LTE. In a countrywide network, only 5 percent of the physical resource blocks (PRBs) were typically used for data traffic, while around 95 percent of the PRBs were empty. The high static energy consumption and low average traffic resulted in very low load dependency in the network, where the additional energy increase due to the traffic in the network was well below 2 percent of the total energy used [3].

In order to improve the load dependency, we introduced micro-sleep transmission, an energy-saving feature that deactivates and reactivates PAs in microseconds. Micro-sleep transmission is effective at all times when there are no transmissions from the base station. However, it could be even more effective if the LTE standard was not filled with mandatory and always-on reference signals that cannot be turned off, even when there is no traffic. The obvious solution to this was to redesign

> OUR RESEARCH SHOWS THAT IT IS POSSIBLE TO MOVE FROM TODAY'S SITUATION TOWARD A MORE ENERGY-LEAN TOMORROW

the physical layer so that only a minimum of signaling would be necessary when there is no data to transmit, but this required a new standard.

When 5G NR was developed, we ensured that its physical layer had an ultra-lean design [4]. This design makes features such as micro-sleep transmission much more efficient, and the effect on live networks is already significant [5]. Micro-sleep transmission has had a major impact over the course of the past decade by dramatically decreasing the energy consumption in the analog radio parts of base stations.

5G NR is also designed for massive MIMO (multiple-input multiple-output), and it supports wider bandwidths than 4G LTE. A typical LTE base station has two transmit and receive branches, 20MHz of spectrum, and the digital processing time in the base station is 1ms (corresponding to one TTI). Early NR products have 64 antenna branches, support 100MHz of spectrum and have a TTI of 0.5ms. This implies that digital processing in radio base stations needs to increase 320 times compared with early LTE products.

In today's NR products, the energy that digital components use can be as large, or larger, than the power used by the analog components (mainly PAs). The additional bandwidth and additional antennas were also introduced while keeping the radio frequency (RF) power spectral density constant at around 2W/MHz. While a typical 20MHz LTE base station could deliver 40W RF output power, a 100MHz NR base station can be capable of delivering 320W RF power.

More bandwidth and more RF power increased the peak power usage of analog components, while more antenna branches and more digital processing increased the power consumption of digital components (see the middle part of Figure 1). This massive upscaling of capabilities has resulted in an increasing energy consumption trend in the industry. Fortunately, it is possible to alter this trend with new building practices. Our research shows that it is possible to move from today's situation (the middle part of Figure 1) toward a more energy-lean tomorrow (the right part of Figure 1) by taking the following actions:

1. Improve the load dependence of components (especially digital ones) by significantly improving sleep modes in terms of idle-mode energy usage, reactivation times and granularity. One example of a technique that is currently underutilized in many digital designs is dynamic voltage and frequency scaling (DVFS).
2. Relax the requirements for analog components dynamically (when possible).
3. Implement more power management functions in the network nodes (requires more load-dependent components to be effective).
4. Use artificial intelligence (AI) techniques to achieve network-level optimizations (requires more load-dependent network nodes to be effective).
5. Consider removing support for legacy technologies such as WCDMA (Wideband Code Division Multiple Access) and GSM (Global System for Mobile Communications).

This last point is particularly important, as supporting legacy technologies results in requirement creep and consumes development and testing resources, thereby reducing the resources available for the implementation of power-saving functions.

Successful execution of these actions would not only optimize energy performance; it would also enable operators to reduce total network energy consumption by adding additional capacity as traffic increases, as more capacity results in more idle mode operation [6].

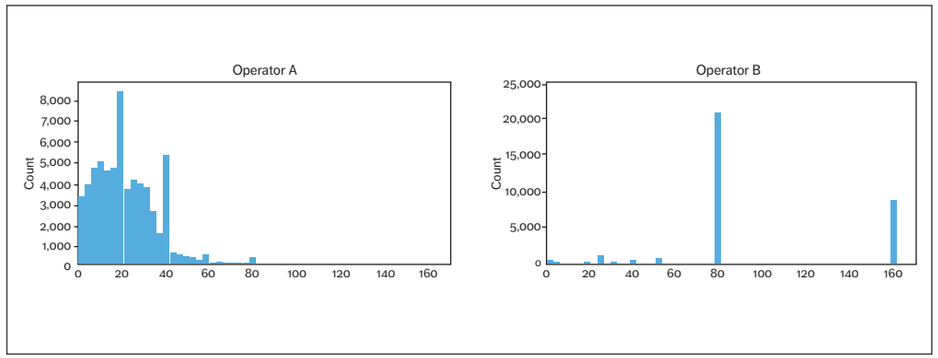

Figure 2 Histograms of configured maximum transmission power for two European operators

Insights from field data

While some operators run their networks with the power settings of every base station fixed at the maximum level, others have more optimized and diverse configurations. *Figure 2* provides histograms of configured maximum transmission power for two European operators. Operator A, on the left, optimized the power levels in its network, while Operator B, on the right, used default maximum power settings.

One study on this topic [7] indicates that the total network energy usage can be reduced by around 10 percent without compromising performance simply by tuning the output-power levels, as shown in Figure 2. There are similar differences in how existing energy-saving features such as MIMO sleep mode, booster carrier sleep, cell deep sleep and the low energy scheduling solution are activated (or not) in different networks. This indicates a clear need for improved tools to optimize the energy performance in existing networks.

Meaningful reference points

Some energy-saving solutions are associated with a negative key performance indicator (KPI) impact. However, we would argue that this often depends on the reference KPI chosen to compare with, or perhaps even more importantly, what target KPI is expected to be fulfilled.

Figure 3 illustrates the total downlink traffic volume (purple) and the mean user throughput (blue) in a real network for one day. Due to variations in traffic volume, the user throughput differs by about a factor of two between the peak traffic hour and the minimum traffic hours. The question is, when evaluating the throughput obtained with an energy-saving feature turned on, should we relate that to the performance during peak hour or during the minimum traffic hour? The large KPI variations that we normally observe in the network are not always desirable, and an energy-saving feature might even be an attractive way to both obtain energy and cost savings, as well as more predictable throughputs in a network throughout the day.

It is not necessary to set the desirable KPI for an energy-saving feature to the minimum KPI obtained at peak hour. It can also be set to other targeted KPIs for a specific network, such as a KPI for which the network has been dimensioned. The network dimensioning KPI target can be regional and/or service dependent. In some areas such as city centers or indoor factories, the requirements may need to be higher, while in other areas they can be more relaxed. In the end, this should be an operator policy decision. One could, for example, allow a fraction of the performance above the minimum performance requirement to be used to reduce energy consumption.

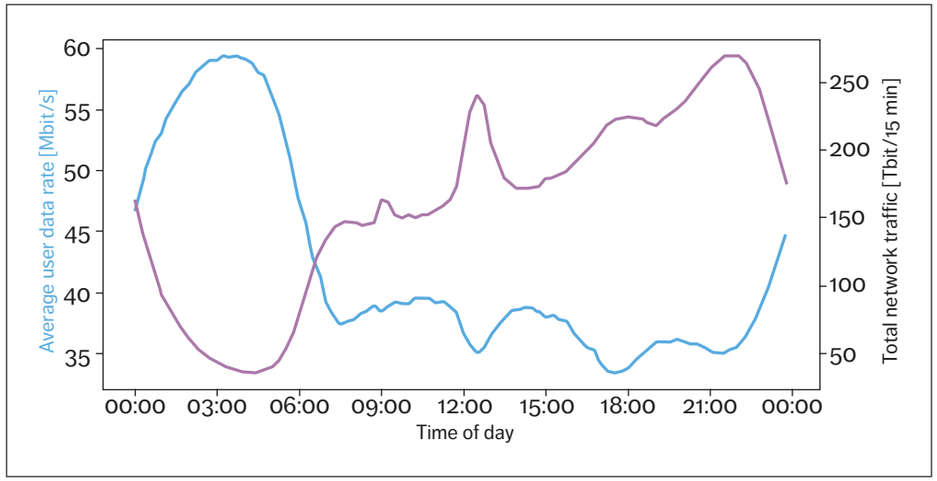

Figure 3 Variations in traffic volume (purple) and user throughput (blue) over a day in a real network

Centralized versus distributed radio access networks

It is often argued that a centralized RAN (C-RAN) is inherently better than a distributed RAN (D-RAN) deployment for enabling low-network power-consumption solutions. As mobile networks are dimensioned for peak traffic, resources are always overprovisioned at some places in the network. As peak traffic occurs at different times of day in different locations of a network, a C-RAN can exploit pooling gains when processing resources can be shared in a common pool. In addition to the processing-pooling gain, C-RAN deployments can also utilize more efficient cooling, power-supply and energy-storage solutions. Based on this, considerable gains have been reported [8].

It is important to understand which parts of a network that can realistically be centralized in a C-RAN deployment and under what circumstances. C-RAN deployments are typically assumed to operate on off-the-shelf hardware, which opens up the possibility to add value with custom-made silicon that can deliver higher load dependence and lower fixed energy costs. Increased latency is a consideration that will limit the centralization of the more time-critical and power-consuming lower-layer processing, such as the digital frontend. The analog (radio) parts must also remain distributed in a C-RAN deployment, as the users will still be distributed.

Traffic: RAN for mobile broadband and beyond

There are two main approaches to reducing energy usage in RAN: rush to sleep, and rate adaptation. The rush-to-sleep approach aims to transmit the data as fast as possible in order to maximize the time in sleep mode. Rate adaptation instead aims to adapt the transmission rate to instantaneous requirements and thereby enable energy savings by under-clocking or deactivating some components while transmitting data.

When looking at traffic statistics in real networks, it is evident that both of these approaches are needed. About 95 percent of all data sessions are small (less than 1MB), and the one percent largest sessions contribute to almost three quarters of the total data volume. For small data sessions, rate adaptation is an effective way of reducing the energy

needs, as large bandwidth and many antenna ports are not required for most sessions. In contrast, the best way to handle large sessions from an energy performance perspective is by combining high data rates with effective mechanisms for the equipment to rush to sleep once the transmission is finalized.

Video traffic is currently estimated to account for 69 percent of all mobile data traffic, a share that is forecast to increase to 79 percent in 2027 [9]. To deliver a satisfactory user experience, a network must be able to play video content on demand, without delays or stalling. Video streaming uses buffering to smooth out throughput variability. As a result, video is a service that can easily be made compatible with power-saving features in the RAN.

The ability to avoid stalling requires rate adaptation on a relatively slow timescale (seconds). Ensuring a sufficiently short time-to-play for video services and time-to-content for web services requires an almost instantaneously available bitrate in the order of 20Mbps [10]. As acceptable time-to-content and time-to-play numbers can be in the order of one to four seconds, "almost instantaneously" needs to be significantly smaller than this (100ms, for example).

For web and video services, it is therefore sufficient for a base station to provide about 20Mbps to any user within 100ms if additional capacity to accommodate the new service can be made available within one second. This would result in around 1.5 seconds time-to-content, which is classified as excellent in the 2025 scenario in a recent Ericsson Mobility Report [9].

Critical machine-type communication and other low-latency services can also be made compatible with RAN-power saving features if the low-latency requirements are managed with care. Low-latency service requirements must be broken down into two additional categories: service-registration latency and protocol ramp-up latency. The activation of a low-latency service requires a quality of service guarantee from the RAN, which includes a setup procedure that can be allowed to take 100ms or more.

In addition, low-latency requirements need to be contained regionally and per band. It is not necessary to provide countrywide support for ultra-low latency to enable factory automation in very limited areas. All the bands, cells and nodes that support low-latency services are not needed all the time. Even the most critical low-latency service does not require a network that is constantly in a high-alert state that prohibits the use of energy-saving rate adaptation or capacity adaptation in the RAN.

In short, it is possible to support all services today and in the future (as far as we know) while deactivating close to 100 percent of the excess capabilities in a base station if the following two conditions are met:

1. At least 20Mbps can be provided to an additional user within 100ms.
2. Any additional required capacity and rate can be made available within one second.

Any active but unutilized spare capacity beyond this is a waste of energy.

Standardization and 6G

System design by standardization is the foundation that enables energy-efficient design of an entire network. But having a good standard is not enough – components and products must utilize the potential that the standard provides. Network management that includes the application of AI tools for load balancing and energy consumption minimization is also essential. To be effective, network management requires a good standard that enables low-energy and load-dependent operation, as well as products that utilize the energy savings that the standard enables.

NR is a lean standard that is already a powerful enabler of low network-energy usage. The most important thing from an energy performance perspective moving toward 6G is to maintain and extend the lean properties on which NR is based, such as enabling up to 160ms of transmission-free periods. For 6G the concept of lean design should be extended to better support network densification and even larger massive MIMO antenna arrays without requiring excessive spatial repetition and beam sweeping of idle-mode signals regarding synchronization, system information and paging.

Today's NR specifications provide signaling support for idle-mode user equipment (UE) to see the full set of beams, bands and nodes that are configured and that can be made available in active mode. The need for this transparency in standards and products is questionable, given the energy cost of associated transmissions and shorter transmission-free periods, as well as reduced sleep possibilities.

In the discussions about next-generation technologies, there is a tendency to allow requirements from one area to propagate into other areas, or even into all areas. The support for extreme capabilities – such as extremely high data rates with corresponding extreme transmission bandwidths, extremely low and predictable latency and extreme reliability – is important to enable the wide range of use cases envisioned for 6G.

At the same time, such capabilities come with a cost in terms of network energy consumption. It is crucial that this cost is limited to the situations where and when the specific capabilities are required. One way of doing this is to prevent requirements for active mode from also applying to idle mode. We would argue that it is worth considering a stricter separation between active and idle mode in future networks.

In addition, efforts should be made to design functionality so that it is self-contained, refraining from the reuse of signals specified for one functionality to support other. This may sound counterintuitive, but experience shows that the associated dependencies between different functionalities often prevent desirable sleep-mode possibilities. A prime example of this is the cell-specific reference signals in LTE that are used for both active-mode demodulation of data and for idle mode UE cell search and mobility. This reuse results in a very high cost for transmitting signals in idle mode.

One of the improvements we would like to see in 6G is the ability to avoid the overhead cost of obtaining channel-state information (CSI) when there is no data to transmit. A more preamble-based design of the physical radio links, where synchronization signals and reference signals for CSI acquisition are transmitted together with the data bursts, would make the cost of all supporting signals (synchronization, CSI acquisition and so on) explicit. More opportunistic scheduling of such supporting signals would become more natural as a result.

The availability of more shared and pooled infrastructure can contribute to further lowering total network energy consumption. Examples of this include multi-radio-access technology (RAT), multi-operator and/or multi-band operation. Although there already are several standardized solutions for operations in such scenarios, further enhancements are worth investigating.

With regard to the role of AI in energy performance improvements, the vast majority of AI functionality is expected to be related to implementation rather than specification. However, standardized AI-supporting functionality related to observability and control that targets energy optimization at network level would be helpful.

As the work on 6G progresses, it is important to be mindful of the risks associated with introducing new functionality into later releases of technology specifications. The potential of lean design can easily be compromised both in standards and implementation, depending on how new functionality and the associated signaling is added.

To summarize, 5G is already a very good standard that enables the implementation of low energy-consuming behavior in mobile systems. Lean design can be further optimized and enhanced in 6G by including additional domains (frequency bands, beams, nodes, RATs, slices and so on).

Conclusion

Lean design is the foundation for improving radio access network energy performance. The lean design of New Radio (NR) standard was a major improvement compared with Long Term Evolution (LTE), enabling unprecedentedly low energy consumption in live 5G networks. It is of the utmost

importance that this progress is not compromised in future standardization and products.

Looking ahead, the main energy performance challenge will be scaling processing with traffic to meet the digital processing needs of high-performing networks with larger antenna arrays and bandwidths, and shorter processing requirements. In the 6G timeframe, we are advocates for a further reduction in fixed idle-mode energy usage and peak power requirements – changes that will enable the use of smaller, lighter products and support novel deployment solutions in 6G networks.

References

1. **Ericsson, Sustainability and Corporate Responsibility Report, 2021, available at:** *https://www.ericsson.com/en/about-us/sustainability-and-corporate-responsibility/sustainability-report*
2. **EARTH Deliverable 2.3, Energy efficiency analysis of the reference systems, areas of improvements and target breakdown , December 2010, available at:** *https://cordis.europa.eu/docs/projects/cnect/3/247733/080/deliverables/001-EARTHWP2D23v2.pdf*
3. **IEEE Xplore, 2014 IEEE 79th Vehicular Technology Conference (VTC Spring), Assessment of Alternatives for Reducing Energy Consumption in Multi-RAT Scenarios, 2015, Frenger, P; Ericson, M, available at:** *https://ieeexplore.ieee.org/document/7022838*
4. **Ericsson, IMT-2020 self-evaluation: Radio Network Energy Performance, 3GPP TSG-RAN WG1 #91, R1-1720954, Reno, US, November 27-December 1, 2017, available at:** *https://www.3gpp.org/ftp/tsg_ran/WG1_RL1/TSGR1_91/Docs/R1-1720954.zip*
5. **Ericsson blog, 5G energy consumption: what's the impact of 5G NR in real networks?, October 8, 2021, Frenger, P; Jading, Y; Bengtsson, J, available at:** *https://www.ericsson.com/en/blog/2021/10/5g-energy-consumption-impact-5g-nr*
6. **Ericsson, More capacity and less power: How 5G NR can reduce network energy consumption, 2019, Frenger, P; Tano, R, available at:** *https://www.ericsson.com/en/reports-and-papers/research-papers/how-5g-nr-can-reduce-network-energy-consumption*
7. **Ericsson Technology Review, Ensuring energy-efficient networks with artificial intelligence, April 13, 2021, Vandikas, K; Hallberg, H; Ickin, S; Nyström, C; Sanders, E; Gorbatov, O; Eleftheriadis, L, available at:** *https://www.ericsson.com/4972d5/assets/local/reports-papers/ericsson-technology-review/docs/2021/ensuring-energy-efficient-networks-with-ai.pdf*
8. **IEEE Xplore, Green Mobile Networks for 5G and Beyond, 2019, Masoudi, M et al, available at:** *https://ieeexplore.ieee.org/document/8786138*
9. **Ericsson Mobility Report, November 2021, available at:** *https://www.ericsson.com/4ad7e9/assets/local/reports-papers/mobility-report/documents/2021/ericsson-mobility-report-november-2021.pdf*
10. **Ericsson blog, Who cares about peak download speeds in 5G?, January 18, 2022, Ludwig, R, available at:** *https://www.ericsson.com/en/blog/2022/1/who-cares-about-peak-download-speeds-in-5g*

THE AUTHORS

Cecilia Andersson

◆ joined Ericsson in 2007 and currently works as a system designer specializing in RAN energy performance. She is a leading promotor of integrating and incorporating the network aspects of energy performance to reduce the total energy consumption of RANs. Andersson holds a Ph.D. in physics from Uppsala University, Sweden.

Greger Byström

◆ is a section manager for radio product development, with extensive experience working with functional systemization of several of the energy performance functions within Ericsson radio products. Before joining Ericsson in 2011, he studied engineering physics at Umeå University, Sweden.

Jonas Bengtsson

◆ is a principal developer specialized in RAN energy performance. One of the drivers behind making Ericsson modems the most energy-efficient in the market 10 years ago, he is now the technical lead of a highly skilled energy performance team that has produced multiple innovations and proof of concepts in the energy performance area. Before joining Ericsson in 2002 he studied computer science at Lund University, Sweden.

Pål Frenger

◆ joined Ericsson in 1999 and currently holds an expert position in RAN energy performance at Ericsson Research. He has worked with radio-network energy performance for more than 10 years, has filed more than 300 patent applications and received the Ericsson Inventor of the Year Award in 2017. Frenger holds a Ph.D. in electrical engineering from Chalmers University of Technology, Gothenburg, Sweden.

Ylva Jading

◆ joined Ericsson in 1999. She currently works as a senior specialist in network energy performance at Ericsson Research. She is responsible for coordinating and leading the early energy performance activities in Development Unit Networks, as well as researching low energy aspects of future 6G networks. Jading holds a Ph.D. in physics from the Johannes Gutenberg University Mainz in Germany.

My Nordenström

◆ is a system developer within Business Area Networks. She joined Ericsson in 2013, and in her current role she primarily focuses on reducing the energy consumption and improving the energy performance for Ericsson's RAN compute products. Nordenström holds an M.Sc. in design and product realization from KTH Royal Institute of Technology, Stockholm, Sweden.

Further reading

- **Ericsson blog, Breaking energy utilization with 5G, available at:** *https://www.ericsson.com/en/blog/5/2022/breaking-energy-utilization-with-5g*
- **Ericsson blog, A holistic approach to address RAN energy efficiency, available at:** *https://www.ericsson.com/en/blog/2021/12/a-holistic-approach-to-address-ran-energy-efficiency*
- **Ericsson, Smart, slim and sustainable 5G networks, available at:** *https://www.ericsson.com/en/ran/elevate-5g*
- **Ericsson blog, Achieving sustainability with energy efficiency in 5G networks, available at:** *https://www.ericsson.com/en/blog/3/2021/1/achieving-sustainability-with-energy-efficiency-in-5g-networks*

Energy-efficient packet processing in 5G mobile systems

Communication service providers around the world are looking for new opportunities to reduce energy consumption in their networks. As a complement to other approaches, we have evaluated a promising new method: applying micro-sleeps in packet processing nodes. Micro-sleeps can be opportunistically applied in all idle periods, thereby saving energy across a wide range of traffic conditions.

LEIF JOHANSSON,
PER HOLMBERG,
ROBERT SKOG

There are several ways to improve the energy efficiency of mobile communication systems and reduce the amount of energy it takes to provide mobile broadband services.

■ As part of its efforts to reduce network energy consumption, a communication service provider (CSP) may choose to upgrade to new hardware (HW) that consumes less energy than the previous generation. It is also possible to cut energy consumption with the help of an advanced scale-in/scale-out mechanism that reduces the amount of HW in use when traffic is low, such as during the night. In these conditions, some servers can be turned off, which results in energy savings.

Regardless of the energy-saving mechanism, however, it is essential to ensure there is no negative impact on real-time characteristics such as jitter and packet latency. It is also important to note that the correlation between Moore's law and power consumption savings for each new HW generation has decreased.

Our latest research indicates that CSPs could save additional energy in their data centers by applying micro-sleeps in packet processing nodes. This approach has both lower overhead and lower latency than existing methods, and makes it possible to reduce power consumption even at high load, without degrading application performance. Because power management is implemented in communications libraries, it is easy to integrate into existing applications.

Packet processing in communication networks

Packet processing in communication networks involves the application of different functions and algorithms to ensure that each packet can run through the network efficiently. It includes steps such as packet identification, inspection and manipulation.

The network packet is a fundamental component in a packet-switched network. It consists of three main parts: the header, the payload and the trailer. The header contains information such as the sender address, the destination address, the length of the packet and the packet sequence number. The payload contains the transferred data – that is, one part of a video, e-mail or other type of content. The trailer marks the end of the packet and can also include error detection and correction information.

The different packet processing functions and algorithms range from fairly simple to more complex ones. A basic routing function is a good example of simple packet processing. More complex packet processing functions involve the application of different policies, charging and manipulation of the packets.

User plane packet processing nodes

While micro-sleeps can be applied to all types of packet processing, our research indicates that they are particularly impactful when used in the user plane function (UPF).

According to its definition in 3GPP, the UPF is a user plane packet processing node in mobile systems that links the RAN to the internet (or similar data networks) [4]. Simply put, the main purpose of the UPF is to forward packets to and from the internet. This is often done in combination with different traffic optimization functions that range from pure Transmission Control Protocol (TCP) optimization to more advanced video optimization in combination with intelligent traffic congestion logic.

> WHILE MICRO-SLEEPS CAN BE APPLIED TO ALL TYPES OF PACKET PROCESSING, THEY ARE PARTICULARLY IMPACTFUL WHEN USED IN THE UPF

The UPF includes several complex processing functions such as GPRS Tunneling Protocol User Plane encapsulation and decapsulation. Other UPF functions include access control, bearer lookup, QoS mapping and marking. It also follows rules for guaranteed bitrate and maximum bitrate. Packets passing through the UPF are also subject to online/offline charging – that is, the application of different charging policies.

Instructions about how to process packets for different user equipment come from the session management function/policy control function. The processing of packets from different users occurs independently for the most part. All the packet processing functions and algorithms must meet the system requirement of real-time characteristics such as jitter and packet latency.

Kernel packet processing

The built-in, native networking support in computers is implemented as part of the operating system (OS) kernel and uses the POSIX (Portable Operating System Interface) socket as its standard application programming interface (API).

Terms and abbreviations

API – Application Programming Interface | **CPU** – Central Processing Unit | **CSP** – Communication Service Provider | **DPDK** – Data Plane Development Kit | **HW** – Hardware | **NIC** – Network Interface Card | **OS** – Operating System | **RTD** – Round-Trip Delay | **SW** – Software | **UPF** – User Plane Function

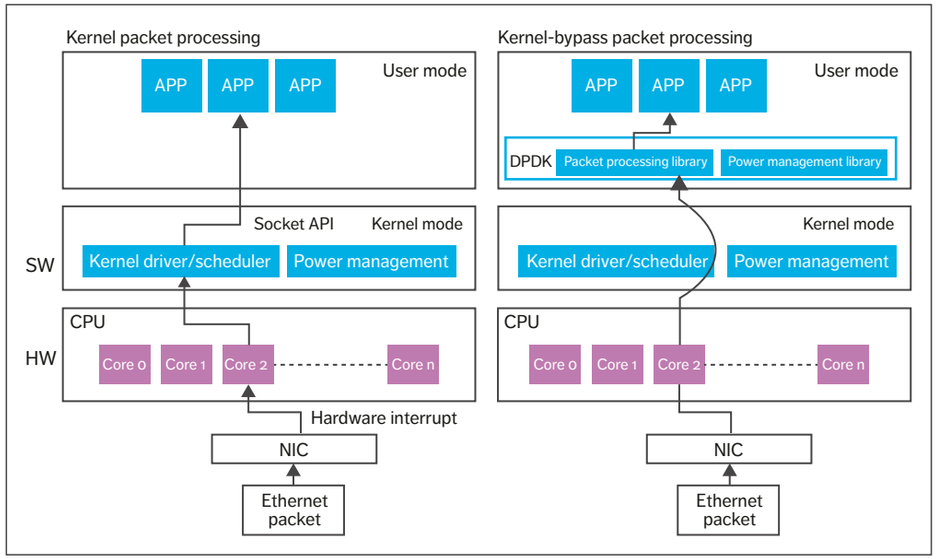

Figure 1 Kernel packet processing versus kernel-bypass packet processing

In kernel packet processing, user-mode application programs use the POSIX socket API to send and receive packets, while the kernel driver/scheduler handles the interaction with the network interface card (NIC). If the network is not ready (no packet has arrived) the kernel can decide to block that application while waiting.

The left side of *Figure 1* illustrates how kernel packet processing works, with "App" representing user-mode applications. In this model, the OS is aware of the idle time, and OS power management can work to save power.

Kernel packet processing has multiple disadvantages. Most importantly, the high overhead of the OS calls and the copying of packets to/from the kernel space makes it hard to scale to high networking speeds and high packet rates. For example, the time between 1,534-byte packets on a 200Gbit Ethernet is 61ns. This corresponds to the execution time for performing a system call, which does not leave sufficient time to perform the packet processing itself. The speed of today's network cards makes kernel-packet processing challenging at best, and impossible at worst.

Power management in the kernel can be problematic at high networking speeds. The OS kernel monitors and saves energy based on core utilization. This is a much slower mechanism that does not have the ability to follow the rapid changes in high-speed network traffic, resulting in queue buildups, delays and even packet drops.

Kernel-bypass packet processing

Kernel-bypass packet processing eliminates the kernel execution overhead by moving packet processing directly to the user-space, as shown on the right side of Figure 1. In this scenario, the OS can dedicate a network interface to an application such as the Data Plane Development Kit (DPDK), which can program the HW from the user-space.

When DPDK is used, packets are received directly in user-space memory without kernel interaction, and all network-related interrupts are disabled. It is up to the application to make sure that packet queues and ring buffers to the NIC are checked frequently for new packets.

To avoid packet drops and reduce packet latency, the DPDK-based application is designed to check the packet ring buffer in busy-waiting mode, where the complete core is assigned to the application thread. This enables packet processing without any context switching or HW interruptions and minimal cache pollution, as the application thread is the only user of the core. Measurements using DPDK as a kernel bypass for packet processing show that millions of packets can be received in a single core and pipelined to other cores for further packet processing.

Kernel-bypass packet processing solves the issue of how to handle packet processing at speeds of 200 Gigabit Ethernet and beyond, but the busy-waiting technique of packet reception comes at a cost. No energy savings will be achieved if all the packet processing cores are 100 percent utilized in busy-waiting mode.

In kernel-bypass packet processing, ethernet packets are transferred directly to the user-mode application memory. Power management needs to be performed in the user-space within the DPDK library. It is important to note that DPDK interacts with the kernel power management when changing core frequency [1].

Energy-efficient packet processing using DPDK

Higher data rates in packet processing have made it necessary to change packet processing to use DPDK with dedicated cores, user-mode execution and poll-mode drivers to remove the high overhead from context switches and system calls. Unfortunately, though, this solution is not as energy efficient as modern HW allows it to be. There are three main issues that need to be addressed:

1. The busy-waiting mode consumes more power than necessary when waiting for work.
2. OS power management is slow and cannot utilize idle time between packets and bursts of packets for high-speed interfaces.
3. The busy-waiting mode in DPDK always makes the core appear to be 100 percent utilized, with no information upon which to base power management decisions such as scaling down at periods of low traffic.

> "WHEN THE EVENT OCCURS (OR THE MAXIMUM WAIT TIME EXPIRES) THE PROGRAM SIMPLY WAKES UP AND CONTINUES EXECUTING"

To solve these issues, power management actions must be fast, controlled directly from the packet processing application and executed in user mode. Two developments have made this possible. Firstly, server processors now have support for entering power-saving sleep states directly from applications in user mode. Secondly, the latest release of the DPDK library includes an option to utilize these sleep states while waiting for new networking events, making it possible to avoid the problems associated with the busy-waiting loop [1].

User-mode sleep states

Processors have specific instructions for entering sleep states, waiting for events and waking up. Previously these instructions were only available in the OS in supervisor mode, but new versions of these instructions have become available that are safe to use for implementing user-mode micro-sleep states. They are much faster than the old instructions and can be implemented without OS support.

The latest instructions allow a user-mode program to set up an event to monitor (that is, specify the update in the ring buffer that the program will wait for), define the maximum allowed wait time and then enter sleep state. When the event occurs (or the maximum wait time expires) the program simply wakes up and continues executing. There is no

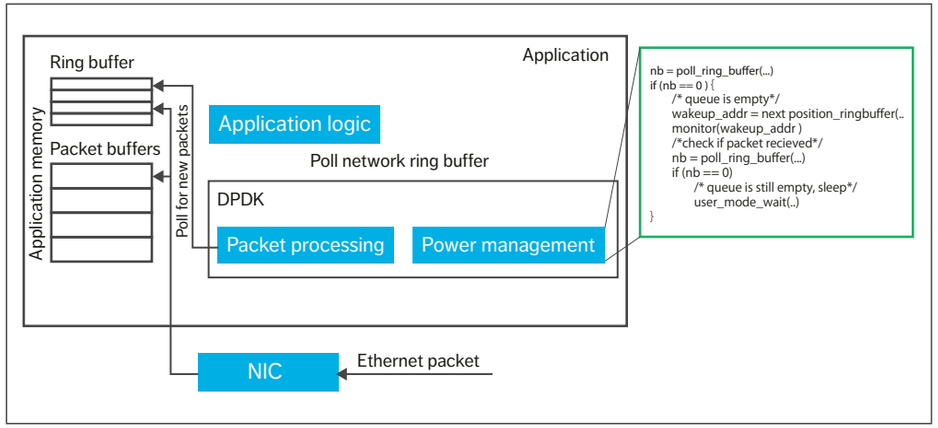

Figure 2 Power management integrated into the DPDK

overhead to enter OS kernel, to do context switches and so on.

Examples of such user-mode instructions for setting up an event to monitor and for entering sleep state are UMONITOR and UMWAIT in Intel processors [2] and MONITORX and MWAITX in AMD processors [3]. These instructions make it possible to save energy for much shorter idle periods than was previously possible. They can also be efficient at higher communication speeds.

DPDK integration

Support for the new user-mode sleep-state mechanism has recently been added to the DPDK library, which makes it possible to enter energy-saving sleep states directly in the poll loop doing busy-waiting of DPDK descriptor ring or rings.

If the polls of the descriptor ring (or rings) are empty, an event monitor is set up and a final poll is done before user-mode sleep state begins. Relevant updates – such as an NIC or other central processing unit (CPU) core updates related to the monitored position in the ring – will be detected by processor HW, which will then wake up the CPU core to restart execution of the next instruction, continuing the poll. The poll loop will then detect the update.

Figure 2 illustrates how power management is integrated into the DPDK. In this setup, the network device allocates the ring buffers and packet buffers in the application user-mode memory and updates the ring buffer with the new packet buffer point at packet reception. User-mode sleep state is entered when the ring buffer is empty (that is, the queue to the NIC is empty). Wake-up is triggered when the NIC updates the ring buffer. It is important to note that the proposed power management does not impact the application logic.

Observability

Processors include a measurement functionality for time spent executing and time spent in different sleep states. User-mode sleep states make the CPU thread utilization visible for this measurement functionality, and the CPU thread utilization can also be observed for the DPDK application. CPU thread utilization can also serve as input to the control loops for other types of power management (such as frequency scaling) that may be used as a complement to user-mode sleep states, potentially leading to further improvements in energy efficiency.

Hardware offload

HW offload is a technique that transfers the handling of selected flows from the processor (the slow path) to the NIC (the fast path). This approach frees up the CPU core and makes micro-sleep technology more effective because a larger number of CPU cores have the opportunity to micro-sleep. HW offload also improves the efficiency of the Peripheral Component Interconnect bus and offers other potential benefits with regard to improving throughput, efficiency and latency in the system.

Processor core voltage/frequency scaling

Scaling down CPU core voltage and frequency is a complementary method for achieving the best efficiency over longer periods of lower processing. There is a trade-off with user-mode sleep state, as scaling down also means shorter idle times.

Since power consumption increases proportionally to the square of the voltage, it is favorable to scale down the voltage and frequency when the packet rate drops for a period of time. A fast and reliable monitoring function is needed to detect increased packet rates and, most importantly, packet bursts. The faster and more reliable the detection is, the more aggressively the voltage and frequency can be scaled down without causing negative effects and degrading KPI due to queue buildups that increase latencies, timing jitter and the associated risk of packet drops.

DPDK provides interfaces that enable fast control of CPU core voltage/frequency scaling. In our implementation, we use a fast packet-burst detection in DPDK together with a control loop that maintains a performance margin that handles even the worst-case burst scenarios. When there is a margin, the user-mode wait states offer an alternative method of energy saving, and we can avoid any negative impact on packet processing without compromising energy efficiency.

Processor uncore voltage/frequency scaling

Uncore refers to the parts of the processor outside the actual cores – that is, the interconnection between the CPU cores (and other functions) and the interfaces on the processor that are shared and used by all CPUs. Scaling down the voltage and frequency of the uncore is a complement to energy-saving mechanisms implemented in the core. As in the core, it is favorable to scale down the voltage and frequency in the uncore when the packet rate drops for a period of time.

> THE RESULTS OF THESE EXPERIMENTS INDICATE THAT THE POWER-SAVING METHODS ARE STABLE AND EFFICIENT

As the uncore is shared by all CPU cores, frequency needs must be coordinated between all cores and applications, including those outside DPDK, making uncore changes a slower mechanism than core frequency changes.

Experimental validation and results

We ran two sets of lab experiments to validate the user-plane micro-sleep design. In the first set, we generated a workload that we consider to be realistic and applied it to a UPF application. The UPF measurements showed that processor energy consumption decreased from 190W to 61W at low traffic load and from 190W to 145W at maximum traffic load. Beyond quantifying the potential energy savings in a realistic use case, our results verified that power-saving methods work well with existing communication applications.

Our second set of experiments used synthetic workloads that were designed to simulate extreme conditions. The results of these experiments indicate that the power-saving methods are stable and efficient, with negligible impact even in worst-case conditions.

Energy-efficient packet processing applied on a user plane function

In the first set of experiments, we applied the DPDK-based power management with micro-sleep and frequency scaling on an Ericsson UPF application. To measure the energy gain when using HW offload, the UPF is designed to off-load packet flows with the highest bandwidth to the network device.

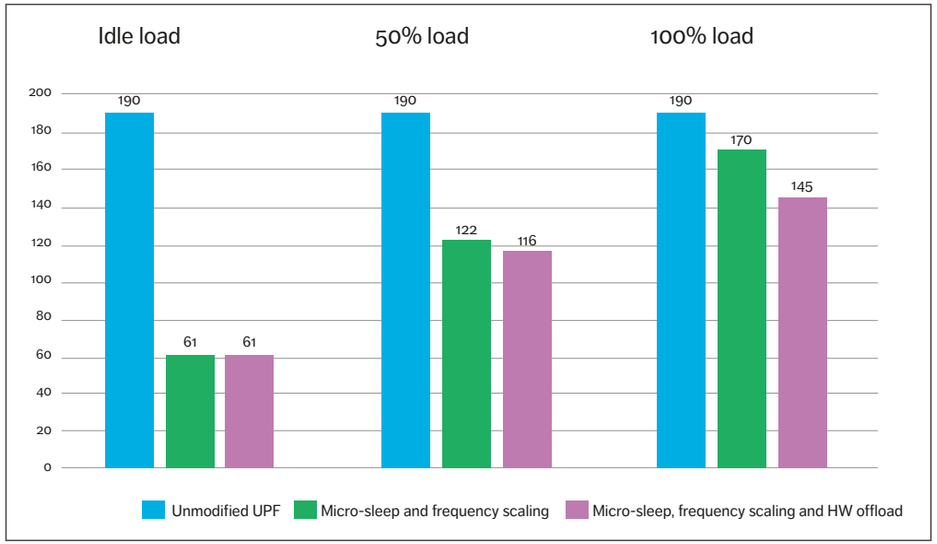

Figure 3 CPU power consumption at idle load, at 50 percent maximum traffic load and at 100 percent maximum traffic load

As a traffic model, we used a traffic mix from our internal stability test. We generated the traffic using an external traffic generator that simulates the surrounding network elements. *Figure 3* presents the results. The blue bar shows CPU power consumption when running the DPDK-based UPF unmodified at idle load, at 50 percent of the maximum traffic load and at 100 percent traffic load. The green and purple bars show the energy consumption when running the same traffic load with power management without HW offload, and with both power management and HW offload.

The measurements show a 68 percent processor power reduction at idle load when using power management (the green bar on the far left in Figure 3). The bars for 50 percent load and 100 percent load with enabled power management show the reduction in processor power even as traffic load increases.

It is difficult to perfectly balance the processor load and thereby enable the cores that are not fully loaded to sleep and save energy. In our case, 10 percent of the energy could be saved at maximum traffic load (the green bar on the far right in Figure 3). This result will vary depending on the application and traffic mix.

NIC HW offload technology enables further opportunities for micro-sleep, leading to an overall 24 percent reduction at maximum load (the purple bar on the far right in Figure 3). The amount of power that can be saved using HW offload will depend on the ability of each application to utilize it and thereby save processor resources.

Synthetic benchmarks

Our proposed power management solution using micro-sleep and frequency scaling targets packet processing with an extremely low impact on packet latency and a fast reaction to traffic bursts. For example, in low-latency power management, traffic bursts must not cause any packet drops. The new user-mode sleep states enable the core to wake up within a few 100ns and should have minimal impact on the total packet latency.

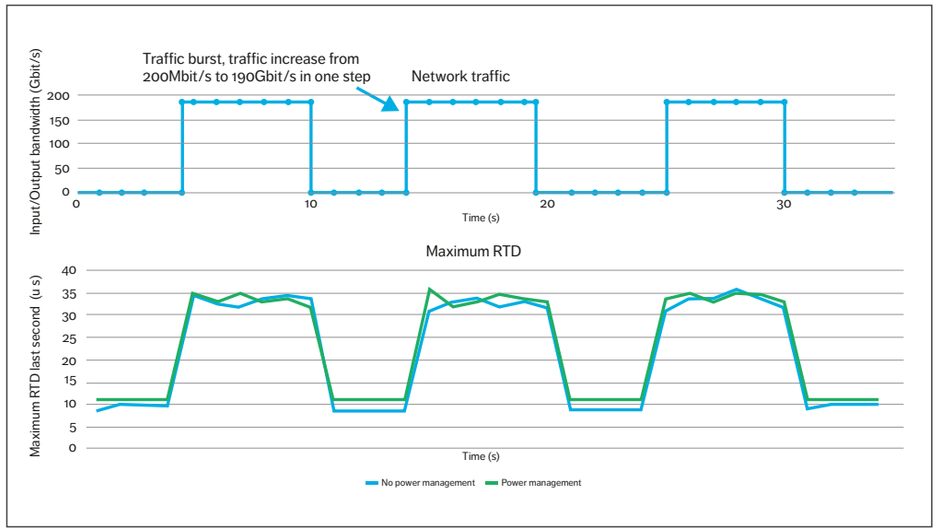

Figure 4 Measured maximum RTD when applying extreme traffic bursts, showing no increased delay when using power management

To measure the packet latency and the speed at which the power management functionality can react to traffic bursts, we used a synthetic benchmark where the application's impact on the packet latency is reduced as much as possible when receiving network traffic. This enabled us to measure the packet latency impact when using user-mode sleep states and frequency scaling. The synthetic benchmark is also used to measure how fast the system can scale up the performance when receiving traffic bursts.

The synthetic benchmark program fetches Ethernet packets using DPDK, swaps the media access control address and sends the packet back to the original source. The source machine measures the time from the sending of the packet to the receiving of the packet (known as the round-trip delay (RTD)) including packet drops. User-mode sleep is used when queues to the network device are empty and frequency scaling is triggered when network traffic increases/decreases.

Figure 4 shows the maximum measured RTD down to the last second, measured on the external traffic generator. At low traffic, the RTD is slightly higher when using power management (the green line in the bottom half of Figure 4) due to reduced core/uncore frequency. When traffic increased from 27Mbit/s to 190Gbit/s, the core/uncore frequency changed to maximum speed using the implemented DPDK burst detector. The power management impact of the packet delay is negligible at maximum load (190Gbit/s).

Our results indicate that the measured worst-case RTD or maximum packet jitter is not impacted, even if the RTD value is slightly higher at low traffic load. No packets were dropped during the measurements.

Conclusion

Our research clearly demonstrates that is possible to reduce energy consumption while simultaneously meeting the requirements of latency-sensitive applications. The latest generation of server processors enables fast power-saving transitions by entering sleep states directly from applications in

user mode, with no impact on important KPIs like jitter and packet latency. With no overhead to enter the OS kernel and no context switches, the central processing unit core simply wakes up and continues executing when packets arrive during user-mode sleep state. Because power management is part of the user-mode packet processing library, energy savings are easy to integrate and validate on existing Data Plane Development Kit-based applications.

The combination of micro-sleep, hardware offload and frequency scaling has the potential to lead to significant energy savings. Our experiments on an Ericsson user plane function show that processor energy consumption decreased from 190W to 61W at low traffic and from 190W to 145W at maximum traffic. We also carried out experiments using a synthetic benchmark, which demonstrate that our proposed power management solution has no measurable impact on worst-case round-trip delay and worst-case packet latency.

References

1. **Data Plane Development Kit, Programmer's Guide, available at:** *https://doc.dpdk.org/guides-22.03/prog_guide/index.html*
2. **Intel, Intel® 64 and IA-32 Architectures Software Developer's Manual, available at:** *https://cdrdv2.intel.com/v1/dl/getContent/671110*
3. **AMD, AMD64 Technology, AMD64 Architecture Programmer's Manual, available at:** *https://www.amd.com/system/files/TechDocs/24594.pdf*
4. **3GPP specifications, available at:** *https://www.3gpp.org/specifications*

Further reading

- **Ericsson, How to break the energy curve, available at:** *https://www.ericsson.com/en/about-us/sustainability-and-corporate-responsibility/environment/product-energy-performance*
- **Ericsson blog, 5G energy consumption: what's the impact of 5G NR in real networks?, available at:** *https://www.ericsson.com/en/blog/2021/10/5g-energy-consumption-impact-5g-nr*
- **Ericsson, Network architecture domains, available at:** *https://www.ericsson.com/en/future-technologies/architecture/network-architecture-domains*
- **Ericsson, Network intelligence and services, available at:** *https://www.ericsson.com/en/networks*
- **Global5G.org, 5G and Energy Efficiency, available at:** *https://global5g.5g-ppp.eu/sites/default/files/BookletA4_EnergyEfficiency.pdf*

THE AUTHORS

Leif Johansson

◆ is an expert in the characteristics of open systems. He joined Ericsson in 1996 after receiving an M.Sc. in physics from Uppsala University, Sweden. Johansson has more than 20 years of experience evaluating and applying new technology in Ericsson's product portfolio.

Per Holmberg

◆ is a computer architect. He holds an M.Sc. in electrical engineering from KTH Royal Institute of Technology in Stockholm, Sweden, and has worked at Ericsson since 1985 designing processing solutions primarily for communication equipment. Holmberg has held specialist and expert positions in computer architecture and processing systems, been lead architect for several processor and computer designs and is responsible for more than 50 inventions.

Robert Skog

◆ is a senior expert in the field of service architecture. After earning an M.Sc. in electrical engineering from KTH Royal Institute of Technology in 1989, he joined Ericsson's two-year trainee program for system engineers. Since then, he has mainly worked on end-to-end solutions and traffic optimization for everything from the first WAP solutions to today's advanced user plane solutions. In 2005, Skog won Ericsson's prestigious Inventor of the Year award.

Ensuring energy-efficient networks

WITH ARTIFICIAL INTELLIGENCE

Finding ways to make networks more energy efficient without negatively impacting QoE is critical to network operators for both cost and sustainability reasons. To assist in these efforts, we are exploring the potential of using artificial intelligence (AI) techniques to recommend energy-efficient configuration settings for network nodes.

KONSTANTINOS VANDIKAS, HELENE HALLBERG, SELIM ICKIN, CECILIA NYSTRÖM, ERIK SANDERS, OLEG GORBATOV, LACKIS ELEFTHERIADIS

Our estimates indicate that the cost of the energy required to power networks represents between 10-30% of the network operating expenses of a communication service provider (CSP), depending on the specificities of its local energy market. In total, this expenditure adds up to approximately 25 billion USD per year [1].

■ Despite the many energy-efficiency solutions already implemented in mobile networks, energy consumption continues to rise in response to the rapid growth of both network traffic and data volumes. Our research indicates that additional energy-efficiency gains can be achieved by using machine-learning (ML) techniques that enable higher levels of automation.

ML is a type of artificial intelligence in which models learn patterns from data without being explicitly programmed. By recommending configuration settings that can be applied to base stations and other equipment, ML-based techniques make it possible to reduce energy consumption in network elements without impacting QoE.

Access nodes are at the center of our work to improve energy efficiency in networks. An access node (often simply called a node) refers to the relationship between connected user devices and the network elements to which those devices are connected. Access node configuration settings strongly influence node energy consumption and potentially many observable network performance QoS metrics.

As configuration settings rarely change at a particular node, the ability to regulate node energy consumption requires a mechanism that enables the generation and evaluation of new configuration settings to explore their impact. To avoid generating configurations that may negatively impact existing key performance indicator (KPI) levels, new configurations need to be bounded by different KPIs that constrain the search.

Given that some configuration settings may require more time than others to take effect, there is a need for accurate predictive models that make it possible to foresee when such changes can be applied ahead of time. This would then minimize any potential disruption to the network's operation. An ideal solution would identify as many potential ways to reduce energy consumption on the current functionality of the network elements.

With all of this in mind, we have developed a concept for end-to-end (E2E) energy optimization that encompasses everything from the power system to the nodes to the network level. *Figure 1* illustrates our concept, highlighting the energy recommendation engine that is at its core.

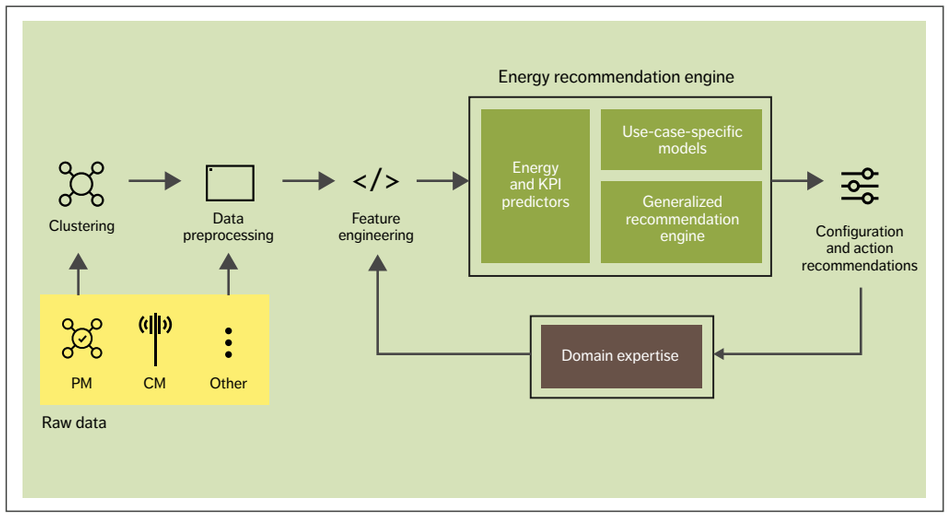

Figure 1 End-to-end energy optimization from power system to node to network

Data set for the energy recommendation engine

All the data used in our work on the energy recommendation engine was collected from a live network. We primarily used performance management (PM) and configuration management (CM) data as measured in the base station, where

Terms and abbreviations

CM – Configuration Management | **CVAE** – Conditional Variational Autoencoder | **DL** – Downlink | **E2E** – End-To-End | **GNN** – Graph Neural Network | **KPI** – Key Performance Indicator | **ML** – Machine Learning | **PM** – Performance Management | **PRB** – Physical Resource Block | **PSU** – Power Supply Unit | **UL** – Uplink

> TO ENSURE ZERO NEGATIVE IMPACT ON QoE, WE USED KPIs TO CONSTRAIN OUR MODELS

energy measurements are already part of the collected dataset within the PM. As a result, there is no need to deploy new hardware to get the data. The radio network performance counter data sets contain observations on cell performance such as the activity count in downlink (DL) and uplink (UL) directions, the utilization in the cell and units.

To ensure zero negative impact on QoE, we used KPIs to constrain our models. While it is technically possible to include additional and/or alternate KPIs based on operator preferences, we selected these five based on how they affect energy consumption:

1. Number of connection attempts to a cell
2. Average number of users in a cell
3. Throughput
4. Latency
5. Interference.

When telecom networks are installed, they are typically configured with certain parameters such as the number of cells and the hardware unit types (indicating frequency bands) and so on. Possible reasons for a reconfiguration could be a problem such as a software issue or a hardware failure after the installation of new parts that may come from different vendors. Over time, subtle changes and different tuning may lead to different energy consumption levels, where, for the same amount of traffic, this may in some cases be positive (less energy consumption) while in others it is negative (more energy consumption).

The CM data set that we used consists of hundreds of configuration attributes of the sector, including the settings of each radio cell (such as frequency in DL and UL directions) and installed hardware types. We used this information to be able to recommend multiple configuration changes at once, rather than focusing on one at a time.

The output of the energy recommendation engine consists of a set of configuration attribute changes for a corresponding node. The output captures the interplay between different nodes and configurations rather than focusing on isolated fine tuning on a per-node level, which may have an effect on other nodes.

Methodologies used in the energy recommendation engine

A content-based recommendation model requires a good representation of the configuration settings in the embedded (also known as latent) space. Such a representation can be hard to obtain manually due to the high number of configuration attributes.

Conditional variational autoencoders [2] (CVAEs) and graph neural networks [3] (GNNs) are two of the most suitable and complementary techniques for our purposes, as they are both fueled by the success of neural networks. While a CVAE is generative and adversarial, helping to explore large spaces in bounded conditions, a GNN can act as a critic (or discriminator) that can suppress abnormal recommendations, especially when the recommendation model diverges too much.

In addition to GNNs and CVAEs, we also used conditions to confine the new configuration settings under specific KPIs that should not be broken while new configuration settings are created. Such conditions may originate from domain expert engineers, while others can be universal or specific to the network that is being examined.

Graph neural networks

GNNs offer a straightforward way to learn from relational data. We use them – and graph convolution in particular – in two ways in this project: to generate conservative recommendations for energy efficiency based on historical information and to generate multi-site predictive models for different KPIs. Multi-site predictive models are essentially enhanced forecasting models that help operators predict performance based on KPIs that serve as measures of how well a network is behaving.

One of the weaknesses of conventional forecasting techniques such as long-short term memory is the inability to account for the spatio/temporal relationship that exists between nodes. This is problematic because nodes are positioned strategically in various parts of urban or rural spaces and configured accordingly to serve the requests made by user equipment in their vicinity. Information about their spatio/temporal relationships is highly relevant in forecasting.

We have studied the possibility of combining graph convolution with regular 2D convolution. This approach combines two inputs: the adjacency matrix that represents the network's topology, and the time series of each node for a different KPI. The network's topology is constructed using geographical information per node. This information is then represented as an adjacency matrix, which captures the temporal aspects of each node.

Time-series information per node is preprocessed to produce the corresponding input and output predictive window. In this case, we use 12 hours and learn to predict the next 12. Graph convolution and regular 2D convolution is interleaved to build a combination of the two. Our results indicate that this approach increases model performance in multi-site prediction.

The second way that we have used GNNs in this project is to generate conservative recommendations for energy efficiency based on historical information – that is, we used GNNs to create a recommendation engine.

To understand these changes, we represent the relationship between different nodes in a telecommunication network and their configuration sets as a graph. In graph theory, a graph is a structure that contains a set of objects (or nodes) that are connected to each other through links. A link between two nodes means that the two are associated. In this context, we consider a heterogeneous graph, as the objects that we connect are of different types. More specifically, we associate nodes with configuration sets, and the association is represented by a link between them.

In addition, each link is labeled based on the efficiency of that association from an energy perspective as well as on the number of connected subscribers. For each of these elements of the graph we learn a representation of that, driven by its features such as the hardware installed or the different types of parameters in each configuration set.

As a result, the problem is transformed to a link-attribute prediction problem, where we train a model that learns to predict how energy efficient that association was. Even though this system is not capable of generating new configuration sets, we see that it is capable of matchmaking similar nodes with potentially similar configuration sets that can have a lower energy impact than the one currently used.

Conditional variational autoencoder

The high-level definition of a generative model infers generalized distribution of observation features that are confined by the constraint conditions. The CVAE generative model-based recommendation engine consists of multiple components including an encoder, decoder, prediction model for the target KPIs and a prediction model for the energy-consumption target.

The encoder model compresses the representation of the raw CM dataset into a low dimensional matrix, which is referred to as latent space. Latent space consists of points, where each point in a 2D latent space represents a full CM configuration setting that is constrained with an energy consumption and a KPI value.

Due to the nature of the CVAE, the CM configurations with the same energy and KPI constraint categories are located close to each other in the latent space. The decoder reconstructs the complete configuration file, which could consist of many CM attributes (configuration parameters and settings), from the embedding representation in the latent space. The latent variables that represent the targeted KPI and energy-consumption levels are drawn by uniform random selection from the many that represent the same category of the targeted KPI and energy levels. The decoder then uses the input to generate new configuration attributes.

In deployment, we provide constraints as input to the decoder of the CVAE model bounded by the

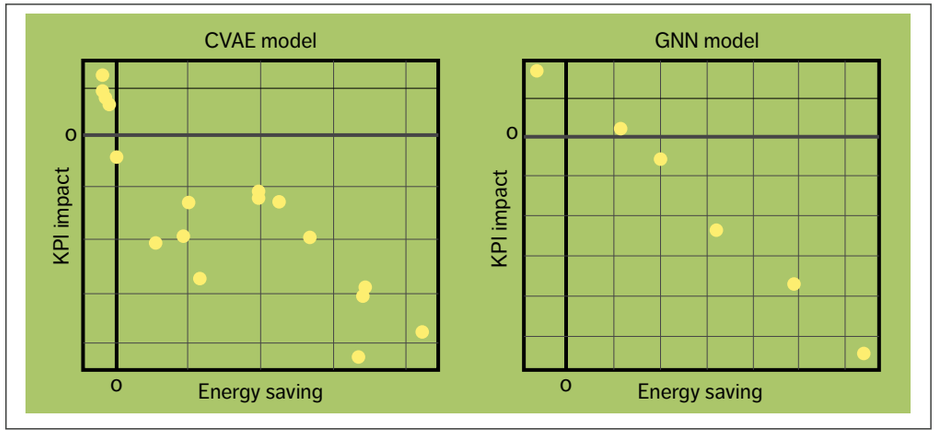

Figure 2 Energy savings and KPI impact according to the CVAE model (left) and the GNN model (right)

required KPI and energy values. We can use multiple constraints and they can be customer specific. The constrained energy value is set slightly less than the predicted energy consumption, as an energy-efficient synthetic CM file is expected as the output of the generative model.

At the same time, we aim to sustain the KPI value. The predicted KPI value at a network node is therefore given as input to the decoder. There are two different prediction models for KPI and energy models:

1. A CM-based prediction model that only uses CM attributes in tabular form as input
2. A PM-based prediction model that only uses PM counters in time-series form as input.

The time-series PM-based prediction models perform 24 hours in advance to give enough time to deploy the desired configuration to a corresponding network element. Prediction results from CM and PM models for energy consumption and KPI (connected users) values performed well. This is important for the accuracy of the generative model output, as the outputs of PM-based models are used as inputs to the CVAE generative model together with the selected latent variables mentioned earlier.

It is important to quantify the amount of energy saved with the generated CM configuration set compared with the existing planned configuration set. For that reason, we used a pretrained CM-based energy model, which only takes CM attributes as input features, and this model predicts energy consumption using purely CM attribute value combinations. This means it can contribute to estimating a mean base energy consumption value given a configuration.

First, the configuration generated by the recommendation model using the actual predicted energy as a constraint is given as input to a pretrained CM-based energy model. Next, the configuration generated by the recommendation model using a value that is lower than the predicted energy as a constraint is given as input to the same CM-based energy model. Finally, the difference between the two energy predictions is computed, and an indicative potential saving is quantified.

We repeat the same steps on a pretrained CM-based KPI model to quantify the KPI impact. This allows us to obtain a KPI versus energy trade-off curve as shown in *Figure 2*. In general, these curves tend to show that higher energy savings yield a higher negative impact on the KPI.

Comparison of the conditional variational autoencoder and graph neural network

A comparison of the CVAE and GNN reveals that, given its generative nature, a CVAE produces new configuration settings. In contrast, a GNN only identifies potential energy savings as marked by a rating function. In this case, the rating function contains six distinct categories, each represented by a dot on the right in Figure 2. (While a GNN is non-generative in the context of our research, the literature indicates that it can also be used in a generative context.)

The comparison also reveals that the efficacy of the CVAE-based recommendation model is dependent on the accuracy of multiple CM- and PM-based KPI and energy prediction models, which means that all the predictive models need to be accurate simultaneously. As a good side effect, this modular structure potentially makes it easy to troubleshoot during the maintenance of the model performance.

Meanwhile, in the case of GNNs, there is only one graph-based model and it is limited by the number of different configurations that are available or have been applied to that network. The use of a single model simplifies its maintenance process.

As both CVAEs and GNNs are predictive recommendation models, it is possible to omit any recommendations that are predicted to impact network performance and QoE in advance. Both models yielded configuration recommendations that are predicted to achieve up to 10% energy savings when applied.

Use-case examples

To further increase the efficacy of our energy recommendation engine, we enhanced its ability to improve energy efficiency by applying recommendations in two specific use cases: radio signal interference detection and PSU load utilization, as shown in *Figure 3*. Both use cases follow the same three steps highlighted in the middle of the figure:

1. Identification (localization) of nodes with the potential for improvement
2. Modeling of the selected nodes to understand their behavior and predict their states
3. Actions to improve energy efficiency (implementation).

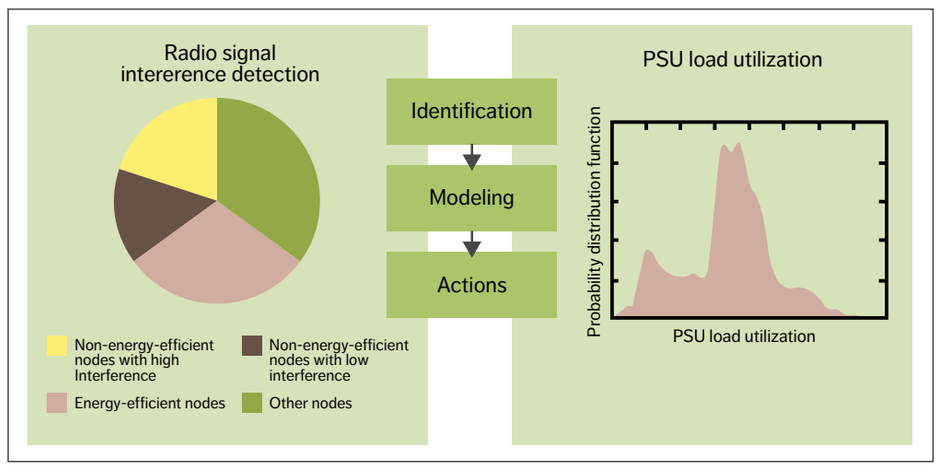

Figure 3 Two use-case examples – radio signal interference detection and PSU load utilization

Use case No. 1: radio signal interference detection

Interference is created by a range of factors such as changes to the environment or having too many users in the same cell, especially if they are located close to the cell edge. It is also possible for cells to interfere in frequency with each other. Interference has a significant impact not only on QoS but also on power consumption. While current radio equipment is designed to handle interference in a way that avoids unwanted emissions and variable techniques are used to limit the interference in the network, radio signal interference has proven to be a formidable challenge to overcome completely.

> A HOLISTIC APPROACH IS THE BEST WAY TO ACHIEVE OVERALL ENERGY SAVINGS

By introducing clusters and dividing the problem into sub-problems, our energy recommendation engine makes it possible to identify the nodes that have high interference and low energy efficiency. The left side of Figure 3 shows a breakdown of radio signal interference into four categories: non-energy-efficient nodes with high interference, non-energy-efficient nodes with low interference, energy-efficient nodes and other nodes. The possibility of improving the nodes classified as other nodes will be addressed in the future.

By modeling the network traffic load, consumed energy and interference-related KPIs, we can save energy by recommending action on the interference cells (such as locking them) to avoid the high energy-consumption state. The cell interference parameters are therefore unique and connected to the network activity and consumed energy.

Use case No. 2: PSU load utilization

A radio base station consists of several PSUs supplying power to the radio units. It is not uncommon for PSUs in radio base stations to be underutilized, as illustrated by the graph on the right in Figure 3. Active but underutilized PSUs may not be working efficiently within the operational range of the unit and consume more power that necessary due to power dissipation. It should be noted that PSU efficiency depends on the load.

To identify nodes for PSU load utilization improvement, we clustered them according to the PSU load, the number of PSUs, relation to the radio network activity (such as the relative radio resource usage), the number of active users, PRB utilization and the number of connected users. We focused on PSUs with less than 50% utilization. Our research show it is possible to propose dynamic power-supply control such as putting underutilized PSUs in sleep mode or turning them off.

PSU efficiency is highly dependent on the load that is applied on the PSU output. Setting one of several PSUs in a system to sleep mode enables savings of 1%. At the same time, the improved utilization and operational efficiency of the PSUs that remain active can provide an additional 1% in savings.

Conclusion

One of our core goals at Ericsson is to continuously improve the energy efficiency of networks. A holistic energy optimization approach is the best way to achieve overall energy savings because it ensures that improvements achieved at one level are not canceled out by increased energy use at another. Our end-to-end (E2E) energy optimization concept is driven by an energy recommendation engine that is powered by artificial intelligence. This solution has great potential for automation, with the help of specific interfaces that can tune the nodes directly without human intervention. It can be fully software based, without the need for additional hardware.

The energy recommendation engine analyzes relevant data to figure out how node configurations can be fine-tuned to reduce energy consumption without impacting QoE. Our research indicates that the total E2E efficiency gains (including radio configurations) generated by our approach can be up to 10% for radio cells and up to 2% for PSU optimization.

On top of building a generalized energy recommendation engine, we are also developing use-case-specific recommendations for different challenges that can be onboarded to the generalized engine at a later stage. In the two use cases we have studied so far, we used predictive models to find the cases where PSUs are underutilized and to detect interference that may cause unnecessary energy usage.

> USE-CASE-SPECIFIC RECOMMENDATIONS CAN BE ONBOARDED TO THE GENERALIZED ENGINE AT A LATER STAGE

References

1. **Ericsson, Network energy performance, available at:** *https://www.ericsson.com/en/about-us/sustainability-and-corporate-responsibility/environment/product-energy-performance*
2. **Advances in Neural Information Processing Systems (NIPS 2015), Learning Structured Output Representation using Deep Conditional Generative Models, Sohn, K; Lee, H; Yan, X, available at:** *https://proceedings.neurips.cc/paper/2015/hash/8d55a249e6baa5c06772297520da2051-Abstract.html*
3. **IEEE Transactions on Neural Network Systems (2021), A Comprehensive Survey on Graph Neural Networks, Wu et al., available at:** *https://ieeexplore.ieee.org/abstract/document/9046288*

Further reading

- **Ericsson, AI operations and optimization, available at:** *https://www.ericsson.com/en/ai/operations*
- **Ericsson, AI in networks, available at:** *https://www.ericsson.com/en/ai-and-automation*
- **Ericsson, Energy Infrastructure Operations, available at:** *https://www.ericsson.com/en/managed-services/energy-infrastructure-operations*

THE AUTHORS

Konstantinos Vandikas
◆ is a principal researcher at Ericsson Research whose work focuses on the intersection between distributed systems and AI. He has been at Ericsson Research since 2007, actively evolving research concepts from inception to commercialization. Vandikas has 23 granted patents and more than 70 patent applications, and is the author or coauthor of more than 20 scientific publications. He holds a Ph.D. in computer science from RWTH Aachen University, Germany.

Helene Hallberg
◆ is a senior specialist in energy efficiency radio systems at Ericsson. She joined Ericsson in 1988, first working on the engineering of power and backup systems for fixed and mobile telecom equipment. Hallberg is active in energy-related regulatory discussions and standardization activity, and has filed patents in the energy-efficiency area.

Selim Ickin
◆ joined Ericsson in 2014 and works as a senior specialist in AI at Ericsson Research. His current research interests are distributed ML and intelligent software prototyping. Ickin has developed numerous data-driven ML solutions in diverse domains. He has contributed to numerous international conferences, written several journal articles, and holds patents in various subareas of ML within the scope of mobile networks. Ickin holds a Ph.D. in computing from the Blekinge Institute of Technology, Sweden.

Cecilia Nyström
◆ is a program manager for data and analytics within Business Area Managed Services. She joined Ericsson in 2016, and in her current role she is primarily focused on data strategy and the development of new AI solutions. Nyström holds an M.Sc. in engineering physics from KTH Royal Institute of Technology, Sweden.

Erik Sanders
◆ is a product manager for AI and automation within Business Area Managed Services. He joined Ericsson in 2005 as an engineer in 3G RAN and continued with hardware development for radio base stations. In his current role, he drives innovation programs for Business Area Managed Services in the ML and reasoning area. Sanders holds an M.Sc. in mobile communication from Linköping University, Sweden.

Oleg Gorbatov
◆ joined Ericsson Research in 2020 as a senior researcher. His research focuses on modeling and optimization of complex systems. Gorbatov holds a Ph.D. from KTH Royal Institute of Technology, Sweden.

Lackis Eleftheriadis
◆ joined Ericsson in 1998 and is currently a senior specialist in sustainable AI operations within Ericsson Research. Prior to his current role, he had been involved in the development of power products, including functionality for radio access and site infrastructure. Along with several patents in his area of AI, power and energy efficiency, Eleftheriadis holds an M.Sc. in electrical engineering from Uppsala University, Sweden.

LEVERAGING LTE AND 5G NR NETWORKS FOR

Fixed wireless access

Globally, there is a huge underserved market for broadband connections, with more than one billion households still unconnected. The growth in high-speed mobile broadband coverage enabled by LTE and 5G New Radio is opening up much more commercially attractive opportunities for operators to use fixed wireless access to deliver broadband services to homes and small and medium-sized enterprises.

HÅKAN OLOFSSON, ANDERS ERICSSON, FREDRIC KRONESTEDT, SVEN HELLSTEN

Unlike the country-wide decisions typically made for mobile broadband (MBB), decisions about fixed broadband and targeted fixed wireless access (FWA) deployments tend to be made at the local market level, and operators have a critical role to play. A number of different drivers govern local market attractiveness, as outlined in Ericsson's recently published FWA handbook [1].

■ We have organized the FWA market opportunities into three distinct segments that we call 'Wireless Fiber', 'Build with Precision', and 'Connect the Unconnected'. Each of these has different characteristics mainly based on the offering, the availability of fixed access and the corresponding average revenue per user (ARPU) that can be expected from customers [1]. The Wireless Fiber segment consists of those cases in which there is a need for very high-rate offerings and capacity as a direct alternative to high-end fixed broadband. The ambition is to provide fiber-like speeds and handle households' TV needs, matched with a correspondingly high ability to pay. Typical sold date rates are 100 to 1,000+ Mbps and monthly ARPU levels of USD 50-100. The FWA sweet spot for this segment is typically suburban environments.

The Build with Precision segment is comprised of those cases where there is competition from performance-limited fixed broadband alternatives, such as xDSL. Here, the need is for high data rate and capacity, with a corresponding level of ARPU. Typical sold data rates are 50 to 200Mbps and monthly ARPU levels are around USD 20-60. The FWA sweet spot for this segment is in suburban or rural villages or towns that are currently underserved. Some more sparsely populated areas are also addressable.

The Connect the Unconnected segment is made up of cases in which fixed broadband competition is virtually non-existent, and smartphones that use MBB are the dominant way of accessing the internet. User expectations of access speed are relatively low. Typical sold data rates are 10 to 100Mbps and monthly ARPU levels are around USD 10-20. Even though ARPU levels are limited in this segment, it has a FWA sweet spot that stretches from urban environments to rural villages, due to limited investment needs.

Subscriptions, data rates and consumption

The paradigms for fixed broadband and MBB are different, both in terms of subscription offerings and dimensioning. Fixed broadband subscriptions tend to focus on maximum data rates that are achieved under normal circumstances – that is, at low to medium load. The user traffic is often shaped so that it does not exceed the sold data rate. Hence, for fixed broadband, the sold data rate is the normal value that household subscribers relate to.

By contrast, for MBB, peak rates are sometimes used for marketing, and normally the network transmits the maximum rate that the mobile device can handle. Monthly data buckets dominate the subscription paradigm, and additional monetization is achieved through upgrades to larger data buckets, all the way to unlimited data. Hence, for MBB, monthly data buckets are the normal subscription value that mobile subscribers relate to.

Terms and abbreviations

ADSL – Asymmetric Digital Subscriber Line | **ARPU** – Average Revenue per User | **CAT** – Category (in LTE) | **CPE** – Customer Premises Equipment | $\mathbf{d_{av}}$ – Average Busy-hour Data Consumption | **DL** – Downlink | **DSL** – Digital Subscriber Line | **FDD** – Frequency Division Duplex | **FWA** – Fixed Wireless Access | **MBB** – Mobile Broadband | **MIMO** – Multiple-input, Multiple-output | **mmWave** – Millimeter Wave | **NR** – New Radio | $\mathbf{R_{min}}$ – Minimum Data Rate | **TDD** – Time Division Duplex | **Tx/Rx** – Radio Transmitter/Radio Receiver | **WCDMA** – Wideband Code Division Multiple Access | **xDSL** – DSL family (e.g. ADSL)

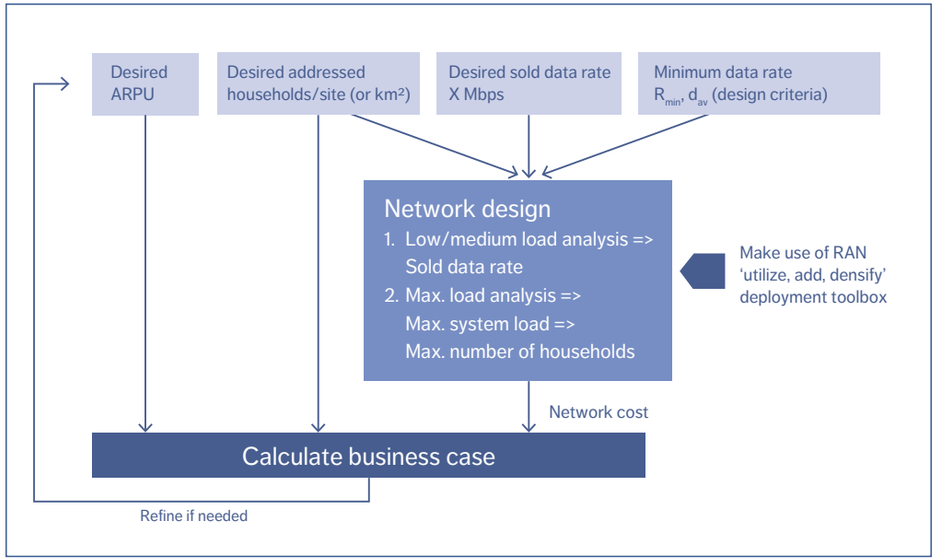

Figure 1: FWA deployment analysis flow

It is important that both consumers and operators (fixed and mobile) understand this crucial difference. Our view is that FWA will inherit the subscription paradigms of fixed broadband rather than those of MBB. That is, households should pay for FWA on the basis of data rate and not be concerned about data consumption.

Last-hop dimensioning

In FWA the last hop is wireless, so all the characteristics of a wireless network apply to the dimensioning. Unlike fiber, but similar to digital subscriber line loop length, there will be varying connection quality to different households. And, unlike fixed broadband overall, the last hop is radio and therefore shared, which means that speeds will degrade with increasing network load. All these characteristics must be taken into account when dimensioning an FWA network. Further, since Ericsson promotes the sharing of assets with MBB (when available), we recommend that FWA is brought into general RAN dimensioning.

Note that for fixed broadband, FWA and MBB alike, there is transport aggregation above the last hop, which is dimensioned according to standard principles and can also contribute to a varying user experience.

In short, while FWA will inherit the subscription paradigm of fixed broadband, due to the radio properties of the last hop to households, it must use modified dimensioning methods and terms from the MBB paradigm.

Figure 1 illustrates a typical FWA analysis flow. It starts with input on the subscription and offering, including dimensioning criteria, which triggers a selected, maximally efficient network design that depends on the offering ambitions and network starting point. A business case can be calculated by balancing the resulting cost items of the deployment with the extra revenues foreseen.

FWA toolbox

An existing mobile radio network, normally designed for voice and MBB, is an excellent base for offering an FWA service. Depending on the radio network starting point and the operator's ambitions for FWA, there is a toolbox available to make the network capable of handling a combination of voice, MBB and FWA.

These tools fall into three main categories: utilize, add and densify. The particular needs of each local situation can be met by deploying a well-planned mix of these tools.

Utilize existing radio network assets

The ability to utilize existing radio network assets is a fundamental advantage that sets mobile operators apart from start-ups or greenfield competitors in the FWA market. However, the advantage is only fully realized if all relevant RAN assets are efficiently combined for voice, MBB and FWA. If the operator chooses not to utilize existing assets built for voice and MBB, the number of economically viable local areas for FWA will be smaller, and the operator risks facing unnecessary competition with standalone FWA providers.

The radio network assets that should be utilized include existing radio sites, spare capacity in deployed spectrum (including associated equipment), and acquired but undeployed spectrum. Existing radio sites are critical assets, whether they are operator-owned or rented. The 'tool' of utilizing existing sites is not used by itself, but in combination with other actions to make those more cost-efficient. Spare capacity in deployed spectrum and associated deployed radio, baseband and transport network equipment is quite common in FWA target areas, and making use of it requires no new capital expenditure. Acquired but undeployed spectrum is also common in FWA target areas, which makes radio deployment in new bands possible without the cost of acquiring new spectrum. The geographical fit for FWA is excellent, since FWA targeted areas are often suburban and rural, where unused spectrum is most prevalent.

Add radio network capabilities

In an MBB RAN, radio capabilities are continuously added to handle more traffic, more customers and better app coverage. To handle FWA as an extra service, some of these additions may have to be made sooner to achieve a combined network with sufficient capabilities.

An existing mobile operator has the significant advantage of being able to add the following radio network capabilities and co-finance them for MBB and FWA:

- Spectrum – upcoming wide spectrum bands in 3-6GHz and millimeter wave (mmWave) open up potential for providing high data rates and capacity, benefiting both MBB and FWA
- Higher-order modulation, multiple-input, multiple-output (MIMO) and beamforming – offering the potential to squeeze out the most from each spectrum band
- FWA-tailored software features – to enhance performance for FWA users and to provide adequate quality to MBB and FWA in shared deployments
- Additional sectors on existing sites
- 5G New Radio (NR) access – designed for low latency and for wide spectrum bands, creating an excellent overall network together with LTE.

> THE ABILITY TO UTILIZE EXISTING RADIO NETWORK ASSETS IS A FUNDAMENTAL ADVANTAGE

Densify the radio network grid

When the 'utilize' and 'add' tools have been used to their full potential, densification can offer further gains. In these cases, MBB enhancements tend to be necessary as well, so the upgrade needs of MBB and FWA should be considered together and the densification of the network should be co-financed.

The two options for densifying the radio network grid are macro site densification and small cell site

> QUALITY ACROSS BOTH SERVICES IS ENSURED THROUGH EXISTING SOFTWARE FEATURES

densification on poles. Macro site densification is an opportunistic approach: where new macro sites can be found, such opportunities can be taken. Small cell site densification on poles may be necessary if the macro grid is sparse and performance requirements are high.

Spectrum sharing across MBB and FWA

Sharing spectrum across FWA and MBB enables significant gains in overall spectral efficiency because higher utilization is possible with one 'bigger pipe'. This is explained by the trunking gain effect, which has been known and used in mobile systems since their infancy, all the way from voice channel capacity to LTE carrier aggregation for MBB. It is also applicable to FWA.

The logical consequence of this is that spectrum assets should be shared as one pool, employing carrier aggregation for LTE and dual connectivity for LTE/NR to ensure that all resources are utilized to the maximum, while securing good user experience for both MBB and FWA. Quality across both services is ensured through existing software features such as RAN slicing.

By contrast, any artificial split of spectrum resources for different services would result in under-utilization of the spectrum assets.

Performance differences of FWA CPE types

Using FWA to deliver broadband services requires new FWA customer premises equipment (CPE), from simple indoor nomadic devices to fixed outdoor-installed units, provisioned through standard device retail or new methods. A CPE management system is likely to be needed to manage CPE in the fixed broadband sense – enabling the operator to log in to the devices, configure them and check status remotely. Converged operators have the choice of reusing the fixed access CPE management system or deploying a separate one for FWA. Both CPE and CPE management systems are separate network entities that generally have limited integration with cellular networks, meaning that the operator can acquire best-of-breed products and expect them to work using standard protocols. The biggest difference between the CPE alternatives is the ability to achieve promised service levels, especially during busy hours.

An outdoor CPE provides the best performance, as it has a built-in directional antenna (3.5GHz, 10-14dBi) and is installed with a predictable radio link quality to the selected base station. The typical antenna configuration has two Rx antennas, but devices with four Rx antennas are also available. The normal transmission mode is rank-2 MIMO, as the modem is expected to be installed with good line-of-sight. Most outdoor LTE devices support CAT 6 and 20+20MHz carrier aggregation but more advanced devices up to CAT 16 support are also available. Inter-band carrier aggregation between FDD and TDD is especially useful, as services can be started on existing FDD bands and later expanded as FWA subscribers and traffic increase.

A correctly installed outdoor CPE is directed to the best-serving cell, leading to a lower path loss and increasing the value of mid-band and mmWave TDD spectrum. The large gain in signal quality is a result of the 10dB difference in antenna gain and the avoidance of 10-15dB in wall/window attenuation losses suffered by indoor devices. Another contributor to signal attenuation for indoor devices is the deep indoor loss, as the device is likely to be placed in a hidden location or to provide optimum Wi-Fi coverage. This could contribute another 5dB in path loss.

Whereas an indoor CPE is comparable to a smartphone in terms of spectrum efficiency, an outdoor CPE is two to three times more efficient. To put it another way, for the same data consumption, around two to three times as many households can be served using outdoor rather than indoor units – or two to three times as much spectrum would be needed to serve indoor-only FWA households. A final advantage of outdoor CPE is that the relative performance difference between the best, median and worst five-percentile users is significantly lower.

In terms of performance, indoor CPE units normally start with CAT 6 capabilities of up to 300Mbps. More advanced devices could support CAT 16 up to 1Gbps and offer rank-4 MIMO. More advanced CPE architectures are also being discussed, such as a split design, where an outdoor window antenna is connected to an indoor unit via induction through the window glass.

Case study: the country town

The country town example represents a market within our Build with Precision segment, characterized by relatively mature LTE MBB and decent fixed broadband offerings, complemented by terrestrial or satellite broadcast services to meet households' linear TV needs. The typical monthly ARPU for MBB is around USD 20, and the predicted willingness to pay is USD 40 for a dedicated household FWA internet service with a sold rate of 50-200Mbps and unlimited data.

The operator uses the following as the basis for dimensioning the system:

- The network should be designed to be able to connect at least 30 percent of households. In contrast to the extensive upfront investments required in a fiber deployment, the ability to design and invest for a limited market share from the beginning and expand later as the subscriber base grows is a useful property of FWA.
- There is no ambition to offer IPTV over FWA, as household TV needs are assumed to be served by satellite or terrestrial access.
- The dominant use case is meeting all the households' internet needs.
- For video streaming support, households should, when needed, experience at least a minimum data rate (Rmin) of 10Mbps even during busy hours. This corresponds to one high definition TV video stream, with some margin, or a combination of multiple standard definition TV streams.
- Based on the operator's experience from similar FWA areas, the average household's consumption during busy hours is 0.9GB/h, corresponding to an average data flow of 2Mbps during busy hours. With the assumption that 10 percent of data is being consumed during busy hours, this would correspond to 270GB per month.

Network starting point

Coverage is provided by a macro network with three-sector sites and an inter-site distance of about 1km. The operator has access to six FDD bands: three bands below 1GHz (typically 700, 800 and 900MHz), and three bands in the 1-3GHz range (typically 1,800, 2,100 and 2,600MHz). The MBB traffic in this area is handled using a subset of the available bands. The majority of smartphones are LTE-capable, and there is also GSM and WCDMA coverage to handle simpler phones. A typical macro site has two LTE carriers (800 and 1,800MHz) as well as a WCDMA carrier in the 2,100MHz band, and a few GSM carriers in the 900MHz band.

High-level analysis [2] has shown that the deployed LTE capacity in western and central Europe is less than 40 percent utilized, given the LTE smartphone subscriber density in the area. This means that there is spare radio capacity that can be utilized by FWA.

> THE ABILITY TO DESIGN AND INVEST FOR A LIMITED MARKET SHARE... IS A USEFUL PROPERTY OF FWA

Overall solution

We recommend utilizing the existing sites, radios and baseband deployed to provide MBB, and sharing these resources across FWA and MBB users. Current deployments have spare capacity both in LTE carriers and in baseband units.
In addition, we recommend utilizing the acquired

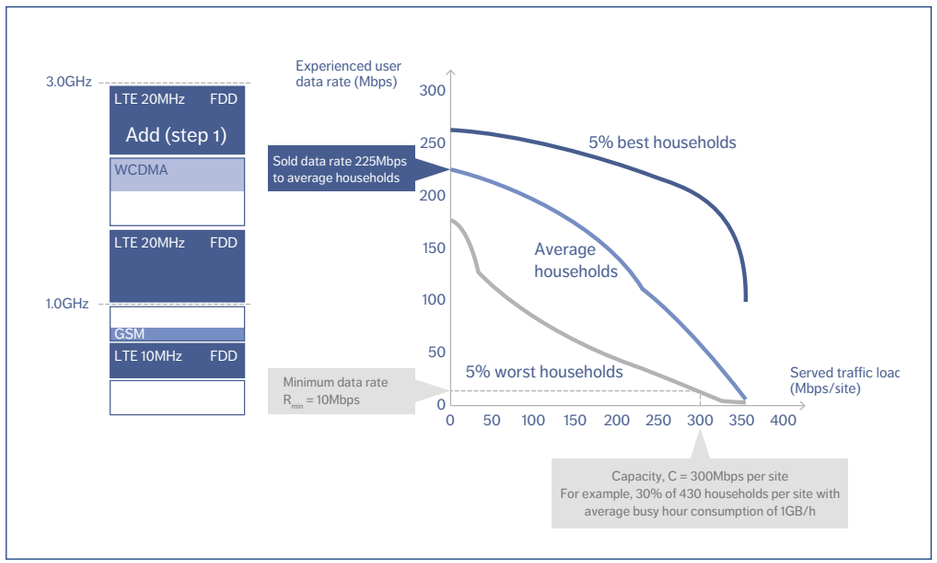

Figure 2: Performance and spectrum use of FWA deployment step 1

but undeployed band below 3GHz (such as 2,600MHz), with a new 2 Tx/Rx radio, together with the existing LTE bands by means of carrier aggregation for both FWA and MBB. Carrier aggregation improves peak speeds as well as coverage for both services. The left side of *Figure 2* shows the spectrum use of the FWA deployment at this first step deployment. A RAN slicing feature can be applied to ensure that there is no negative impact on MBB services (and vice versa) during peak loading as a result of FWA and LTE users sharing the same carriers.

There is no need to densify the network in this case. With regard to CPE choices, we suggest using high-end 4 Rx outdoor (roof-top mounted) CPE, as FWA speeds need to be high in this case to compete with xDSL services in the area. Indoor CPE may be deployed as a complement for households where their performance is acceptable.

Performance analysis

Although MBB and FWA services share spectrum in the country town case, to simplify the presentation of the performance analysis, our evaluation only shows FWA. Further, we have chosen to focus on the downlink (DL) because the FWA traffic (and broadband traffic in general) is DL-heavy and so capacity is DL-limited.

The performance is illustrated in Figure 2. The experienced DL data rate for a specific household depends on its location, as with xDSL services, and may be up to 270Mbps in this scenario. An average household would experience around 225Mbps at low system load. This could be used as the sold data rate to a typical customer.

Note however that, unlike MBB, where users move around and experience both good and bad radio environments, in this scenario the CPE is fixed and variation in the radio environment is smaller,

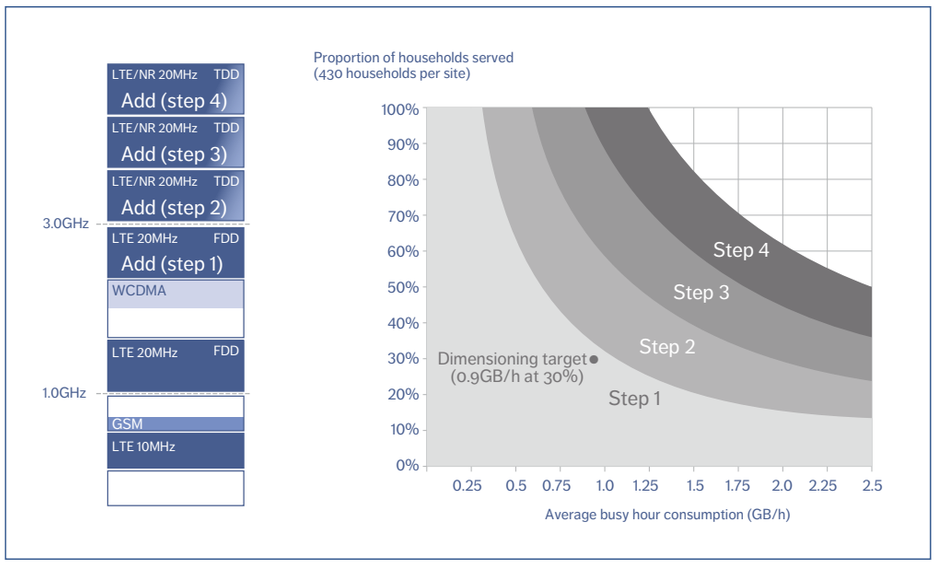

Figure 3: FWA deployment solution evolution to steps 2-4: spectrum use and performance

meaning that households with worse radio environments will likely always have worse than average data rates. In this scenario, the five percent worst-performing households experience close to 175Mbps at best. Therefore, it may be worth considering having different subscription categories; it may not be possible for all households to subscribe to the higher service level.

To dimension the system, the Rmin is set to 10Mbps. This means that the five percent worst-performing households should experience at least 10Mbps DL data rate during busy hours. This results in a capacity of 300Mbps, or 135GB/h, per site. As long as the total traffic in all three sectors does not exceed 300Mbps, the Rmin requirement will be fulfilled.

Assuming there are 500 households per square kilometer, and an inter-site distance of 1,000m, an FWA market share of 30 percent corresponds to some 130 households per site. At 135GB/h capacity, this market can be served with an average busy hour consumption of slightly above 1GB/h – that is, above the dimensioning target of 0.9GB/h (2Mbps). In addition, MBB will benefit from the additional 20MHz spectrum, for example in terms of increased peak rates.

Solution evolution

It is important that the solution is future-proof and can evolve to handle more connected households and higher demand per household over time. To provide higher capacity and cope with greater demands, operators can acquire and add a new TDD band above 3GHz (such as 3.5GHz) using 8 Tx/Rx advanced antenna system radios. The multi-user MIMO feature can be activated to provide additional capacity. *Figure 3* illustrates how additional capacity can be provided in several evolution steps.

Initially, the system is dimensioned to serve 30 percent of households with an average busy hour consumption of 1GB/h. The area of the graph in Figure 3 marked as Step 1 indicates the possible combinations of percentages of households and average busy hour consumption for this solution.

The area of the graph that is marked as Step 2 indicates the capacity provided by an additional 20MHz. This shows that the system can serve a customer base of 30 percent with an average busy-hour consumption of 1.9GB/h. Alternatively, the higher capacity can be used to serve an increased market share (up to 58 percent) with an unchanged average busy hour consumption.

Increasing the bandwidth with another 20MHz of TDD spectrum provides a system capacity represented by the area marked Step 3 in the graph. This will serve 30 percent of households in the area with an average busy-hour consumption of 3GB/h. Again, the higher capacity could instead be used to serve an increased market share with an unchanged average busy-hour consumption, or a combination of increased market share and increased average consumption.

Finally, Step 4, the darkest grey area of the graph in Figure 3, indicates what can be achieved when a total of 60MHz of TDD spectrum is added beyond Step 1. Assuming a 30 percent market share, an average busy-hour consumption of up to 4.1GB/h can be met (outside graph range).

In summary, by using the FWA toolbox and limited initial investments, and then adding TDD spectrum as needed, the chosen deployment is able to support high data rates and consumption immediately at launch. Then, through a series of smooth solution evolution steps, capacity can grow to more than four times the initial offering.

Conclusion

The large number of underserved households around the world represents a profitable FWA growth opportunity for current 3GPP operators. Mobile-only operators can explore a new business opportunity with FWA, and converged operators can add FWA as a complement to their fixed broadband strategy for certain locations as a more cost-efficient solution with faster time to market. Segmented solutions are needed, with subscriptions and dimensioning based on fixed and mobile paradigms. We believe that the best way to deliver future-proof broadband solutions is based on the evolution of LTE and 5G NR, and that the most promising approach is shared investment using the same ecosystem, assets and spectrum bands for both MBB and FWA.

References

1. **Ericsson, Fixed Wireless Access Handbook (extracted version), available at:** *https://www.ericsson.com/assets/local/narratives/networks/documents/fwa-handbook.pdf*
2. **Ericsson Mobility Report, November 2017, available at:** *https://www.ericsson.com/assets/local/mobility-report/documents/2017/ericsson-mobility-report-november-2017.pdf*

Further reading

- **Ericsson Technology Review, Fixed wireless access on a massive scale with 5G, 2016, Furuskär A; Laraqui, K; Nazari, A; Skubic, B; Tombaz, S; Trojer, E, available at:** *https://www.ericsson.com/en/reports-and-papers/ericsson-technology-review/articles/fixed-wireless-access-on-a-massive-scale-with-5g*
- **Ericsson ConsumerLab, Connected homes, June 2015, available at:** *https://www.ericsson.com/assets/local/news/2015/6/ericsson-consumerlab-connected-homes.pdf*

The authors would like to thank the following people for their contribution to this article: Tomas Dahlberg, Hani Elmalky, Bo Göransson, Henrik Johansson, George Jöngren, Michael Kühner, Per Lindberg, Staffan Lindholm, Reiner Ludwig, Claes Martinsson, Björn Möller, Richard Möller, Per Arne Nilsson, Christoph Schrimpl-Rother, Sibel Tombaz, Henrik Voigt, David Waite and John Yazlle.

THE AUTHORS

Håkan Olofsson

◆ has worked in the mobile industry for 25 years, with a particular focus on its RAN aspects. He joined Ericsson in 1994 and has served in a variety of capacities, mostly dealing with strategic technology development and the evolution from 2G all the way to 5G. He is currently head of the System Concept program at Development Unit Networks. Olofsson holds an M.Sc. in physics engineering from Uppsala University, Sweden.

Anders Ericsson

◆ joined Ericsson in 1999 and is currently working as a system designer in Development Unit Networks. During his time at Ericsson, he has worked at Ericsson Research and in system management, as well as heading up the Algorithm and Simulations department at Ericsson Mobile Platforms/ST-Ericsson. He previously worked at the Swedish National Defense Research Establishment (FOI). Ericsson holds a Licentiate Eng. in automatic control and an M.Sc. in applied physics and electrical engineering from Linköping University, Sweden.

Fredric Kronestedt

◆ joined Ericsson in 1993 to work on RAN research. Since then, he has taken on many different roles, including system design and system management. He is currently working at Development Unit Networks, where he is focusing on radio network deployment and evolution aspects for 4G and 5G. Kronestedt holds an M.Sc. in electrical engineering from KTH Royal Institute of Technology, Stockholm, Sweden.

Sven Hellsten

◆ joined Ericsson in 1993 and over the years he has worked with radio technologies ranging from analogue AMPS, GSM, WCDMA, to LTE and 5G/NR. His main focus has been on product management of base stations, but he has also worked with signal processing design and systems management. Hellsten holds an M.Sc. in physics engineering from Uppsala University.

ISSN 0014-0171
284 23-3369 | Uen

Ericsson
SE-164 83 Stockholm, Sweden
Phone: +46 10 719 0000