GSMA Intelligence

Decarbonising compute: from the ground up

February 2024

GSMA

The GSMA is a global organisation unifying the mobile ecosystem to discover, develop and deliver innovation foundational to positive business environments and societal change. Our vision is to unlock the full power of connectivity so that people, industry, and society thrive. Representing mobile operators and organisations across the mobile ecosystem and adjacent industries, the GSMA delivers for its members across three broad pillars: Connectivity for Good, Industry Services and Solutions, and Outreach. This activity includes advancing policy, tackling today's biggest societal challenges, underpinning the technology and interoperability that make mobile work, and providing the world's largest platform to convene the mobile ecosystem at the MWC and M360 series of events.

We invite you to find out more at www.gsma.com
Follow the GSMA on Twitter: @GSMA

GSMA Intelligence

GSMA Intelligence is the definitive source of global mobile operator data, analysis and forecasts, and publisher of authoritative industry reports and research. Our data covers every operator group, network and MVNO in every country worldwide — from Afghanistan to Zimbabwe. It is the most accurate and complete set of industry metrics available, comprising tens of millions of individual data points, updated daily.

GSMA Intelligence is relied on by leading operators, vendors, regulators, financial institutions and third-party industry players, to support strategic decision-making and long-term investment planning. The data is used as an industry reference point and is frequently cited by the media and by the industry itself.

Our team of analysts and experts produce regular thought-leading research reports across a range of industry topics.

www.gsmaintelligence.com

info@gsmaintelligence.com

Authors

Tim Hatt, Head of Research and Consulting

Shiv Prashant Putcha, Director, Consulting

This report was authored by GSMA Intelligence with support from Arm.

Arm

Arm technology is building the future of computing. Our energy-efficient processor designs and software platforms have enabled advanced computing in more than 280 billion chips and our technologies securely power products from the sensor to the smartphone and the supercomputer. Together with 1,000+ technology partners, we are enabling artificial intelligence to work everywhere, and in cybersecurity, we are delivering the foundation for trust in the digital world – from chip to cloud. The future is being built on Arm.

www.arm.com

Contents

Executive summary

5G, digitisation and AI feed into demand for compute resources

Between 2022 and 2030, data traffic is projected to rise sixfold. One of the biggest implications is the need for a corresponding increase in compute resources. As 5G network deployments increase in volume and scale globally, many operators are also investing heavily in network automation to handle increasingly complex network management workloads.

At the heart of most automation strategies is the introduction of artificial intelligence (AI) and machine learning (ML) into the network. Telecoms networks already have a number of use cases for AI, starting with generative AI in operations and call centres, and progressing to AI/ML deployed in the RAN. Utilising AI/ML in the RAN requires a significant upgrade to the current baseband chipsets deployed, with a higher number of cores as well as AI engines featured in the designs. AI in the network will therefore drive a wave of demand for computing resources. This will manifest across the board – from on-device to edge to central clouds.

Increased compute requires improvements in semiconductor designs and manufacturing processes to keep pace with surging digital traffic. While significant advances have been made in this field, there are constraints in terms of the supply and cost of the energy they consume with current processes. Moreover, new chipset designs must plug into new networks and IoT grids, and be capable of reducing energy consumption compared to current rates.

New chipset designs must be made with sustainability as a bedrock principle. It is not sufficient to only make improvements to existing processes to drive energy efficiencies. Rather, it must be done from the ground up – and this starts with the process of decarbonising compute.

A multiplicative effect

Compute gains can be conferred on multiple levels of the stack through more efficient chipsets. This feeds through to overall reductions in network energy usage. This starts at the chipset level and extends to the cloud and ultimately devices, whether consumer (e.g. smartphones and tablets) or industrial (e.g. drones).

RISC and CISC (X86) architectures have prioritised energy efficiency in successive designs. Arm designs, in particular, have made strong advances by tailoring chip designs and processing algorithms to the specific devices they run on, whether a smartphone, cloud data centre or electric vehicle (EV) charge point. RISC has scaled immensely off the back of Arm reference designs and is now included in a strong majority share of smartphones, IoT devices and data centre/cloud infrastructure. The scale underpins the flow-through effect of energy efficiency gains (up to 40% in RISC) up the compute stack. This is why choosing the right compute architecture is so important. Arm's partnership with RedHat and its OpenShift containers using Neoverse CPUs explicitly targets energy efficiency as a unique selling point (USP) in targeting telecoms operators and other buyers of private wireless solutions.

The question, then, is how much compute efficiencies can translate into energy reductions in these infrastructure categories? By extension, the same goes for carbon emissions. According to GSMA Intelligence modelling, the amount is material.

Telecoms access networks (fixed and mobile) and the cloud each account for just over 1% of global energy usage, equivalent to around 300 terawatts (TW) each. From this starting point, if energy efficiency improved by 5% (enabled by more power-efficient chipsets), overall energy consumption by mobile networks would drop by 9 TW per year. The comparable figure for data centres is 17 TW. If compute energy efficiency improved by 25%, the same figures on an annualised basis would be 47 and 86 TW. There are caveats, of course, such as the changing share of renewables in the overall energy mix and the growth rate of data traffic. The projections provide a sense of the order of magnitude, rather than a precise forecast.

With the same gains in energy efficiency (e.g. 5%, 10% and so on), compute performance gains in access networks and data centres would, on their own, contribute 0.5% of the total worldwide CO2 reductions needed by 2030. This rises to 1.5% and 2.5% for 15% and 25% gains, respectively, in energy efficiency. This does not include the emissions savings in other industries that are enabled by 5G connectivity (e.g. for IoT or other B2B applications such as robotics) and cloud computing by increasing productivity and reducing waste.

Outlook: redressing the paradox

Net-zero carbon by 2050 is the goal, and commitments to net-zero targets are being embraced by the operator community around the world. Commitments have been made by operators that together cover a third of global market share, rising in the period following COP26. Commitments centred on the 2050 target date imply the need for CO2 reductions of 50% in each successive decade until then. Some operator groups have set more aggressive targets, including Vodafone (2040), Verizon (2040), Telefónica (2030) and Telia (2030), enabled by rapid moves to renewable energy in place of fossil fuels, especially in Europe.

These commitments and the projected surge in mobile data traffic mean there is a clear need to keep investing in additional computing resources. The incremental resources will power an increasing diversity and volume of connected end points, as well as data centres in central and edge locations of networks in a software-first environment. Cloud-native 5G RAN will undoubtedly carry through to 6G. We therefore expect to see flexibly provisioned compute architectures become the norm, helped by AI optimising resources for where and when they are needed.

It is telling that 50% of operators see themselves as competitively weak on sustainability in their product design and marketing, compared to only 30% for competitively strong, despite the consensus that energy efficiency is now a key purchasing criteria of enterprise tech buyers. The implication for chipset providers is to work at pace with operator partners, as well as the broader compute ecosystem, to redress this paradox.

1 Squaring the soaring demand for compute with carbon reductions

The 21st century is shaping up to be a digital century, with sustained growth in communications and internet connectivity, as well as rapid digital transformation of enterprise and government sectors across the world. These trends have led to continued investment in the construction of telecoms networks, including mobile networks and fixed line broadband access networks, optical-fibre transport networks and even subsea cable systems that transport the bulk of internet traffic around the world.

These developments have established the centrality of telecoms as a key enabler and driver of the 'digital age' that will accelerate during the century. However, the other (often countervailing) theme of the century is climate change and mitigation. This has spurred an increasing focus on reducing greenhouse gas emissions and shrinking carbon footprints at all levels, as well as a strong emphasis on building sustainable operating practices that extend through the supply chain.

1.1 Why sustainability is an imperative now

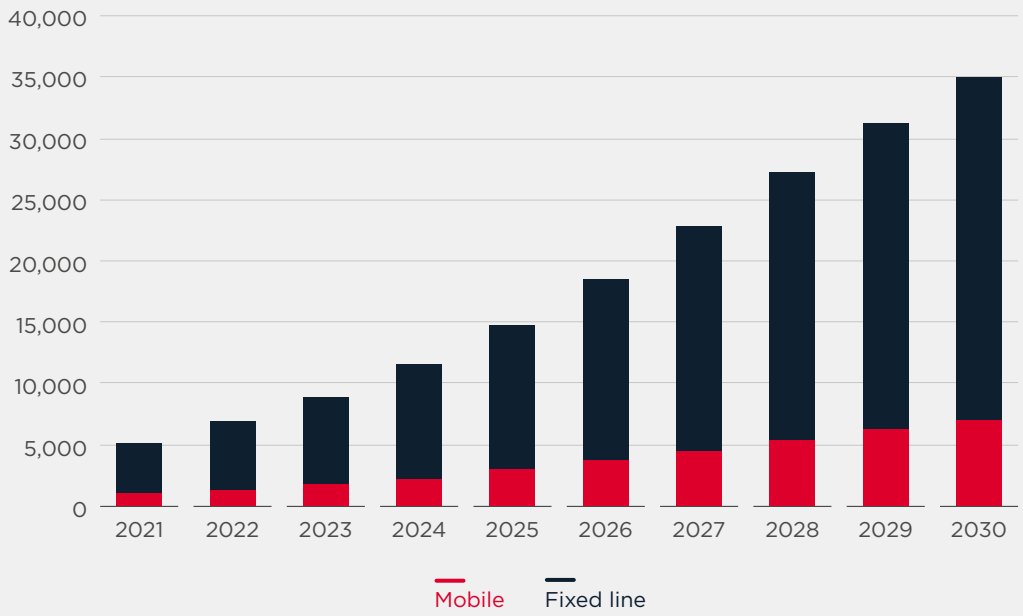
Today, there is significant momentum in shaping public policy towards meaningful change that drives sustainable practices. The 2021 UN Climate Change Conference (COP26) was an important moment in determining the progress of governments and businesses in addressing the impact of climate change over the previous five years. The International Telecommunications Union (ITU) has predicted that the ICT industry must cut carbon emissions by a minimum of 45% by 2030 to meet UN climate change goals. The telecoms sector is unique in the sense that its output (e.g. data traffic and number of served connections) is always increasing, while the services it offers (e.g. quality of service and service coverage) continue to improve rapidly. However, the resources consumed to sustain this growth are also increasing, forcing attention on reducing carbon emissions and driving sustainability.

Sustainability is now an imperative for governments and enterprises around the world, as we are at a critical juncture. The strong progress with connectivity and democratisation of internet access around the world is happening at the same time as increasing pressure is placed on resources – natural and generated. With surging growth in data traffic transmitted over mobile and fixed line networks, there is a clear need to keep investing in additional computing resources to power an increasing diversity and volume of connected end points, as well as data centres in central and edge locations of networks in a software-first environment.

Figure 1 offers a glimpse into the sheer scale of the growth in internet traffic from connected end points, growing sixfold by 2030.

Figure 1 Mobile and fixed (mostly fibre) traffic will rise sixfold by 2030

Exabytes per year



Source: GSMA Intelligence, Ericsson

One of the biggest implications of the projected rise in data traffic is the need for a corresponding increase in compute resources to process and route the traffic from the end points through to data centres and back. Increased compute requires improvements in semiconductor designs and manufacturing processes to keep pace with the surging digital traffic. While significant advances have been made in this field, there are constraints in terms of the supply and cost of the energy they consume with current processes. Moreover, new chipset designs must plug into new networks and IoT grids and be capable of reducing the energy consumption compared to current rates.

1.2 The rationale for the sustainability focus

Beyond the environmental imperative, an increasing focus on sustainability offers other benefits to the telecoms industry. Telecoms is at the heart of the digital revolution transforming daily lives and jobs. As such, decarbonisation changes made in the network will generate positive externalities within the telecoms domain, will have a positive cascading effect on other industries and verticals, and ultimately help with the greater goal of fighting climate change.

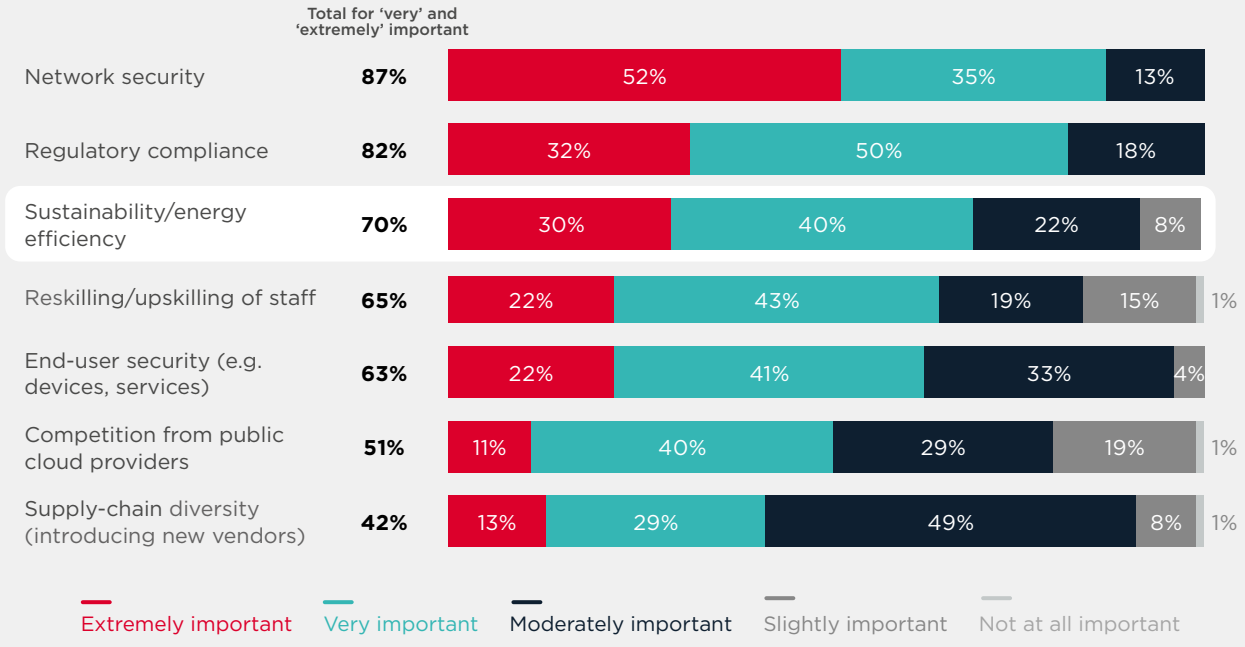
Benefits from decarbonisation will stem from a number of areas:

- **Financial benefits** – Telecoms operators today are faced with rising capex for 5G network deployments and input costs, especially energy. At the same time, top-line revenues remain flat, with consumer adoption of 5G yet to generate the monetisation opportunities initially touted. The 5G-Advanced features that will drive enterprise adoption are not yet mainstream. With this scenario expected to continue in the short term, any savings that accrue from lower input costs due to reductions in energy consumption will have a direct, beneficial impact on the bottom line.
- **Reputation and external relations** – With the existential challenge of climate change gaining mindshare around the world, companies can no longer afford to be on the sidelines or opt out of the debate around climate change and sustainability, as popular sentiment and official government policies have noticeably shifted in this regard. Burnishing sustainability credentials and driving change through the ecosystem with key stakeholders is increasingly important, whether with customers, employees, suppliers, investors or regulators.
- **Corporate sustainability goals** – Many operators have publicly made statements on their sustainability goals and net-zero approaches. There is increasing scrutiny from investors and regulators on their ability to hit these goals. With the majority of operators public companies, success will be rewarded with tangible benefits to the business. Even for private companies, their sustainability commitment and ability to deliver against these targets will have a significant impact on their ability to finance new capex and network deployments and more.

A growing number of operators have indicated that sustainability and decarbonisation are important priorities for them. See Figure 2.

Figure 2 Sustainability ranks among top network priorities for operators in 2023

How important are the following business priorities as a part of your current network transformation strategy?



Source: GSMA Intelligence Operators in Focus: Network Transformation Survey 2023

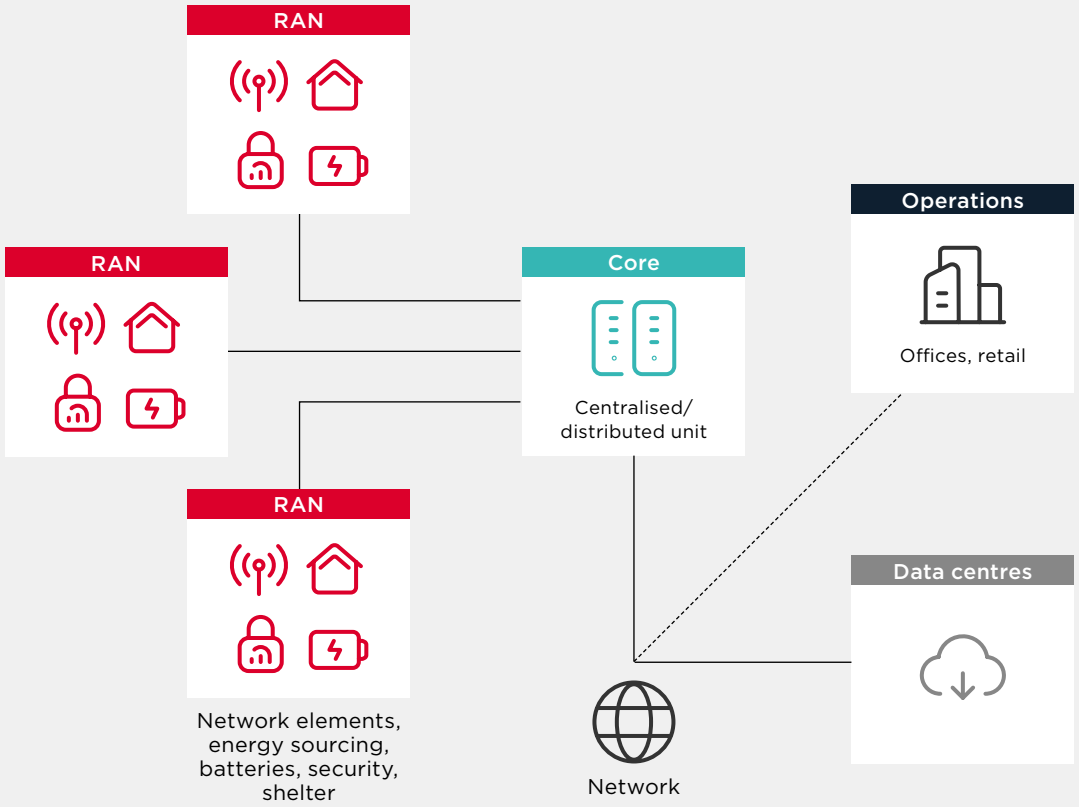
2 Decarbonising compute: the bedrock of ICT

2.1 Rising demand for compute

As networks expand to eliminate coverage gaps, the need to cover a much larger footprint will require an increase in energy consumption. Data provided by mobile operators as part of GSMA Intelligence's Mobile Energy Efficiency Benchmarking research provides a view of where energy is used across a mobile network. See Figure 3.

The majority of electricity (87%) is consumed in the RAN. Providing coverage across thousands of square kilometres, transforming energy into radio waves, and receiving and processing incoming signals are still energy-intensive functions. At the same time, the number of connections (both consumer and IoT) supported by the network is increasing as connection density grows. This means each 'base station', whether macro or small cell, will need to be much more capable in terms of capacity. This requires enhanced silicon in the baseband, which invariably consumes more energy. The remaining consumption is by data centres and the core network (12%) and operations (1%).

Figure 3 Where mobile operators use energy in their network operations



Source: GSMA Intelligence

There is also significant energy consumption in the passive infrastructure elements of a telecoms network. The role of passive infrastructure is to support, defend and supply the active network elements. There are significant variations between mobile sites, the regulatory and physical environments they operate in, and the traffic load experienced, based on country or location. Improving the energy efficiency of passive infrastructure can therefore be a complex and labour-intensive task. Depending on the climate and the quality of the electricity grid, passive infrastructure (especially air-conditioning) can be responsible for a significant part of operators' energy use, meaning the stakes can be high.

Meanwhile, there is a pronounced shift towards distributed architecture rather than today's highly centralised networks. New distributed elements, coupled with an increasing footprint, will require new techniques to drive energy efficiencies. Otherwise, the aggregate increases in energy consumption will drown out any per-unit improvements in energy efficiency.

2.2 Chipsets: the base of the pyramid

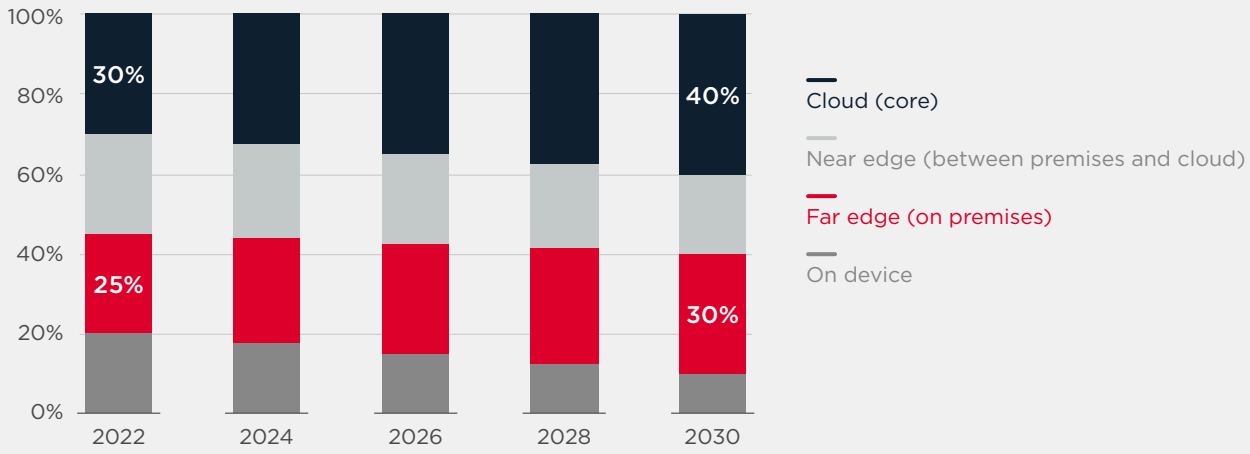
A key driver of internet traffic over telecoms networks is the global transition to 5G. This offers enhanced mobile broadband for consumers and new spectrum bands. These will not only offer increased capacity for consumer services; they will also cater to a range of new end points by enabling IoT connectivity for enterprise networks. Indeed, one of the major aspects of 5G network deployments is that the technology is capable of subsuming a number of segments previously working on disparate network technologies. IoT is a case in point; IoT devices currently running on multiple LPWAN technologies will eventually converge on 5G RedCap. This applies to end points deployed over private 5G networks, and includes connected end points and IoT modules that can slot or be retrofitted into legacy machinery and equipment.

The number of connected end points will rise sharply, extending well beyond the traditional smartphones used by consumers. With so many new end points coming online, operators will need to extend network coverage and add significant capacity across the network footprint to handle the expected surge in data traffic.

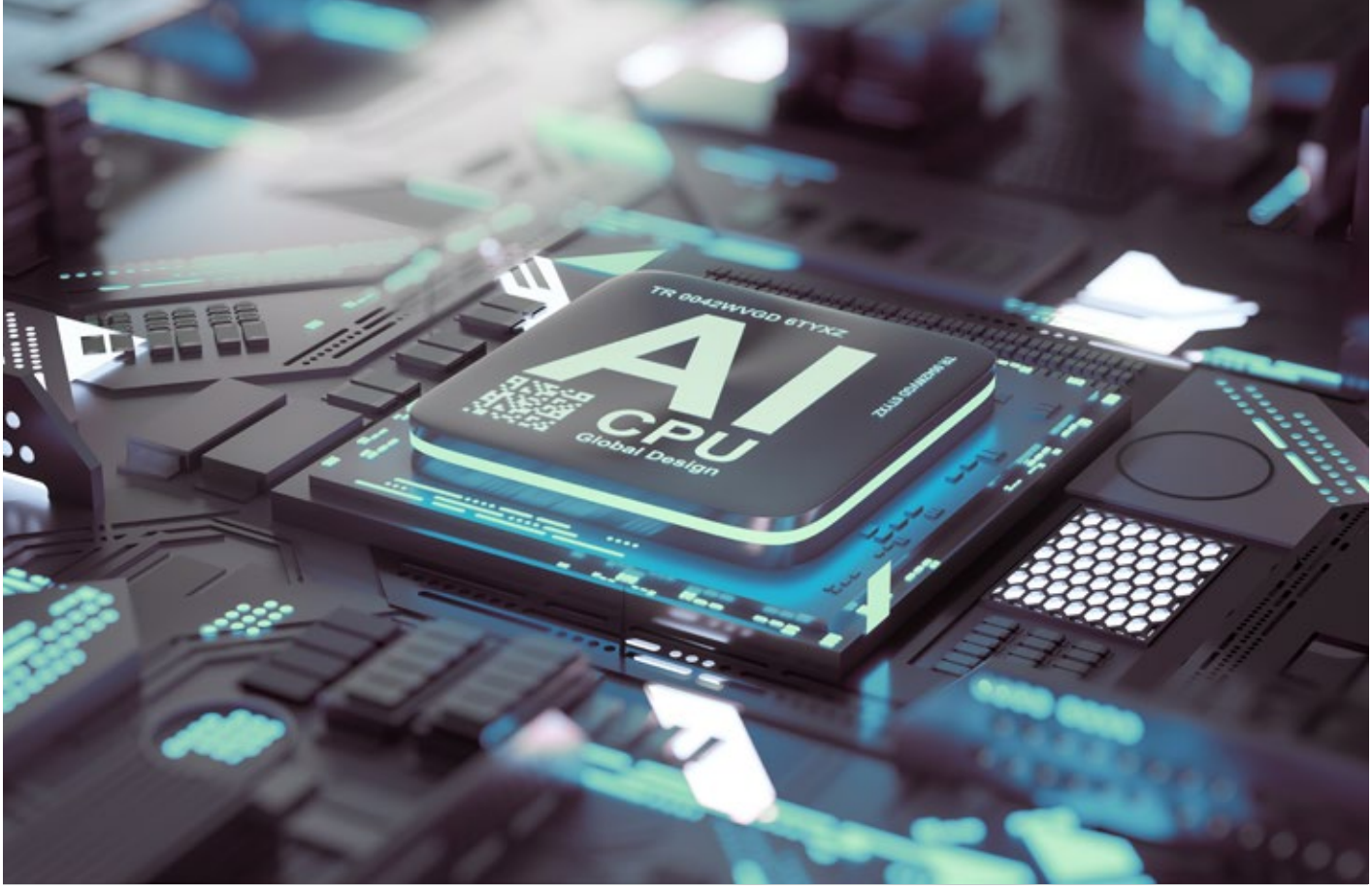
The net result of all the new end points is an aggregate increase in demand for compute resources, with both the cloud and on-premise edge soaking up the majority of incremental traffic. That said, even individual endpoint categories such as smartphones are seeing a sharp rise in compute power on-device, as they handle more and more applications and use cases (even if the total share drops as cloud and edge see larger incremental rises in traffic processing). See Figure 4.

Figure 4 Compute processing will be needed both in the cloud and at the network edge

Percentage of enterprise mobile traffic



Source: GSMA Intelligence



2.3 AI to the fore

As 5G network deployments increase in volume and scale globally, many operators are also investing heavily in network automation to handle increasingly complex network management workloads. At the heart of most automation strategies is the introduction of AI and ML into the network. Mobile networks already boast of a number of use cases for AI, starting with generative AI in their operations and call centres, and progressing to AI/ML deployed in the RAN. Utilising AI/ML in the RAN would require a significant upgrade in the current baseband chipsets deployed, with a higher number of cores as well as AI engines featured in the designs. AI in the network will drive a wave of demand for computing resources. This will manifest across the board, from on-device to edge to central clouds.

Another incremental gain from AI is that compute resources can be more flexibly provisioned. This is particularly important in the data centres that process over half of global internet data, given that traffic distributions are often uneven and disproportionately borne by a small number of points of presence (PoPs). Shunting traffic from 'hot' data centres to underutilised ones based on real-time analytics of compute workloads, and having flexible cores and accelerators, are examples of how to help maximise workload efficiency.

3 Impact pathways

3.1 Understanding electricity usage

To assess the potential impact of improving the energy efficiency of compute power and resources, it is first important to understand the global dashboard for electricity usage. See Table 1.

Mobile and fixed data traffic is either processed along the edge continuum – such as on the device or on-premises – or in data centres that make up the public cloud. Translating this activity into actual electricity consumption, operator access networks and the cloud each account for just over 1% of global energy usage. These industries have a higher use of renewables than the broader economy, which means their CO2 footprint is proportionately lower. In absolute values, access networks account for around 114 Mt of CO2 output per year (as of 2022), while data centres account for just under 130 Mt (excluding bitcoin mining).

The challenge is to mitigate the inexorably rising growth of global internet traffic. This now stands at 8,000 exabytes, with volumes rising around 40% per year, underpinned by 5G and fibre rollouts. Mobile networks handle around 20% of this, with fixed line (copper and the various fibre variants) taking the majority at 80%. The shift to renewables and energy efficiency are two sides of the same coin in getting energy usage down even as traffic rises. However, the focus of this analysis is the energy efficiency dimension up and down the compute stack.

Table 1: ICT accounts for 3–4% of global electricity usage, with access networks and data centres accounting for the bulk

2022	Electricity usage			CO2 footprint	
	Terawatt hours	Percentage of global total	Energy efficiency (kWh per GB)	Megatonnes CO2e	Percentage of global total
Mobile networks (excl. operator data centres)	168	0.6%	0.17	64	0.2%
Fixed line networks	132	0.5%	0.03*	50	0.1%
Total mobile and fixed line networks	300	1.1%		114	0.3%
Data centres	338	1.3%	0.16	128	0.3%
Global total (all industries)	**26,799**	**100%**		**37,857**	**100%**

*Average of different fixed broadband technologies, including copper, ADSL, VDSL and other fibre variants
Source: GSMA Intelligence, Aalto University, Finland

3.2 Translating compute efficiencies into energy reductions

Compute gains can be conferred on multiple levels of the stack through more efficient chipsets. This starts at the network level and extends to the cloud and ultimately devices, whether consumer (e.g. smartphones and tablets) or industrial (e.g. drones). The two largest (as reflected in the electricity consumption numbers) are networks and data centres. RISC and CISC (X86) architectures have prioritised energy efficiency in successive designs. Arm designs have, in particular, made strong advances by tailoring chip designs and processing algorithms to the specific devices they run on, whether a smartphone, cloud data centre or EV charge point. The company's partnership with RedHat and its OpenShift containers using Neoverse CPUs explicitly targets energy efficiency as a USP in selling to operators and other buyers of private wireless solutions. Similar energy efficiency gains are the core objectives of work with other cloud and IT groups such as Wind River, AWS, Microsoft and Google.

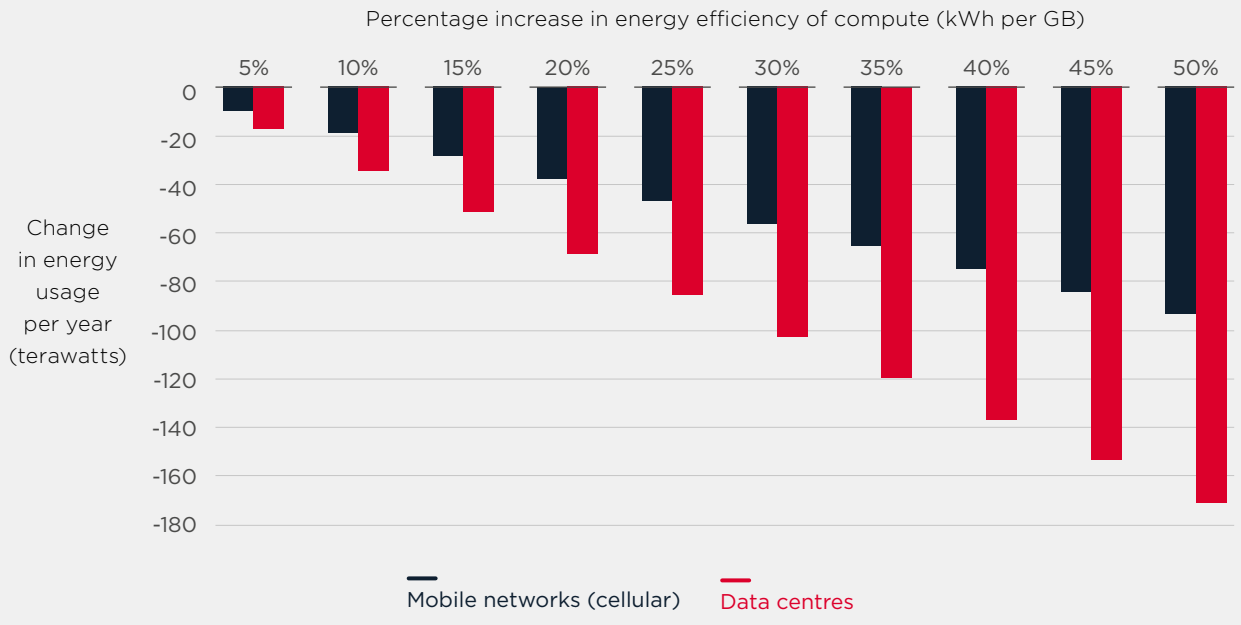
The question, then, is how much compute efficiencies can translate into energy reductions in these infrastructure categories? By extension, the same goes for carbon emissions. Based on GSMA Intelligence modelling, the answer is significant – see Figure 5.

This analysis works by assuming a certain level of improvement to the energy efficiency of data transfer at different levels of the network stack – again, only focusing on access networks and data centres. Starting from the baseline global energy usage that each of these two industries recorded in 2022, Figure 5 shows how much less energy usage would be on an annual basis at each level of compute efficiency gain.

For example, if energy efficiency improved 5%, the overall energy consumption of mobile networks would drop by 9 TW, with the comparable figure 17 TW for data centres. If compute energy efficiency improved by 25%, the same figures on an annualised basis would be 47 and 86 TW

There are caveats to this of course, such as the changing share of renewables in the overall energy mix and the growth rate of data traffic. The projections are indicative rather than a precise forecast. It is the direction of the trend that matters. The implied cost savings are also significant. For mobile operators, energy still accounts for 15–20% of opex. Of this, the RAN accounts for 90% of consumption. Any technology gains to bring energy usage down will, by default, feed through to EBITDA.

Figure 5 How much energy could be saved by improving the power efficiency of networks and computing



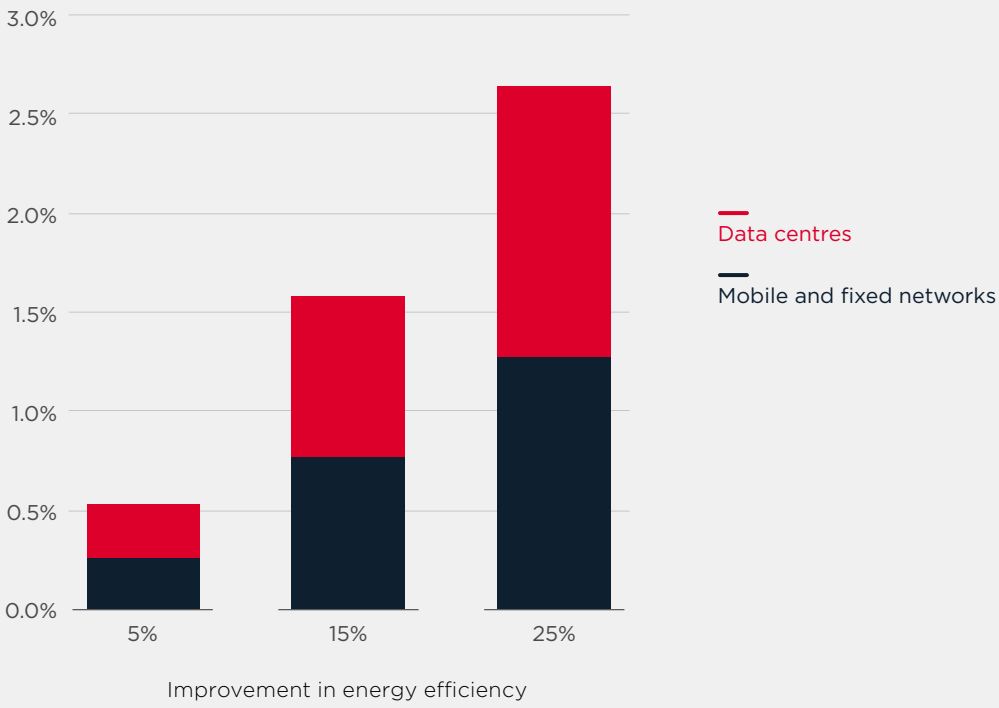
Source: GSMA Intelligence

The effect on CO2 emissions is equally important, as global reductions of 45–50% on 2021 levels are needed by 2030 just to stay on track for net zero by 2050. This makes the 2020s the hardest decade of the net-zero journey.

Playing through the same gains in energy efficiency (e.g. 5%, 10% and so on), compute performance gains in access networks and data centres would, on their own, contribute 0.5% of the total worldwide CO2 reductions needed by 2030. This rises to around 1.5% and 2.5% for 15% and 25% gains, respectively, in energy efficiency (see Figure 6). This does not include the emissions savings in other industries that are enabled by 5G connectivity (e.g. for IoT or other B2B applications such as robotics) and cloud computing by increasing productivity and reducing waste. This is the so-called enablement impact. On that level, GSMA Intelligence analysis suggests 40% of the global CO2 reductions needed by 2030 can be enabled by digital technology (see Industry Pathways to net zero).

Figure 6 Improving compute efficiency by 25% would contribute around 2.5% of the needed CO2 savings by 2030 for all industries combined

Percentage of required CO2 savings by 2030



Source: GSMA Intelligence

4 Outlook

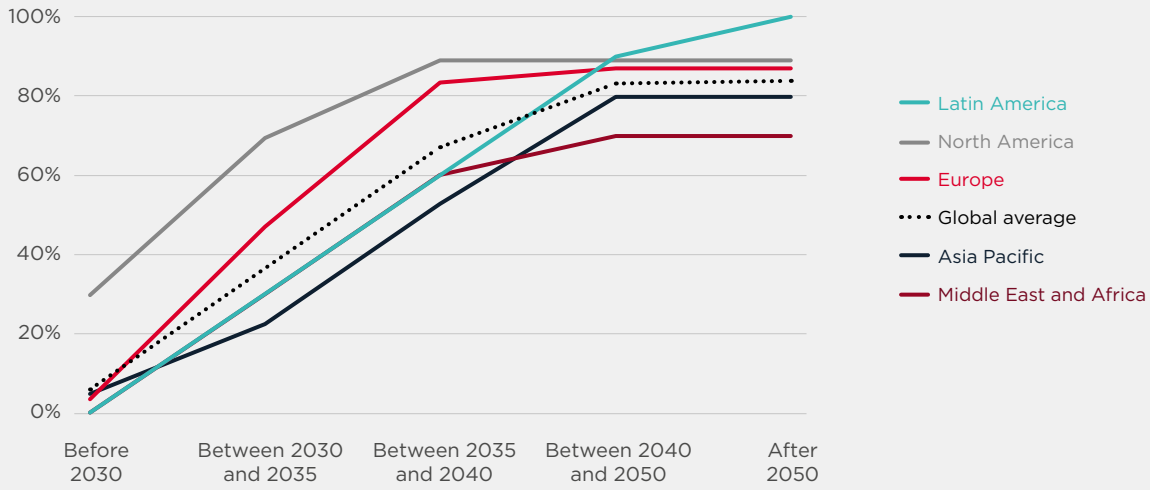
4.1 Sustainability as a bedrock principle

The interconnected nature of the network stack means new chipset designs must be made with sustainability as a bedrock principle. It is not sufficient to only make improvements on existing processes to drive energy efficiencies. It must be done from the ground up. This starts with the process of decarbonising compute.

The goal is net zero by 2050. Commitments to net-zero targets are being embraced by the operator community around the world. Net-zero commitments have been made by operators that together account for more than 50% of market share, with the recent COP28 in Dubai a further way marker. Commitments centred on the 2050 target date imply the need for CO2 reductions of 45–50% in each successive decade until then. Some operator groups have set more aggressive targets, including Vodafone (2040), Verizon (2040), Telefónica (2030) and Telia (2030), enabled by rapid substitution of renewable energy in place of fossil fuels, especially in Europe. In general, the rest of the industry has only expressed intent, with most expected to eventually adopt a 2050 target in line with the Paris Agreement.

Figure 7 Most operators have committed to net zero; many are on aggressive glidepaths ending sooner than 2040

Cumulative share of operators committed to net zero by timeframe



Note: the actual share of operators who have committed to net zero based on disclosures to CDP may differ from these figures.
Source: GSMA Intelligence based on survey of telecoms operators in June/July 2023 (N=100)

The 50% reduction in each decade means the window is considerably shorter than a 30-year horizon would suggest. This will accelerate competitive activity among chipset designers and producers to improve power efficiency. It will likely draw in R&D support from their own industry clients. In practice, that means hyperscalers (e.g. AWS, Microsoft Azure, Google), enterprise IT vendors (e.g. Dell, HPE) and device OEMs, given the strategic importance of getting it right. It is telling that 50% of operators see themselves as competitively weak on sustainability in their product design and marketing, compared to only 30% as competitively strong – despite the consensus that energy efficiency is now a key purchasing criteria among enterprise tech buyers. The implication for chipset providers is to work at pace with operator partners (as well as the broader compute ecosystem laid out above). It starts from the ground up.

GSMA
Intelligence

gsmaintelligence.com