Electronic Devices & Networks Annex
EDNA

Intelligent Efficiency for Data Centres and Wide Area Networks

MAY 2019

Intelligent Efficiency

For Data Centres & Wide Area Networks

Report Prepared for IEA-4E EDNA

May 2019

Electronic Devices & Networks Annex
EDNA

The Technology Collaboration Programme on Energy Efficient End-Use Equipment (4E TCP), has been supporting governments to co-ordinate effective energy efficiency policies since 2008.

Fifteen countries have joined together under the 4E TCP platform to exchange technical and policy information focused on increasing the production and trade in efficient end-use equipment. However, the 4E TCP is more than a forum for sharing information: it pools resources and expertise on a wide a range of projects designed to meet the policy needs of participating governments. Members of 4E find this an efficient use of scarce funds, which results in outcomes that are far more comprehensive and authoritative than can be achieved by individual jurisdictions.

The 4E TCP is established under the auspices of the International Energy Agency (IEA) as a functionally and legally autonomous body.

Current members of 4E TCP are: Australia, Austria, Canada, China, Denmark, the European Commission, France, Japan, Korea, Netherlands, New Zealand, Switzerland, Sweden, UK and USA.

Further information on the 4E TCP is available from: **www.iea-4e.org**

The EDNA Annex (Electronic Devices and Networks Annex) of the 4E TCP is focussed on a horizontal subset of energy using equipment and systems - those which are able to be connected via a communications network. The objective of EDNA is to provide technical analysis and policy guidance to members and other governments aimed at improving the energy efficiency of connected devices and the systems in which they operate.

EDNA is focussed on the energy consumption of network connected devices, on the increased energy consumption that results from devices becoming network connected, and on system energy efficiency: the optimal operation of systems of devices to save energy (aka intelligent efficiency) including providing other energy benefits such as demand response.

Further information on EDNA is available at: **http://edna.iea-4e.org**

This report was commissioned by the EDNA Annex of the 4E TCP. It was authored by Anson Wu of Hansheng Ltd, Paul Ryan of EnergyConsult Pty Ltd and Terence Smith of Mississippi Consulting Pty Ltd (with assistance and review provided by Henry Wong of E3HS IT Consulting). The views, conclusions and recommendations are solely those of the authors and do not state or reflect those of EDNA, the 4E TCP or its member countries.

Table of Contents

Glossary

ADSL	asymmetric digital subscriber line
AI	artificial intelligence
CDN	content delivery network
CPE	consumer premises equipment
CPU	central processing unit
DC	data centre
DRL	deep reinforcement learning
DSLAM	digital subscriber line access multiplexer
FAN	fixed access network
FTTC	fibre to the cabinet
FTTH	fibre to the home
GEPON	gigabit ethernet passive optical network
GPON	gigabit passive optical network
GPU	graphics processing input
ICT	information and communication technology
IoT	internet of things
IP	internet protocol
KPI	key performance indicator
MEC	mobile edge computing
MIMO	multiple-input and multiple-output
ML	machine learning
MPLS	multiprotocol label switching
MSAM	multi-service application module
NFV	network functions virtualisation
NFV-MANO	NFV management and orchestration
NG/XG	next generation
NIEE	network infrastructure energy efficiency
OLT	optical line termination
ONU	optical network unit
PON	passive optical network
PSTN	public switched telephone network
PUE	power usage effectiveness
QoE	quality of experience
QoS	quality of service
RAN	radio access network
SDN	software defined network
SEE	site energy efficiency
SEEM	server energy effectiveness metric
SLA	service level agreement
TWDM	time- and wavelength-division multiplexing
UPS	uninterruptible power supply
USA	United States of America
VDSL	very-high-bit-rate digital subscriber line
WAN	wide area network

Executive Summary

This report discusses the energy consumption of data centres (DC) and the wide area network (WAN) which connects computers and other devices together on the internet. These are complex, interconnected systems whose energy consumption is determined not simply by the sum of the hardware and products but also by the manner in which they interact and can be controlled. Opportunities to influence their energy efficiency are possible, and in particular new opportunities are arising from emerging "intelligent efficiency" techniques that can actively monitor and manage workloads and equipment. While energy consumption is not expected to rise substantially, research shows that theoretical energy savings opportunities of up to 75% exist in parts of the DC/WAN system (Section 4.4) if these techniques are deployed. Policies are recommended to realise as many of these opportunities as possible.

The WAN can be broken down into three parts, the Radio Access Network (RAN) which connects mobile devices to the internet including 4G networks, the Fixed Access Network (FAN) connecting homes and offices including broadband, and the high-speed core network which connect regions together (Figure 1). Data centres provide the business and consumer end user applications which run on a platform installed on ICT equipment. To ensure reliable service, DC/WAN are designed with redundancy so operation continues in the event of equipment failure. This includes infrastructure for environmental and electrical control. The DC/WAN can be further classified into three generations, legacy, modern and next generation. Each new generation is distinguished by new technology that cannot be used as a simple 'drop-in' replacement for older equipment but often requires a combination of new skills, new hardware and/or new software to integrate into, or replace, existing networks and data centres.

Metrics have been developed with the involvement of many stakeholders nationally and internationally, to measure the efficiency of different parts of the DC/WAN. The primary metrics can be found in the ITU L.13xx recommendations for networks, and the ISO/IEC 30134 standards for data centres. Due to the differences between DC/WANs, no metric is able to compare different systems and instead (Key Performance Indicators) KPIs cover a number of different metrics and are targeted primarily as tools for operational tracking and management of efficiency over time. The most common metric is the Power Usage Effectiveness (PUE) which measures the efficiency of the infrastructure as a proportion of the total energy consumed. Metrics and testing methods for equipment are the second most common type of metric, and typically measure the average efficiency at a number of utilisation points including at idle. Measuring at multiple utilisation points is needed to better represent actual use and because power is not perfectly proportional to utilisation - at idle equipment can consume at 30-70% of the peak power. In the area of IT efficiency, much work remains to develop a metric that can assess the utilization over time of IT capacity (CPU, memory, I/O and storage) and workload delivered per unit of energy consumed in a way that is simple and effective. This is a difficult problem, the solution for which will depend on using data collected by automated systems in a 'simple' form (not a research project) to give a meaningful but not overly complex assessment of capacity utilization. Work also needs to be done to understand the true limit of 'maximum' efficiency for different types of data centres.

Data demand between data centres and end users is projected to grow at over 20% a year for the near future (Cisco, 2018). This is primarily driven by media consumption, video, VR (virtual reality) and gaming. If efficiency were to remain unchanged, energy consumption would rise at the same rate. However, historically efficiency has improved at a similar rate and maintaining this balance in the future will require efficiency to continue to improve. A number of new technologies for DC/WANs are becoming more common and could have the potential to reduce energy consumption. However, current experience suggests they are being used to create new services and not to improve efficiency. In addition, software developers have a very significant influence on the amount of data generated as well as how and where data is stored and processed. This strongly influences the efficiency of the service they are providing (e.g. video streaming) and is not the responsibility of the DC or WAN operators.

- The Internet of Things (IoT) includes sensors to monitor and actuators to control equipment and systems. The very rapid growth of IoT (Cisco, 2018) is expected to increase energy consumption, improved control of equipment can increase the energy efficiency of systems such as building management systems and industrial processes, more than offsetting any consumption of the sensors and control software.
- 5G is the next generation of mobile network (RAN) that will have higher download speeds and faster response. It will enable a massive increase in the number of devices and services such as autonomous driving, virtual reality (VR) and various health and safety services.
- Software Defined Networking (SDN) enables fine grained control and management of the data travelling through the WAN and the equipment in real time from a central controller. This would make it possible, for example to shutdown underutilised equipment at night and reroute data through other equipment and save energy.
- Network Function Virtualisation (NFV) uses virtualisation technology to replicate network services on the WAN that have traditionally required individual and specialised hardware, and easily create new services. Like server virtualisation, this can reduce the hardware requirements and improve efficiency.
- Heterogeneous computing refers to systems with one of more type of processor to perform different tasks. By using specialised hardware, the processor can perform calculations much more rapidly, and more efficiently. This is already common for artificial intelligence (AI) applications where speed, rather than efficiency is the main driver. Currently, this requires the software to be developed specially to take advantage of the new hardware, which means older software will not benefit from this.
- Content delivery networks, edge computing and fog computing are methods of distributing data and processing power geographically around the WAN so it is closer to the end user and can connect faster. By 2022, 72% of all data is expected to be served by CDNs (Cisco, 2018). The distribution means that more equipment is required and data and processing is being replicated which could increase energy consumption. However, careful choice of location and management can reduce the amount of traffic in the WAN since each connection is closer which can save energy overall.
- Artificial intelligence, specifically deep reinforcement learning, is a method for analysing data and identifying patterns and predicting outcomes that are too complex for ‘smart’ algorithms defined by humans, especially in non-linear systems. While current ‘smart’ systems can

already reduce cooling energy in DCs by 20-30%, AI has shown the ability to save even more. In the context of energy savings, AI can replace 'smart' algorithms such as the cooling infrastructure as long as there is sufficient data and the equipment can be controlled centrally. However, the analysis can be very energy intensive and slow depending on the complexity. In addition, it may behave unpredictably in the event of unexpected events which could risk the reliability of the DC/WAN operation.

A review of case studies and research papers identified new intelligent efficiency techniques that can take advantage of existing and new technologies to improve efficiency of modern and next generation DC/WANs (see Figure 11). The following technical requirements were identified for implementation of intelligent efficiency:

- Energy aware hardware and equipment that can report power consumption and be managed remotely.
- Interoperability of the equipment that enables it all to be managed by a single controller.
- Data and monitoring of the system for analysis by AI or other smart controllers
- Information accessibility where the data is produced by one company but influences the operational efficiency of another.
- Software to take advantage of the energy efficiency opportunities available.
- Automation and AI to monitor and control the system continuously and in real time.

Opportunities to save energy were found in every part of the DC/WAN, by reducing the total workload and dynamically moving workloads between underutilised equipment, allowing operation at higher utilisation and efficiency and shutting down idle equipment. This was most pronounced at night due to a common diurnal pattern of high utilisation during the day and low utilisation at night. While some techniques are commercially available, such as using machine learning to optimise server virtualisation, and can be taken forward directly by policies, removing barriers and creating the general conditions for intelligent energy techniques to be widely implemented and targeting the parts of the DC/WAN with the highest potential savings is the main goal.

Modelling of the DC/WAN showed that energy consumption is relatively flat for DC/WAN, with the exception of rapidly falling historic energy consumption of 2G networks. Legacy DC/WAN consumes energy far out of proportion to the amount of work done - over 30% of DC energy (Figure 12) for 8% of the work (Figure 14)Figure 14. Modern networks are much more efficient but the volume of work means they consume the highest amount of energy. Within the WAN, the core network consumes only a small fraction (13%) of energy (Figure 13). The FAN is currently the largest consumer but will be overtaken by the RAN in future.

To understand how much additional energy is consumed due to equipment underutilisation, the model developed for this study calculates the energy that the DC/WAN would consume if the equipment could theoretically be operated continuously at the peak efficiency. This gives an indication of the potential energy savings that could be available through intelligent efficiency techniques, although never fully attainable under real operating conditions. The greatest potential for energy savings were found in modern (cloud) data centres (Figure 16) and the FAN (Figure 19). Large energy savings were also found in the legacy DC/WAN but are not considered suitable for these

techniques to be applied due to lack of future investment. Instead plans to shut down and transfer the legacy workloads may be the most effective solution.

The results of the research show that improved efficiency could substantially reduce the global energy consumption of the DC/WAN. The technology and techniques exist but it is unclear how much implementation will occur and how aggressively. Maximum energy savings will impact the service level being provided and a balance must be made between the two. Discussions with industry and stakeholders will probably need to be initiated by Government but are a necessary step. From the modelling and case studies, the following priorities and policy implications were identified:

- Raising the priority of energy efficiency in technology development with education, engagement with industry, feasibility studies and new research. The software development community is a key stakeholder and needs much greater engagement.
- Integrate intelligent efficiency into modern DC/WANs including minimum monitoring requirements including utilisation, energy aware equipment, and promotion of commercially available solutions for DC infrastructure and virtualised servers. In addition, energy management must become a more attractive job for AI experts and data scientists.
- Developing standardised detailed reporting for efficiency/power testing rather than over simplified metrics which have limited use when selecting equipment. Since every DC/WAN is different, the equipment will operate under different conditions, and limits the use of standardised metrics designed to emulate ‘typical’ conditions.
- Ensure next generation DC/WAN integrate intelligent efficiency. Equipment lifetimes are long and retrofitting equipment is expensive and unlikely. Ensuring interoperability from the outset and requiring efficiency reporting increases motivation and energy saving opportunities. This should also include addressing possible privacy and security concerns arising from the data collection that may be required.
- Data hierarchy and metadata standards can help manage the growth in data from IoT as well as improve security. Engaging with industry, particularly those operating the major consumer IoT platforms, could help this development.
- Shutting down legacy DC/WAN and supporting the transfer of workloads onto next generation DC/WAN through information provision to raise awareness, financial assistance and building trust by certification of businesses which carry out workload transfer services.

1 Introduction

This report discusses energy consumption of data centres (DC) and the wide area network (WAN) which connects computers and other devices together on the internet. These are complex, interconnected systems whose energy consumption is determined not simply by the sum of the hardware and products but also the manner in which they interact and can be controlled. Opportunities to influence the energy efficiency are similarly possible, and in particular the opportunities arising from new and emerging intelligent efficiency techniques that can actively monitor and manage workloads and equipment.

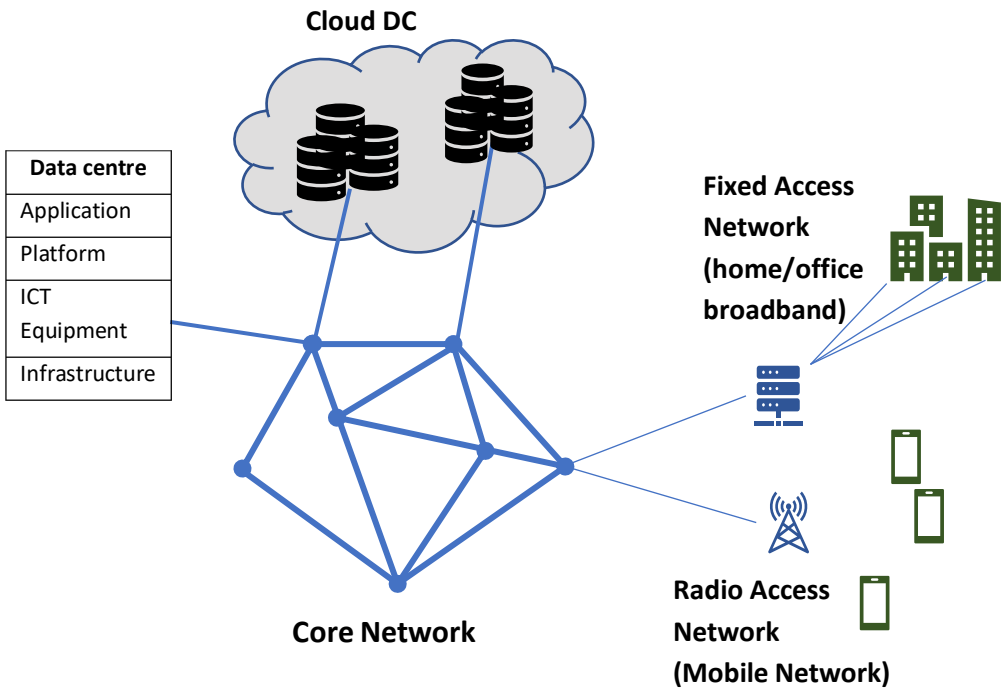
1.1 Description of the wide area network

This section covers the physical systems that exist, mainly the equipment and how it operates in the context of the services it provides and the key performance indicators (KPIs) that tend to govern the service. For example, the resilience of the service and the expectation that it can operate with virtually no interruptions or downtime is one the most important factors which influences the system design and management. There is frequently redundant equipment operating to minimise this. Any energy saving measures must also be aware of the impact it might have on a service, and therefore simply removing or switching off all redundant equipment is not a practical option.

While the internet is considered virtual, the WAN transports data to devices and servers spread geographically around the world. The most important functions for the WAN is to ensure the data reaches the correct destination completely intact.

The wide area network can be split into two parts, the core and the two types of access network (Figure 1) which connect together all the data centres with the premises and end-user devices.

Figure 1 Structure of the WAN



The core network are the main highways of the internet which connects the internet together. These can travel long distances and can carry very high volumes of data. Data centres will also be connected to the core network.

Modern core network transport is achieved almost entirely optically, although some electronic, radio (and satellite) is used. Optical transport use lasers to generate light signals carried along fibre optic cable. The electronic data is typically converted into optical signals which are then aggregated and transported into the core network. Routing equipment is located at various points in the network to make sure the data is heading to the right location. While all new equipment is optical, older networks are still operating in many countries including the backhaul and metro portions for traditional PSTN telephone lines. Core network equipment have long lifetimes and will operate for over 10 years and some legacy network equipment is over 25 years old (Krug, Shackleton, & Saffre, 2014).

The access network connects the individual homes and offices to the WAN. For mobile devices, base stations form the Radio Access Network (RAN). For premises, the connection is commonly made over existing telephone copper lines (ADSL), cable or fibre optics. Fibre optics can carry more data and installing fibre optic cables is cheaper than copper cable but for many regions, the copper cables already exist and therefore have no cost associated with installation. In this situation, physically installing individual fibre cables to every individual premises (fibre to the home, FTTH) creates additional access and equipment costs, particularly if the cable is installed underground and the premises are very far apart. Solutions that gradually bring the fibre closer and closer to the home such as fibre to the cabinet (FTTC) are common solutions to balance cost and service. In all these options, many premises will be connected to a single connection point on the network. These individual premises connections are the slowest on the WAN, however the access points connect and aggregate a large number of them before connecting to the core network. The Wi-Fi network, hotspots, ethernet and customer access devices within the premises are not part of the WAN.

There are currently many generations of radio network simultaneously in operation, 2G, 3G, 4G networks and soon 5G networks. 2G and 3G networks can be considered legacy networks which lack the technology and standards introduced in 4G networks which will be further developed in 5G networks. The legacy networks are still in wide operation, however because these networks share equipment, efficiency gains can continue to occur for these legacy networks. Backward compatibility and the heterogeneity of 5G networks may allow for a more consolidated approach to supporting legacy end point connections, whether through traditional RAN sites or Wi-Fi hotspots. Legacy networks are still in use and most voice communications are still being carried over these networks. Electricity smart meters and other infrastructure can also be dependent on the 2G network.

It is also common to identify an intermediate layer of the network between the core and access network called the metro network. The metro networks are geographically small, often within a metropolitan area and comprise a number of nodes connected together, which transport data between the metro nodes but mainly distribute data from and to the core. They can carry a medium volume of data. Data centres will also be connected directly onto the metro network since they have greater data bandwidth requirements than premises. However, it is becoming harder to distinguish the boundary between the core and metro networks, while the access network tends to have a more distinct role using different technology. Therefore, the metro network is considered as part of the core network for the purposes of this work.

1.1.1 Network services and topology

The WAN does not just carry internet data but a wide variety of services, including emergency communications, voice communications and private connections e.g. between company offices. Each type of service has different requirements and priority levels, with emergency service taking the highest priority. The most important criteria for a WAN connection is the availability and bandwidth. Businesses will often have service level agreements (SLAs) which define how much downtime is acceptable for a connection.

To achieve high availability, multiple routes are created between nodes which means that the data has an alternative path to travel if any equipment fails somewhere. This is typically undertaken using a mesh network.

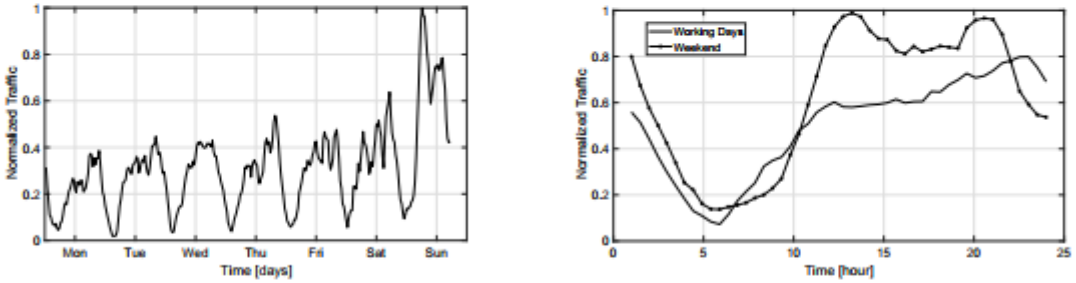
The service also needs to be fast, both in terms of time taken to respond (latency) and the volume of data carried (bandwidth). Particularly for voice/video calls and other real-time services, the connection must also be consistent and stable to allow for clear and uninterrupted communications.

Security and privacy of data is also extremely important. Though real-time services for voice and video streaming can be secured at the end points during the session, network and computing system intrusions and prevention techniques are becoming a greater source concern in providing secured uninterrupted services. Additional security mechanisms may reduce bandwidth and availability, increase latency, and increase energy consumption. As new end user services are being envisioned, data traffic and security requirements have increased the demands of the network both at the core and aggregation and edge sites. The expanding service requirements will determine how effective and efficient networks will need to be designed and operated.

Minimising the distance travelled minimises the latency and improves stability by avoiding travel through intermediate nodes and equipment. This can be achieved by creating as many direct connections as possible to avoid routing through intermediate nodes. However, this is also expensive and is generally only used when there are very high volumes of traffic. Where multiple routes do exist, the system will try to route it as efficiently as possible using a variety of techniques and algorithms.

In older networks, routing equipment have limited knowledge of the network, and may only be aware of the closest neighbours. As a result, the data would be sent sequentially through many nodes which analyse the data's destination and forward it to the next router before finally reaching its destination. This routing is one of the most energy intensive aspects of the network. Newer network routing equipment is more capable of communicating and coordinating amongst themselves. This means that it is possible to map a much larger part of a route initially and bypass the subsequent routers using a cross-connect. However, as the network gets bigger, there are limits to how much network knowledge each router can store and the optimal configuration and route becomes an increasingly difficult problem to solve.

Figure 2 Weekly and daily LTE traffic patterns

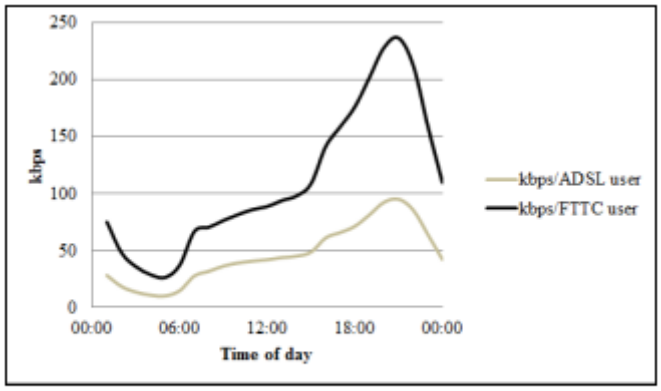


(c) Traffic per day. (d) Average daily traffic: weekdays vs. weekend.

Source: (Trinh, 2017)

Analysis of network utilisation show clear diurnal use patterns with consumption dropping very significantly during the night (Figure 2, (Trinh, 2017)). This is aggregated from a number of different nodes since the utilisation pattern is expected to vary depending on whether it is primarily residential or in a commercial zone. The diurnal pattern is expected to become more pronounced over time as the peak traffic increases faster than the average demand when higher bandwidth internet such as fibre to the cabinet (FTTC) replaces older ADSL (Figure 3). The equipment must be designed to be able to transport the peak level of data travelling through the node, and most likely with additional capacity for increasing data demands in the future. As a result, during off-peak times, next generation equipment is expected to be operating at even lower utilisation levels.

Figure 3 Average daily traffic patterns from UK residential broadband users



Source (Krug, Shackleton, & Saffre, 2014)

1.1.2 Business models

The WAN is largely privately owned and operated and there are many businesses, and sometimes Governments operating different parts of the core and access network, often in competition. Due to the importance of the services and the very high barrier to entry for physical infrastructure and radio frequencies, the company services tend to be highly regulated. This can include requirements to share and lease infrastructure capacity to reduce costs and help to maintain competition.

Infrastructure is generally expensive to build and electricity costs are a significant operating expense. Nokia estimate that electricity costs represent around 15% of operating expense for a mobile network operator (Nokia, 2016). As capacity and demand increases, these costs are expected to rise unless there are improvements in operation and technology, or reductions in energy price. The companies are therefore already incentivised to increase efficiency. However, there are trade-offs between capital investment and energy costs which may result in higher energy consumption, especially if highly efficient technology is significantly more expensive.

Each entity operates and manages their own networks and equipment but data will frequently need to pass between networks to reach the desired destination. To ensure data travels freely across networks, peering agreements between operators specify routing responsibilities and costs. The WAN operating businesses are largely separate from the end-user services but there can be significant overlap between the telecoms companies and data centre operators. Large Internet companies will operate, or jointly operate, parts of the WAN, while telecoms companies can own and operate a large number of data centres.

1.2 Data centre

Data centres (DC) provide the end user services and functions on the internet such as websites, email, streaming media, etc. This is performed in conjunction with the end user device by the ICT equipment housed in the data centre. The DC ICT equipment is most commonly split into network, storage and computer servers, although the distinctions are increasingly blurred. A platform such as the virtualisation layer and operating system is installed on the ICT equipment (servers), which manages the application software that provides the end user service. ICT equipment tend to be replaced relatively quickly, with lifetime estimates for servers of around 3-8 years (ASHRAE, 2016) and tend to be longer for legacy applications and shorter for cloud.

The infrastructure includes the building envelope, power and environmental controls which securely house the equipment, provide a reliable power supply and ensure a suitable operating environment. The infrastructure has a longer lifespan than the ICT equipment, around 10-15 years between retrofit and any major design changes. Figure 4 illustrates the simplified data centre stack.

Similar to the WAN, the end user requires availability, security, and speed from the data centre services. Ensuring availability is the primary role of the infrastructure as it controls the air temperature, humidity and cleanliness to reduce the risk to the servers from overheating, damage from rapidly fluctuating temperatures, electrostatics, condensation and contamination. Redundant systems such as multiple chillers, fans and computer room air conditioners are used to ensure that the system remains operational 24/7. There are multiple power routes and backup power systems (Uninterruptable Power Supplies, generators) to ensure no interruption of power in the event of equipment failure or planned maintenance. UPS can also manage the quality of the power to protect against sudden fluctuations in voltage and other problems. The building itself provides physical security and an envelope for all the systems.

Resilience can also be designed into the operating platform and application by spreading the service over many servers or locations and using a mechanism known as failure detection and recovery which routes the services to whichever server is available. Failure detection and recovery to resume services are key availability strategies used by large computing service providers and are increasingly being demanded by customers. In the case of a failure, parallel sites help support demand until more facilities

come back online. Coordinating between sites requires high bandwidth and multiple routing options until recovery or an alternative site can provide support. This is highly dependent on the resilience and speed of the network. The various options for meeting high levels of availability mean that there is a trade-off between resilience and efficiency in both the DC and WAN.

Figure 4 Simplified data centre stack

	User IT Service/business process		
	Applications		
Platform	Operating System, Virtualisation		
ICT Equipment	Server	Storage	Networking
Infrastructure	Physical space to install ICT equipment	Power	Cooling
	Data Centre		

Environmental control (mainly cooling) is the largest energy consuming component of the infrastructure. To minimise risk, older data centres tended to use lower temperatures than recommended by current operating guidelines such as ASHRAE (ASHRAE, 2016), which greatly increased the energy consumed. This was considered necessary because there was a lack of understanding regarding safe operating conditions of IT equipment and the poor control of the air flow and temperature. There have been steady efforts to reduce the cooling energy power requirements.

Higher allowable temperature and the use of 'free-cooling' (using cooler outside air rather than mechanical refrigeration) means that newer data centre infrastructure is more efficient, with some data centres infrastructures consuming less than 10% of the of the energy used to supply the ICT equipment (compared to over 100% in the past). To achieve this, better design and operation is required, for example the outside air will fluctuate in temperature and humidity and the system must respond. In addition, if air temperature is allowed to increase too high then server fan speeds and energy consumption will increase, negating potential energy savings and increasing risk of equipment failure. Optimal efficiency of different equipment also varies greatly and do not follow simple linear relationships between power and efficiency. Automation of control systems with finer levels of control of equipment and sensors reporting environmental status around the data centre are used. However, older data centres still use old operational techniques and equipment.

1.2.1 Data centre business models

Data centre businesses have many operating models whereby the control of the infrastructure, IT equipment, platforms and software may all be controlled by different entities (Table 1). For example, a colocation data centre will operate the building infrastructure, providing cooling and power and space for the ICT equipment with availability as agreed by contract. The ICT equipment could then be owned by another business who also manages and provides cloud platform services. The final software will then be developed and run on that platform by another business which the end user will interact with.

Since the colocation DC service provider will be providing space to many businesses and the cloud platform to many software developers, and the software provider to many users, each level cannot be fully optimised for any particular client. This situation presents a compromise to accommodate all the different requirements of the various businesses and clients. However, the large scale of the operations and level of expertise may mean they will still operate more efficiently than a small, single occupant data centre which may not have the resources to optimally operate their own DC. Similarly, for a cloud platform, it may not be as highly optimised for one specific user, but by having many users it can maximise utilisation, including through techniques such as elastic compute and spot prices that vary computing cost with demand.

Table 1 Data centre operating and ownership models

Ownership	Data centre	IT equipment	Application	Other services
Wholesale (enterprise DC)	Freehold (owned by enterprise)	Customer	Customer	None
Colocation	Long term leasehold	Usually customer	Customer	Low end hardware support
Hosting (managed service)	Mix leasehold/rent	Usually provider	Website, email	Low end tech support
Generic managed services	Mix leasehold/rent	Provider	Specific areas such as database, storage	Low end tech support
Specialised managed services	Mix leasehold/rent	Provider	More customised and mission critical applications	Professional services
Full outsourcing	Mix leasehold/rent	Provider	Full suite applications	Taking over staff from user

The largest internet companies tend to be more vertically integrated which means they own and operate the data centres, infrastructure, IT equipment and end user services. This is now even extending to designing their own specialised processors in the IT equipment. In general, this has allowed them to achieve very high efficiencies faster than the rest of the data centre market including very low PUEs and operating in the cloud. However, large companies do not apply a single ownership model and will use a mix of ownership to meet the business and operating costs targets. This further increases the complexity of managing the efficiency across all the data centres.

The efficiency of the end-user services is a function of all the equipment in the WAN and data centre. For example, a mobile user is watching a video. The software receiving the request that is operating on a server will use the data centre network to access the video from the storage and then send it to the WAN, which converts it to an optical signal and routes it between the nodes. Finally, the base station receives the data, which it broadcasts via radio signal and received by the laptop, tablet or smartphone.

Increasing efficiency must therefore consider all the parts and the interactions. For example, if the data was closer to the user, it would travel a shorter distance and through fewer routers. However, the

decision to run the application software on a particular server and data centre location is not made by the telecom companies but most likely by the software developer. In addition, if there are multiple users accessing the video, the best location must consider many different routes, greatly increasing complexity.

2 Standards and metrics

The DC and WAN industries have been revising existing standards and developing many new standards and metrics which are often interlinked and very closely related. These standards also continue to be revised and new standards developed. This can make it confusing to map but reduces the number of competing metrics and extends the applicability of the metric to suit a wider range of purposes. This section summarises the metrics prepared by international standards organisations and should be accurate at the time of writing (July 2018). While the metrics can be relatively simple, the test methods and measurement requirements are extensive and are not covered in this report. There are three main types of efficiency metrics: system, infrastructure, and equipment efficiency.

2.1 Standards and Standards bodies

Various standards, recommendations, guidelines and metrics have been developed by a wide range of stakeholders globally, including national and international standards and industry bodies including ISO/IEC, ITU, JEITA, JDCC, ETSI, CEN/CENELEC, The Green Grid, and ANSI. A list of standards has been compiled in Annex 1. The two most significant international standards and recommendations have been developed by the International Telecommunication Union (ITU) and the International Standards Organisation with the International Electrotechnical Commission (ISO/IEC). However, this does not preclude the use of national standards and guidelines which have been developed with the local situation in mind such as the JDCC Guidelines (JDCC, 2016).

2.1.1 ITU L.13xx recommendations

Two related series have been developed by the ITU, ITU-T L.13xx for energy and L. 14xx which covers other environmental and life cycle aspects. These are still being developed and improved, L.1332 for example was published in April 2018. It is important to note that these are technically not classed as standards but as recommendations. The ITU has also provided a framework for visualising the network and energy efficiency (see

Annex 3: Networks and virtualisation).

The recommendations include general best practices for operation as well as metrics. In addition, L.1301 sets minimum datasets and communication interface requirements for DC energy management. The remaining standards cover efficiency metrics themselves such as L. 1332: Total network infrastructure energy efficiency metrics.

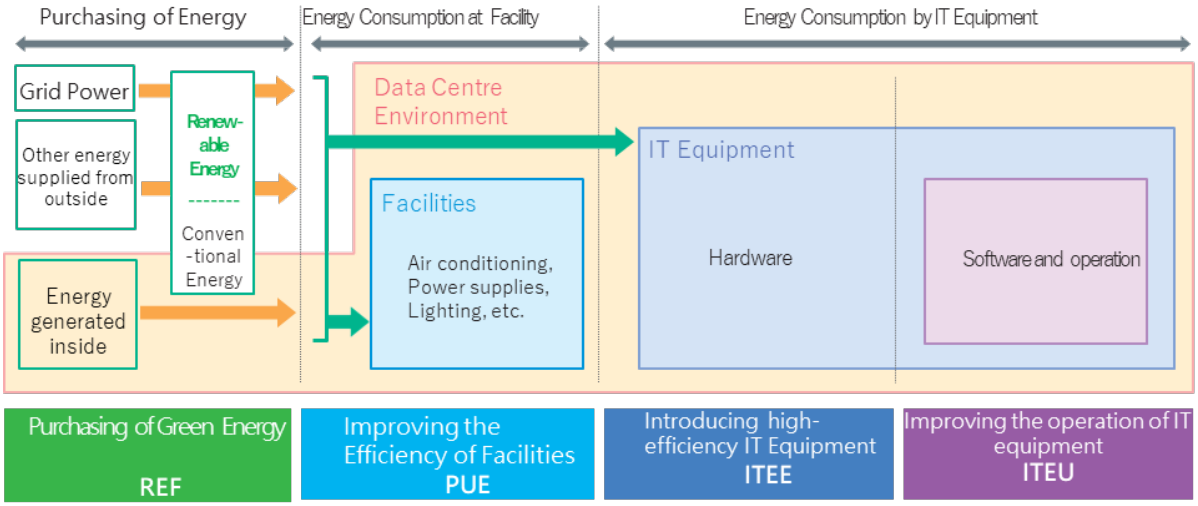
2.1.2 ISO/IEC 30134 and related standards

ISO/IEC (JTC1 SC39) is preparing both series 30314-n and 22237-n series; the latter started by proposing the EN 50600 series at international level. ISO/IEC 30134 is a comprehensive series of metrics covering data centre efficiency (Figure 5) whose relationship is illustrated in Figure 6. Global harmonization of metrics is desirable because the equipment and DC owners operate globally and inconsistencies can hinder this (JEITA, 2014). In addition, ISO/IEC 22237 addresses DC energy efficiency in the context of DC design, operation and assessment. Rather than using metrics, KPIs are defined which refer to the ISO/IEC 30134 series.

Figure 5 ISO 30134 and related DC energy efficiency KPIs

Data centre resource efficiency	
ISO/IEC TR20913:2016	Holistic Approach
ISO/IEC 30134-1:2016	General requirements
ISO/IEC 30134-2:2018	Power usage effectiveness (PUE)
ISO/IEC 30134-3:2018	Renewable energy factor (REF)
ISO/IEC 30134-6 (under development)	Energy reuse factor (ERF)
ISO/IEC TR 21897(under development)	Impact of ISO 52000 standards for energy performance of buildings
ISO/IEC TR 23050 (under development)	Data centres – excess electrical energy (XEEF)
Server Equipment	
ISO/IEC 30134-4:2017	IT Equipment energy efficiency for servers (ITEEsv)
ISO/IEC 30134-5:2017	IT equipment energy utilization for servers (ITEUsv)
ISO/IEC21836 (under development)	Server energy effectiveness metric

Figure 6 Boundary of Data centre Performance Per Energy (DPPE)



2.1.3 DC key performance indicators (KPIs)

Each data centre and network differs in terms of design, operation and service being provided, and therefore creating a single, holistic efficiency metric which encompasses all facets of the data centre fairly is not possible. Instead, KPIs help to put the metrics into the context of the individual data centre's operations and allows the data centre to identify and focus on the most important factors. ISO/IEC

22237 provides an approach which can be highly tailored to the data centre but can be harder to interpret by a third party. In contrast, ISO/IEC 20913-5:2016 is a simpler and more standardised approach for data centres which allows the ISO 30134 metrics to be visualised through the use of spider web (or radar) charts with each KPI placed on its own axis. More information on this and a comprehensive assessment method is also available from JEITA (JEITA, 2012).

2.2 Metrics - Infrastructure

Infrastructure metrics compare the total energy consumed by the site including the power and cooling infrastructure against the ICT and telecommunications equipment considered to be doing useful work.

$$Infrastructure\ efficiency = \frac{ICT\ and\ telecoms\ equipment\ energy\ consumption}{total\ site\ energy\ consumption}$$

In the case of Power Usage Effectiveness (PUE), the equation is inverted. This means an efficient data centre approaches a PUE=1, but an inefficient data centre PUE can exceed 3+.

The main infrastructure efficiency metrics include:

- (Data centre) Power Usage Effectiveness (PUE) (ISO 30134-2:2018, L.1302)
- Network infrastructure energy efficiency (NIEE) (L.1332)
- (Base station) Site Energy Efficiency (SEE) (L.1350)

The boundaries under test vary in range and scope, such as ICT equipment undertaking specific tasks, all ICT equipment in a data centre, base stations, telecommunications sites and entire networks but all follow the same basic principles. They might also specify the type of test equipment, the length of time to test, how to include multiple energy sources such as renewables, diesel backup generators, and reuse of waste heat.

These metrics are for actual operation and less suited for comparing facilities because they each have unique operating conditions. However, given its flexibility, simplicity, and applicability to various facilities housing IT equipment, PUE levels are being applied in some regions as a regulatory minimum energy performance requirement including China and frequently used in voluntary standards.

2.3 Metrics- Equipment efficiency

Equipment also never operates at 100% utilisation for any length of time and can spend long periods of time at low or no utilisation. However, the power consumption is not directly proportional with the utilisation, i.e. some power will be consumed at 0% utilisation, usually around 30-70% (Jalali, Hinton, Ayre, Alpcan, & Tucker, 2016) of the peak power. This means that efficiency is much lower at typical, low utilisation conditions.

Equipment efficiency metrics can be classified into three types, peak efficiency, variable efficiency and extended idle metrics (Kharatinov, 2012). These tend to be measured under strictly controlled laboratory conditions and are well suited for comparing products. However, if the test conditions do not reflect actual use, they may not be informative.

Equipment efficiency standards in the ITU-T are based on the Alliance for Telecommunications Industry Solutions (ATIS) 01600015 standard series and ETSI standards.

2.3.1 Peak efficiency

Peak efficiency is the simplest of metrics and is a measure of peak performance against peak power consumption. For network equipment, this is generally measured in data throughput per watt, or bits/joule which are mathematically equivalent. For telecom equipment these have been superseded.

ISO/IEC 30134-4 (ITEEsv) is a similar metric for servers but does not specify the test or benchmark to use for quantifying performance. This is because servers can have many different functions and the most appropriate test will depend on the function.

2.3.2 Variable efficiency

Peak efficiency is generally achieved at maximum utilisation, which does not reflect actual use. For virtually all equipment, the power does not scale perfectly with performance, this means at low utilisation the efficiency is also lower.

Variable efficiency metrics measure power and performance at different utilisation rates, typically three for network and telecoms equipment, and these are designed to be indicative of the actual load in use. The overall efficiency is then weighted based on the time spent at the load level.

$Efficiency = \sum \frac{performance_i}{power_i} \times time_i$ where *i* represents a utilisation rate

Some metrics do not sum the performance level and use only the maximum performance. While this changes the absolute value of the efficiency it has no impact on the relative efficiency of one piece of equipment to another when both measured under the same metric.

Variable efficiency is the most common type of metric and includes L.1310 which itself references:

- Routers and ethernet switches (ATIS- 0600015.03.2013/ETSI ES203136)
- DSLAM. MSAM, GPON, GEPON (ETSI EN 303 215)
- Mobile base stations (ETSI ES 202 706-1)

For servers, ETSI 303470 and ISO/IEC 21836 Server Energy Effectiveness Metric measures the variable efficiency across a set of 13 different, standardised worklets and up to 8 different utilisation levels for each worklet. The worklets are classified into CPU, Memory and Storage. One significant difference with this metric is the use of the geometric mean rather than the arithmetic mean shown in the equation above and most typically used. There are a number of advantages both in theory and practice. This includes simplifying the metric by avoiding the issue of weighting the workloads within a classification against each other and defining the weighting of different utilisation levels.

ISO/IEC 21836 also introduces a concept of data centre scaling to determine the efficiency of deploying racks of servers as opposed to a single machine. Given that each machine has overhead just to be operational and additional overhead for operating system and services, scaling based on a single machine's performance per watt will likely show an inappropriate result for IT facilities provisioned for

more compute than a single machine. ISO/IEC 21836 provides a deployed power analysis that validates whether the KPI actually selects those systems which would be more efficient at a data centre level.

2.3.3 Extended idle metrics

Extended idle metrics are very similar to variable efficiency but recognise that some equipment can be placed in different operating states with different performance levels. For example, under low utilisation, a proportion of the network interfaces could be put in sleep mode. This limits the maximum data throughput but increases the achievable power savings. Extended idle metrics are not widely used but as energy aware networks and equipment is deployed, it will become increasingly important.

However, a problem with idle metrics is that it encourages equipment manufacturers to develop low power idle modes that are rarely used in current operating environments. To determine energy consumption when data centre equipment is not working on active workloads it may be more appropriate for some equipment to measure power levels at minimum, non-zero utilization levels. The power levels at these conditions are more representative of the equipment in low utilization in a live environment. To determine this power value, one can use a linear interpolation from two low utilization points, e.g. 10% and 20%, to a 0% intercept in a power verses utilization level plot. The workload employed in the assessment should be highly active such as the SPECPowerssj2008 for servers, or other workloads for the equipment which exercises a majority of the circuitry and can adjusts the workload percentage.

For long (greater than seconds) resume low power modes, it's recommended to determine the recovery time for full operation verses the power savings achieved. Recovery time assessment of low power modes may be employed in a dynamic provisioning environment such as platooning a set of equipment to be ready in support of peak conditions including the diurnal IT demand cycles. The energy savings would be the difference between cumulation of energy of the full equipment deployment under low or near zero utilization, and a smaller set of equipment at higher utilization combined with the energy of the remaining equipment in these extreme low power modes. There are however, no dynamic hardware provisioning energy opportunity metrics available today.

2.4 Metrics - Utilisation

The efficiency of most equipment depends on the utilisation level so monitoring utilisation is a key tool for maximising efficiency. Utilisation metrics for routers and switches were identified in ITU L.1310 and for servers in ISO/IEC 30134-5:2019 which uses the server CPU utilisation to measure the server utilisation level. Measuring the utilisation level of network equipment may be less complex than servers.

2.5 Metrics - System efficiency

The system efficiency seeks to assess the efficiency of the service provided. This can depend on many factors and is therefore significantly more complex. As discussed in the KPI subsection, there are very few system efficiency metrics. Work to develop a useful metric continues, one that can measure the energy used per unit workload while recognising the differences in types of DC/WAN and work being delivered.

Mobile network energy efficiency (L.1331) considers a wide range of variables which affect the energy consumed per bit of data transported. This includes the demography, the geographical range covered,

the geographic topology, climate zone and type of data traffic (i.e. circuit switched and packet switched). The efficiency EE_{MN} is then defined as the volume of data over the energy consumed (bit/J). In addition, the efficiency for the area covered (CoA) is also calculated – which is the area covered divided by the energy consumed, while taking into account the quality of the coverage.

$$EE_{MN,DV} = \frac{DV_{MN}}{EC_{MN}} \qquad EE_{MN,CoA} = \frac{\text{CoA\_des}_{MN}}{EC_{MN}}$$

This can be extrapolated for the entire mobile network taking into account the different demographic regions.

While this is valid for current RANs, the ITU has identified that new metrics will be needed for 5G RANs making use of MIMO and small cells.

2.6 Metrics - Renewable energy

Renewable energy is not a measure of efficiency and is described here only for completeness of ISO/IEC 30134. Renewable energy metrics measure the proportion of the energy consumed that is supplied by renewable sources. For electricity, the most common energy source, this may be supplied from the normal electricity grid mix, additional 'clean' energy purchased through the grid or direct renewable supply. For DCs this is covered under ISO/IEC 30134-3:2018

The major problem is this KPI compares final energies and not primary energies and has led to the development of ISO/IEC TR 21897. Additionally, this approach is based on the average renewable energy generation for grid supplies, and it may be more appropriate if the renewable energy generation matches the consumption pattern of the DC or network and that this was taken into account.

2.7 Metrics - Energy reuse

Energy reuse is not a measure of efficiency and is described here only for completeness of ISO/IEC 30134. Almost all the energy consumed in the data centre is converted into heat, which is then expelled into the air as waste heat. However, this can be utilised to offset the energy that would otherwise be consumed to heat other buildings or water. The most common form of energy re-use is to connect to district heating, where the heat can be transported for hot water and heating in nearby homes and offices, particularly in the winter. Examples of this include Stockholm and Finland where district heating is commonly used. Energy reuse metrics are currently under development under ISO/IEC 30134-6 for data centres.

Table 2 Summary Metrics – energy efficiency

	WAN (ITU L. 13xx)	DC (ISO/IEC 30134)
System level	L.1331 (mobile networks)	ISO/IEC 22237 KPIs , ISO/IEC 20913-5
Infrastructure	NIEE, SEE	PUE (ISO/IEC 30134-2)
Equipment: Peak		ITEEsv (ISO/IEC 30134- 4)
Equipment: Variable	L.1310	SEEM (ISO/IEC 21836, servers)
Equipment: Extended Idle	L.1310	
Utilisation		ITEUsv (ISO/IEC 30134-3, servers)

3 Emerging trends

The use of the internet and end user services measured in terms of data demand are expected to increase by over 20% every year from 2016-2021 (Cisco, 2018), and this trend is likely to continue into the future.

By far the biggest driver of data growth is media consumption, video is the single biggest driver and is expected to increase further from 75% of data traffic in 2017 to 85% in 2022 (Cisco, 2018). By 2022, most of the data (72%) will be served by CDNs through wireless and mobile connections (71%). The fastest growth areas will be virtual/augmented reality (65% CAGR), gaming (55% CAGR), and IoT (49% CAGR). While they only represent a small proportion of total global IP traffic, the high growth rate suggests they will become more significant beyond 2022.

Without improvements in energy efficiency, this growth in data will have a massive impact on the total energy consumption. Historically, energy efficiency of data centres and WAN have improved over time, but the average efficiency improvement is significantly lower than the achievements of the most efficient data centre operators.

This section describes the most significant and likely changes that will occur in technology and what opportunities and risks this poses to the total energy consumption of the global DC/WAN industry. These trends promise to improve the QoE (Quality of Experience) for the end user with new and better services as well as lower costs and generate new business opportunities for the service providers.

3.1 Internet of Things

The internet of things has a very broad scope and refers to a network of devices with embedded sensors, actuators, software and connectivity that enables them to connect and interoperate through the internet and provide an integrated service to the user. The interoperability can be applied to virtually anything from simple solutions such as home automation controlling lighting and heating to large scale industrial processes, agricultural practices and networks. As such it could grow to tens or hundreds of billions of individual devices across the globe.

The sensors and actuators all connected devices to be controlled by software and respond to whatever the combination of sensors measure. This allows more sophisticated and granular control depending on the control algorithms being applied and potentially increase the efficiency of the system under control. However, to achieve these efficiencies, the sensors can generate greater amounts of data and this must be sent over the network, stored and analysed. With billions of devices, the IoT data is projected to grow at 49% annually, over twice the overall data growth rate, from 2016-2021 (Cisco, 2018) and represent 6% of global IP traffic at the end of that period. If growth continues at similar rates IoT will become increasingly more significant.

While more frequent sensing and data can improve the service, there are likely to be diminishing returns to the user which results in lower overall efficiency. For many systems, particularly consumer devices, the volume of data being generated and the processing occurring is not controlled by the user. How the device and system communicate is determined by the producers and designers of the device, software and service. The volume of data can be controlled by the frequency of communication and how the communication is triggered. It is common for servers to ‘poll’ sensors for data. This means the server connects to the sensor at regular intervals and request updated data regardless if anything has changed. Continuously polling also means a device is unable to enter sleep mode for any extended

period. A simple example may be a temperature sensor controlling a heating system. The control may check the system every minute and adjust the boiler accordingly. While this may be useful when the heating system is first switched on, when the temperature stabilises, the polling is wasting energy. However, if a high network available product is able to enter low power states within seconds or milliseconds, energy savings can still be achieved.

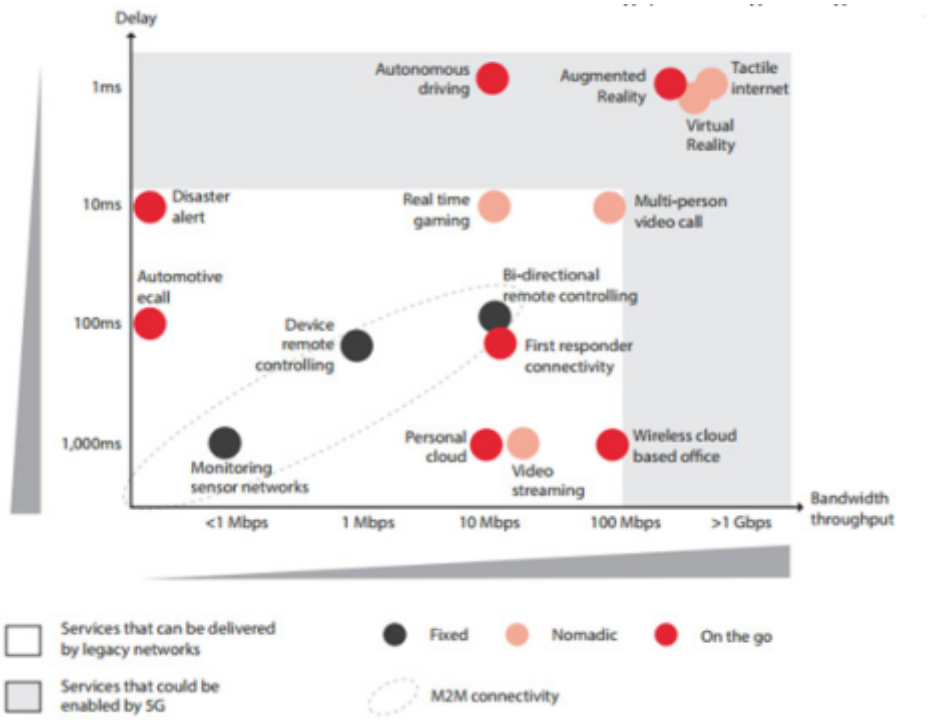
There are also many competing service platforms for IoT, and while a single platform is not suitable for all situations, efficiency is reduced since they are not directly interoperable. Services exist that can allow multiple platforms to be accessed, monitored and devices triggered across platforms but this adds another layer of data and computing with associated energy costs.

The variety of IoT service platforms and increase number of IoT devices and its associated data give rise to the prediction of increasing demand of compute services and energy. The diversity in the data service infrastructure and unstructured addition of IoT devices and data simply forces more data across the networks and uses centralized IT facilities to restructure, organize, and manage both the data as well as determining the resulting structured information. Despite the need for low latency for IoT systems, the existing custom platforms and methods pushes much of the unstructured data into centralized computing sites, driving increased burden on the network, longer latencies and energy expenditures that scale directly to increased number of IoT devices. Energy expenditures in this method also includes the computing resources to restructure the data based on re-assembling the context that was already known at the IoT location. Latency issues (e.g. on autonomous cars, drones, emergency services, etc..) have prompted a demand for more localized data aggregation and computing sites. Localized sites, however, also offers significant reduction in bandwidth and central computing needs, reduce energy consumption across the network, and improved security and privacy to each of the IoT networks, in addition to improved service via reduced latency for IoT and edge devices,

3.2 5G mobile networks

5G is the next generation of mobile network, currently being piloted in US, Europe and Asia and wide scale deployment is expected to start in 2019. It is envisioned that the operating environment will become more heterogenous with different levels of technology and service to expand connectivity, services and capabilities. These include fast speeds and access covering all locations, including those currently with poor reception such as remote locations, dense urban environments and travelling at high speed. New use cases being discussed include such as AR and tactile communications, ultra-reliable and critical and emergency services such as medical and disaster support as well as broadcast services with high bandwidth and low latency requirements.

Figure 7 Bandwidth and latency requirements of potential 5G network use cases (GSMA Intelligence)



To support such services the NGMN White paper (NGMN Alliance, 2015) prioritises the following requirements:

- Higher bandwidth and lower latency with a flexible and scalable network
- Consistent service and experience
- Flexibility to support a wide range of services which can be dynamically allocated in response to demand
- Cost and energy efficiency
- Innovation.

While initially there are only small differences compared to the latest 4G network technology, more advanced technology will need to be deployed as demand increases to provide the services envisioned. Since the network must cope with the peak data rate, this means the average utilisation is expected to be even lower than currently. If similar technology with the same efficiency to 4G networks were to be used, the higher data and number of antennas would increase the energy consumption dramatically.

The main technology differences of 5G networks are the use (higher frequency) millimetre wave radio and more sophisticated MIMO. Millimetre wave radio can carry more data for a given amount of radio spectrum. Since the higher frequency has a shorter range for the same amount of power, antennas will need to be more densely installed. More dense antennas installation also has the effect of further reducing power consumption for a given coverage area. These small cells will be connected with a base station and can be switched on and off as demand requires. In addition, the signals from many antennas can be used in concert through multiple-input, multiple-output (MIMO) which further enhances the capacity of the network. It is estimated that an antenna may be required at every road junction to

provide optimal coverage. The NGMN white paper states the goal is to halve energy consumption while improving capacity 1000x, resulting in a 2000x increase in efficiency.

3.3 Software Defined Networking

Software defined networking facilitates network management and enables the network to be configured using software to improve monitoring, performance, and efficiency. This is achieved primarily by creating a centralised intelligence layer and routing of data between the equipment in the network (control plane) that is independent of the main data transport layer (data plane) carrying the user data. SDN can be applied to DC, WAN and mobile networks.

SDN potentially offers a solution to many of the requirements identified by the NGMN including energy efficiency. The ability to reconfigure the network, for example allows equipment to be shut down if utilisation is low and the data rerouted via another node or cell tower. Since it is controlled by software, increasingly sophisticated management algorithms and techniques can be more easily implemented across the network from a centralised site rather than needing on-site reconfiguration or changes to the hardware itself. SDNs could also balance the traffic around a network, avoiding heavily congested nodes at peak times and therefore maximising utilisation of available equipment rather than requiring higher power/performance. Network overlays applied on older equipment are also considered to be a type of SDN. However, since these do not control the equipment directly, they are not capable of providing the monitoring and routing functions needed.

3.4 Network functions virtualisation (NFV)

NFV uses IT Virtualization technology to consolidate many network equipment types into standard high-volume servers, switches and storage (network functions virtualisation infrastructure), which could be located in data centres, network nodes and the end user premises. NFV decouples the hardware and software of telecom appliances and recreates the network function typically provided by specialist hardware in software (virtualised network functions). This reduces the TCO, enhances the system flexibility and accelerates the pace of innovation. An NFV management and orchestration architectural framework (NFV-MANO) manages the collection of functions blocks and nodes to ensure the network services are provided.

The NFV standard has already been finalised and is supported by a number of vendors.

By using standard ICT equipment, costs are reduced but equipment efficiencies may reduce since they are not purpose design. However, the additional flexibility may enable many functions to be performed on one piece of equipment, increasing utilisation and offsetting the efficiency loss.

Current SDN and NFV are being applied purely to increase the ability of commercial services offered rather than for efficiency. It is not clear if current SDN/NFV equipment has the energy savings capabilities discussed. Though there may be some opportunity for energy savings if constructed as part of a dynamic hardware provisioning scheme as discussed previously, SDN and NFV offers primarily the flexibility of adjusting services as opposed to changing equipment. SDN and NFV without dynamic hardware provisioning and extreme low power modes, may only provide some opportunity (energy)-cost savings by means of a more generic hardware deployment. More details on NFV are provided in Annex 3: Networks and virtualisation.

3.5 Heterogeneous computing

Heterogeneous computing refers to systems which use more than one type of processor to perform different tasks. As the quantity of services grow and demand increases, it is more efficient to perform specific types of processing on specialised processors or platforms. Heterogeneous computing primarily offers performance advantages but energy efficiency is also a very strong driver for its application. A growing use for heterogenous computing is machine learning and AI which perform the same calculations in parallel in huge quantities. This has led to the use of GPUs and now specialised custom designed processors to process the data rather than use general purpose CPUs. These chips are claimed to be approximately 30-80X more power efficient than contemporary CPUs and GPUs[1].

Another example could be the processing of photos and videos taken on a mobile phone. For simple editing, processing can occur on the relatively low performance, general purpose phone CPU but for extensive changes sending the data to the data centre for processing on dedicated hardware and then sending the edited image back again is faster and more efficient.

To maximise efficiency, the software must be specifically designed to make use of the different processors to perform different tasks. However, research is being carried out to try to automate this type of allocation of the tasks across the internet to maximise energy efficiency (Ruiu, et al., 2016). Ideally, the decision would also consider the energy consumption of the network itself. However, there is also a risk that too much specialisation and computing capacity will lead to underutilisation if it is not managed and designed well, especially if the use if unpredictable or peaky, leading to inefficiency in the overall network and system.

More generally, the networks and storage are also becoming more heterogenous as well as being located in more places within the network which creates similar issues and opportunities.

3.6 Content delivery networks, edge computing and fog computing

CDNs, Fog and Edge computing all distribute some or all of the computing resources geographically and logically across the network. The main benefit is that the resource is located closer to the user which improves latency (lower and more consistent) and reduces the amount of data travelling over the network. Improved latency provides a better service and can make new services feasible while reduced network traffic and can lower utilisation and/or reduce the peak bandwidth capacity of the network and as a result can also improve energy efficiency and reduce capital costs required for higher performance equipment.

Content delivery networks have existed for a long time on the internet and serve a large proportion of the data. There are numerous commercial CDN services such as Akamai, as well as private CDNs such as Open Connect which is used by Netflix. These systems supply content to the user and offer faster download speeds and reliability by hosting the data at numerous locations globally. This means that a user can be automatically connected to the nearest location and the file downloaded or streamed from there. For data accessed by many people across the world, by reducing the distance travelled by the data, latency is reduced, energy is saved and utilisation of the core network is reduced.

The costs of the CDN is highly related to operating ICT equipment at multiple locations from which the data can be served. Therefore, the choice of what and how much data to place on the CDN, at how many and which locations have a significant impact on the energy efficiency. These decisions are

[1] https://cloud.google.com/blog/big-data/2017/05/an-in-depth-look-at-googles-first-tensor-processing-unit-tpu

generally made by the CDN service provider and not the network or DC operator. The efficiency of the data centre, ICT equipment and WAN also influence the overall energy efficiency of the service delivery.

Edge computing and mobile edge computing (MEC) places compute resource specifically at the edge of the access network near the source of data. This can include data processing resources as well as storage. The 'edge' can include the premises side or at the ISP or mobile carrier, mobile edge computing could be located at every major mobile network cell site, which may imply a huge number of locations compared to a CDN.

Fog computing is similar to edge computing but with greater distribution of resources within the network. This means the compute resource could be located anywhere within the network, not just the edge and cloud but also on user premises and regionally. Instead of a vertical view of communication to/from the cloud to the user, cross communication with other resources along the edge, or premises is also possible. Another benefit of fog/edge is the ability to dynamically allocate compute resources to the services and tasks being provided. Dynamic allocation could take advantage of devices with low utilisation and select the computing resource with the highest efficiency. Maximising utilisation of existing resources and optimal allocation have energy efficiency benefits. In the same way as heterogeneous computing, the software application would need to know where to carry out the processing or some automated system needs to exist.

Edge and fog are still being developed and defined and it is not clear how they will evolve. One of the questions is with the ownership of the resources and how they interoperate and are accessed. CDNs and clouds tend to run on closed, competing systems with their own data centres and locations; moving a software application from one to another is not possible. However, the greater geographical distribution of edge and fog means it may become prohibitively expensive to operate at every edge. Interoperable standards and multi-access edge computing standards are being developed under ETSI to provide interoperability as well as through the OpenFog Consortium. Fog computing for a single service may then end up operating on resources owned by multiple entities, and a system for paying for the use of this resource will also need to be developed. This creates additional security and privacy concerns since your data is stored across multiple networks.

It appears the telecoms operators may be in one of the best positions to provide the computing resource, especially for Mobile Edge Computing (MEC). If this is the case, they may have a lot of influence in determining how the edge computing will operate, including efficiency.

3.7 Blockchain

Blockchain is a distributed ledger technology meant to confirm and make unique each transaction made. Blockchain is a non-centralized way to distribute the work of confirming and notification of a transaction. This is achieved by distributing work and information across a collection of unreliable and untrusted nodes and using cryptographic techniques which enable verification of the work being done, while assigning value against the work and payment that can be automatically transferred after verification. Blockchain is being applied to virtually every possible transaction, including fog and cloud computing and storage. Blockchain's resiliency is attributed to the distributed network, and therefore individual nodes and hardware do not need to be resilient which may increase efficiency. However, this is achieved by locating multiple copies of data, replicating work over many nodes, and cumulative unique identifier calculations which reduces efficiency. The blockchain itself can also be energy intensive. In terms of overall energy efficiency and consumption, it is too early to determine the

magnitude of Blockchain's energy impact. However, components of Blockchain are expected to increase energy consumption in addition to increasing the demand on computing and networks (DataCenterDynamics, K. Hunt, April 2018). Although Blockchain is not discussed in further detail in this report, the technology could have major implications to energy consumption and demands further investigation.

3.8 Summary and predicted data growth

The data growth is expected to continue to rise at around 20% a year and new technologies are being developed to cope with it. There are some very ambitious stated targets for energy efficiency (2000x improvement in the NGMN) and many of the same technologies offer opportunities to save energy by managing equipment and data with dynamically and with greater control. Improvements in equipment efficiency is certain to occur, and their energy profiles will become more proportional. Good resource management will also be a key factor in realising energy savings and the ability to scale resource with demand, i.e. sleeping equipment when overall network utilisation is low is likely to be essential.

However, many of the most significant factors are determined by the software and end user service provider, not the telecom or data centre operator. Management is also becoming increasingly sophisticated and complex, and therefore increasingly hard to manage efficiently. In a situation when service quality and efficiency must be balanced, it is most likely that efficiency is sacrificed.

Taking both of these seemingly contradictory statements together indicates the significant range in uncertainty that exists in future energy consumption, but also the potential for large efficiency improvements. The following section reviews current and new techniques that enable sophisticated management of equipment and systems in order to optimise energy efficiency and how much potential energy savings might exist.

4 Intelligent efficiency options

There is a large volume of literature describing new algorithms and techniques to manage networks which have been tested using modelling. However, there are limited actual case studies because the necessary conditions, such as SDN are not deployed widely. The research often describes highly mathematical models and solutions which is beyond the scope of this study. The aim of this section is therefore to identify research and case studies to highlight the approaches and opportunities available for applying intelligent efficiency techniques and give an indication of the magnitude of energy savings that could be achieved.

4.1 Deep reinforcement learning (DRL), machine learning (ML), and artificial intelligence (AI)

Many of the case studies investigate the use DRL/ML/AI. It is helpful to understand how the technology and approach differs from more common 'smart' technology and its advantages. It is also important to understand the current limitations to this technology and therefore how it might influence future policy.

DRL/ML/AI is developing rapidly and its application is already being discussed within ITU focus group 'ML5G for Future Networks'. The three terms are subsets of each other with DRL describing a specific approach being used in most commercial AI work, but the terms can be considered mostly interchangeably in the context of this report.

DRL systems generally work with a large dataset that describes the state of the system (inputs and outputs) under different scenarios. This can then be used to train and emulated neural network through an iterative process which tries to determine the heavily interlinked relationships between the inputs and outputs and as a result make accurate predictions. By telling the system the desired outputs, the AI can then control the inputs to achieve this even under changing conditions. By contrast, conventional smart systems require a human to identify the relationships and tell the system how to behave under each situation. In the case of a data centre infrastructure control, for example, given a simple data set of energy consumption, chiller temperature and server temperature, it would quickly be able to predict higher energy consumption and lower server temperatures arise from lower chiller temperatures.

The advantages of AI with regards to networks have been described by (Xu, et al., 2017) in three areas:

- The ability to learn from large amounts of data the characteristics of data traffic, management and controls with higher accuracy than humans
- Ability to make predictions based on the network state where relationships between factors are unclear and mathematical models cannot be developed to accurately describe the system, or are too complex to be solved in real-time, which are required for human engineered 'smart' management systems
- Ability to collaborate and manage entire network(s) as the size, scale and complexity of the network grows rather than managing small regions independently.

While AI systems can be small and basic, greater benefits are gained with larger datasets (more data sources such as measuring CPU utilisation over a longer period of time and at higher frequency to accumulate patterns that occur daily and annually) and sophisticated AI design to enable very fine control and minimise risks of unexpected behaviour. This means the effectiveness of AI to achieve

energy savings is less dependent on energy managers but instead depends on sufficient data and the data scientists and engineers developing the AI software.

One issue with DRL is that the internal functioning is not clear and the predicted behaviour might not be correct. This is especially likely when an unexpected or emergency situation that has never occurred in previous data set used to train the model (e.g. what would happen if the DC power failed and had to switch to backup generators). For risk averse industries whose main job is to withstand unexpected emergencies and with contractual requirements for availability there is likely to be reservations.

In addition, the iterative training process can be time and energy intensive and grows as the dataset and complexity increases. It is also difficult to transfer a trained network from one system to another without having to retrain the network with data from the new system. There is therefore a trade-off between the amount of training which must be done and the energy consumed to create the neural network for each DC/WAN against the energy savings being made. However, the energy consumption is being mitigated by using specialised, highly efficient hardware to perform the DRL neural net training. Additionally, true neural networks are being developed and this could address the energy consumed for training and the unpredictability of the network.

4.2 Mobile edge computing in 5G heterogeneous networks

Problem

Mobile edge computing provides access to computer servers close to the user primarily to increase the performance and service for the end user. Each server and cell tower serve multiple users who have different devices and are performing different tasks. The goal is to minimise the overall MEC energy consumption by offloading the computation between the mobile device and edge while considering the network, MEC and device efficiency.

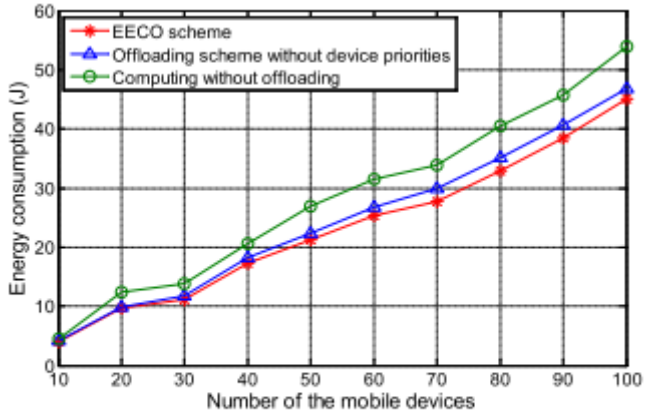
Solution

An algorithm was developed which was able to assess the efficiency of different network routes and the computing efficiency of the MEC and device. This could also prioritise work by taking into account the delay and number of devices connected to the cell station. The resources were then allocated that minimised energy consumption.

Results

Since the MEC is more efficient than the device, energy consumption was reduced by around 13% (Figure 8). By taking into account the relative efficiency of the network data transmission, an additional 2% could be saved. These energy savings will be highly dependent on the relative efficiency of the device and the edge computing (Zhang, et al., 2016).

Figure 8 Energy consumption of different offloading schemes



Source: Zhang, et al., 2016

4.3 Fog Networks

Development of fog networks is ongoing and unconstrained growth could result in higher energy consumption. Research by Baccarelli (Baccarelli, et al., 2017) identifies three areas of research for energy efficient fog networks:

- Energy efficient transport protocols which take into account different network options
- New network layers to take into account the different roles and hierarchical structure of different devices
- Distributed and adaptive orchestrators that perform energy and delay efficient allocation of the work across all available computing nodes

Furthermore, they identify the importance of security and trust which will be needed for a migration to fog. The research however, does not quantify the energy savings that could potentially be achieved.

4.4 Energy aware SDN networks

Problem

The core WAN consists of many network nodes with multiple connections which the data will travel through to reach the final destination. Each connection consumes energy and may be underutilised. Reducing the number of connections between the nodes could reduce energy consumption but may impact the performance of the network since the data may need to travel a longer route. The goal is to maximise energy efficiency and understand the impact on the performance.

Solution

Reducing energy consumption in a core SDN network takes advantage of the ability to control and route data by a centralised controller and therefore allow nodes and links to be put in sleep or lower power modes.

Numerous proposals have been made for routing algorithms in the SDN and optical networks. Researchers have proposed a static method of network pruning which reduces the number of interconnections between nodes (Fernández-Fernández, et al., 2017). The algorithm (SNetCA) is designed to maximise efficiency by removing as many interconnections as possible while maintaining acceptable resilience. This is achieved by iteratively removing links, starting with the most highly

connected nodes and continually checking for connectivity across all the nodes in the network. A dynamic energy aware routing algorithm is then applied to the network to route traffic

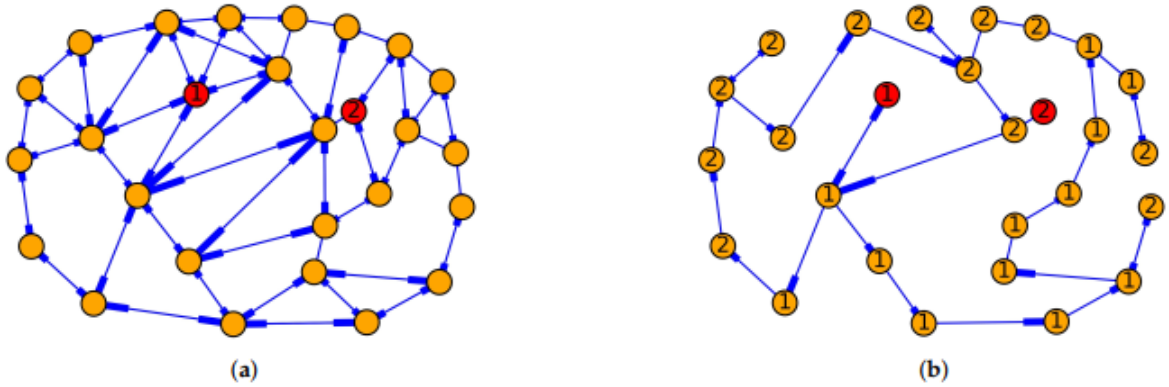
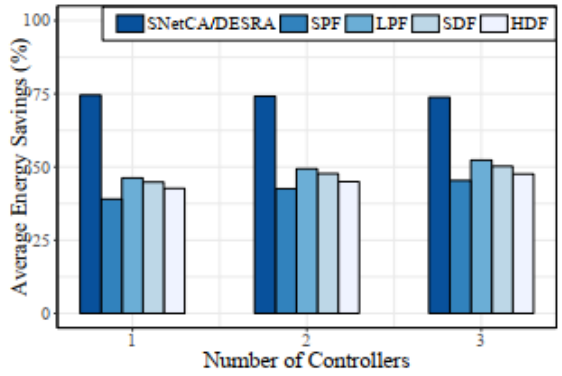


Figure 3. (a) Original Norway graph. (b) Resulting Norway graph after applying SNetCA.

Results

Simulated results for New York, Geant and Norway showed average energy savings of 50-75% (Figure 9). Other energy saving approaches saved approximately 50-25%. However, network performance was impacted and latency increased. The energy savings and network impacts were highly dependent on the original topology of the network.

Figure 9 Energy savings for New York using different algorithms



4.5 Hybrid networks

Problem

How to maximise the effectiveness of networks with a mix of traditional and SDN equipment by selecting the equipment for upgrade which has maximum impact and enables traffic engineering.

Solution

Nodes were selected which had the most connections since it was found other options, such as picking nodes with most traffic made no difference. By replacing 20% of nodes, it was found that 90% of traffic would pass through an SDN and enabled the use of routing approach that prioritised the least utilised route first (Hong, et al., 2016).

Results

Traffic congestion was reduced by an average of 32% and with only 6% of the equipment replaced and operating in an SDN. This provides a route to gradual replacement of legacy equipment but does not directly relate to energy savings, but reducing congestion can reduce the need for more powerful equipment. However, if the replacement is halted at only 20%, energy savings from low power states, shutting down links and equipment are potentially lost (Hong, et al., 2016).

4.6 Energy aware SDN and scheduling for fixed access passive optical network (PON)

Problem

The FAN is one of the least efficient parts of the network while one of the most energy consuming. TWDM PON[2] has been identified as one of the most energy efficient next generation FAN technologies (Lambert, et al., 2014). Since utilisation is low, energy savings can be increased by improving the energy proportionality in the FAN equipment (OLT).

Solution

Passive optical networks are one of the most popular technologies for high speed broadband to the home. The OLT uses multiple power consuming transceivers to send and receive data. When utilisation is low, the number of transceivers can be reduced and energy is saved.

SDN can centralise the control of the number of active transceivers to ensure the minimum number is required that does not impact the Quality of Service (Pakpahan, et al., 2017).

In addition, new scheduling protocols have been proposed based on minimisation of the number of scheduling voids rather than wavelength minimisation (Dutta, et al., 2018)

Results

Simulations show that SDN control can minimise any QoS effects while 25% more energy can be saved at high load and is very close to energy proportional by using minimisation of scheduling voids. However, this scheduling technique has an effect on the latency and there is no research combining the two techniques.

4.7 Data centre cooling with DRL

Problem

DC cooling infrastructure can have multiple cooling circuits and be controlled from multiple points in the system. Finding the optimal settings for energy efficiency can be difficult when conditions are changing and the system is non-linear. The goal is therefore to use new techniques to reduce energy consumption of DC cooling infrastructure where cooling can be supplied from five different systems

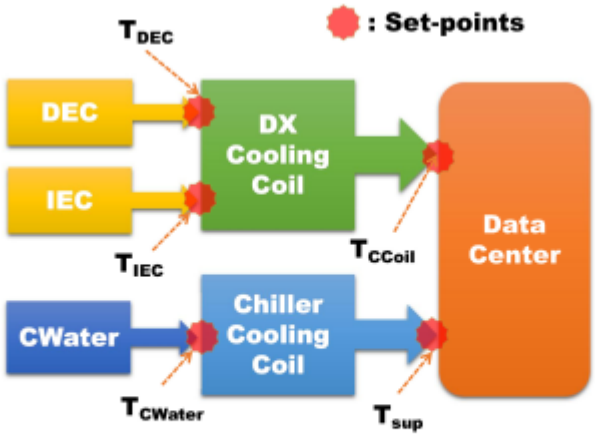
Solution

A deep reinforcement learning end-to end cooling control system was created that controls the temperature at 5 points in the systems where cooling is supplied:

[2] Time and wavelength division multiplexed passive optical networks. A technique for FTTH broadband.

- Direct Evaporative cooler (DEC)
- Indirect evaporative cooler (IEC)
- DX cooling coil (Ccoil)
- Chilled water input (Cwater)
- Chiller cooling coil (sup)

Figure 10 Main components of cooling system and control points



Source: Li, et al., 2018

Results

10% energy reduction for cooling compared to current optimal approaches which use a human built model of the cooling system behaviour, and a simulated 14% energy savings based on real data from the Singapore National Super Computing Centre (Li, et al., 2018).

4.8 Netflix Open Connect CDN

Problem

Maximising the value of a CDN by ensuring high utilisation of storage and compute resources.

Solution

Netflix (Berglund, 2017) uses data of historical viewing patterns based on each file using data science and a combination of time series forecasting, constrained optimization, and high-level network modelling to predict what content will be popular on a daily basis and in which regions, the content is then distributed to the CDN at night when network utilisation is low and before the demand is expected. Netflix has different files for the same title which are tailored for different languages and playback devices. This minimises the amount of data transferred compared to streaming every language audio and 4K video to every viewer but increases storage. They also try to minimise the amount of content that is updated day over day to minimise data and costs. Caching efficiency measures the amount of data streamed to a region from the local CDN node against all the data streamed.

Results

Netflix achieves a cache efficiency of 50% and in combination with other measures, in 2014 the energy consumed was 0.013kWh per streaming hour delivered[3]

4.9 Google AI management of DC infrastructure

Problem

PUE is dependent on a high number of variables in a large data centre. Many datacentres have automated 'smart' controls to manage the efficiency of the cooling system which respond to the conditions in the data centre. These 'smart' systems are based on human specified control algorithms which dictate how the system should respond. However, since there are so many possible points of control and changing environment, the infrastructure does not respond linearly to changes being made and in some cases in unexpected ways.

Solution

Google monitored over 1200 state variables and enabled 20 actions to be taken (Gao, 2017). Two years of data was used to train the AI including:

- Incoming IT load
- Power meters
- Pressure sensors
- Temperature sensors
- Water flow meters
- Pump and fan speeds
- Fault alarms
- Weather condition
- Number of cooling towers
- Number of chillers
- Number of pumps
- Temperature setpoints
- Pressure setpoints
- Flow setpoints
- Valve positions

Results

The AI prediction showed very close agreement with actual results. A 40% reduction in cooling and a 15% reduction in overall building energy from an already highly efficient building.

4.10 Virtualisation machine learning

Problem

Virtualisation can increase utilisation by running multiple applications on the same physical server. However, with hundreds of different cloud and virtualisation services available and the need to run many different applications using different resources and at different times, finding the optimal solution to minimise costs, and energy consumption is time consuming and requires specialist knowledge. The goal is to maximise utilisation of virtualised environments and reduce costs.

Solution

Commercial Automated consolidation tools and services, such as, Densify, TSOLogic and Baselayer, monitor the CPU, memory and I/O utilisation and distribution of images. Data is collected for around a

[3] https://medium.com/netflix-techblog/netflix-streaming-more-energy-efficient-than-breathing-57658d47b9fd

year to understand variations in workload and the large amount of data enables highly accurate predictions and optimisations to be made, including new workloads. This data is compared against the business policies for deployed workloads and assesses if the HW if well utilised. It can then analyse options to optimise the workload. The analysis uses a mix of machine learning, algorithms and human expertise, depending on the goals and the commercial provider. The workloads are then redistributed to optimize deployment. The reduction is primarily aimed at reducing costs.

Results

On average, Densify increased VM density by 48% and reduces hardware by 33% (Densify, 2018). It is expected this also reduces energy consumption relative to the hardware reduction.

4.11 Data management and transmission metrics

Effective data transmission rates are not always published and determined by a number of factors such as content size, bit transmission rates, line rates, transmission line quality, error correction and others. Efficiency and data integrity are however, critical needs of the system and eventually the value of the network employed. For these reasons, a more data centric view is needed to ensure we maximize the efficiency and effectiveness of the network of systems employed. Line rates and bit transfer rates are too arcane and merely address an already standardized portion of the factors. If metrics are to encourage improvement in the effectiveness, one must consider the full transmission, the energy consumed to complete the full transaction, and standards and best practices that would further technologies along those lines. Data centric considerations and approach provide a logical way to maximize the effectiveness and efficiency of secured data management and transmission.

Problem

Existing data transmission metrics and assessments drive several issues. Transmission rates are usually quoted as line rates or sometimes advertised as cumulative line rates, which one will almost never be able to achieve. Hidden amongst these bits-per-second quotes are issues such as resend occurrences and drop rates. When a single bit error occurs, is it repaired or is the packet or transmission resent? What about dual bit errors? What's the equipment non-drop rates? Full mesh or line capability? As speeds increase, the number of these occurrences (per second) will increase (a statistical certainty) driving an even greater amount of wasted energy. Obviously, if one can repair single and dual bit errors, without retransmission, the network and end point demand and energy is reduced. If one only needs to resend a small packet as opposed to restarting a transmission, the energy savings for items like video transmission can also be very significant.

Solution

Viewing the data needs points another efficiency and effectiveness set of issues around unstructured data, aka Big Data. With the influx of IoT devices, all with different applications and data constructs, the amount of unstructured data to hit the network has be staggering. Current methods send the unstructured data to a remote computing site, where platform or services aggregates and restructures the data so it may become useful. Restructuring includes authentication, context, validation, authorization and other information. Not only is the restructuring of data at a remote centralized location inefficient, this method is high latency, prone to error, a high security risk and high privacy risk. The re-structuring of data should be systematic, validated and secured as close to the source and near the useful response as possible. Centralized IT should be used to set policy, coordinate between

accepted and validated sites, and organize information (not raw data) to provide intelligence for the remote sites to execute functions relevant to their proximity.

4.12 Summary

Figure 11 summaries the various intelligent efficiency techniques identified in the case studies and their application in relation to the DC/WAN as well as the emerging technologies which support it. It is clear that opportunities have already been identified in every part of the system and might reduce energy consumption by up to 75%. The most important enablers of the energy savings are energy aware equipment, SDNs and a smart or AI management system that can track and optimise energy efficiency or other related targets. These case studies represent a snapshot in time of the techniques available and it is certain that other techniques are being developed. For example, beyond just energy efficiency, new intelligent techniques have been developed which can also take into account renewable energy supplied to the equipment and coordinate traffic accordingly (Sheng, et al., 2015). When considered together it strongly suggests that a lot of potential efficiency improvements are still to be implemented.

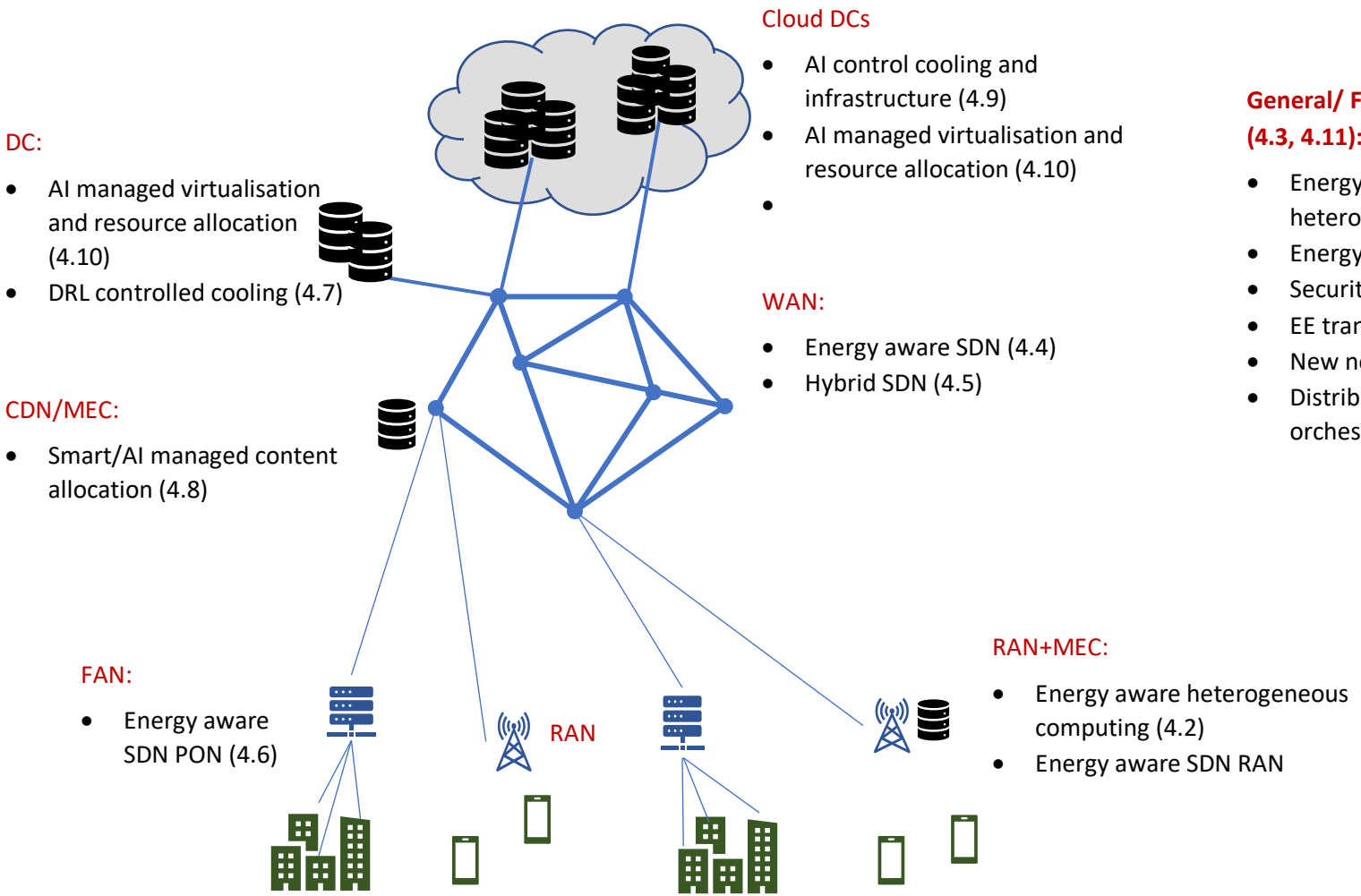


Figure 11 Summary of case studies applied to DC/WAN

5 Analysis of the new trends and case studies

The cases studies show how new technologies and emerging trends can enable new ways to improve efficiency through more intelligent use and allocation of the ICT equipment and resources. They also demonstrate commercial demand in some sub-markets already exists. The next step is to therefore understand how policies can encourage both new techniques and more widespread implementation of existing techniques to maximise energy savings that are available. By their very nature, case studies cannot be comprehensive and so the policy development is focussed on creating the conditions required for intelligent efficiency to flourish rather than implementation of any particular techniques. This starts with an analysis of the case studies to identify the necessary conditions.

5.1 Requirements for intelligent efficiency

In general, DC and WAN are well suited to intelligent efficiency techniques because they are complex interacting systems and all have unique characteristics which defy a one-size fits all solution. A prescriptive approach to technology is therefore unlikely to work and could result in policies that hinder efficiency and innovation, and may also impact the quality of the service required under agreed service level agreements (SLAs). Instead, by distilling the essential elements needed for intelligent efficiency to be deployed, policies and regulatory bodies can work to define the framework for these, to ensure efficiency is achieved and help new efficient technologies and innovations to continue.

Based on the case studies and research papers, the following principles can be extracted:

- **Energy aware hardware**

  Most techniques depend on the ability to move tasks between equipment to increase utilisation and overall efficiency. This is most easily achieved when the equipment is able to report energy consumption. Equipment may also need to be able to report utilisation and efficiency characteristics such as peak efficiency and proportionality.

- **Interoperability**

  The equipment must all be controllable by the same management and orchestration system to take advantage of automated systems.

- **Data collection, energy-, geography-, network topology-, consumer demand-awareness**

  Predictive software shows the highest potential for saving energy and this means that more data is required to describe the service and customers. Realtime information has the highest potential. More information generally enables greater energy savings to be made although there must be diminishing returns since the analysis of the data will itself consume energy. It is not clear where the optimal balance lies.

- **Information accessibility**

  The information collected must be accessible to whatever system is making the efficiency decisions. This may cross business types and owners, e.g. from data centre operator to the software application.

- **Software capabilities**

Software must be able to take advantage of the efficiency opportunities available. This includes design that takes into account of how much data is generated and how and where it is being stored and processed. In addition, when taking advantage of next generation technologies, the application software may need to be rearchitected to take advantage of new techniques such as specialised processors. In the future, software languages and operating systems may be more capable of understanding what an intended function of the application software is and allocating the most efficient resources.

- **Automation and AI**

  The constant changes to the operating conditions means that automation is needed to continuously monitor and adjust the system. Because system efficiency is non-linear and as the volume of data increases, AI automation is more effective even with relatively small datasets. As more AI is used, it becomes less necessary for the equipment to report efficiency characteristics since these can be implicitly determined by the AI.

Networks can be divided into two types of systems, Operational Technology (OT) and Information Technology (IT). Typical users and professionals are aware of LAN or WAN connections targeted for IT. These are the data services the average user is familiar with. OT describes the network and systems providing administrative and machine services. One can think of OT as an inhouse private network for command and control (C&C) of the systems. Obviously placing C&C on IT is a higher security risk than placing those on a dedicated OT network. For convenience and ease of use, many networks have physically combined IT and OT to share the same physical network, while creating logical ports to serve as the transport and isolation for OT. With the expansion of industrial IoT systems in addition to consumer IoT devices, the need for separation is even greater, to address security, privacy and efficiency. As any single intrusion on OT systems can cause irreparable harm to the underlying systems and data, security of OT networks and systems can be a relatively small energy and financial cost as compared to the data systems recovery, remediation and end user liabilities resulting from a compromise.

5.2 Barriers to implementation of intelligent efficiency techniques

A variety of technical, operational and business reasons may prevent the conditions identified in the previous section and thus hinder the implementation of intelligent efficiency techniques. This section applies knowledge of the existing market to the principles identified and identifies what these barriers are, and therefore what policies might want to tackle.

A general observation is that it seems that effective efficiency policies are becoming harder to isolate from other factors. This requires more discussion and analysis to understand how to assess and identify the optimal point between efficiency, end user services and ease of their development.

5.2.1 Legacy equipment

The biggest issue is the number of legacy data centres and the legacy WANs which do not have compatible, energy aware equipment and are highly inefficient. Modelling suggest that every Core WAN consumes approximately the same amount of energy and by shutting down 1 or 2 of the 3 to 4 networks still in concurrent operation could itself reduce energy consumed by 25-50%.

There are economic incentives to shut down and save energy, but the capital costs of new equipment mean that accelerated replacement of working equipment is not effective. In addition, without specific incentives to save energy, research suggests than that benefits of the new technology to improve service e.g. traffic engineering to improve QoE from SDN can be realised with just 20% of the network using SDN. This means that the large 50% energy savings demonstrated from SnetCA would not be realised.

There is also an inertia and lack of knowledge, especially for smaller companies, in how to move systems into the cloud. This may also have implications when trying to switch from the cloud to newer technologies. While we expect large companies will continuously redevelop and rewrite their software to make use of the latest technology, there are many smaller cloud software companies developing small bespoke cloud applications for small and medium enterprises (SME) clients which may need to be rewritten to run on the new technology. The energy savings are achieved by the relatively small individual savings accumulating over a very number of applications. Due to lack of budget, lack of knowledge and lack of incentives the individual client might not have the application updated to run on the latest technology.

5.2.2 Data availability

Some of the data required may not be available since it is not monitored, or no metric is able to express it. For example, it is difficult to quantify the efficiency of a particular server for performing a specific task. In addition, the efficiency varies depending on the utilisation level of the equipment at the time it will be performed. Standardised metrics designed to represent efficiency with a single figure may be insufficient to meet the demands of intelligent efficiency services.

5.2.3 Data sharing

The principles also require access to detailed information for the most efficient decision-making. This information crosses over many domains, for example energy efficient MEC needs to be able to quantify the efficiency of the various computing options as well as the network to compare and decide which is the best option. Since the ownership of the software application, network and computing resource may all be different, the data must be shared between these boundaries. In some cases, this is also considered to be commercially sensitive since it constitutes a competitive advantage.

5.2.4 Infrastructure sharing

Multiple networks increase duplication of hardware and lower utilisation and efficiency. Sharing infrastructure and hardware can increase efficiency and lower capex but this has implications for competition, especially when it has been demonstrated multiple networks can improve competition e.g. coverage for mobile networks.

5.2.5 Interoperability

There are consortia developing open source solutions to enable interoperability for SDN, Fog, NFV through OpenFlow, OpenFog, OPNFV and other standards which appear to be widely supported by manufacturers. However, there are also a small number of competing standards and as the standards continue to develop it may not be possible to achieve backwards compatibility. This creates a risk of

incompatible and non-interoperable equipment being deployed on the same network due to the private ownership of different parts of the network.

5.2.6 Very high service level requirements

There is a trade-off between service level and achieving maximum energy savings. When contracts and SLAs demand very high service levels, implementing some options to save energy may result in a breach of contract. Inconsistencies in services over a period of ten minutes will carry associated SLA penalties or lost revenue, which could dwarf any energy savings and result in disabling any energy minimization system.

5.2.7 Security and Privacy

Sharing data also raises security and privacy issues. For example, storing photos and social media content at the MEC of a person's home and work address clearly requires long term tracking of each person's frequently visited locations to find common trends.

Privacy rules may prevent the data collected from being used to optimise efficiency, for example, the EU GDPR requires a lawful basis for processing data, and the user would most likely need to give explicit consent for its use to improve efficiency.

Some of the activities being suggested could also have security implications for critical infrastructure, if network equipment can be remotely shut down, it might also be possible for a hacker to shut down the entire WAN.

5.2.8 Support by software and application developers

Software and applications developers have a big influence on efficiency, and they are able to decide how much data is created, transported and processed. Discussions with industry suggest that many have no interest or incentive to improve efficiency. Furthermore, software efficiency is not a mature subject and more work needs to be done before implementing metrics, standards and policies. Some services lend themselves better than others to efficiency metrics, e.g. kWh per hour of video for media streaming services is relatively easy.

5.2.9 Resistance to unproven techniques including AI/DRL

Networks and data centres may be unwilling to take the risk of deploying AI just in case it behaves incorrectly, resulting in service failure. A standardised process to validate the efficacy of AI for the network or data centre may need to be developed. Similarly, energy aware SDN management that deliberately slows or shuts down equipment will need to win the confidence of the network operator before deployment.

5.2.10 AI/DRL expertise

Data science expertise is required to analyse and implement automated management. This is a relatively new field, and there is huge demand for data scientists in all industries. Energy management may be a less attractive field compared to those which may pay higher such as finance/marketing or hold more interest such as self-driving cars.

6 Efficiency roadmap

6.1 Energy consumption modelling

As part of this study, a model of the global energy consumption of DC and WAN was developed based on the latest available data. In general, very limited efficiency data for WAN was identified and most information is contained in academic research with a narrow scope. This contrasts with consumer electronics which are well addressed by product energy performance regulations and have been modelled extensively. In some areas, the data showed large variations in historic and future projections between data sources, often the result of different boundaries when measuring as well as differences between geographical regions. Consolidating this data to give a representative global value expert judgement from the research team was required.

The model projects energy consumption under business as usual (BAU) conditions to 2030, however there is very weak evidence available to support such long-term projections beyond simple extrapolations. **Projections beyond 2021 are therefore highly speculative and these should be interpreted with extreme caution**. As a result of this, the outputs of the model were normalised against previous IEA projections, for consistency. Because projections are extrapolated and changes accumulate over time, a small annual percentage increase e.g. in the predicted data demand, will be compounded into a very large increase in energy by 2030.

Although it was not possible to model the uncertainty, this energy projection would lie at the lower range of possibilities as there is a much higher upper limit compared to the lower limit. This is because there are more possibilities for how data demand could increase even faster, rather than situations which would reduce growth. However, it should also be noted that increases in data demand do not occur independently of the efficiency improvements but must be enabled by the technology so the increasingly data intensive end-user services are technically and economically viable. It is therefore highly unlikely that energy consumption would ever increase at the same ~20% rate as the data demand as this suggests equivalent annual growth in capital and operating costs.

Future work is encouraged based on a sensitivity analysis to understand the possible range of scenarios, which is beyond the scope of this project. The model has therefore been structured to enable this analysis to be undertaken with relative ease. A more detailed explanation of the model, data sources used and some qualitative discussion of the uncertainty can be found in Annex 2.

The model is divided into the constituent parts of the data centre and WAN, i.e. ICT equipment, DC infrastructure, Core network, RAN and FAN. Where possible, three generations of technology are modelled, the legacy, modern and next generation. The generational divide is distinguished by new technologies that are incompatible with older hardware and require new knowledge to be applied. For example, to achieve maximum speeds the 5G network will require many additional small cells and new mobile devices while virtualising applications requires learning how to operate virtualised environments as well as updating the applications themselves. In general, legacy technology efficiency will improve slowly, while modern and NX have much faster improvement rates. Data demand is aggregated to the global level which means it must be interpreted carefully with respect to a specific region. Different regions and countries have very different states of network maturity and data growth rates. In addition, the share of the network between fixed/radio access networks and legacy/modern

generations also varies widely. Regional Governments or organisations are encouraged to apply more granular data to the model to gain a better understanding of their situation.

Table 3 describes the key categorisations of the DC/WAN components.

The model is based on the energy consumed per byte of data transferred in the DC and WAN. This means data demand is the proxy for all DC/WAN activity including processing, storage and transport. An implicit assumption must therefore be made that at the global level the data demand scales predictably with the energy required for processing and storage. The model therefore does not explicitly account for changes in the complexity of the processing performed on the data, which has a very significant impact on the CPU energy consumed. For example, AI and deep learning involves repeating calculations on the same dataset potentially millions of times but the data might only be transferred a few times. In addition, the effect of the server and CPU processing efficiency and the impact of Moore's Law to drive efficiency improvements is not directly addressed. This could be a more accurate approach for estimating data centre energy consumption but was not possible within this report and could be investigated in future work.

Data demand is aggregated to the global level which means it must be interpreted carefully with respect to a specific region. Different regions and countries have very different states of network maturity and data growth rates. In addition, the share of the network between fixed/radio access networks and legacy/modern generations also varies widely. Regional Governments or organisations are encouraged to apply more granular data to the model to gain a better understanding of their situation.

Table 3 DC WAN energy model categorisation of components

	Legacy/Traditional	Modern	Next Generation
Core WAN	PSTN core, metro and backhaul	Fibre optic	SDN, NFV
RAN	2G, 3G networks	4G networks	5G networks
FAN	ADSL	VDSL, cable, PON	XG PON
DC ICT	Small data centres, low virtualisation ratios	Cloud, highly virtualised	Fog, heterogeneous, MEC
DC Infrastructure	Low efficiency, poor proportionality	DCIM, efficient	Highly efficient, very proportional

6.2 Energy consumption trends

The overall energy consumption graphs (Figure 12, Figure 13) show that the energy consumption is expected to drop relatively rapidly in the near future and slowly rise gradually in the future. The fall is dominated by the projected energy consumption of the 2G network which is discussed later. After this, there tend to be more gradual changes, while energy consumption is driven up by increasing amounts of data, this is mitigated by the replacement of older technologies by newer and more efficient technologies.

Figure 12 Data centre energy consumption by category

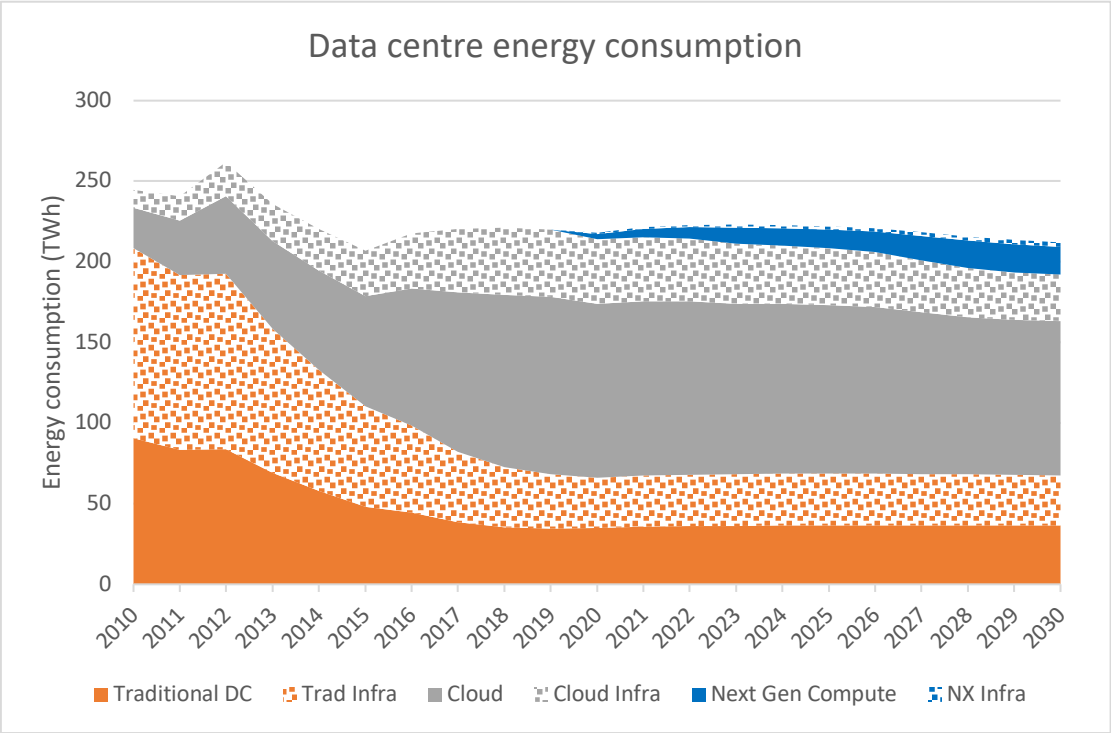


Figure 12 shows that data centre energy consumption is projected to be relatively flat from 2017 to 2030. The IEA *Digitisation and Energy* report (see figure 5.1) (IEA, 2017) projects a small increase in energy consumption by data centres from 2014 to 2020. Both this study and the IEA report assume that large efficiency gains are made as the technology transitions from traditional to cloud data centres, although this study looks further into the future (to 2030) than the IEA report.

Figure 13 WAN energy consumption by category

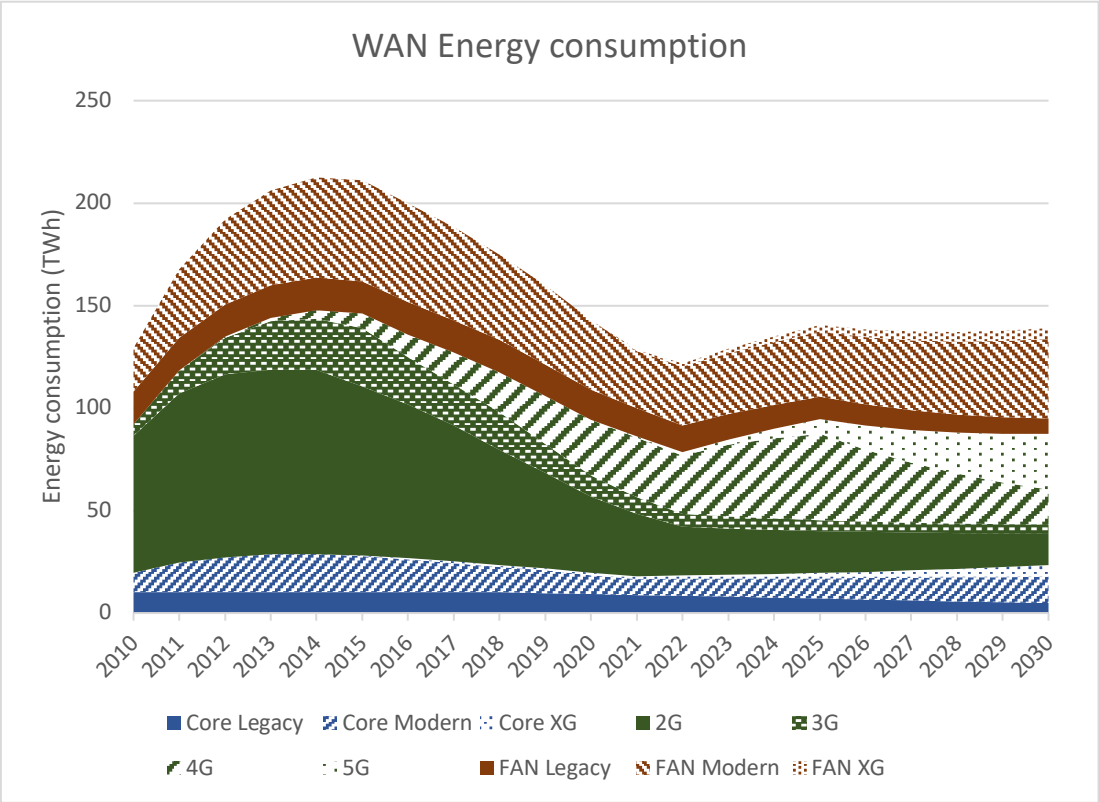


Figure 13 projects that WAN energy consumption decreases over the period 2014 to 2022 to 120 TWh and then increases slowly to 140 TWh. In the IEA *Digitisation and Energy* report (see figure 5.3) (IEA, 2017) a range of projections is provided for the period 2015 to 2021: the "Moderate efficiency scenario" predicts energy use to increase from 190 to 310 TWh (2015 to 2021) and the "High efficiency scenario" predicts energy use to decrease from 190 to 160 TWh (2015 to 2021). This EDNA study projects a lower energy consumption outcome in 2021 compared to the IEA report "High efficiency scenario". This suggests that the assumptions used in this study assume greater efficiency gains than the IEA report, although this study looks further into the future (to 2030).

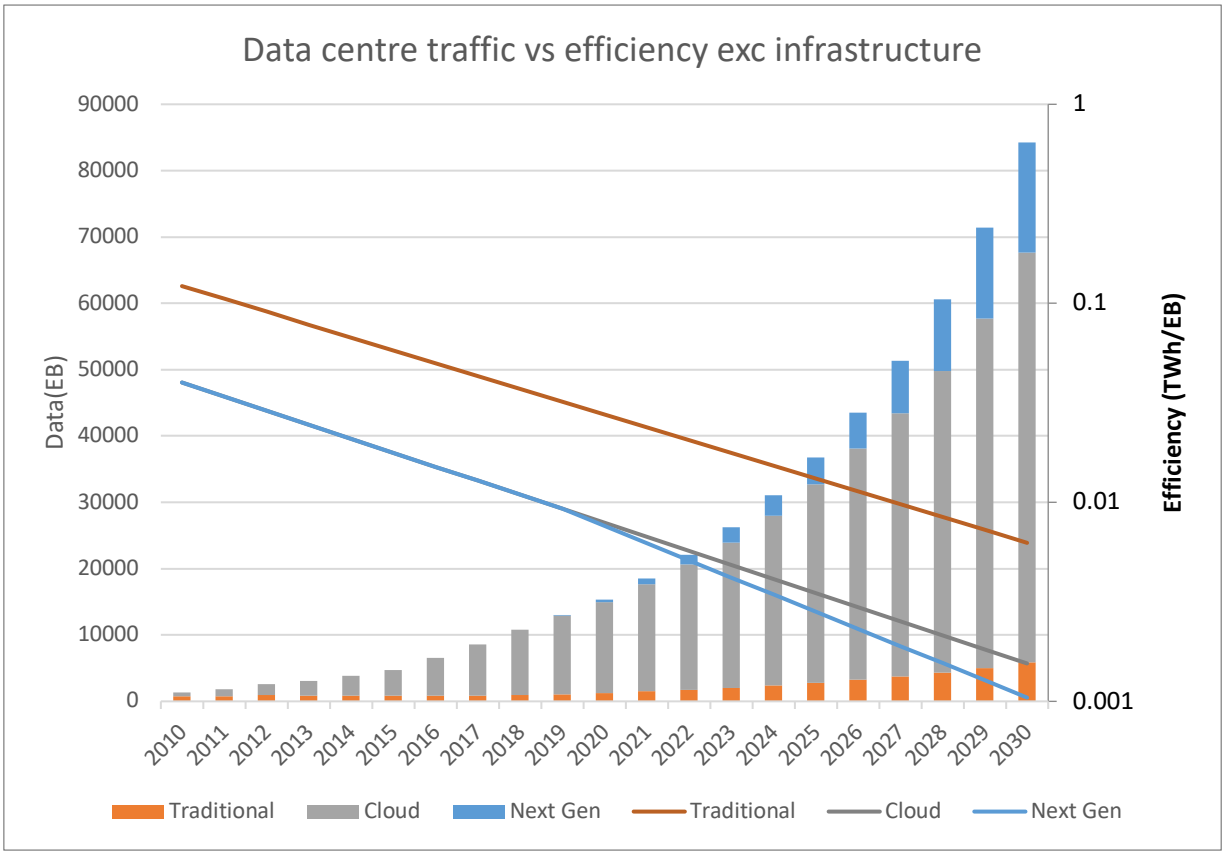
Both the DC and WAN projected energy consumption to 2030 are subject to much uncertainty, as discussed in the earlier paragraphs, and the projections resulting from this model appear to be at the higher end of the efficiency improvement range. More detail on the results of the modelled energy consumption by category of technology is provided in the following sections.

6.2.1 Data centres

DCs in 2017 consumes approximately 220TWh, about 17% more energy than the WAN. The energy consumed by the cloud is the largest proportion (140TWh) and will peak in 2019 before very gradually falling again. This trend contrasts sharply with the rapid growth in data (Figure 14) and is only achieved by the continued and stable efficiency improvements which arise from efficiency improvements of the software, platform and hardware (See Annex 2 for more discussion on data assumptions). Legacy data centre energy will fall slightly to 2020 then remain stable. The legacy data centre represents only 8% of the total data but consumes over 30% of the energy. This is because the efficiency is much lower than

cloud and projected to increase more slowly since only hardware efficiency improvements occur with no further development of the platform or software. Conversely, next generation DCs are projected to represent approximately 20% of the data by 2030 but consume only 12% of the energy. This is due to the efficiencies gained from the new technologies and best practices which allow efficiencies to continue to improve. However, only a relatively conservative additional efficiency improvement is projected due to uncertainties about the extent to which these intelligent efficiency techniques are implemented.

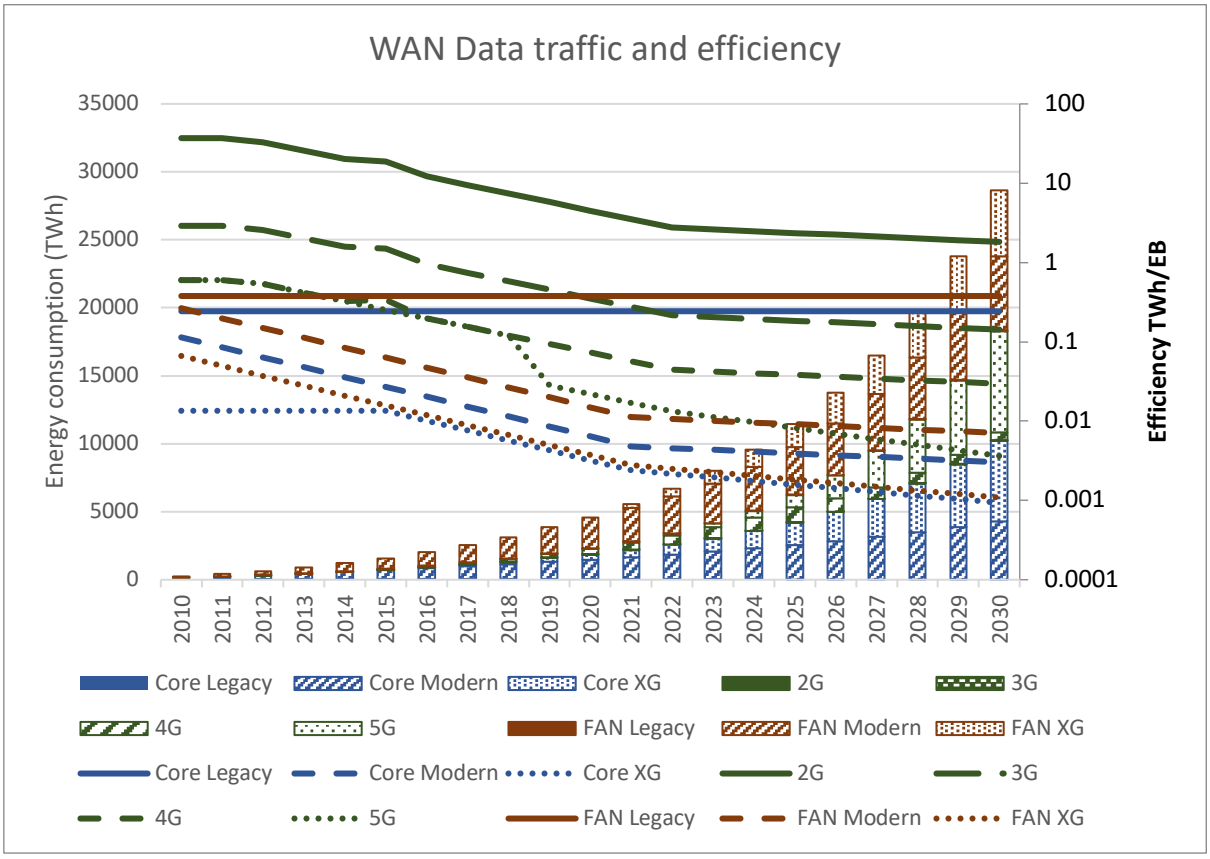
Figure 14 DC traffic and efficiency



6.2.2 WAN

Figure 15 shows the WAN data traffic and efficiency by category of technology considered in this study.

Figure 15 WAN data traffic and efficiency



Since the WAN has more constituent parts it is more difficult to break down. Generally, the energy consumed by the core is very low (~13% of total WAN). However, again, the legacy networks contribute a large proportion of the overall core network energy consumed. Because the equipment energy is not proportional (Jalali, 2016), even if the data traffic falls, energy consumption remains the same until it is shut down.

The projected RAN traffic (mostly 5G network) is growing very fast, over 40% a year compared to around 20% for FAN, although both are declining in growth rate. Since FAN represents a much larger proportion of the total data, the overall data growth of the access networks is projected to remain relatively stable at 21% growth. The RAN however, is relatively less efficient which can be seen in Figure 15 (noting the logarithmic scale on the right-hand axis).

There are widely varying projections on the efficiency gains from networks and, compared to computing, there appears to be more concern that networks will approach theoretical limits (Ellis, 2014) before growth in data demand slows which restricts efficiency improvement techniques. From 2021, a relatively conservative efficiency improvement of only 5% for 2G, 3G, 4G networks is projected. It is generally estimated that a 5G network is 10 times more efficient than 4G network, however, assuming the initial rollout is used to create broad coverage and with very similar technology, then efficiency will start at a similar level to 4G. Only when there is sufficient data demand to justify the investment will small cells, MIMO and millimetre wave radio be deployed. As a result, 5G networks efficiency improves at a higher rate, 20% a year. The energy consumption of these RANs suggests that the 2G and 3G networks are projected to consume a similar amount of energy as the 4G and 5G networks, despite very low use.

FANs are the largest energy consumer because of the high volume of data compared to RAN, and its inefficiency compared to the Core network. It is not clear what future technologies and efficiency improvements are expected, and therefore it is assumed that a 25% improvement is expected until 2021 in line with historic trends and thereafter a conservative 5% efficiency improvement is expected. As a result of these assumptions and the very different rates of efficiency improvement, 5G RANs is more efficient than FAN beyond 2025. This may not be realistic and may require more analysis.

The results shown do not address the energy consumption of the consumer premises equipment (CPE) including the modems and mobile devices. This could change the net efficiency of the network considerably because a FAN CPE can have a higher energy consumption than a RAN device. However, the Total Energy Model[4] includes the CPE and LAN equipment.

6.3 Efficiency and utilisation

The previous sections have shown how efficiency of a DC or WAN is the result of the efficiency of the individual pieces of equipment (both peak efficiency and energy proportionality) as well as how they are managed together. Many intelligent efficiency techniques take advantage of this to reduce energy consumption by increasing utilisation and reducing the amount of equipment operating, particularly during periods of low utilisation.

The model calculates the energy consumption during high and low utilisation periods which are representative of the common day/night pattern found across data centres and WANs. Inputs allow the average peak efficiency and energy proportionality for the equipment to be adjusted, as well as the percentage of the equipment active, to calculate and simulate the *overall* efficiency. The model also calculates the theoretical peak efficiency of the DC/WAN (if it was possible to operate at 100% utilisation). This is important because the maximum efficiency can only be improved by developing more advanced technologies which might not continue indefinitely and at the same rate as data growth. With these inputs and outputs, it is possible to understand how targeting changes to individual equipment or system management can achieve an improvement to the overall efficiency.

[4] A related report on the total energy use of connected devices calculates the energy consumption of the LAN and additional energy used by connected devices see *Total Energy Model for Connected Devices*, (EDNA, 2019a)

Figure 16 Projected energy consumed in data centres

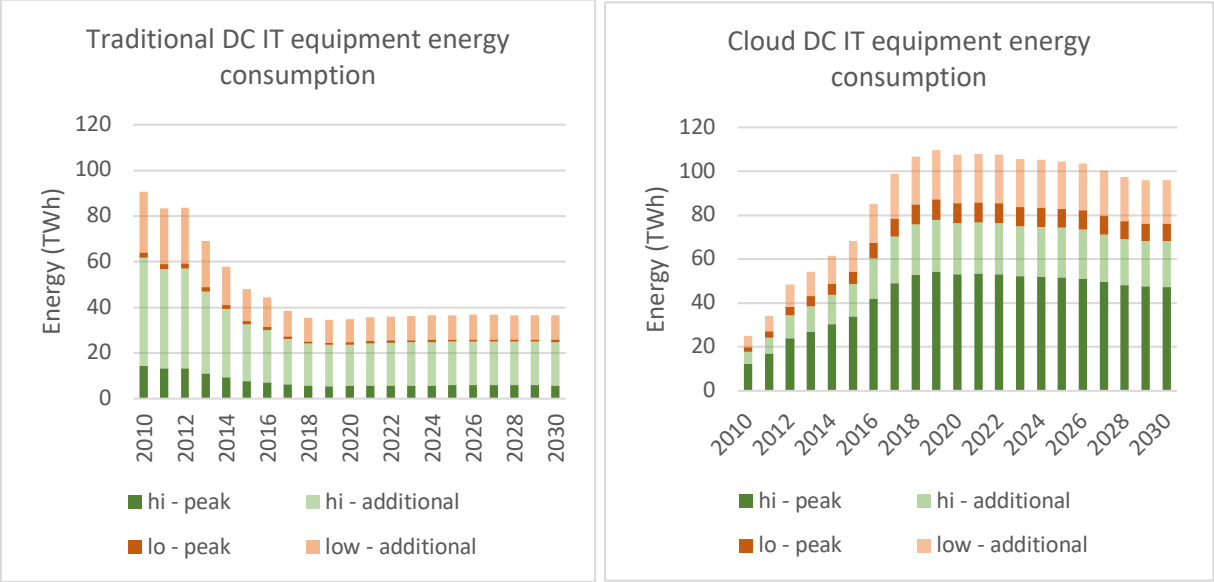
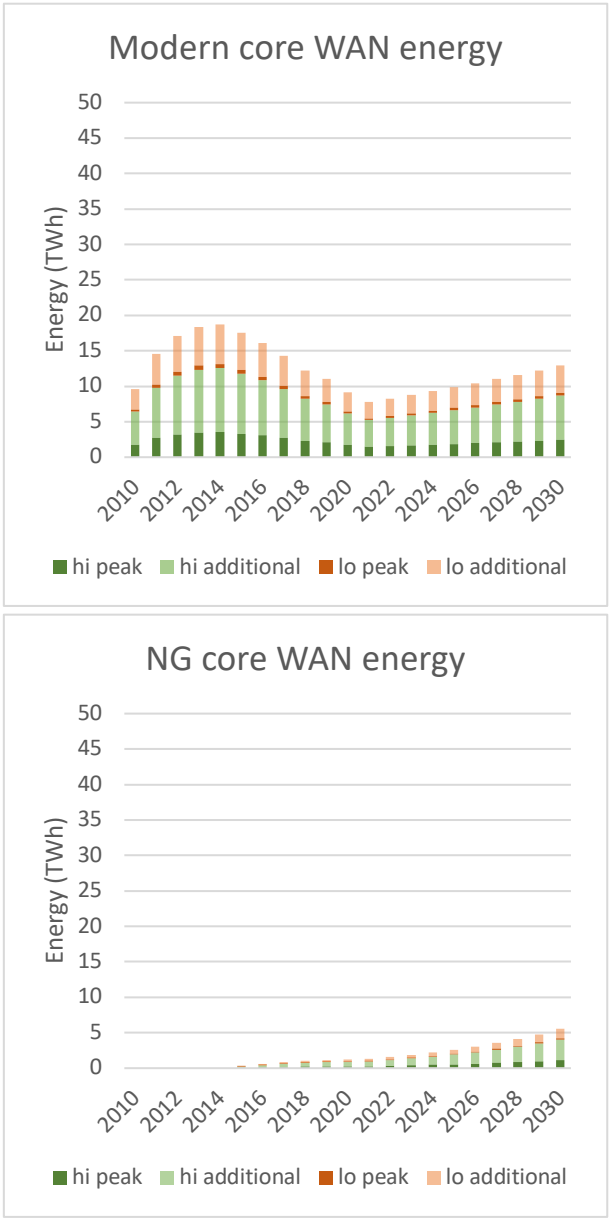


Figure 16 shows the energy consumed by the IT equipment in traditional and cloud data centres and excluding the DC infrastructure. The energy consumed is split by the diurnal pattern or high utilisation during the day (green) and low utilisation at night (orange). As expected, this shows that most of the energy (approximately 70%) is used during the high utilisation period.

In addition, each utilisation level is then broken down into 'peak' and 'additional', distinguished by the lighter shading. The peak energy represents the amount of energy that would be consumed during the processing of the data if the equipment was theoretically able to operate at the calculated peak efficiency level. The additional energy being used is therefore due the DC/WAN not perfectly scaling power proportionally with utilisation and indicates potential energy savings available from intelligent efficiency techniques. For legacy data centres, the additional energy is 80% of the total energy consumed and significantly higher than the theoretical peak energy. As a proportion of total energy, this is reduced greatly by cloud and similar modern DCs to 40%, although it is still far from eliminated. In addition, because the total energy consumed by cloud is higher than traditional data centres, the additional energy in absolute terms is still larger than traditional data centres and should remain a priority.

Figure 17 Projected energy consumed by core WAN at high and low utilisation



A similar analysis of WAN shows that the additional energy is very high for the core even next generation WAN, although total energy consumption is low (Figure 17). 4G and 5G networks by comparison are expected to relatively efficient and minimise the additional energy although total energy consumption remains high. This means it should continue to be monitored to ensure the networks are achieving the level of efficiency expected.

Figure 18 Projected energy consumed by 4G and 5G WAN at high and low utilisation

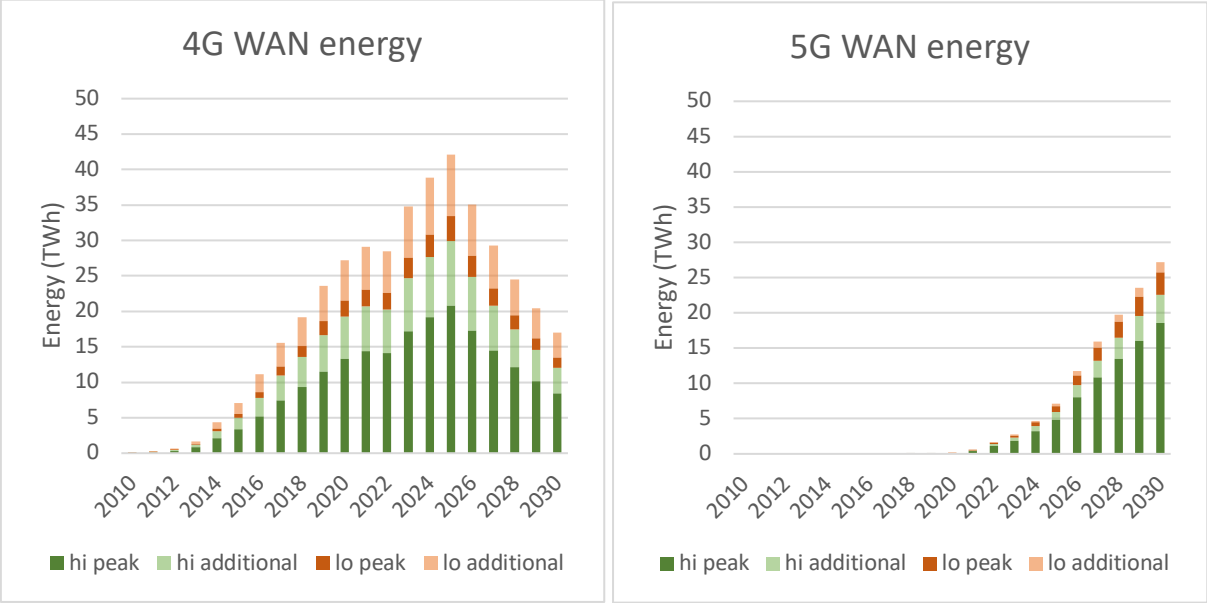
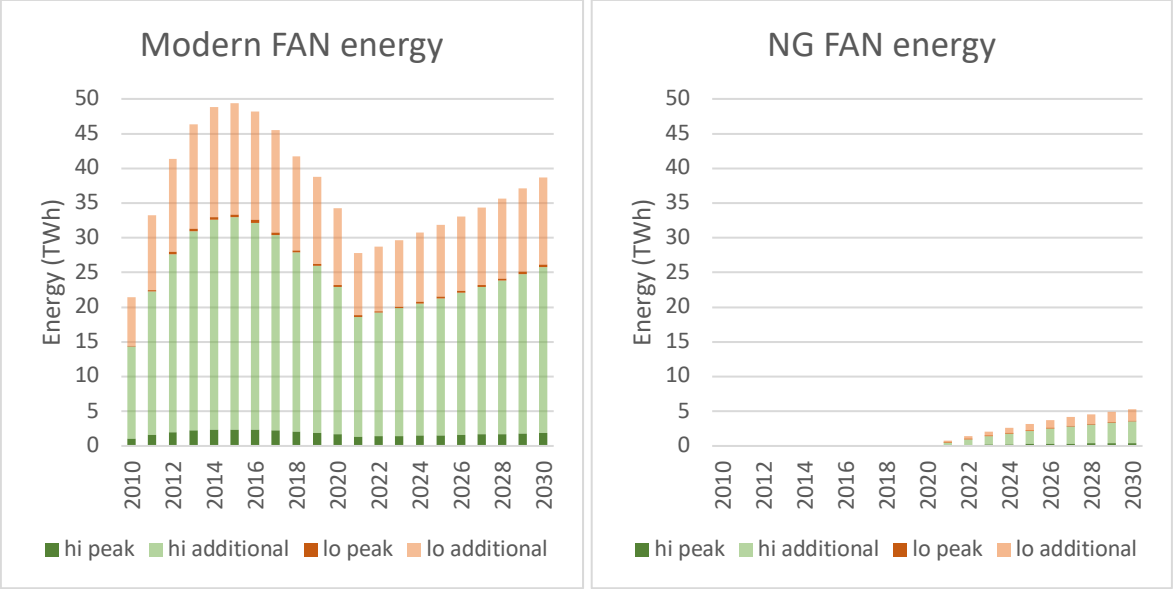


Figure 19 Projected energy consumed by FAN at high and low utilisation



The additional FAN energy is very high (Figure 19) due to the very low average network utilisation level, only 5%. While some modern FAN equipment has different power levels for different utilisation levels, the average is too low for this to have a significant effect. Next generation FAN equipment is expected to be much more energy proportional and in the near term have low total energy consumption. Given the high total energy consumption and additional energy consumption, the modern FAN should be a focus for energy savings.

6.4 Intelligent efficiency opportunities and priorities

There are still many unknowns about how technology will develop, and a prescriptive technology path is unlikely to be suitable. New technologies hold the possibility of making major gains in efficiency but it remains to be seen whether businesses will pursue them aggressively without strong motivation.

Based on the energy modelling and best practices, we can prioritise actions which will create the greatest energy savings due to efficiency improvements or current energy share, which are necessary steps for enabling energy savings to be made. From these, policies and incentives can be developed to accelerate and improve the application of intelligent efficiency technologies.

6.4.1 Legacy DC/WAN

Intelligent efficiency techniques cannot be applied to current legacy data centre and network equipment since they are not energy aware and interoperable. Solving this would require additional investment which is highly unlikely. Networks and systems containing a mix of old and new is inevitable but some technologies such as circuit switching can have severe impacts on overall efficiency. While some energy savings can be achieved by improving the power and cooling infrastructure, pursuing a policy to assess legacy operations then switch off and migrate WANs and DCs to modern or next generation solutions whenever environmentally and economically feasible is the most effective solution.

6.4.2 ICT equipment utilisation

Maximising utilisation by optimising virtualisation and switching off unused equipment offers some of the biggest immediate energy savings, especially since this also reduces the infrastructure energy consumption. Commercial AI services are already available that offer this, provided the necessary information, particularly utilisation data is available and centrally monitored.

6.4.3 Data centre and telecoms Infrastructure

New power and cooling infrastructure efficiency has already improved greatly compared to ten years ago. However, managing the infrastructure with AI appears to be relatively simple, and worthwhile efficiency improvements are achievable even with relatively limited controls in the infrastructure.

6.4.4 Next generation DC

Heterogeneous computing, MEC and fog in theory can increase efficiency if utilised well. This will require software support and for current software to be redeveloped to take advantage of it. In addition, it will require the efficiency data of the networks and computing to be made available to the software, or some other mechanism, to ensure the most efficient computing resource is being used. Since these are still in development, work to ensure standards consider efficiency which can easily be utilised by software developers is currently the most important factor.

6.4.5 FAN

Fixed Access Networks are a major energy consumer and carry a high proportion of the data. There are a number of techniques to reduce the energy consumption of the equipment at low utilisation levels which can provide reduced energy consumption at lower utilization levels while still maintaining required connectivity and latency This is particularly important because FAN average utilisation is

already low and expected to fall. Modern FANs have low power modes but these can be extended further. Some research suggests that FAN energy can be reduced by 90% though a mixture of low power modes and cross connecting PON links, however, this does not take into account the power consumption of the equipment in the premises which have a much higher energy consumption.

While one specific technology has been identified (TWDM PON) as highly efficient, the additional cost of installing new FTTH cabling compared to using existing cables for G.Fast and Cable means it may not be used. Almost all energy efficiency research is focused on PON with very limited information on alternatives.

6.4.6 4G, 5G RAN

RANs have shown the biggest efficiency improvements over time. It is expected that 5G networks will continue this trend, although the very high data growth rate will overwhelm these gains and lead to rising energy consumption. Requiring mobile operators to report efficiency could help ensure that higher efficiency levels are being targeted. In addition, ITU is investigating the use of AI in the network, and efficiency should be made a high priority.

4G networks are less efficient than 5G networks but is expected to continue in tandem. Switching off the 4G once 5G networks is rolled out could improve overall efficiency.

6.4.7 Core network

The core network consumes very little energy compared to the rest of the WAN and DCs. However, this potentially has energy savings of up to 75%. To maximise this, SDN management and orchestration will need to have explicit energy efficiency rules, which would probably most effectively be achieved with AI. However, the biggest energy savings will have an impact on the service level and this must be taken into consideration to find the optimum balance.

6.4.8 Reducing traffic

Traffic can be reduced by good design of the software and service. Though network and DC operators believe IT traffic management is out of their control, it is not. The services and tools can be architected and integrated as part of the service and logical platform. There exists a need for software developers, network architects, hardware developers and service providers to develop and integrate routing schemes to minimize the hops and reroutes of targeted services.

Traffic can also be reduced and optimized with a data centric approach implied previously in the big data example. A standardized data hierarchy may not only reduce traffic to only relevant and optimized data constructs, but also offer privacy, security, low latency, reduced errors, and efficiency. A data hierarchy would localize big data, provide opportunities for multi-level access control, localized context and data validation, and anonymization of privacy information to share with different access control and authorization lists.

7 Policy implications

The outcomes of this report and the energy modelling provide significant background information on the relevant contributions of WAN and DC to energy use. The modelling also shows the changing energy use over time with different generations of equipment. The model does not indicate runaway energy consumption will occur, although more sensitivity analysis is required to confirm this. However, this does not mean that no action is necessary because as the case studies show, significant energy savings are still available.

From a technical perspective, achieving the maximum energy savings requires meeting the principles needed such as energy aware hardware, interoperability etc. The barriers however, are more related to the businesses, how they operate, incentivised and the services being provided. Governments must be more engaged with industry in every sector from the RD&D stage through to deployment and operation of DCs and WANs.

Energy savings do not necessarily result in a reduced service, but the greatest energy savings are likely to require some compromise based on the current and expected new technologies. Given the renewed urgency to address climate change, there may be no alternative and so finding the balance between energy consumption and end-user service must start to be discussed with industry, consumers and public bodies. The opportunity for most potential energy savings and minimum impact is during low utilisation at night and could serve as a starting point for discussion. The initiative will most likely have to come from Government who can provide an open platform for starting these discussions with industry and stakeholders.

This study has also identified that the most effective actions will depend on the particular characteristics of the network, platform, business requirements and regional policies. The energy model shows the average global consumption but the different types of networks operating in a region can be very different. This will change the priorities and therefore recommended actions undertaken.

This section describes six key activities which could be promoted and supported by policies which are expected to have the greatest impact on the application of intelligent efficiency and subsequent energy savings.

7.1 Raise the priority of energy efficiency

There are many areas for improving energy efficiency which are not implemented. While EE is a concern for some industries its priority needs to be raised in others. Governments can encourage these through education and facilitating industry programs that raise awareness of the opportunities and metrics such as the Liaison Group of Japanese Industries (JEITA, 2015). Since many of the case studies provided have not been tested for commercial applications this includes supporting and pushing for new techniques through feasibility studies and new research and development. In general, Government and NGOs need to be more engaged with the industry at every level to understand how energy efficiency is integrated into all aspects of operations.

Voluntary actions could be effective in parts of the industry because there are relatively few companies involved such as mobile network operators. It is unclear if national telecoms regulatory bodies have a mandate to maximise energy efficiency but this could be an effective body to coordinate voluntary, or regulatory activities, starting with comprehensive energy assessments. The EU Data Centre Code of

Conduct or the EU Broadband Code of Conduct are examples where industry commits to certain actions to improve energy consumption.

Software applications and platforms are one of the biggest influences in the overall energy consumption. Hardware manufacturers are already engaged through policy and voluntary actions, but in contrast, software developers are poorly engaged in energy savings activities. More engagement could help steer the design of new security, transport protocols and hardware. Similarly, energy efficiency advocates must be more engaged in the standards activities around security, transport etc.

7.2 Integrate intelligent efficiency into modern DCs (and WANs)

Significant energy efficiency gains can be made by integrating the intelligent efficiency techniques into the modern DCs and WAN. Governments can assist industry integration of this by promoting case studies, reducing barriers to their implementation and encouraging businesses to cooperate with open standards. The opportunities that have been identified in this study for consideration are:

- For all DC/WAN encourage the installation of energy aware equipment with the capability needed for efficient management. This is a requirement for virtually every case study reviewed. For WAN this must be implemented as soon as possible due to the long equipment lifetimes.
- Promote more advanced AI/DRL management of cooling infrastructure in DC/WAN by settings targets and programs (See section 4.9, 4.7)
    - Minimum sensor/control/data collection requirements for new DCs
    - PUE targets based on size/age of data centre or implementation of AI/DRL management system
- Promote advanced management of virtual servers (section 4.10)
    - Monitoring and reporting of utilisation levels
    - Switching off under-utilised servers
    - Setting targets for under-utilised server
    - Use of AI to optimise virtualisation
- Encouraging software development by
    - Publishing End-user service efficiency metrics, e.g. kWh per hour video streamed
    - Awards for energy management to attract data scientists/AI experts.
- Develop performance indicator to drive efficient and effective transmissions to include full security management, minimized resends and retries, and optimize complete transmissions instead of just number of bits moved. More engagement with industry, including security technical committees is needed to understand how to progress this.

7.3 Develop detailed, standardised equipment efficiency reporting and metrics

Because utilisation varies so much, equipment metrics that describe how the power varies over utilisation are most informative. This is already occurring but at only a few utilisation levels which are then aggregated into a single figure, similarly to energy labels found on domestic products.

Having a single number to express efficiency is helpful for marketing, and necessary for consumers who will not have the expertise or time to study efficiency data. However, these are commercial and industrial products, not sold through a typical consumer shopping experience. Maximising efficiency

can only be achieved by comparing the product efficiency at the range of utilisation levels which the device is expected to operate in, and this is most easily determined with knowledge of the complete efficiency curve, rather than a single number. A standardised reporting format, JSON and/or API to access the power and performance data for each type of equipment would be much more useful for making detailed comparisons. The ease of reporting and comparing large amounts of data electronically also means that there is less need to reduce the efficiency behaviour to a single figure. Any standardised report and metric should also include testing at a minimal, non-zero utilisation level.

Many Governments already have extensive experience engaging industry on developing test standards and metrics. Focussing on reporting detailed data can help reduce the efforts spent on the difficult and sometimes contentious task of developing metrics.

7.4 Ensure next generation DC/WAN integrate intelligent efficiency at the outset

There is a considerable opportunity to promote integrated intelligent efficiency in the fabric of the next generation DC/WAN systems. This opportunity can be enhanced by:

- Interoperable, open standards which include efficiency as a primary issue
- Governments and industry working together to:
    - Assess trade-off between energy consumption and service levels (particularly at night)
    - limit devices with high impact on the network, e.g. critical health, to prevent devices claiming unnecessary and inefficient operating conditions
    - Network efficiency reporting
- Addressing the security and privacy issues that limit the sharing of data

7.5 Data Hierarchy and metadata standards

Common data standards metadata, converting, anonymizing, and restructuring data for optimized use across the network would reduce traffic across various networks. The hierarchical access control would limit unauthorized access and work with authentication and data validation schemes to further reduce traffic. This requires more industry engagement and discussion. Since consumer IoT is expected to be dominated by a few global platforms, and smart equipment designed to operate within these ecosystems, engaging these service providers could be the first useful step.

7.6 Transfer existing services from modern to next generation networks and DCs

When the efficient next generation DC/WANs start operating, the modern networks will become the new, inefficient legacy. A huge number of services, hardware and software will exist on the modern networks and some will never shift. These will continue to consume energy and resources for many years. The difficulty in moving people and services to modern technology has been a significant failure in the current situation. Policies that support and accelerate the shift of services to the latest technology can therefore have significant long-term impact. This may include awareness raising, financial assistance or some sort of certification scheme which recognises businesses and services with the expertise to transfer their services.

Annex 1: Standards

ITU Standards

L.1300	Best practices for green data centres
L.1301	Minimum data set and communication interface requirements for data centre energy management
L.1302	Assessment of energy efficiency on infrastructure in data centres and telecom centres
L.1310	Energy efficiency metrics and measurement methods for telecommunication equipment
L.1315	Standardization terms and trends in energy efficiency
L.1320	Energy efficiency metrics and measurement for power and cooling equipment for telecommunications and data centres
L.1321	Reference operational model and interface for improving energy efficiency of ICT network hosts
L.1325	Green ICT solutions for telecom network facilities
L.1330	Energy efficiency measurement and metrics for telecommunication networks
L.1331	Assessment of mobile network energy efficiency
L.1332	Total network infrastructure Energy efficiency metrics
L.1340	Informative values on the energy efficiency of telecommunication equipment
L.1350	Energy efficiency metrics of a base station site
L.1360	Energy control for the software-defined networking architecture
L.1400	Overview and general principles of methodologies for assessing the environmental impact of information and communication technologies
L.1410	Methodology for environmental life cycle assessments of information and communication technology goods, networks and services
L.1420	Methodology for energy consumption and greenhouse gas emissions impact assessment of information and communication technologies in organizations
L.1430	Methodology for assessment of the environmental impact of information and communication technology greenhouse gas and energy projects
L.1440	Methodology for environmental impact assessment of information and communication technologies at city level

ISO/IEC Standards

ISO/IEC 19395:2015	Sustainability for and by Smart data centre resource monitoring and control
ISO/IEC TR 20913:2016	Data centres -- Guidelines on holistic investigation methodology for data centre key performance indicators
ISO/IEC CD 21836 [Under development]	Data Centres -- Server Energy Effectiveness Metric
ISO/IEC PDTR 21897 [Under development]	Data Centres-- Methods and tools to assess and express energy production, storage and consumption at data centre level in reference to primary energy
ISO/IEC TS 22237-1:2018	Data centre facilities and infrastructures -- Part 1: General concepts

ISO/IEC TS 22237-2:2018	Data centre facilities and infrastructures -- Part 2: Building construction
ISO/IEC TS 22237-3:2018	Data centre facilities and infrastructures -- Part 3: Power distribution
ISO/IEC TS 22237-4:2018	Data centre facilities and infrastructures -- Part 4: Environmental control
ISO/IEC TS 22237-5:2018	Data centre facilities and infrastructures -- Part 5: Telecommunications cabling infrastructure
ISO/IEC TS 22237-6:2018	Data centre facilities and infrastructures -- Part 6: Security systems
ISO/IEC TS 22237-7:2018	Data centre facilities and infrastructures -- Part 7: Management and operational information
ISO/IEC PDTR 23050 [Under development]	Data centres -- Excess electrical energy XEEF
ISO/IEC TR 30132-1:2016	Information technology sustainability -- Energy efficient computing models -- Part 1: Guidelines for energy effectiveness evaluation
ISO/IEC PDTR 30133 [Under development]	Data centres -- Guidelines for resource efficient data centres
ISO/IEC 30134-1:2016	Data centres -- Key performance indicators -- Part 1: Overview and general requirements
ISO/IEC 30134-2:2016/Amd 1:2018	Data centres -- Key performance indicators -- Part 2: Power usage effectiveness (PUE)
ISO/IEC 30134-3:2016/Amd 1:2018	Data centres -- Key performance indicators -- Part 3: Renewable energy factor (REF)
ISO/IEC 30134-4:2017	Data centres -- Key performance indicators -- Part 4: IT Equipment Energy Efficiency for servers (ITEEsv)
ISO/IEC 30134-5:2017	Data centres -- Key performance indicators -- Part 5: IT Equipment Utilization for servers (ITEUsv)
ISO/IEC CD 30134-6 [Under development]	Data centers -- Key performance indicators -- Part 6: Energy Reuse Factor -- ERF

ETSI Standards

ETSI EN 303 470 methodology V1.1.1 (2019-01)	Environmental Engineering (EE); Energy Efficiency measurement and metrics for servers
ETSI EN 303 215 V1.2.1 (2017-10)	Environmental Engineering (EE); Measurement methods for energy efficiency of router and switch equipment
ETSI EN 303 215 V1.3.1 (2015-04)	Environmental Engineering (EE); Measurement methods and limits for power consumption in broadband telecommunication networks equipment
ETSI EN 305 174-1 V1.1.1 (2018-02)	Access, Terminals, Transmission and Multiplexing (ATTM); Broadband Deployment and Lifecycle Resource Management; Part 1: Overview, common and generic aspects
ETSI EN 305 174-2 V1.1.1 (2018-02)	Access, Terminals, Transmission and Multiplexing (ATTM); Broadband Deployment and Lifecycle Resource Management; Part 2: ICT Sites

ETSI EN 305 174-5-1 V1.3.0 (2018-03)	Access, Terminals, Transmission and Multiplexing (ATTM); Broadband Deployment and Lifecycle Resource Management; Part 5: Customer network infrastructures; Sub-part 1: Homes (single-tenant)
ETSI EN 305 174-8 V1.1.1 (2018-01)	Access, Terminals, Transmission and Multiplexing (ATTM); Broadband Deployment and Lifecycle Resource Management; Part 8: Management of end of life of ICT equipment (ICT waste/end of life)
ETSI EN 305 200-1 V1.1.0 (2018-04)	Access, Terminals, Transmission and Multiplexing (ATTM); Energy management; Operational infrastructures; Global KPIs; Part 1: General requirements
ETSI EN 305 200-2-1 V1.1.1 (2018-02)	Access, Terminals, Transmission and Multiplexing (ATTM); Energy management; Operational infrastructures; Global KPIs; Part 2: Specific requirements; Sub-part 1: ICT Sites
ETSI EN 305 200-2-2 V1.2.0 (2018-04)	Access, Terminals, Transmission and Multiplexing (ATTM); Energy management; Operational infrastructures; Global KPIs; Part 2: Specific requirements; Sub-part 2: Fixed broadband access networks
ETSI EN 305 200-2-3 V1.1.0 (2018-02)	Access, Terminals, Transmission and Multiplexing (ATTM); Energy management; Operational infrastructures; Global KPIs; Part 2: Specific requirements; Sub-part 3: Mobile broadband access networks
ETSI EN 305 200-3-1 V1.1.1 (2018-02)	Access, Terminals, Transmission and Multiplexing (ATTM); Energy management; Operational infrastructures; Global KPIs; Part 3: ICT Sites; Sub-part 1: DCEM
ETSI EN 305 200-4-4 V1.1.1 (2018-01)	Integrated broadband cable telecommunication networks (CABLE); Energy management; Operational infrastructures; Global KPIs; Part 4: Design assessments; Sub-part 4: Cable Access Networks
ETSI ES 201 554 V1.2.1 (2014-07)	Environmental Engineering (EE); Measurement method for Energy efficiency of Mobile Core network and Radio Access Control equipment
ETSI ES 202 336-12 V1.1.1 (2015-06)	Environmental Engineering (EE); Monitoring and control interface for infrastructure equipment (power, cooling and building environment systems used in telecommunication networks); Part 12: ICT equipment power, energy and environmental parameters monitoring information model
ETSI ES 202 706 V1.4.1 (2014-12)	Environmental Engineering (EE); Measurement method for power consumption and energy efficiency of wireless access network equipment
ETSI ES 202 706-1 V1.5.1 (2017-01)	Environmental Engineering (EE); Metrics and measurement method for energy efficiency of wireless access network equipment; Part 1: Power Consumption - Static Measurement Method
ETSI ES 203 184 V1.1.1 (2013-03)	Environmental Engineering (EE); Measurement Methods for Power Consumption in Transport Telecommunication Networks Equipment
ETSI ES 203 237 V1.1.1 (2014-03)	Environmental Engineering (EE); Green Abstraction Layer (GAL); Power management capabilities of the future energy telecommunication fixed network nodes
ETSI ES 205 200-2-1 V1.2.1 (2014-03)	Access, Terminals, Transmission and Multiplexing (ATTM); Energy management; Global KPIs; Operational infrastructures; Part 2: Specific requirements; Sub-part 1: Data centres
ETSI ES 205 200-2-2 V1.1.0 (2018-03)	Access, Terminals, Transmission and Multiplexing (ATTM); Energy management; Global KPIs; Operational infrastructures; Part 2: Specific requirements; Sub-part 2: Fixed broadband access networks

ETSI ES 205 200-3 V1.0.0 (2017-01) *On Approval*	Access, Terminals, Transmission and Multiplexing (ATTM); Energy management; Global KPIs; Operational infrastructures; Part 3: Global KPIs for ICT sites
ETSI ES 205 200-2-4 V1.1.1 (2015-06)	Integrated broadband cable telecommunication networks (CABLE); Energy management; Global KPIs; Operational infrastructures; Part 2: Specific requirements; Sub-part 4: Cable Access Networks
ETSI GS OEU 001 V2.1.1 (2014-12)	Operational energy Efficiency for Users (OEU); Global KPIs for ICT Sites
ETSI GS OEU 008 V1.1.1 (2013-09)	Operational energy Efficiency for Users (OEU); Global KPI for Information and Communication Technology Nodes
ETSI GS OEU 001 V1.2.4 (2014-10)	Operational energy Efficiency for Users (OEU); Global KPIs for Data Centres
ETSI GS OEU 012 V1.1.1 (2015-10)	Operational energy Efficiency for Users (OEU); Technical Global KPIs for Fixed Access Networks
ETSI TR 102 489 V1.4.1 (2015-10)	Environmental Engineering (EE); European telecommunications standard for equipment practice; Thermal management guidance for equipment and its deployment
ETSI TR 103 117 V1.1.1 (2012-11)	Environmental Engineering (EE); Principles for Mobile Network level energy efficiency
ETSI TR 103 229 V1.1.1 (2014-07)	Environmental Engineering (EE); Safety Extra Low Voltage (SELV) DC power supply network for ICT devices with energy storage and grid or renewable energy sources options
ETSI TR 105 174-6 V1.1.1 (2015-03)	Integrated broadband cable telecommunication networks (CABLE); Broadband Deployment and Energy Management; Part 6: Cable Access Networks
ETSI TS 102 706 V1.3.1 (2013-07)	Environmental Engineering (EE); Measurement method for energy efficiency of wireless access network equipment
ETSI TS 105 174-1 V1.2.1 (2014-09)	Access, Terminals, Transmission and Multiplexing (ATTM); Broadband Deployment and Energy Management; Part 1: Overview, common and generic aspects
ETSI TS 105 174-2 V1.2.1 (2017-01)	Access, Terminals, Transmission and Multiplexing (ATTM); Broadband Deployment and Energy Management; Part 2: ICT sites

ATIS standards

ATIS-0600015.2013, May 2013	Energy Efficiency for Telecommunication Equipment: Methodology for Measurement and Reporting – General Requirements
ATIS-0600015.09.2015, July 2015	Energy Efficiency for Telecommunication Equipment: Methodology for Measurement and Reporting of Base Station Metrics

ATIS-0600015.01.2014, November 2014	Energy Efficiency for Telecommunication Equipment: Methodology for Measurement and Reporting -- Server Requirements
ATIS-0600015.07.2013, May 2013	Energy Efficiency for Telecommunication Equipment: Methodology for Measurement and Reporting – Wireline Access, Asymmetric Broadband Equipment
ATIS-0600015.10.2015, July 2015	Energy Efficiency for Telecommunication Equipment: Methodology for Measurement and Reporting DC Power Plant – Inverter Requirements
ATIS-0600015.04.2017, December 2017	Energy Efficiency for Telecommunication Equipment: Methodology for Measurement and Reporting DC Power Plant – Rectifier Requirements
ATIS-0600015.11.2016, March 2016	Energy Efficiency for Telecommunication Equipment: Methodology for Measurement and Reporting DC/DC Converter Requirements
ATIS-0600015.05, April 2010	Energy Efficiency for Telecommunication Equipment: Methodology for Measurement and Reporting Facility Energy Efficiency
ATIS-0600015.03.2016, August 2016	Energy Efficiency for Telecommunication Equipment: Methodology for Measurement and Reporting for Router and Ethernet Switch Products
ATIS-0600015.08.2014, October 2014	Energy Efficiency for Telecommunication Equipment: Methodology for Measurement and Reporting for Small Network Equipment
ATIS-0600015.13.2017, February 2017	Energy Efficiency for Telecommunication Equipment: Methodology for Measurement and Reporting of 802.11xx Wi-Fi Access Points
ATIS-0600015.12.2016, June 2016	Energy Efficiency for Telecommunication Equipment: Methodology for Measurement and Reporting Power Systems – Uninterruptible Power Supply Requirements
ATIS-0600015.02.2016, March 2016	Energy Efficiency for Telecommunication Equipment: Methodology for Measurement & Reporting – Transport & Optical Access Requirements

Annex 2: Energy Model

At the most basic level, the model is based on the data demand multiplied by the energy efficiency of the DC/WAN measured in TWh/EB. It uses the same approach as previous models such as Andrae (2015) and updated with newer data where available.

The figure at the end of this annex shows the complete model structure from the inputs, calculations and to the outputs. Where a box describes a calculation as 'Hi-Lo' this indicates there are two separate values being calculated for the high utilisation period and low utilisation. The shaded blue area shows the inputs used which are connected to the calculations via blue arrows. The intermediate calculations are shown in the pink area. Due to the number of interlinking calculations there are numerous interconnecting and overlapping arrows and for visual clarity these have been coloured in red and grey. The energy consumption outputs are shown in the green area, connected to the inputs or calculations by green arrows.

Model Inputs

The energy model is built on the following inputs.

Data demand

This is the annual data traffic consumed through the network and in the data centre.

The Cisco sources provide a comprehensive and coherent projection which are updated annually and used for the time period covered by the reports. Additional projections were based primarily on extrapolation with Andrae (2015) modelling and adjusted based on personal communications. Cisco sources were also used to breakdown the data across different network and data centre types and supplemented with Andrae (2015) where this was not available. For next generation DC/WAN beyond 2021, conservative projections of growth were assumed. It is possible that data growth will be higher beyond 2021 and this would have a significant upward impact on the energy consumption projections.

Other sources show higher historic growth rates

WAN Sources: (Cisco, 2017) (Andrae & Edler, 2015) (Andrae, 2018-05-02)

DC Sources: (Cisco, 2018) (Andrae & Edler, 2015) (Andrae, 2018-05-02)

DC and Network efficiency

The average DC/WAN efficiency is the second critical input data for the model.

WAN Sources: (Andrae & Edler, 2015) (Andrae, 2018-05-02) (Krug, et al., 2014) (Nokia, 2016) (Malmodin & Lundén, 2018) (Aslan, et al., 2017) (ITU, 2014)

There is a wide range of efficiencies and therefore a large uncertainty in the efficiency. IEA provided upper and lower bounds for 2015 which were averaged to set the absolute values. Efficiency estimates

for other years were taken primarily from Aslan (2017) and Krug (2014) and the data fitted to a curve to produce the efficiency estimates. This was then adjusted for future projections based primarily on Andrae (2015, 2018).

DC Sources: (Andrae & Edler, 2015) (Andrae, 2018-05-02) (EC, 2018), (Naffziger & Koomey, 2016), original research

There was only one source identified, Andrae (2015) to provide an absolute value. This produced a relatively high energy consumption and was therefore normalised to agree with previously published IEA reports which generally agree with other reports from the US EPA and EC. Relative changes over time and between types of DC were then based on the additional sources (Naffziger & Koomey), (EC, 2018) and primary research and analysis of the server efficiency data from the ITI based on the latest server efficiency metrics.

The DC and WAN models differ in their treatment of the infrastructure energy consumption. For WAN the infrastructure is taken into account by the average efficiency. The DC model calculated the Infrastructure energy separately based on the PUE levels from EC (2018).

DC and Network average utilisation and high/low utilisation

The overall network utilisation is a measure of the amount of data traffic over the maximum possible traffic capacity. Different parts of the network have different utilisation levels.

WAN Sources: (Nokia, 2016) (Trinh, 2017) (Krug, et al., 2014) (ITU, 2017) (Lorincz, et al., 2012) (Debaillie & Desset, 2014)

The average Core WAN and FAN network utilisation is based on ITU reference utilisation levels that are used for the efficiency metrics (ITU, 2017), additional sources are used for the RAN include Nokia. It is assumed that there is no change in the average utilisation level, although Krug (2014) shows that average utilisation has fallen in the past and Cisco (2018) projects utilisation will fall.

For the diurnal utilisation pattern, there is strong agreement between the sources, although variations can be seen between residential and commercial areas, network types as well as rural environments. Nokia (2016) was the main source used to fit the model and was applied across all networks. From this, it is estimated that 16 hours are spent in high utilisation and 8 hours in low. For simplicity, no shoulder period is modelled between the transition from high to low and the utilisation is completely flat within each period.

DC Sources: (Cisco, 2018), (EC, 2018)

The average utilisation for traditional and cloud data centres is based on the number of virtualised servers per physical server estimated in Cisco (2018) using the methodology from EC (2018).

The diurnal utilisation pattern is assumed to be the same as the WAN.

Network proportion active

Because utilisation is significantly below 100% there is the possibility to shutdown parts of the network to intelligently save energy. Reducing the proportion of the network active assumes that the energy of the inactive portion is zero and the utilisation (and power) of the remaining network increases.

There are no sources so it is assumed that the network is 100% active for legacy and current generation DC/WAN at all times. Next generation core network and RAN is assumed to switch off 25% and 50% of the network during low utilisation periods.

Product proportionality

This describes how the power changes with utilisation for the DC/WAN equipment. It assumes there is linear relationship between utilisation and power. At 100% utilisation, 100% power is consumed and at idle it is some proportion of the maximum power. 0% would represent perfect scaling of power and performance while 100% means the power does not change at all.

WAN Sources: (Jalali, et al., 2016) (Yan, et al., 2016) (ITU, 2014)

DC Sources: original research, (BladeRoom, unknown)

Model Outputs

Total energy consumed

The energy consumed is the product of the network efficiency and the data consumed.

Low and high utilisation energy consumed

This is the energy consumed at the two utilisation levels. This is calculated from the total energy consumed and split based on the network utilisation and product proportionality.

Peak and additional energy consumed

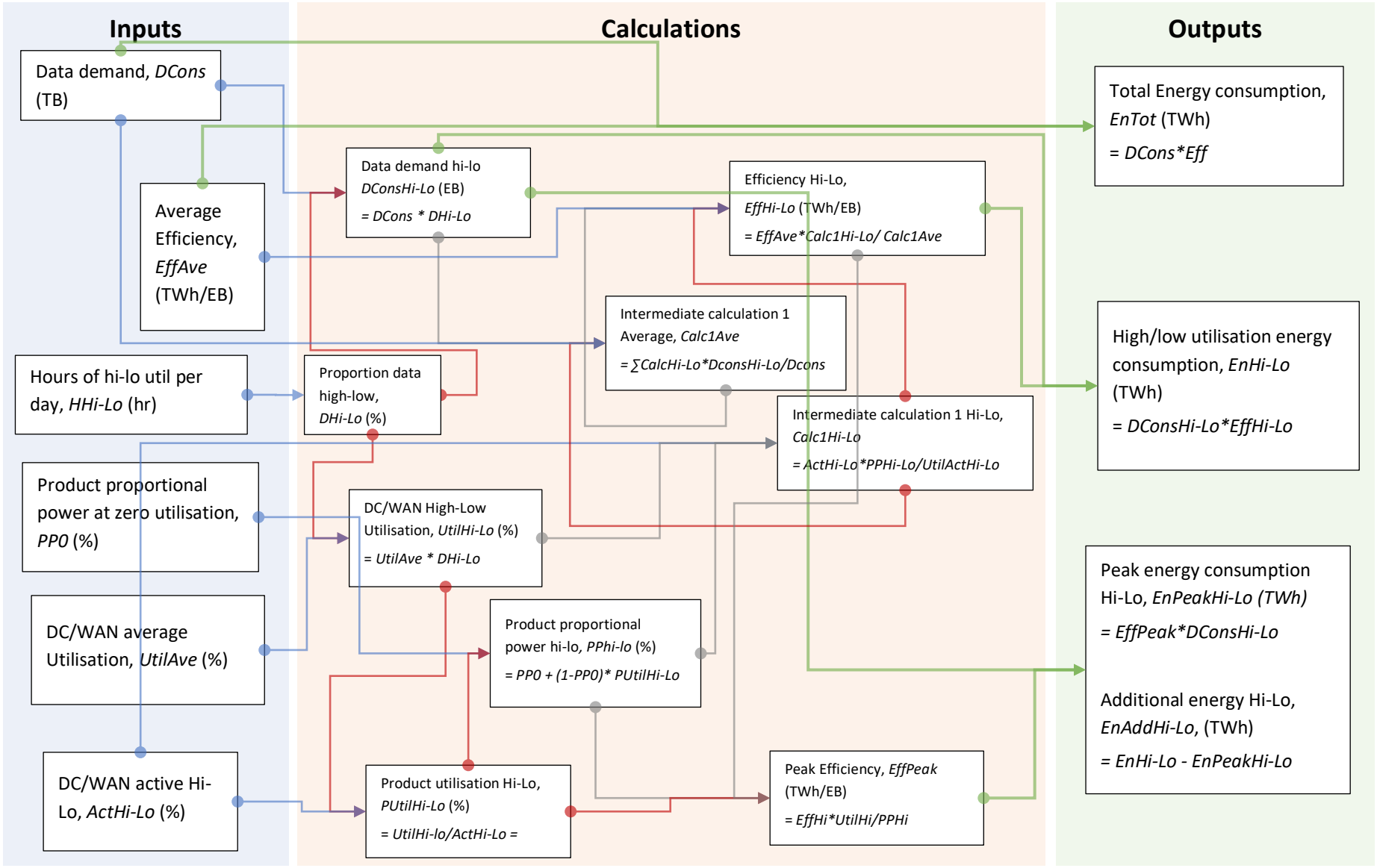
This is theoretical energy that would be consumed if the network could always operate at peak efficiency. The additional energy is the amount consumed to meet the calculated energy based on the calculated efficiency.

Breakdown

To model the evolution of technology over time as well as component parts of the WAN the model is broken down into:

- Core network
  - Legacy (backhaul and metro for PSTN)
  - Modern (Optical, MLPS)
  - Next generation (SDN)
- Radio access (mobile) network
  - 2G networks
  - 3G networks
  - 4G networks
  - Next generation (5G networks)
- Fixed access network
  - Legacy (dial up, ADSL)
  - Modern (cable, VDSL, FTTx)
  - Next generation (XGFTTx)

All the inputs and outputs can be defined in these components.

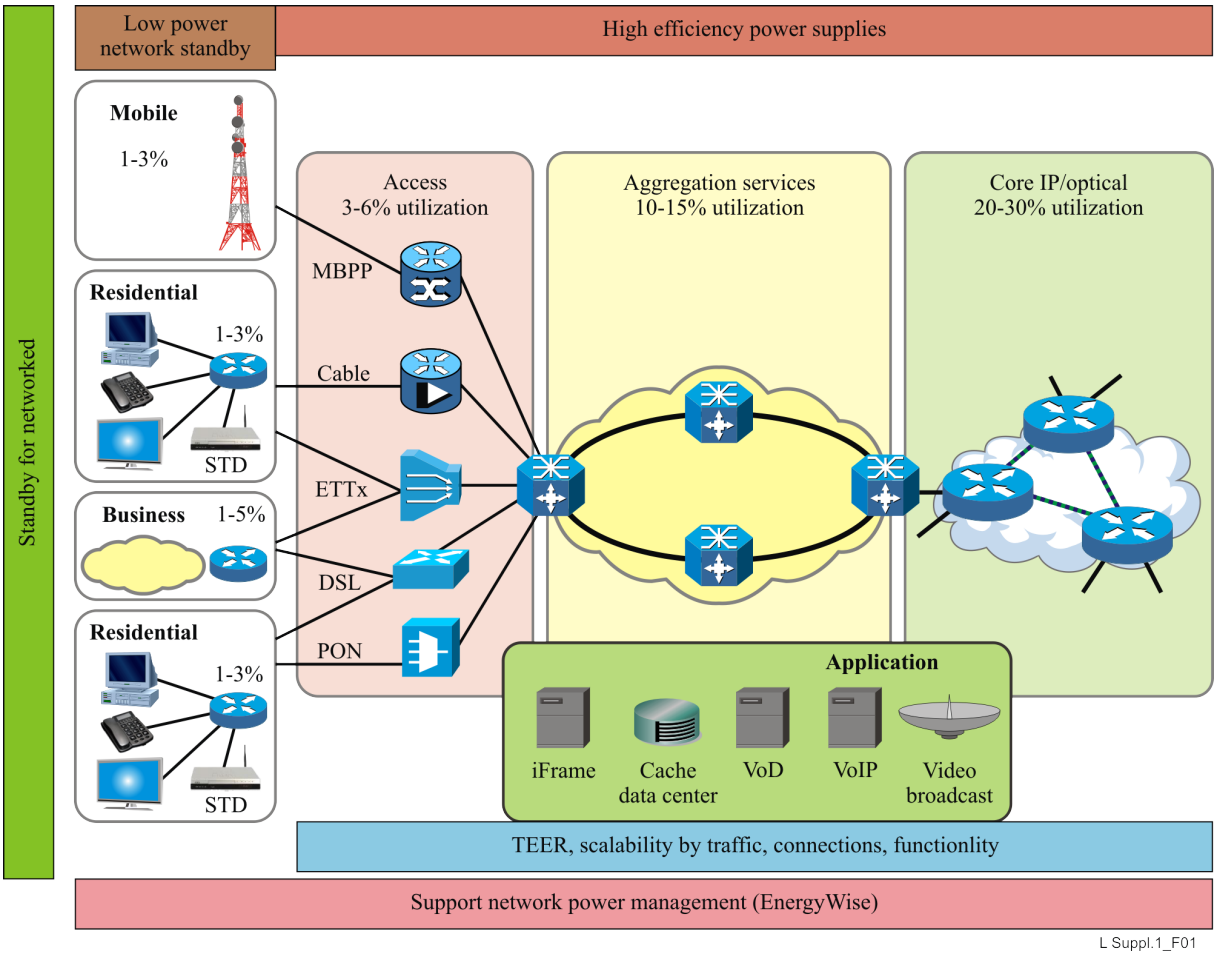

Inputs
Calculations
Outputs
Data demand, DCons (TB)
Average Efficiency, EffAve (TWh/EB)
Hours of hi-lo util per day, HHi-Lo (hr)
Product proportional power at zero utilisation, PP0 (%)
DC/WAN average Utilisation, UtilAve (%)
DC/WAN active Hi-Lo, ActHi-Lo (%)
Data demand hi-lo DConsHi-Lo (EB)
= DCons * DHi-Lo
Proportion data high-low, DHi-Lo (%)
Efficiency Hi-Lo, EffHi-Lo (TWh/EB)
= EffAve*Calc1Hi-Lo/ Calc1Ave
Intermediate calculation 1 Average, Calc1Ave
= ∑CalcHi-Lo*DconsHi-Lo/Dcons
Intermediate calculation 1 Hi-Lo, Calc1Hi-Lo
= ActHi-Lo*PPHi-Lo/UtilActHi-Lo
DC/WAN High-Low Utilisation, UtilHi-Lo (%)
= UtilAve * DHi-Lo
Product proportional power hi-lo, PPhi-lo (%)
= PP0 + (1-PP0)* PUtilHi-Lo
Product utilisation Hi-Lo, PUtilHi-Lo (%)
= UtilHi-lo/ActHi-Lo =
Peak Efficiency, EffPeak (TWh/EB)
= EffHi*UtilHi/PPHi
Total Energy consumption, EnTot (TWh)
= DCons*Eff
High/low utilisation energy consumption, EnHi-Lo (TWh)
= DConsHi-Lo*EffHi-Lo
Peak energy consumption Hi-Lo, EnPeakHi-Lo (TWh)
= EffPeak*DConsHi-Lo
Additional energy Hi-Lo, EnAddHi-Lo, (TWh)
= EnHi-Lo - EnPeakHi-Lo

Annex 3: Networks and virtualisation

ITU network model

For the purposes of this report it is very useful, if not essential, to have an energy efficiency focussed network model showing the key elements: the core, aggregation, access and edge networks. The ITU-T Study Group 5 Recommendation L.1310 – Supplement 1, (ITU-T, 2013), included such a model. The ITU-T energy efficiency network model is shown in Figure 20. Along with the four key network elements, shown is the relative traffic utilization within each element. For example, the network core consisting mostly of data centres connected by optical fibre networks entail an estimated 20 – 30% of the traffic. The ITU-T network utilization figures shown in Figure 20 are indicative of the actuals at the time of publication, 2013, which are likely to have altered since then. However, the trend in falling network utilization with movement towards the edge networks while the current Internet network architecture remains.

Figure 20 The ITU-T network model.



Source: International Telecommunications Union (ITU)

For explanatory purposes, the Access technology terms used in Figure 20 are:

- MBPP — Mobile Backhauling Point to Point
- Cable — Cable TV system, DVB-C/C2 or similar, with DOCSIS or similar
- ETTx — Ethernet to the Home or Business
- DSL — Digital Subscriber Line, ADSL or VDSL.
- PON — Passive Optical Network

While TEER (Telecommunications Energy Efficiency Ratio) is an ITU-T energy efficiency metric defined as the ratio between the total data rate and the power consumption.

Network virtualization

The advent of network virtualization around 2012 marked an inflection point in the development of ICT technology. In general, networks have been built using proprietary hardware running proprietary software. At the heart of Network Function Virtualisation (NFV) is the divorce of software and hardware, the software will be standards based and operate at key layers in the OSI model and the hardware will be commercial off the shelf (COTS) systems. It is the software which manages how and where data is sent over the network and attempts to ensure the data arrives at the intended destination. NFV provides the potential for network hardware to adapt to changing network circumstances. For example, when data traffic demands it, in a virtual network additional resources can be allocated to meet the demand. Then as demand falls those resources may be put into a sleep mode or allocated to another service. NFV is often used in Data Centres to provide for more efficient use of hardware when the demand permits. Rather than supporting one service at a time, a server and network function infrastructure support additional services.

NFV is able to play a major role in improving energy efficiency because potentially it could be used to reduce the amount of network hardware. However, it is conceivable that some NFV applications could equally result in an increase in energy consumption of future networks. In developing energy policy relating to Energy Aware Devices, it is therefore of considerable significance for policies to take account of impact NFV may have on network architecture. Energy efficiency policies will need to adapt to a the rapidly developing field of network architecture design.

In regard to NFV, the European Telecommunications Standards Institute (ETSI) is an organization amongst the leaders developing NFV standards. Its NFV Industry Specification Group (ISG) has developed a large number of NFV standards. Under Mandate M/462, (European Commission, 2010), the European Commission has requested ETSI, the European Committee for Standardization (CEN) and the European Committee for Electrotechnical Standardization (CENELEC), develop standards to "*enable efficient energy use in fixed and mobile information and communication networks*".

In 2014, ETSI published a standard, ES 203 237, on power management capabilities of future energy-aware telecommunication fixed network nodes using the ECONET project proposed Green Abstraction Layer (GAL) software application specifying power management interfaces in network equipment. While the ETSI GAL standard is very much device centric, the ITU-T Study Group 5 is planning to adapt the standard by creating GALv2 integrating it into an NFV network.

Another ITU-T study group, Study Group 13, is working on international standards for the next-generation of networks, (ITU-T, 2018).

Further opportunities for improving energy efficiency lies in the way networks are managed, rather than simply the hardware upon which it runs, by software related to a specific OSI layer. For example, (Bianzino, Chaudet, Rossi, & Rougier, 2012) describe, based on the original (Gupta & Singh, 2003)

proposal, an energy-aware routing scheme which provides for routers in one link could be put into Low Power Idle mode while traffic is directed over another link.

The energy model can be used to assess the potential energy savings that NFV and similar technologies can bring to networks. Similarly, though, the application of NFV and related technologies redefine the network architecture, so will the Total Energy Model need to reflect these changes.

References

Andrae, A. (2018-05-02). Personal Communication. NA: NA.

Andrae, A., & Edler, T. (2015). On Global Electricity Usage of Communication Technology: Trends to 2030. *Challenges*(6), 117-157. doi:10.3390/challe6010117

ASHRAE. (2016). *Data Center Power Equipment Thermal Guidelines and Best Practices.*

Aslan, J., Mayers, K., Koomey, J., & France, C. (2017). Electricity intensity of internet data transmission. *Journal of Industrial Ecology*. doi:10.1111/jiec.12630

Baccarelli, E., Vineuza Naranjo, P. S., Shojafar, M., & Abawajy, J. (2017). Fog of Everything: Energy-efficient networked computing architectures, reasearch, challenges, and a case study. *IEEE Access, 5*, 2169-3536. doi:10.1109/ACCESS.2017.2702013

Berglund, A. (2017). *How Data Science Helps Power Worldwide Delivery of Netflix Content*. Retrieved 07 02, 2018, from https://medium.com/netflix-techblog/how-data-science-helps-power-worldwide-delivery-of-netflix-content-bac55800f9a7

Bianzino, A. P., Chaudet, C., Rossi, D., & Rougier, J.-L. (2012). A Survey of Green Networking Research. *IEEE COMMUNICATIONS SURVEYS & TUTORIALS, 14*(1), 3-20.

BladeRoom. (NA). *Data center efficiency at any load*. Retrieved 03 07, 2018, from http://www.bladeroomus.com/data-center-efficiency-at-any-load.php

Cisco. (2017). *Cisco Visual Networking Index: Forecast and methodology, 2016-2021.* Retrieved 04 19, 2018, from https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/complete-white-paper-c11-481360.html

Cisco. (2018). *Cisco Global Cloud Index: Forecast and Methodology, 2016–2021.* Cisco. Retrieved 02 15, 2018, from https://www.cisco.com/c/en/us/solutions/collateral/service-provider/global-cloud-index-gci/white-paper-c11-738085.html

Cisco. (2018). *Cisco Visual Networking Index: Forecast and Trends, 2017-2022.*

Debaillie, B., & Desset, C. (2014). Power modeling of base stations. *5GrEEn Summerschool* . Stockhom: imec.

Densify. (2018). *Automated optimization for on-premise virtual infrastructure & bare metal clouds.* Retrieved 07 02, 2018, from https://www.densify.com/wp-content/uploads/densify-datasheet-automated-on-premises-virtual-infrastructure-bare-metal-cloud-optimization.pdf?cb=20180201

Dutta, S., Roy, D., Bhar, C., & Das, G. (2018). Online scheduling protocol for energy-efficient TWDM-OLT design. *IEEE/OSA Journal of Optical Communications and Networking, 10*(3), 260 - 271. doi:10.1364/JOCN.10.000260

EC. (2018). *Impact Assessment accompanying Implementing Directive 2009/125/EC of the European Parliament and of the Council.*

EDNA. (2019a). *Total Energy Model of Connected Devices.* Paris: IEA, prepared by EnergyConsult Pty Ltd.

European Commission. (2010, April 30). M/462. Standardisation mandate addressed to CEN, CENELEC and ETSI in the field of ICT to enable efficient energy use in fixed and mobile information and communication networks. Brussels, Belgium: European Commission.

Fernández-Fernández, A., Cervelló-Pastor, C., & Ochoa-Aday, L. (2017). Energy efficiency and network performance: a reality check in SDN-based 5G systems. *Energies, 10*(12). doi:10.3390/en10122132

Gao, J. (2017). *Intelligent Energy - How AI can have a positive impact.*

Gupta, M., & Singh, S. (2003). Greening of the Internet. *Proc. ACM SIGCOMM'03*, (pp. 19-26). Karsruhe, Germany.

Hong, D., Ma, Y., Banerjee, S., & Morley Mao, Z. (2016). Incremental deployment of SDN in hybrid enterprise and ISP networks. *SOSR '16 Proceedings of the Symposium on SDN Research.* Santa Clara, USA: ACM. doi:10.1145/2890955.2890959

IEA. (2017). *Digitalization and Energy.* Paris: International Energy Agency.

ITU. (2014). *L.1340 Informative values on the energy efficiency of telecommunication equipment.* ITU.

ITU. (2017). *L. 1310 Energy efficiency metrics and measurement methods for telecommunication equipment.* ITU.

ITU-T. (2013). *ITU-T L.1310 – Supplement on energy efficiency for telecommunication equipment.* International Telecommunications Union, Telecommunications Sector. Geneva: International Telecommunications Union.

ITU-T. (2018). *ITU-T SG13: Future networks including cloud computing, mobile and next-generation networks (2013-2016)*. Retrieved 2018, from International Telecommunications Union: https://www.itu.int/en/ITU-T/studygroups/2013-2016/13/Pages/default.aspx

Jalali, F., Hinton, K., Ayre, R., Alpcan, T., & Tucker, R. (2016). Fog computing may help to save energy in cloud computing. *IEEE Journal on Selected Areas in Communications, 34*(5), 1728-1739. doi:10.1109/JSAC.2016.2545559

JDCC. (2016). *Japan Data Center Council Guidelines*. Retrieved 12 11, 2018, from http://www.jdcc.or.jp/english/guidelines.html

JEITA. (2012). *Enhancing the energy efficiency and use of green energy data centers.*

JEITA. (2014). *Harmonizing Global Metrics for Data Center Energy Efficiency.*

JEITA. (2015). *Effective Action on Global Warming Prevention by the Japan's Electrical and Electronics Industries.* Retrieved 12 11, 2018, from https://home.jeita.or.jp/eps/pdf/GlobalWarmingPrevention_2016English.pdf

Kharatinov, D. (2012). *Green Telecom Metrics in Perspective.* Retrieved from https://arxiv.org/abs/1208.0577

Krug, L., Shackleton, M., & Saffre, F. (2014). Understanding the Environmental Costs of Fixed Line Networking. *e-Energy, June 11-13*. doi:http://dx.doi.org/10.1145/2602044.2602057

Lambert, S., Lnnoo, B., Dixit, A., Cole, D., Pickavert, M., Montalvo, J., . . . Vetter, P. (2014). Energy efficiency analysis of high speed triple-play services in next-generation PON deployments. *Computer Networks*.

Li, Y., Wen, Y., Guan, K., & Tao, D. (2018, 05 24). *Transforming cooling optimization for green data center via deep reinforcement learning.* Retrieved 07 02, 2018, from https://arxiv.org/pdf/1709.05077v2.pdf

Lorincz, J., Garma, T., & Petrovic, G. (2012). Measurement and modelling of base station power consumption under real traffic loads. *Sensors*(12), 4281-4310. doi:10.3390/s120404281

Malmodin, J., & Lundén, D. (2018). *The electricity consumption and operational carbon emissions of ICT network operators 2010-2015.* Stockholm, Sweden: KTH Centre for Sustainable Communications.

Naffziger, S., & Koomey, J. (2016, 11 28). Energy efficiency of computing: what's next? *Electronic Design*. Retrieved 05 01, 2018, from http://www.electronicdesign.com/microprocessors/energy-efficiency-computing-what-s-next

NGMN Alliance. (2015). *5G White Paper.* Retrieved from https://www.ngmn.org/5g-white-paper/5g-white-paper.html

Nokia. (2016). *5G network energy efficiency White Paper.* Online. Retrieved from https://pages.nokia.com/2396.5G_Network_Energy_Efficiency.html

Pakpahan, A., Hwang, I., & Nikoukar, A. (2017). OLT energy savings via software-defined dynamic resource provisioning in TWDM-PONs. *Journal of Optical Communications and Networking, 9*(11), 1019-1029. doi:doi.org/10.1364/JOCN.9.001019

Ruiu, P., Scionti, A., Nider, J., & Rapoport, M. (2016). Workload Management for Power Efficiency in Heterogeneous Data Centers. *2016 10th International Conference on Complex, Intelligent, and Software Intensive Systems (CISIS).* IEEE. doi:10.1109/CISIS.2016.107

Sheng, M., Zhai, D., Wang, X., Li, Y., Shi, Y., & Li, J. (2015). Intelligent energy and traffic corrdiantion for green cellular networks with hybrid energy supplies. *IEEE Transactions on Vehicular Technology*.

Trinh, H. e. (2017). Analysis and Modeling of Mobile Traffic Using Real Traces. *2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC).* Montreal, Canada: IEEE. doi:10.1109/PIMRC.2017.8292200

Xu, G., Mu, Y., & Liu, J. (2017, Oct 13). Inclusion of Artificial Intelligence in communication networks and services. *ICT Discoveries*(Special Issue No 1). Retrieved from https://www.itu.int/en/journal/001/Pages/04.aspx

Yan, M., CA, C., Li, W., Lin, C., Bian, S., Gygax, A., . . . Nirmalathas, A. (2016). Network energy consumption assessment of conventional mobile services and over-the-top instant messaging applications. *IEEE Journal on Selected Areas in Communications, 34*(12), 3168-3180. doi:10.1109/JSAC.2016.2611978

Zhang, K., Mao, Y., Leng, S., Zhao, Q., Li, L., Peng, X., . . . Zhang, Y. (2016). Energy-efficiency offloading for mobile edge computing in 5G heterogeneous networks. *IEEE Access, 4*, 5896 - 5907. doi:10.1109/ACCESS.2016.2597169