Data centres provide a boost to companies' energy efficiency efforts

Amid the ongoing energy transition, the infrastructure for green electricity supply is under pressure to meet growing demand. This poses a challenge for fast-growing sectors, such as Information and Communications Technology. Fortunately, the data centre sector has avenues to improve efficiency

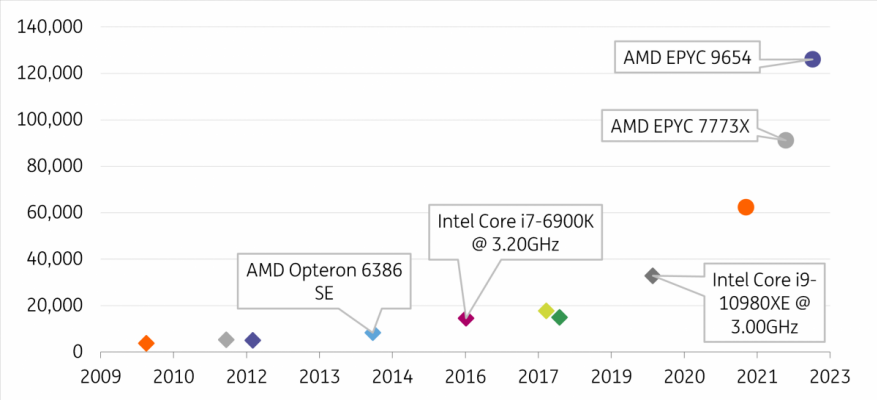
As the focus on companies' environmental footprint increases alongside the digitisation of our economy, the most sustainable data centres will be able to gain a competitive edge

Computing power grows at an exponential rate

In the past decade, the data centre industry has witnessed a decoupling between the growth in compute instances (explained below) and the energy use of data centres. In this article, we want to highlight three main developments:

1. The spectacular increase in computing capacity;
2. The role played by data centres in increasing the proportion of energy use for computing functions. That is to say, a relatively larger share of energy use is consumed by the primary activity of a data centre;
3. An increase in the use of energy by data centres, which is far smaller than one would expect based on the computational output.

Indicative processor speed developments (CPU Mark)



Source: cpubenchmark.net, ING

A compute instance can be seen as a virtual machine, a server resource provided by a third-party cloud service. Until 2020, the number of compute instances increased roughly eightfold, from 100mn to 800mn, according to Masanet et al. This is an impressive increase and can be explained by the rise in the number of servers and the continued evolution in hypervisor-based applications, as well as by the increase in computing power over time, which allows for more virtual machines to be run on servers. An indication of the development in server capacity is shown in the graph above. Server sales have also gone up considerably, according to IDC data. In 2022, 13.8m servers (x86 architecture) were sold, which is up from 11.8mn in 2018.

The servers themselves, which are installed in data centres, have also become more efficient over time. Microchips have become more energy efficient due to innovations in microchip design and manufacturing capabilities. As a result, data centres have been able to perform many more calculations, while keeping the relative power consumption in the DC in check. Also, microchips and software can be designed for specific applications, which can enhance efficiency.

In their 2010 research paper, Koomey et al evaluate computational power efficiency. According to Koomey's Law, the number of computations per joule of dissipated energy doubled every 1.57 years from 1945 to 2000 (100x per decade). After 2000, the efficiency advance rate slowed down but still doubled approximately every 2.6 years. In addition to technological advances in the domain of semiconductors, there is also scope for other efficiencies, such as more efficient software, compression applications, Hadoop and the optimal use of servers.

Masanet et al (2020)

Koomey et al (2010)

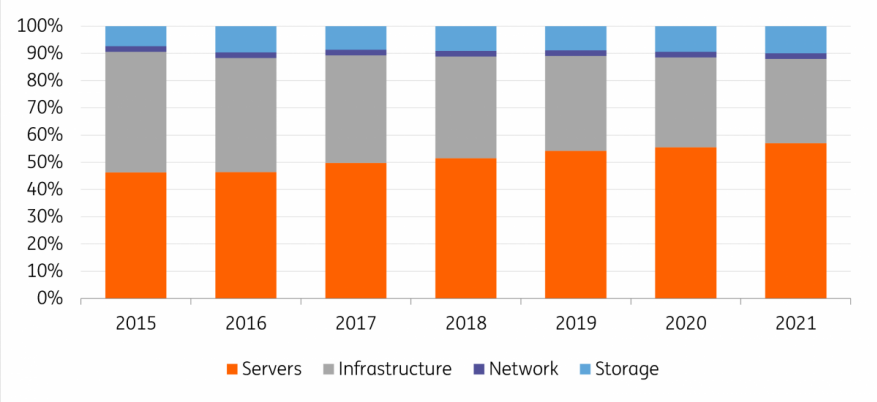
More energy allocated for use by servers (computing)

While the computational efficiency has gone up over time, so too has the efficiency of data centres. We believe it is important to look not just at the total energy consumption of data centres, but also at the way this energy is used. Today, servers, performing the compute function, take a proportionally larger share of the total electricity consumption of data centres, while the share of other components (non-computing) has gone down. We will elaborate on this below, but the main

takeaway is that this change makes data centres more efficient. This gain comes on top of the potential efficiency gains of the servers themselves (as described above).

The three main functions of data centres (storage, distribution, and processing) all require electricity. Yet not all three functions do so in equal parts. **Storage** (saving information on a storage device) is not the most energy-intensive function and has become more efficient over time. Also, the **network function** (or communication function) has seen significant efficiency gains over time so this is not the most energy-intensive in a data centre. Despite the intuitive expectation that an increase in traffic implies more energy use, existing fibre networks often have substantial spare capacity and can often be upgraded to faster standards with a disproportionate increase in energy use. Today, the most energy-intensive data centre function is **processing or computation**. The servers that perform calculations require energy, and energy use increases with the server's capacity. It is on the servers that the actual software runs (from office and banking to social media and AI applications). Finally, the (physical) **infrastructure** of a data centre has to be powered. This function includes the energy costs of cooling the data centre. Notably, the energy demand of the servers as a percentage of total data centre energy demand has increased from 46% in 2015 to 57% in 2021.

Larger proportion energy take-off used directly for servers

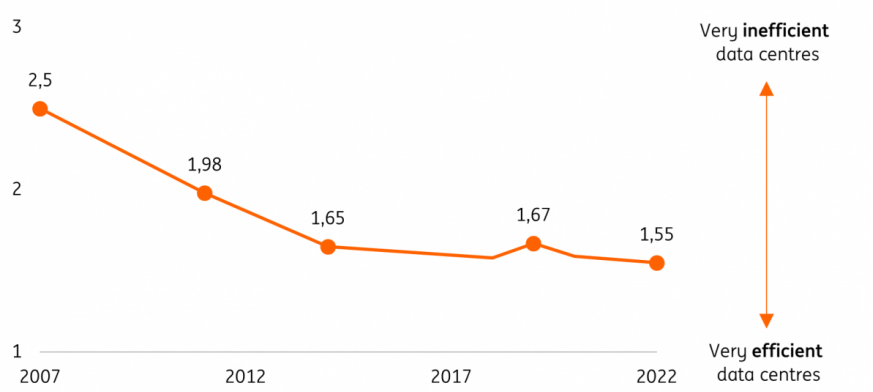


Source: IEA & Masanet et al. (2020)

Efficiency of data centres has improved substantially over the past decade

A frequently used measure to gauge the enhanced energy efficiency of data centres is power usage effectiveness (PUE). PUE compares the amount allocated to non-computing uses (such as cooling) to the amount of energy used to power IT equipment (such as compute and storage functions). As a result of technological progress PUE has gone down, from an average of 2.5 in 2007 to 1.55 in 2022. Large data centre builders (such as Equinix, NTT and Digital Realty) have reported that the PUE of new data centres hovers between 1.2 and 1.1 whereas average PUEs are between 1.6 and 1.4. Google's newest data centres are reported to have PUEs below 1.1.

Average annual power usage effectiveness (PUE) at the largest data centre of operators worldwide



Source: Uptime Institute; A PUE of 2 means that for every watt of IT power, another watt is consumed for cooling and other non-IT equipment. A PUE of 1.0 means that all energy is used for computing.

While PUE has become a standardised metric in measuring data centre energy efficiency, there has been some debate over whether it is effective. Now that PUEs are approaching 1.0 among some companies, the extent to which this metric can measure future efficiency gains becomes questionable.

Moreover, PUE does not measure energy consumption at the rack level, which indicates that the ratio is not as comprehensive. PUE can be impacted by (i) region, (ii) utilisation levels of the DC, (iii) the primary function of a DC – storage, communications, how companies are using their IT infrastructure and (iv) type of DC, as retail colo or hyperscalers differ in terms of infrastructure control and standardisation potential. That said, PUE is still a widely accepted and implementable benchmark that allows companies and investors to make comparisons.

Energy use statistics do not include the switch-off of enterprise data centres

Across the economy, there is also scope for further efficiency gains. The implementation of IT services is moving away from on-premise servers to off-premise data centres, as well as from smaller enterprise data centres to large and (or) efficient data centres, such as hyperscalers.

The proportion of servers housed in efficient hyperscalers went up from 7.7% in 2010 to 38% in 2018 (Masanet et al, 2020). A study commissioned by Microsoft (updated in 2020) finds that companies could potentially save 22-93% of energy when they run applications in the cloud. An earlier, 2013 study, commissioned by Google, found a potential efficiency gain of 87%. However, often, only the energy use of the data centres is measured, not its offsetting impact of obsolete corporate data centres or the energy requirements associated with the actual data transmission. This is shown in the example below.

Example of increased efficiency

An example of the impact on energy use from server migration can be found in the Netherlands. The percentage of the national electricity use of data centres has grown from 1.5% in 2017

(1,65TWh) to 3.3% in 2021 (3,73TWh). However, this increase in electricity is measured because the off-premise data centres now consume more energy, while the declining use of the legacy (on-premise) is not reflected in the figures on data centre energy use. Hence, when institutions move their servers to an off-premise facility, energy use of data centres rises on paper but there may actually be a net saving of energy. This was the case in the Netherlands when the Dutch government moved its servers from 60 on-premises locations to five large off-premise ones, and the energy use of servers by the Dutch government went from 235GWh to 128GWh.

The main challenge for researchers investigating efficiency gains in the data centre sector is that there is sometimes only data on the energy use of data centres within a country, while the decreasing use at the site of companies is not explicitly measured. There is hardly any energy consumption data measuring the uses of servers that are switched off within companies. This likely overstates the increase in electricity demand by data centres, also because data centre services in a country can replace servers in other countries. Finally, there are projects through which data centres reuse waste heat for heating residential and commercial areas, which complicates the measurement of net energy use further.

Data centres consume more energy, but by less than what could be expected

It is hard to find a precise measure of the energy use of data centres. The IEA estimates that the global use of electricity by data centres has gone up by 20%-70% from 2015-2022 to 240-340TWh (excluding the electricity use associated with crypto, roughly 110TWh in 2022) while the total electricity use by Amazon, Microsoft, Google, and Meta more than doubled to 72TWh from 2017 to 2021. According to Masanet et al (2020), the increase in electricity use has been more moderate from 2010-2018, increasing by 6% to 205TWh.

They estimate that the impact of efficiency measures has been tremendous (20% per annum, expressed as energy use per compute instance). The EU has also evaluated energy use and found that “the energy consumption of data centres in the EU28 increased from 53.9 TWh/a to 76.8 TWh/a between 2010 and 2018. This means that in 2018, data centres accounted for 2.7% of the electricity demand in the EU”.

The way we look at the figures above is that the increase in energy use matches the ICT sector GDP growth, of c.6% per annum. This number also matches the increase in annual server sales. Note that these increases are smaller than the increase in computing power, which has risen over time and reflects the substantial efficiency gains made.

How do we look at future energy demand from servers and data centres without strong AI growth?

With the arrival of new technologies such as generative AI and the Internet of Things (IoT), the demand for the computing potential of data centres will likely grow further. Therefore, society needs an increase in computing power. Part of this demand can be met through efficiency gains at the server level, but we also expect that the number of servers will increase as a result of these trends. When looking at the future electricity demand, this is largely driven by the increase in the number of servers. Let’s assume that the number of servers increases at a high single-digit annual growth rate in the medium term. This should (on average) also provide an estimation for the growing electricity demand of the ICT sector. Note that the growing availability of servers implies

an even higher increase in workloads due to faster servers, which is beneficial – but it will also cost more power.

Moreover, the efficiency of the new servers is an important element in determining total electricity demand. We do not have a forecast for this, because it could go both ways. Servers could become more efficient, lowering energy consumption. Higher workloads, however, could also be responsible for increasing power consumption.

How does AI change the picture?

Recent developments in generative AI, such as the launch of GPT-4 and its uptake by the public, have caused an increase in demand for the computing power of data centres. This also means an increase in the electricity use of data centres, as 'compute' is an energy-intensive function. The extent of the increase of generative AI use remains uncertain. Yet, De Vries (2023) worked out multiple scenarios based on NVIDIA's AI servers, as INVIDIA has a 95% market share. The 100,000 servers NVIDIA delivers in 2023 will consume between 5.7 and 8.9 TWh of electricity, which is a large amount of energy but is not very significant given the overall demand of the sector. In addition, challenges in the manufacturing of advanced semiconductors may prevent fast adoption. However, NVIDIA may be able to deliver many more AI server units in the future. According to De Vries (2023), these servers could annually require as much as 134 TWh of electricity in 2027, an extreme scenario. This would of course be a large amount of electricity.

It is important to note, however, that these calculations hinge on a variety of factors that are, as of yet, uncertain, such as the ability to have profitable business models at that scale, and also the absence of further efficiency enhancements (in both hardware and software). Moreover, in our base case, we already assume a mid- to high-single-digit increase in annual server growth. Accelerated growth in generative AI would increase server growth further, but our base case can already account for regular growth in AI.

In short, the expected growth of the data centre sector is a consequence of the increasing digitisation of our economy. However, since many companies must report the environmental footprint from their procured services, environmental impact measures of data centres, as described in sustainable finance frameworks, will become more important over time. This will give the most sustainable data centres a competitive edge. We discuss this in the article about data centres and sustainable finance.

Authors

Jan Frederik Slijkerman
Senior Sector Strategist, TMT
jan.frederik.slijkerman@ing.com

Diederik Stadig
Sector Economist, TMT & Healthcare
diederik.stadig@ing.com

Coco Zhang
ESG Research
coco.zhang@ing.com

Disclaimer