The race to profit from tech's AI revolution

In the following articles, we discuss the important role of supercomputers, proprietary data and the technology race to build the leading AI platform. We expect the ICT sector to show strong growth, which countries must embrace. Over time, prices for AI tools will come down

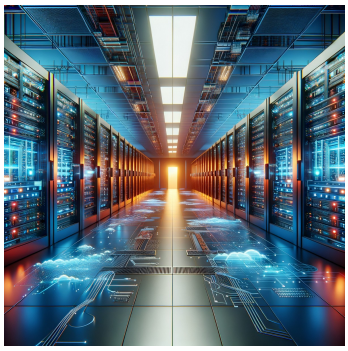
In this bundle

TMT

AI revolution driven by new supercomputers

The AI revolution is driven by spectacular increases in computing speeds. Semiconductors are therefore an important element in the AI value chain

By Jan Frederik Slijkerman



TMT

Data is the new gold for AI development

We expect that the owners of private data will build proprietary solutions, opening a new source of revenue

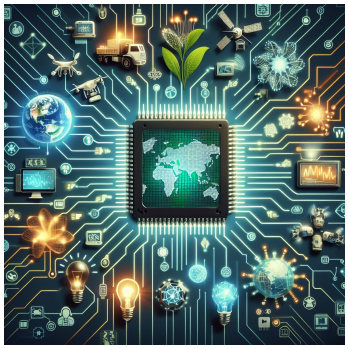
By Jan Frederik Slijkerman

TMT

AI frontrunners will benefit most, with Microsoft in the lead

We expect many sector-specific AI solutions to emerge. Given economies of scale, early leaders like Microsoft will likely benefit most

By Jan Frederik Slijkerman

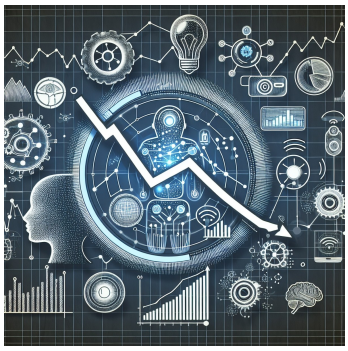


TMT

Countries must embrace AI or risk missing out on strong ICT growth

We think data centre capacity will grow by 10% a year, facilitating growth of the market for AI services

By Jan Frederik Slijkerman



TMT

AI services will become more affordable over time

We expect prices of AI services to come down, in line with historical developments of ICT hardware prices

By Jan Frederik Slijkerman

Article | 24 April 2024 TMT

AI revolution driven by new supercomputers

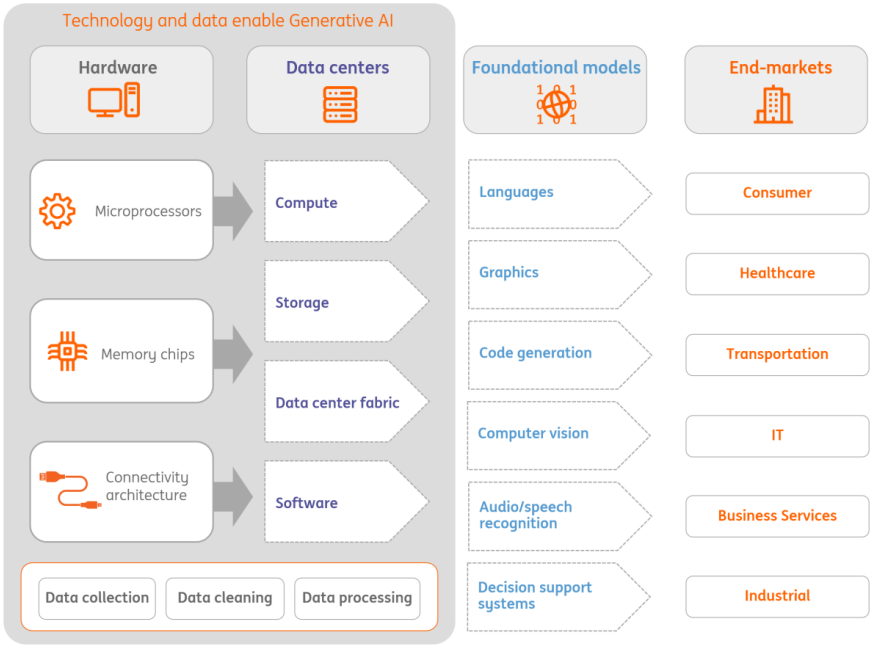
The AI revolution is driven by spectacular increases in computing speeds. Semiconductors are therefore an important element in the AI value chain, along with other innovations applied in supercomputers. Nvidia's market-leading position will not be matched by competitors in the near term, in our view

The current generative AI boom is driven by a spectacular increase in the capacity of microchips. Today, a web of servers can perform calculations on these extremely large datasets. And this is just the beginning. In this article, we discuss the spectacular innovations giving rise to Nvidia as a leading player in the technology domain on multiple fronts. In a related article, we discuss how content plays an important role in these developments.

The need for very large data centres

What is driving the emerging Generative AI sector? Recently, we have witnessed a strong increase in compute power, driven by fast server microchips (central processing unit) and novel computation accelerator cards, with fast processors for calculation purposes: GPUs. The most advanced AI models, such as large language models (LLMs) need thousands of GPUs for the computation of all parameters (the pre-training phase of the models). The GPUs are linked to servers which need to be connected to the other servers in the network. This requires a state-of-the-art data centre fabric, a term describing the data centre architecture of cables, switches and software. It is very important to have the most efficient data centre design as well as optimised software. Efficiency is key to reducing the computational time needed to calculate the AI models. AWS, Microsoft, Alibaba and Alphabet run their own very large data centres, called hyperscale data centres, something which is difficult to replicate for competitors. This is a key competitive benefit for these companies. Interestingly, Nvidia also wants to expand into the data centre market and is cooperating with Amazon Web Services and Equinix.

Semiconductors are an important element of the AI value chain



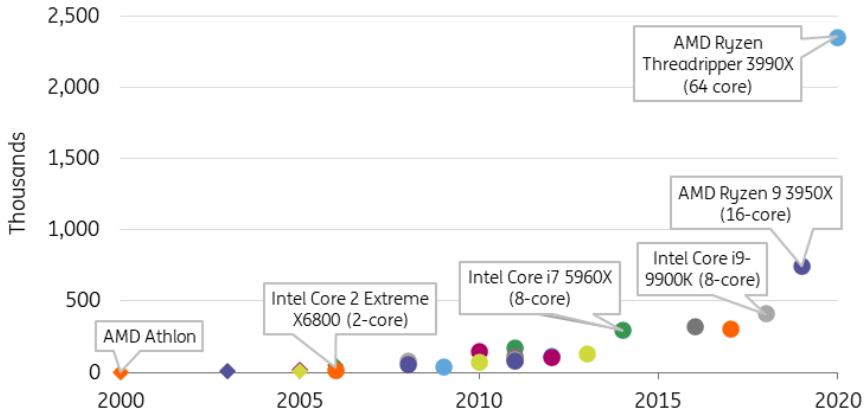
Source: ING

Supercomputers are a key enabler of AI

A typical AI server is built around a central processing unit, memory chips, and a processor designed for fast calculations. In all these areas, there have been many innovations, leading to faster speeds and more computation power. Also, because servers can work together to perform a joint analysis of the extremely large data sets. Notable developments in this field are the Nvidia and, more recently, AMD accelerator cards (GPUs), which include extremely fast processors. These have provided the necessary computer power to be able to calculate the latest AI models. A rule of thumb for technological innovation in the semiconductor industry is Moore's Law, which states that the number of transistors on a microchip (providing computing power) doubles every two years. We therefore expect that further advances in computing speed are on the horizon,

enabling more complex models. More efficient microchips could also lower the power consumption of the existing processing power: promising developments.

Indicative processor speed developments (PCU Mark)



Source: cpubenchmark.net, ING

Today, Nvidia is the leading provider of AI semiconductors. The current, leading configuration for an AI server with accelerator cards is an HGX100 system, which has two server microchips (CPUs) and eight Nvidia GPUs, the H100. Recently, Nvidia announced the launch of a new solution the DGX B200, which is expected to be around three times faster than its predecessor. An interesting feature of the Nvidia product line-up is that individual systems can be combined into a supercomputer, as shown in the table below. Through combining multiple, so called, basepods, organisations can build their own supercomputer.

Scalable solutions from Nvidia build supercomputer

	Solution	Speed
Basepod	DGX A100 DGX H100 DGX B200	10 petaflops 32 petaflops 3x32 petaflops
Hopper superpod	32 x DGX H100 (256 x H100 GPUs)	1 exaflops
Eos supercomputer	18 x H100 SuperPods (576 x DGX H100)	18 exaflops

An exaflop equals 1000 petaflops; speeds based on FP8 precision

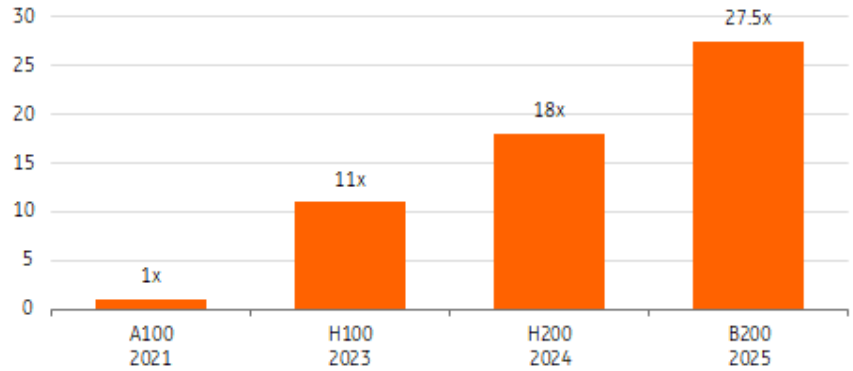
Source: Nvidia, ING

AMD has launched a product that is intended to compete with Nvidia, the Instinct MI300X GPU, which is particularly suited to the training of LLMs. Besides the traditional GPU designers, the large cloud operators are also venturing into this field. Microsoft, Alphabet and AWS are all developing their own microchips. Microsoft is working on the Azure Maia 100 AI Accelerator; Alphabet has a series of AI accelerators, called Tensor Processing Units, of which the TPU v5p is the latest. AWS is developing the Trainium2 chips while Meta is developing its new MTIA chip. Other developers of superfast microchips are Intel (Gaudi) and Cerebras (WSE-3). There are also very different processor designs being developed, tailored to AI applications, such as IBM's Northpole. Nevertheless, the

Nvidia microchips will likely dominate this market for some time, given their leading speeds and high degree of integration within the current systems. This follows from a strategic focus to develop their fast computing eco-system, which included many acquisition.

The performance of accelerator cards (GPUs) is also increasing over time

GPT-3 175B Inference Performance of Nvidia GPUs



Source: Nvidia, the B200 performance figure comes from Tom's Hardware

A critical part of the modern IT hardware infrastructure is high bandwidth memory, as superfast processors work with superfast memory microchips (called high bandwidth memory, HBM). Because of the great demand for AI systems, there is also good demand for these memory chips. Manufacturers are SK Hynix and Samsung. Gartner expects that "HBM revenue will grow from $1.1 billion in 2022 to $5.2 billion in 2027. Between 2022 and 2027, there will be eightfold bit growth for HBM compared to fivefold growth in revenue".

Data centre infrastructure needs to be state of the art

As discussed, the communication requirements between the different functions in a data centre have evolved over time. Within modern supercomputers, the communication within (and among) servers has to facilitate high bandwidth data transfers at a low latency. The drawback of traditional, shared, communication channels is that it is more challenging to enhance communications speeds, as well as the number of devices linked to it.

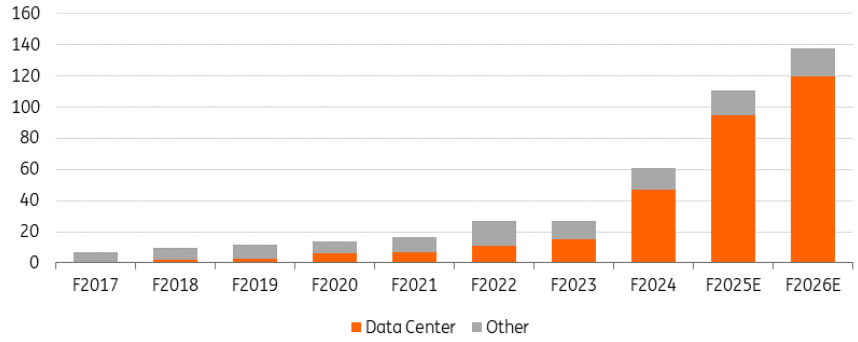
A key feature of modern data centre technologies is that multiple components, such as accelerator cards (with the GPU) can access shared memory directly, bypassing the CPU. This removes a bottleneck, because, in such a setup, there is no longer a shared communication infrastructure (bus-architectures). We are therefore witnessing the implementation of innovative point-to-point architectures, where all functions are linked through a switch architecture. In this field, Nvidia offers a leading solution.

Nvidia is the market leader showing strong sales growth

The developments above show that Nvidia has developed spectacular products used to build supercomputers. Its success can also be seen from the strong increase in its revenues, something that is expected to continue, according to consensus estimates depicted in the graph below. We do

not expect a competitor to match Nvidia's competitive offering, given its technological leadership and the width and depth of its competitive offering.

Nvidia shows strong demand for its data centre products (US$bn)



Source: Company, Refinitiv EIKON, ING. 2025 and 2026 revenue split calculated using previous three year average growth rate for "Other-segment" revenues.

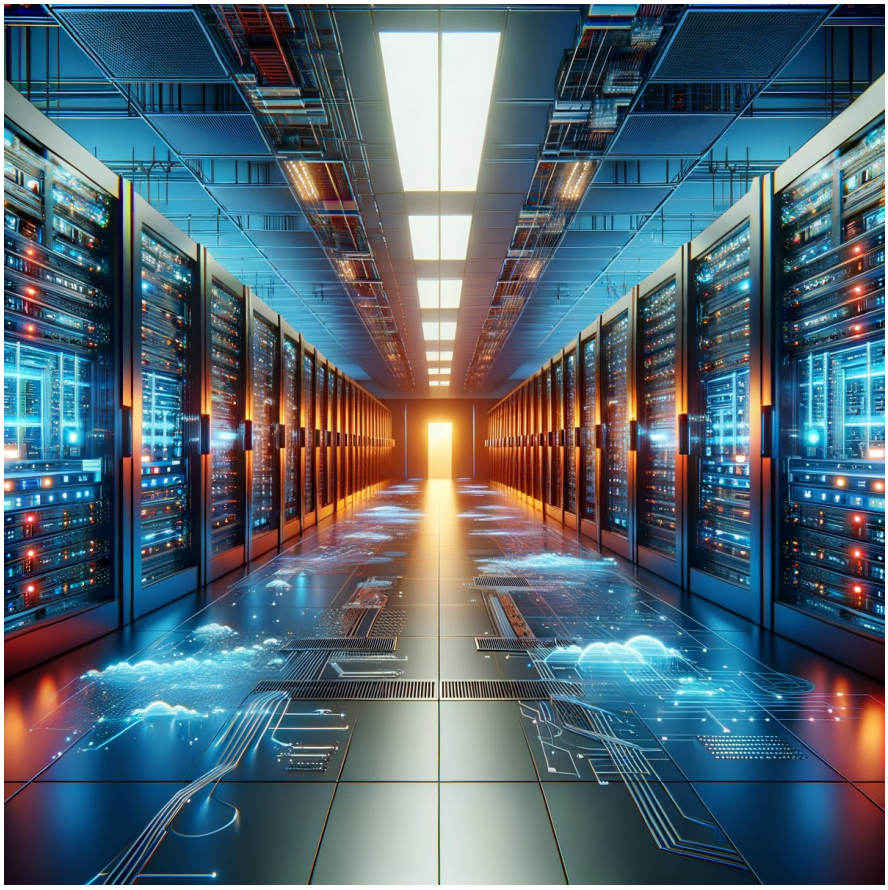
Author

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

Article | 24 April 2024 TMT

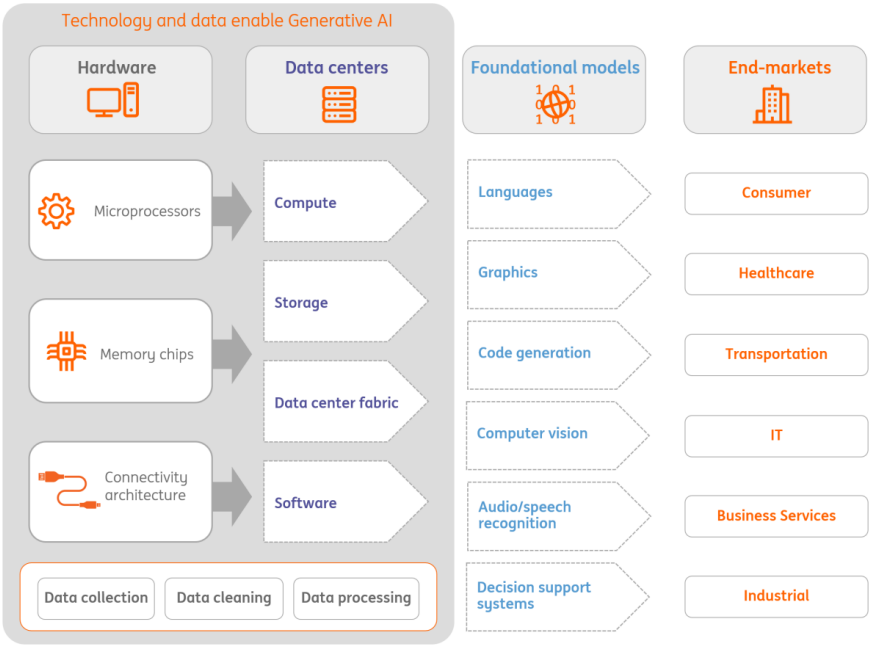
Data is the new gold for AI development

The latest AI models require spectacularly large datasets. In this article, we discuss leading data sources as well as the challenges that come with it. We expect that the owners of private data will build proprietary solutions, opening a new source of revenue. Nevertheless, more clarity is needed around copyright and the rights of content creator



The foundational models need a lot of data to be trained. Companies can use public data to train their models, such as the data that is available on the open internet. Moreover, the quest for the best foundational model may also result in a search for the best data. The AI developments are therefore an opportunity for companies that own, collect, and manage data. Furthermore, solutions that can prove the authenticity of content (or the person we interact with) will also become much more important.

Data is an important input for the AI value chain



Source: ING

Use of data is not without challenges

A feature that has enabled the creation of ever larger models is the availability of relatively good quality data sets that are free to use to train AI models. Although there is quite a lot of training data available, the latest AI models are characterised by billions of training tokens, and therefore require (extremely) large data sets, which are not readily available. As noted in a research paper from Google, "we find that for compute-optimal training, the model size and the number of training tokens should be scaled equally: for every doubling of model size the number of training tokens should also be doubled". The table below shows the increase in size of the training dataset for some models from OpenAI.

Newer models from OpenAI require larger training datasets

Model name	Size training dataset (words / datapoints)
GPT-2 (1.5B)	3.00E+09
GPT-3 175B (davinci)	3.74E+11
GPT-4	4.90E+12

Source: Epoch

Nevertheless, so far, there has been limited disclosure as to the exact data that has been used for the training of some frequently used models. Also, there is litigation on the use of data that is copyright protected, such as newspaper articles and books. Furthermore, many potential business

users of AI technology are reluctant to send their customer data to a public tool, because of potential confidentiality issues. These limitations create an opportunity for the companies that own private data to develop their own services. Or they could sell this data to the large technology platforms if it follows from regulations that the large technology platforms cannot use data sources that are copyright protected.

Datasets are being built to train AI models

A great effort has been made to build datasets, also for scientific purposes. There are quite a few sources that are widely used such as: Common Crawl, Colossal Clean Crawled Corpus (C4), The Pile, BookCorpus and Open Super-large Crawled Aggregated coRpus (OSCAR, based on Common Crawl). Datasets sometimes consist of multiple sources, such as books and crawled web pages or combinations of them. Other sources are, for example, Wikipedia, social media sites, (X, Reddit), software repositories (GitHub), and scientific journals.

Before data can be used, the data has to be cleaned to omit inappropriate words, discriminatory language as well as incomplete sentences. With the cleaning process comes the deletion of a substantial amount of valuable training data. Also, some sources are large but contain vast amounts of text that seems similar, such as some stories in the data included within BookCorpus. There is also a risk that training data has been generated by AI. This may create complications, which need a solution (such as a watermark so that the data can be removed more easily).

Content creators are dependent on copyright protection

The availability of training data is challenged by copyright protection. Many models have been trained with data that may be subject to rights protecting the creators of texts. The most famous example of a rights owner starting litigation is the New York Times. The newspaper has argued in court that its articles cannot be used to train AI. Other interest groups are the Rights Alliance, from Denmark. They have argued that Books3 infringed copyright laws. Moreover, in the USA, 10,000 authors, united in the Authors Guild have written an open letter, asking for financial compensation.

There is, therefore, still debate on whether technology companies can use published materials for training their models, as there is much debate about the interpretation of copyright protection and the fair use of materials. In this context, it is relevant to mention that many authors today mention (in electronic form) that their text cannot be used for the training of AI. Also, work is underway to create a better version of The Pile, dubbed The Pile v2, which aims to improve the data quality over the previous version. Also, the way the data will be cleaned may be improved as companies are more experienced today doing this in an effective way.

Proprietary data becomes ever more valuable

The training requirements make data a valuable input. Companies also collect data when their applications are used to be able to refine their models. This is something to consider when using applications because the input one gives can contain confidential (or private) information. There are therefore substantial risks involved when transmitting confidential (customer) data to applications which are managed outside a company. Because of privacy regulations and reputational risks, companies often take the utmost care to protect their data, which may require more tailored solutions in the form of private AI applications.

Also, proprietary datasets may represent significant value. Monetising owned data sources

through AI tools can constitute one of the new business models enabled by AI. The first solutions that come to mind are in the scientific publishing space, or owners of databases with large repositories of programming code.

Although it is slightly out of the scope of this article, we would like to mention that there are still challenges with the data generated by AI. Is the data copyright protected without significant human involvement? How can the problem of hallucinations be reduced? And will tools solve, at some point, the problem of generating data that is false or contains discriminatory language? We do not even discuss the problems arising from deep fake images. And can we prevent AI-generated data from being used for further training? Including a (mandatory) watermark in AI-generated content would seem to be necessary to solve some of these challenges. Nevertheless, detecting AI-generated content is also a nice opportunity for new software solutions.

More clarity is needed on the ability to use data for training purposes

Over time, we expect to get more clarity on the balance between the rights of content creators and the desire by the technology companies to use data. This is also a task for regulators. We think that a guiding principle should be that content creators must have the choice of whether their data can be used for the training of generative AI models. Nevertheless, the quality of data sets will likely improve over time, while new business models around the management of data will arise.

Author

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

AI frontrunners will benefit most, with Microsoft in the lead

Large technology firms are investing a lot to become AI services leaders. They are active across the AI value chain and will likely work to make their foundational model an industry standard. We expect that many sector-specific AI solutions will emerge. Given economies of scale, early leaders will likely benefit most

The hyperscale data centre market is currently dominated by four large technology companies in the West. These companies are well set to be leading the AI revolution. Moreover, we expect them to try to take control of the full value chain, which will enable them to extract maximum value. We expect the sector to grow by 10% per annum in the near future, in line with their electricity needs. Although we expect to see many different business models, we see Microsoft as the market leader

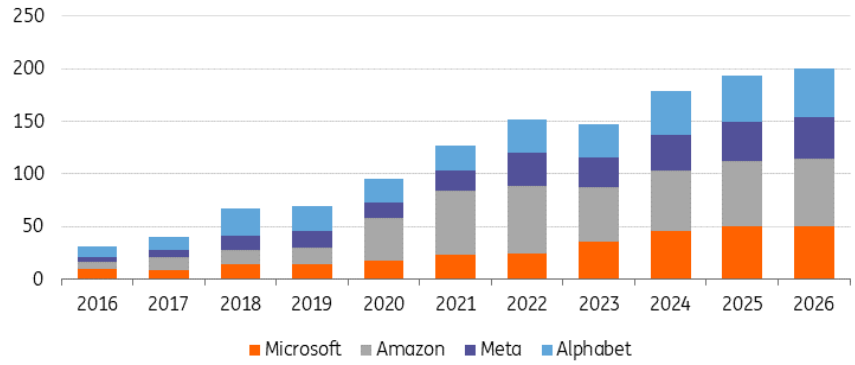
providing actual services. We expect the existing large cloud operators to consolidate substantial market share in the domain of Generative AI. Also, high capital investments provide a barrier to entry for followers. Outside the hyperscale operators, Nvidia also has a very strong AI offering, as explained in a related article.

The market leaders ramp up their asset base

The largest American technology companies should be able to benefit greatly from generative AI. To deploy Generative AI on a worldwide scale, we have recently seen a sharp increase in investment in digital infrastructure by Amazon, Alphabet, Meta and Microsoft. Digital infrastructure investments are a broader category than investments in Generative AI, but they do contain investments in Generative AI because this requires ultrafast microprocessors, and data centres. Counterpoint Research estimates that in 2023, roughly 13% of Microsoft's digital infrastructure spending was specifically for Generative AI. This percentage will likely increase in the future, as companies compete with their AI offerings.

Investments in data digital infrastructure are booming

Investments in digital infrastructure by Amazon, Microsoft, Meta and Google (US$, bn)



Source: Refinitive Eikon, ING (Microsoft capex held constant from 2025 to 2026 because of limited data availability)

Therefore, their investments are expected to keep growing for the foreseeable future. The capital investments by the four large operators of hyperscale data centres exhibit a compound annual growth rate of 11% between 2021 and 2025, where the 2025 number is based on consensus estimates. But note that the hyperscale data centres are expected to grow faster than smaller data centres, as we explained in our article on data centres.

Leading Generative AI developers and examples of AI models

OpenAI	GPT-4 Turbo GPT-4 GPT-3	Mistral	Mistral Large Mistral Medium Mistral Next
Alphabet	Gemini Ultra Gemini Pro Gemini Nano Palm -2	Anthropic	Claude 3 Opus Claude 2.1 Claude
Meta	LLaMA-3 LLaMA-2 LLaMA-1	Hugging Face	BLOOM

Source: ING

Foundational models are a key building block for AI applications

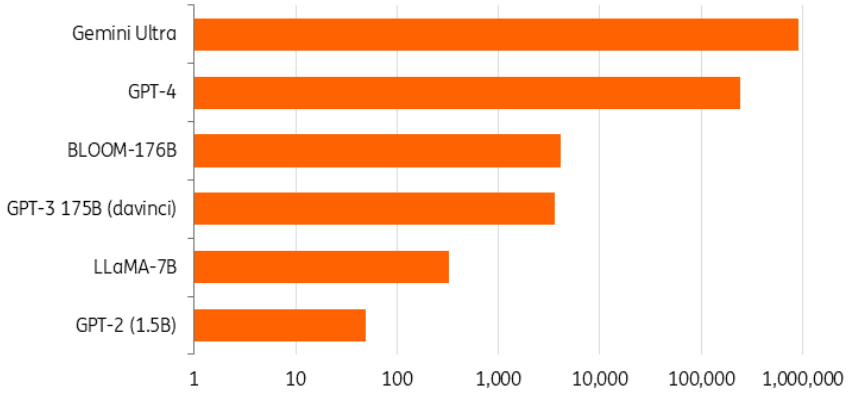
Breakthrough innovations in the field of AI have come in the form of some sophisticated foundational models which are general-purpose technologies. These general-purpose models can be used for a diverse range of use cases. As shown above, the development and training of these models require huge amounts of data and a vast infrastructure. There are quite a few prominent foundational models that can analyse and generate, for example, images, music, and languages. Prominent developers of foundational models are: OpenAI, Alphabet, and Meta, as can be seen in the table above. These companies are well-placed to lead the innovation race. This is important, as the technology sector typically benefits from economies of scale. In short, the company that comes first with the best technology has the best chance to take a leading, dominant, market share. This is because it makes the most money (which it can reinvest), but it also benefits from the interaction with customers and past experiences to further refine the service. This explains the race to have the leading foundational model. Below we show the computational requirements for a selection of models. Notably, the requirements went up from the old GPT-2 model to a more modern version, GPT-4. With it came a spectacular increase in user experience. And we have not seen all of it yet, as larger models will be around soon.

Many AI solutions are scalable

The above mainly discusses the large, breakthrough Generative AI models. Nevertheless, there also exist smaller (legacy) models, while companies are also working on smaller models that can be used for specific tasks. These smaller models can run on standalone servers and smartphones. The advantage of such solutions is that the operator has full control over the data that goes in and out of the system, which makes it more palatable for processing private data. Software providers, such as Alphabet, Meta and OpenAI offer families of models, offering different degrees of complexity. Notably, there are (trained) open-source models that can be used by developers to create their own, tailored, solutions.

Required time to train models increases with complexity

(Petaflop/s-days, exponential scale)



Source: Our World in Data, ING

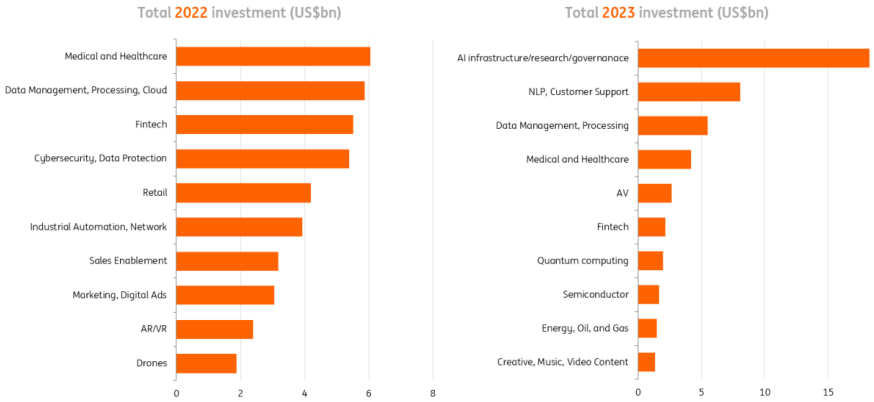
Server manufacturers will also find new opportunities as the AI market segment opens up. Server manufacturers are, for example, Hewlett Packard Enterprise, Dell Technologies and Super Micro Computer. As an example, Dell Technologies and Meta are selling a proposition (including servers and software) that can be used by corporations to run their own, private AI models for software development.

Asking a remote AI application for help is also possible. Today, the capabilities of low-tech hardware can be enhanced through communication with the intelligent cloud, running state-of-the-art AI models. Through a relatively simple device, a customer can send a request to a data centre that can perform more complex calculations or analysis. This operating model can therefore bring AI services to local, smaller, devices. What is needed are excellent, low-latency, communication networks, such as those based on the 5G standards. AI therefore provides another incentive to complete these networks, because they are an enabler of smart Internet of Things applications. Such developments are part of a recent transaction between Microsoft and Vodafone, and we expect many telecom companies and smartphone manufacturers to bring innovative services and applications in the future, including smartphones capable of running AI software.

It has proven difficult in history to alter business models

For companies, it is important not only to focus on the ongoing technical developments that are needed to pursue the full opportunities of AI but also on the business models. As we saw previously, it is sometimes difficult to embrace technological change in the right way. Two examples from a 2019 Forbes magazine article illustrate this. The industrial conglomerate GE experienced failure when it tried to build GE Digital. They wanted to develop an IoT service model around the traditional industrial equipment division, but the project was not managed properly, at a huge cost. Also, automaker Ford created a digital division, separate from its traditional automotive division. This also resulted in disappointment because the digital products were not well integrated. Before investing in new AI-related product developments, companies therefore should focus on the business case of a new, AI-inspired, product value proposition. A first example, with a good chance of success, is Copilot for Microsoft 365, a subscription-based revenue model.

Substantial investments are made to develop sector specific solutions



Source: Quid 2023; Netbase Quid 2022 taken from the Stanford University Artificial Intelligence Index Reports 2022 and 2023 (largest amounts only)

Sector specific applications are being developed

It is very likely that we are going to see the emergence of many AI-driven solutions that will be derived from the foundational models, such as user interfaces, chatbots or dedicated tasks for the business services industry. As can be seen in the figure above, this is becoming a very large industry. Investors see large opportunities in sectors like Medical and Healthcare, Customer Support, Data Management, Processing and Cloud, and Fintech. These sectors have attracted huge amounts of private investments in 2022 and 2023 alone.

Most of these solutions under development will likely not run in the public domain. Companies owning private information may want to develop services to unlock these in a more intelligent (AI-like) way, through proprietary models that could be based on (or combined with) foundational models. Companies, often, cannot share (customer) data with a public cloud model because they lose control over the information. The development and implementation of private (in company) generative AI tools could, therefore, be a solution for privacy and confidentiality challenges. A company like Together.AI specialises in the development of new applications. They can help train AI models for the purpose of clients but also help with the tailoring of datasets and training of the private AI models. A platform like Hugging Face provides tools for building machine learning applications. These tools will become available on the Amazon (AWS) platform.

Leading ecosystems are emerging

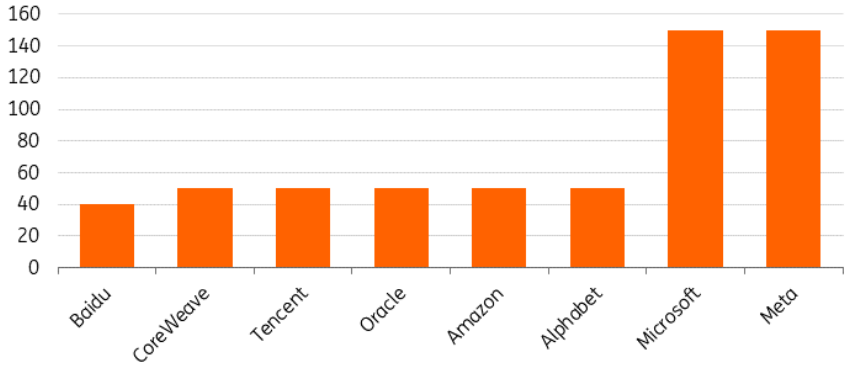
Alphabet, Microsoft and AWS are the leading global providers of data centre services and are among the largest owners of data centre and cloud infrastructure. They are developing the tools and models to provide a full suite of services for their clients that use their infrastructure. Likely, AI will become an important driver of growth in the cloud software market. Through a leading position in AI, a company such as Microsoft or Alphabet, will have the opportunity to solidify (or grow) its cloud service model. IDC expects the spending by businesses on Generative AI to reach $143 billion in 2027.

As described in the previous section, the development, maintenance and roll-out of AI tools require specific knowledge, tools and substantial amounts of capital. As described in the Stanford

University Artificial Intelligence Index Report 2024, the training of Generative AI models is expensive, likely costing millions. But, also, the inference, data, and maintenance of models is costly, especially since the most advanced semiconductors from Nvidia have been in scarce supply.

Nvidia's products were in scarce supply in 2023

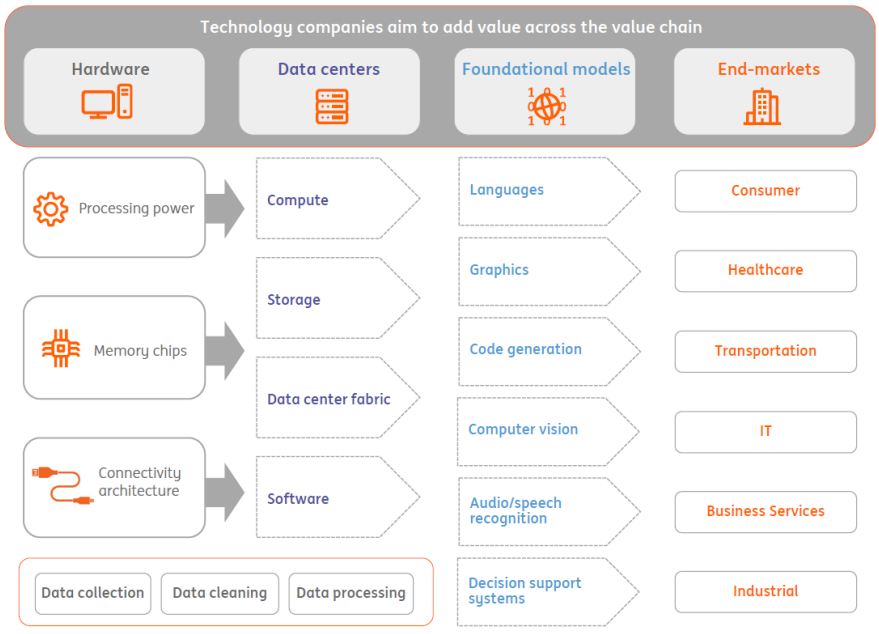
H100 GPU shipments by customer



Source: Omdia Research, through theverge.com

The figure above shows that Microsoft and Meta managed to build the largest portfolios of Nvidia H100 systems, while Amazon and Alphabet are building leading compute capacities. As we have seen, these companies are also the leading data centre operators, with extensive software (development) suites, and a global footprint, which makes them well-connected. Through bundling solutions, Microsoft, Alphabet and (over time) Amazon are well-positioned to provide a strong AI ecosystem, building on an existing strong cloud offering. The challenge to operate across the value chain can be seen in the diagram below.

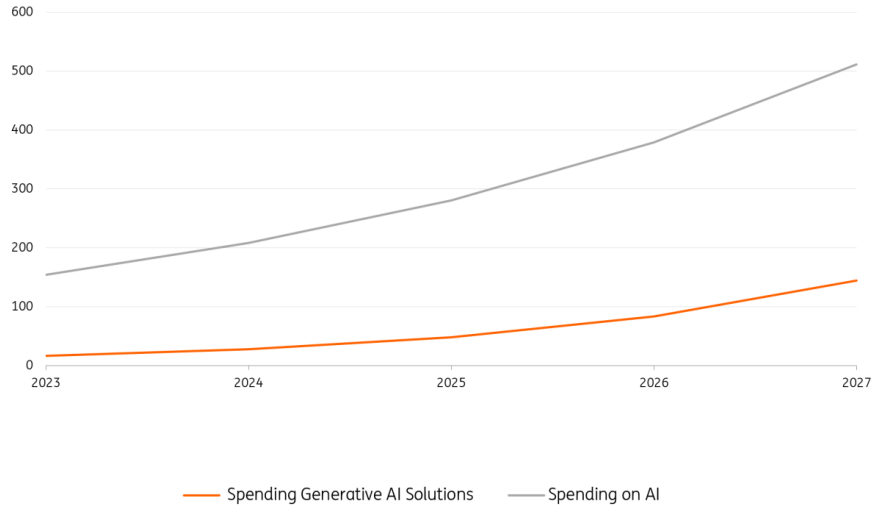
Technology companies aim to sell a full product suite



Source: ING

This will likely have implications for customers buying AI services. Notably, it is not simple for companies to migrate to another ecosystem. And it will be very difficult for new entrants to carve out market share at a later stage because the market leaders benefit from an early mover advantage.

The market for AI services will expand strongly



Source International Data Corporation, ING

Leading cloud operators and Nvidia will dominate the AI landscape for some time to come

We therefore expect that a few AI ecosystems will dominate the landscape for some time, although there is possibly room for vendor-neutral hyperscale data centres, such as the Equinix xScale data centres to join the race now. Equinix, for example, will offer a private cloud service to enterprises which unlocks a private Nvidia DGX AI infrastructure which they can use to build and run custom Generative AI models.

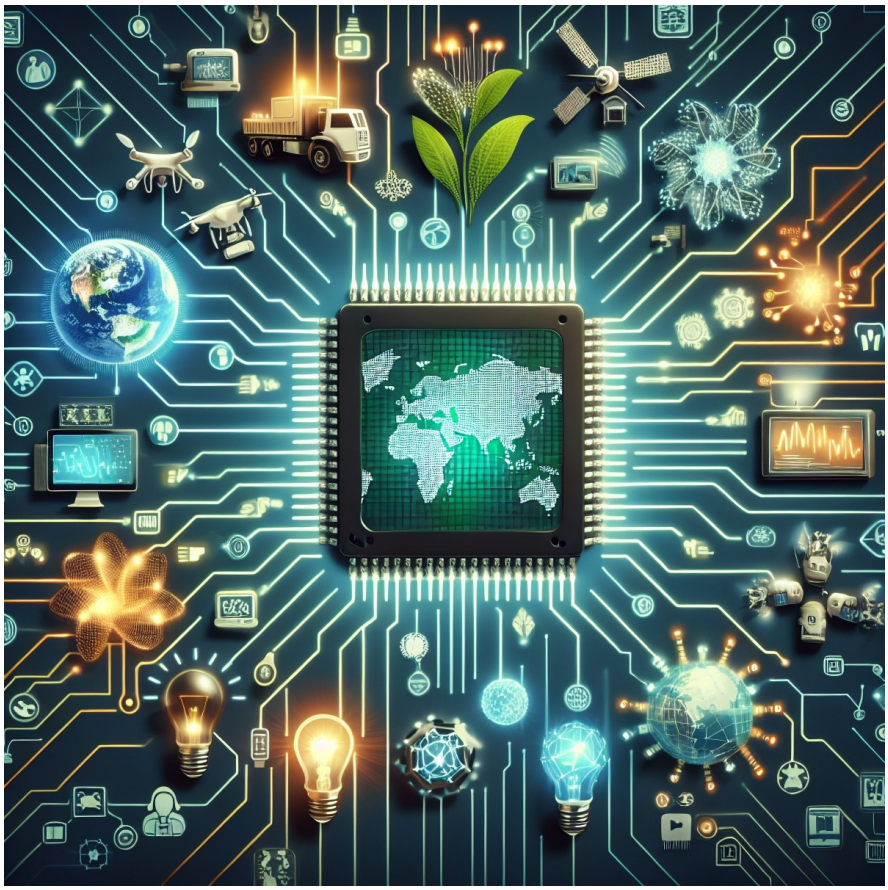
Author

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

Article | 24 April 2024 TMT

Countries must embrace AI or risk missing out on strong ICT growth

We show why we think data centre capacity will grow by 10% a year, facilitating growth of the market for AI services. We do not expect more spectacular growth numbers but do argue that countries should embrace the changes that come with AI and embrace the chance to build AI data centres because the ICT sector continues to be fast-growing



We expect the ICT sector to exhibit strong growth on the back of the development of Generative AI services. Nevertheless, our expected growth figures are somewhat below numbers from Generative AI advocates. However, strong growth from Generative AI will result in substantial electricity demand, fuelling a debate on the desirability of further data centre investments in some countries.

As we have shown in previous articles, the data centre sector and Generative AI will be an important factor in enhancing productivity, while the ICT sector shows above-average growth rates. Therefore, we argue that countries should not miss out on this growing sector.

Industry veterans forecast extreme growth

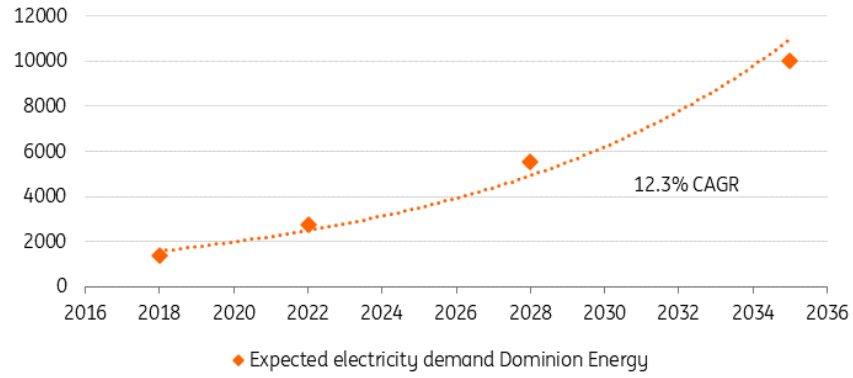
In February 2024, Sam Altman the CEO of OpenAI posted this on X:

According to the WSJ, he has pitched a plan to develop his own semiconductor foundries to meet the global demand for AI microchips with an investment need of US$7 trillion. That view has been contested by Jensen Huang of Nvidia, who thinks the investment needs are overstated because of advances in semiconductor speeds as well as the development of more efficient AI models. According to Bloomberg, Huang expects that the current installed AI data centre capacity is worth US$1tn, which probably has to double to US$2tn in five years' time.

Are investments of up to US$7tn across the value chain really needed? We will evaluate the different segments of the value chain to put this amount into perspective and present our view on future investments.

Utilities expect strong electricity demand from data centres

Dominion Energy plans for strong electricity demand growth in North Virginia



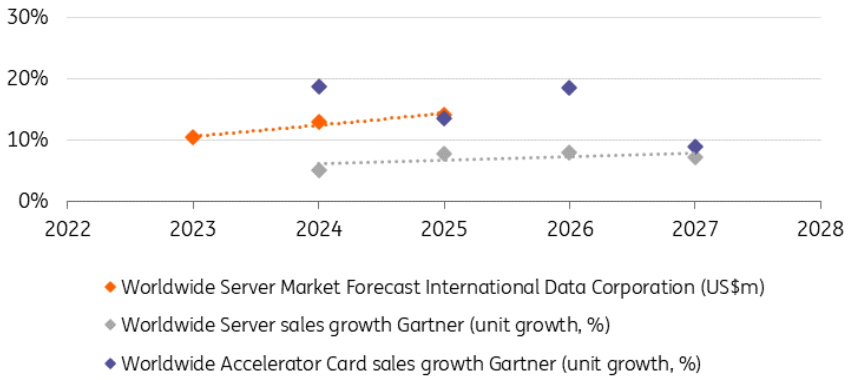
Source: Datacentrefrontier.com, Dominion Energy, ING

When evaluating the capital investments by the four large operators of hyperscale data centres, one finds a compound annual growth rate of 11% between 2021 and 2025 (the last year being based on consensus expectations).

Another growth figure is provided by the expected electricity demand in Virginia (USA), which is home to an agglomeration of data centres. Dominion Energy, the largest utility in Virginia, expects that the electricity demand in the area will grow by 12.3% until 2035. Although this figure is contested by some, it matches the recent growth rate in investment.

Semiconductor demand is expected to show strong growth rates

Strong semiconductor demand shows sector electricity demand will increase



Source: International Data Corporation, Gartner, ING

We think investments in semiconductors point to strong 10% growth

However, when we study the sales of server microchips (CPUs), growth rates are lower. Gartner expects that the worldwide sales of semiconductors will grow by a high single-digit figure, as can be seen in the graph below. Nevertheless, besides the demand for server microchips, data centres will also employ more accelerator cards. Gartner expects a higher growth rate for these accelerator cards (GPUs) than for the server microchips (CPUs). This is in line with our thinking, as companies will rely on the accelerator card for the training and use of Generative AI models. When taking all growth figures into account, we expect the energy consumption of data centres to grow by c.10% in the coming years, although this is a figure surrounded by much uncertainty, because it also implies that the compute instances grow much faster (because chips become more efficient).

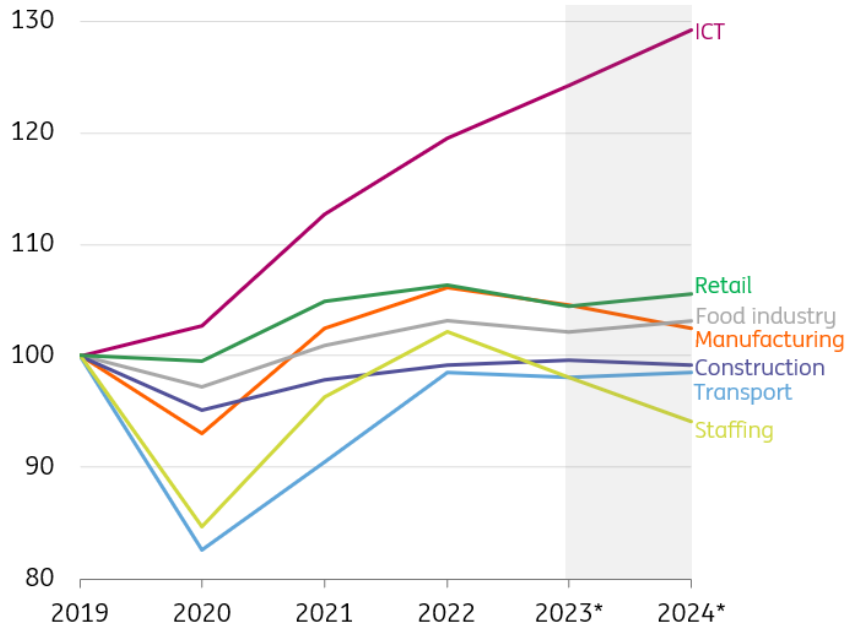
To come back to the debate between Jensen Huang and Sam Altman, we do believe that the sector will witness strong growth but expect it to take a bit more time to reach these figures, as investments will be more moderate in the near term.

The expected growth rate in energy consumption by c.10% reflects an increase from our previous research. In earlier research, we evaluated energy efficiency in the sector (resulting from the scale efficiencies of data centres and Moore's Law), as well as the need to use green electricity. We showed that it is tricky to only focus on the growing electricity demand of the data centre sector because this data does not incorporate a potential reduction in energy use in other sectors.

Nevertheless, it is clear that low energy prices offer a comparative advantage for the industry since energy is an important input for data centres. Besides a favourable climate and good data connections, this will be an important factor driving the attractiveness of investment locations. Moreover, we expect data centres to exhibit a desire to procure green energy and nuclear energy, because the companies and their clients sometimes have to report the CO2 emissions from their suppliers. We discussed the procurement of green energy and sustainable finance in an earlier article.

The ICT sector expected to show continued above average growth

Development production (volume value added) EU sectors (Index 2018=100)



Source: Eurostat, ING Research (* 2023 Estimate & 2024 Forecast)

The ICT sector will continue to show above average growth rates as well

Because AI-related services will likely be implemented across the economy, the Information and Communication Technology sector has the scope to grow volumes faster than the overall economy, as it did in previous years. If companies see an opportunity to lower their cost base through the use of AI services, this increases the total addressable market for the ICT market. Previously, a large part of the advertising market was replaced by digital solutions. A similar development can be seen in the linear TV market, which is developing into a streaming video model. The developments around chatbots, translation tools and the many other AI-based services imply that the technology sector has a revenue opportunity, although these services will be an expense item for other companies. We expect that the overall ICT sector will show an above-average growth rate in the mid-single digit area over the coming five years, at least.

We need to invest in AI data centres because of labour market challenges and the need for technological sovereignty

We have shown in recent articles that AI will have a positive impact on labour productivity and will change the tasks employees will perform. We would argue that it is better for countries to embrace this change and try to get a fair share of the newly emerging business opportunities. Having access to this technology improves technological sovereignty. Of course, the environmental impact of new investments should be considered, as we have shown in our article on sustainable finance principles. However, as the required investment scale is a challenge in itself, we should

better take on the tasks and try to build data centres for AI when opportunities are there.

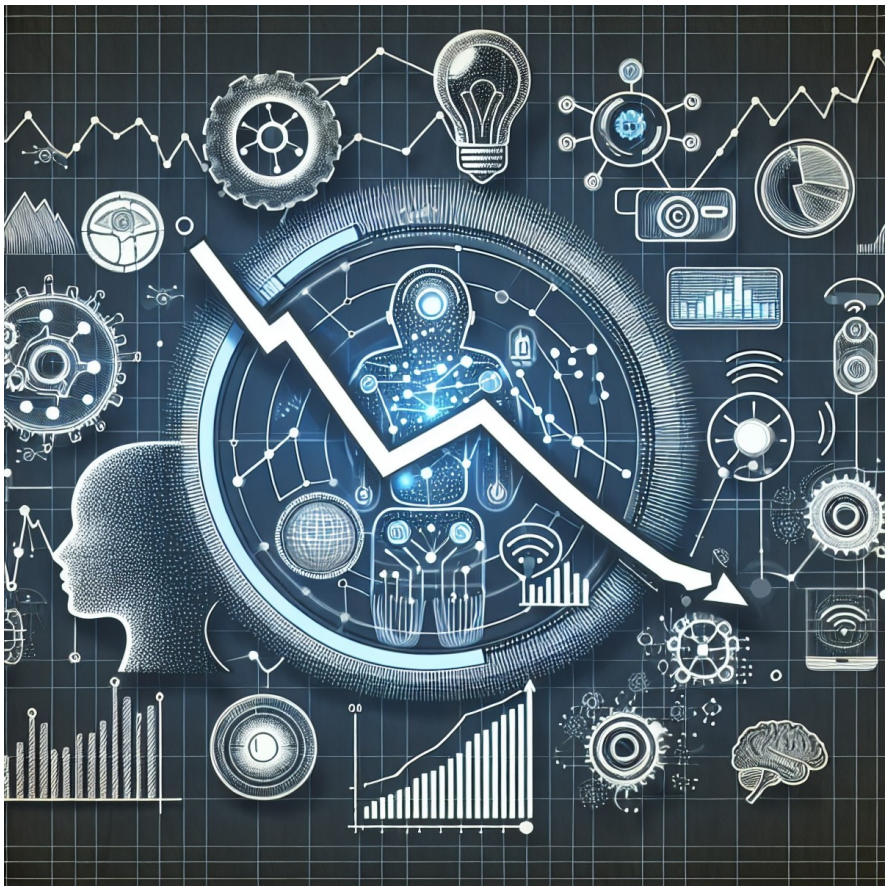
Author

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

Article | 24 April 2024 TMT

AI services will become more affordable over time

Today, some Generative AI services are rather expensive. We expect the prices of these services to come down, in line with historical developments of ICT hardware prices



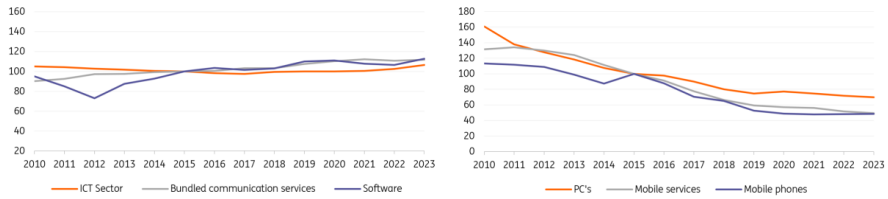
Declining prices are common in the ICT sector

When evaluating the historical evolution of the ICT sector, there is one important aspect that stands out. Prices have come down in substantial segments of the sector. This especially holds true for hardware such as phones and PCs but also for prices of mobile communication. One of the other notable aspects in this context is the provision of free services, through open-source projects or advertising driven models. Nevertheless, the main cause of falling prices is the advancement in semiconductor manufacturing capabilities, in line with Moore's Law. Because of this, semiconductor prices have come down, leading to lower prices for end-products, such as PCs,

tablets, and mobile phones. Because microchips are an important cost component of AI models, we expect the replacement cost of the existing infrastructure needed to run AI models to come down over time as well.

In a November 2023 blog post, OpenAI introduced a new model GPT-4 Turbo. Interestingly, compared with GPT-4, OpenAI will offer "GPT-4 Turbo at a 3x cheaper price for input tokens and a 2x cheaper price for output tokens". Although one should be careful about generalising from one statement (possibly the initial price was too high to attract enough volume), it fits with trends of falling prices in parts of the ICT sector. Moreover, note that the news on GPT-4 Turbo is already outdated because OpenAI now plans to release GPT-5 around the summer of 2024. This shows that companies keep developing better, more complex, and more expensive, models over time. We do therefore expect that spending on AI infrastructure will continue for some time to come.

Prices are falling for some products in the ICT sector



Source: CBS, ING

Expect downward price trend for AI, like many product and software categories

The graphs above show that the costs of computing equipment have come down substantially over time. Although one could argue that a smartphone today is a completely different product than 10 years ago, most of us would agree that there are many more affordable devices around today than in the past, with better capabilities. This will also happen with Generative AI, as the costs to make calculations will come down, but also because engineers will find ways to optimise these. Also, Generative AI will likely bring costs down for software engineering, customer support as well as administrative tasks. Of course, quality will also improve, which may cost somewhat extra. The same holds for products that rely on proprietary data or knowledge. Also, other effects have an upward effect on prices, such as oligopolistic market structures and strategies making it difficult for customers to switch suppliers (vendor lock-in, product bundling). Nevertheless, we believe in a downward trend for many product and software categories.

Author

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

Disclaimer

This publication has been prepared by the Economic and Financial Analysis Division of ING Bank N.V. ("**ING**") solely for information purposes without regard to any particular user's investment objectives, financial situation, or means. *ING forms part of ING Group (being for this purpose ING Group N.V. and its subsidiary and affiliated companies).* The information in the publication is not an investment recommendation and it is not investment, legal or tax advice or an offer or solicitation to purchase or sell any financial instrument. Reasonable care has been taken to ensure that this publication is not untrue or misleading when published, but ING does not represent that it is accurate or complete. ING does not accept any liability for any direct, indirect or consequential loss arising from any use of this publication. Unless otherwise stated, any views, forecasts, or estimates are solely those of the author(s), as of the date of the publication and are subject to change without notice.

The distribution of this publication may be restricted by law or regulation in different jurisdictions and persons into whose possession this publication comes should inform themselves about, and observe, such restrictions.

 ING Bank N.V. is authorised by the Dutch Central Bank and supervised by the European Central Bank (ECB), the Dutch Central Bank (DNB) and the Dutch Authority for the Financial Markets (AFM). ING Bank N.V. is incorporated in the Netherlands (Trade Register no. 33031431 Amsterdam). In the United Kingdom this information is approved and/or communicated by ING Bank N.V., London Branch. ING Bank N.V., London Branch is authorised by the Prudential Regulation Authority and is subject to regulation by the Financial Conduct Authority and limited regulation by the Prudential Regulation Authority. ING Bank N.V., London branch is registered in England (Registration number BR000341) at 8-10 Moorgate, London EC2 6DA. For US Investors: Any person wishing to discuss this report or effect transactions in any security discussed herein should contact ING Financial Markets LLC, which is a member of the NYSE, FINRA and SIPC and part of ING, and which has accepted responsibility for the distribution of this report in the United States under applicable requirements.

Additional information is available on request. For more information about ING Group, please visit www.ing.com.