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

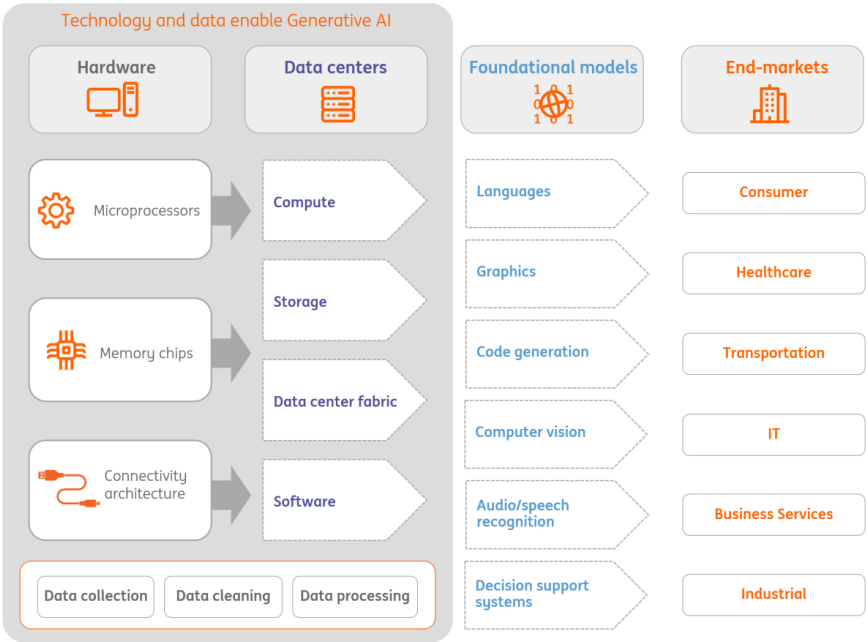
The current generative AI boom is driven by a spectacular increase in the capacity of microchips. Today, a web of servers can perform calculations on these extremely large datasets. And this is just the beginning. In this article, we discuss the spectacular innovations giving rise to Nvidia as a leading 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



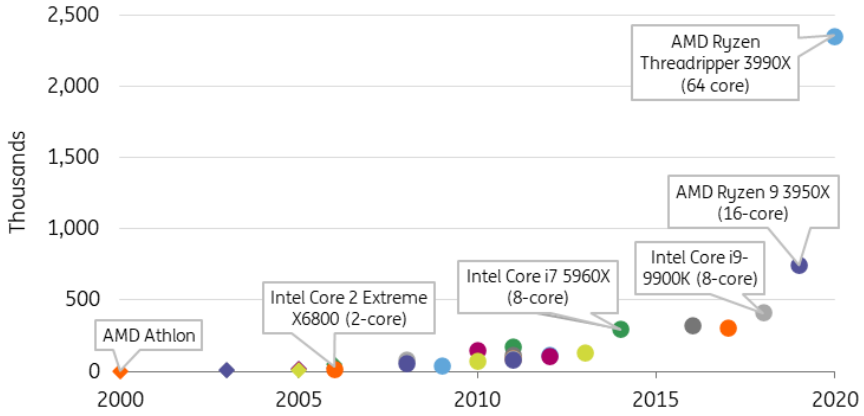
Source: ING

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

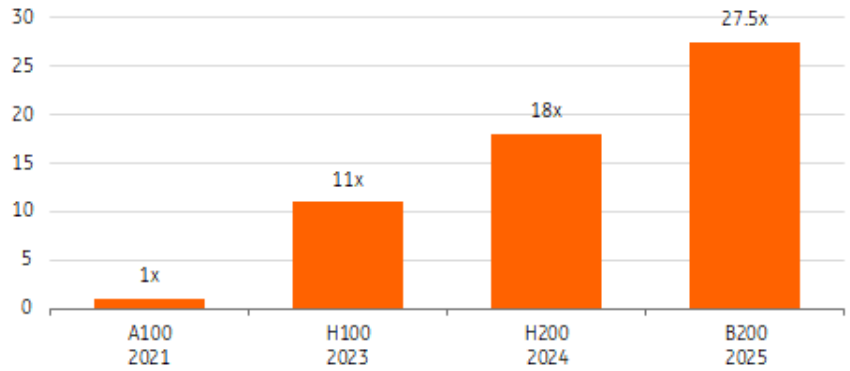
Source: Nvidia, ING

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.

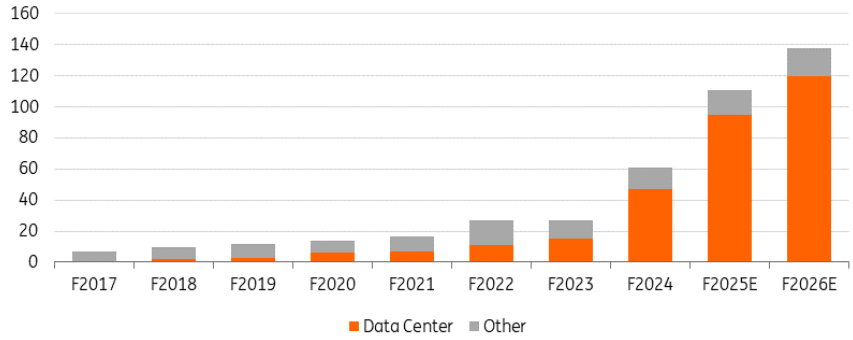
A key feature of modern data centre technologies is that multiple components, such as accelerator cards (with the GPU) can access shared memory directly, bypassing the CPU. This removes a bottleneck, because, in such a setup, there is no longer a shared communication infrastructure (bus-architectures). 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