Roland Berger

GenAI hardware – A trillion-dollar market for GPUs

APRIL 2024

AUTHORS

THOMAS KIRSCHSTEIN
Partner

ALEXANDER FEY
Principal

FALK MEISSNER
Senior Partner

The rise of generative artificial intelligence (GenAI) is paving the way for a mid-term boost to the semiconductor and related equipment industries.

According to our calculations, the market for graphics processing units (GPUs) – the main hardware used for GenAI – could grow from USD 100 billion in 2023 to more than USD 3 trillion by 2040. This translates into an upside for semiconductor equipment companies of between ten and 30 percent compared to current market expectations. Yet the industry faces a sustainability challenge that could put the brakes on its expansion: The energy that will be needed by GenAI datacenters globally in 2040 could surpass the total energy consumption of the United States today. Moreover, the high profits currently earned on GenAI hardware – almost triple those generated by conventional datacenter hardware – create a significant cost of ownership challenge for end users. We examine the current state of the GenAI hardware market, its predicted growth, the limitations it faces and the exciting opportunities for investment and innovation by market players.

A trillion-dollar market

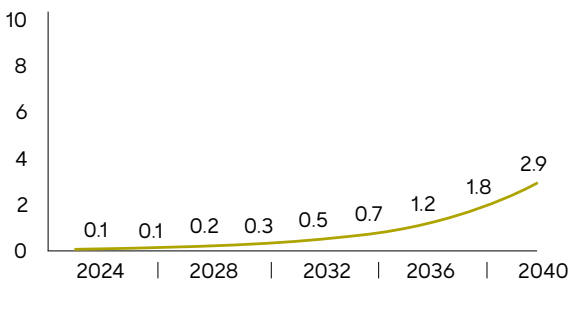
GenAI is transforming industries and creating new value for businesses and society. Its growth has significant implications for a range of industry segments. Typically, GenAI services rely on datacenters, today mainly operated by a few large cloud infrastructure service providers, sometimes known as 'hyperscalers'. A datacenter consists of a system for power supply/conversion, uninterrupted power supply systems and HVAC systems for cooling. The hardware itself is typically stored in racks that consist of a series of GenAI systems, normally themselves comprising a central processing unit (CPU), storage, memory and a large number of GenAI accelerator cards. These GenAI accelerator cards are graphics processing units (GPUs), which, like all electronic circuits and devices, are built around semiconductors.

As GenAI begins to really take off, the underlying global market for hardware and semiconductors is expected to grow substantially. By 2040, driven by increasing demand for AI services and solutions, the market for GPUs for GenAI datacenters could reach somewhere between USD 3 trillion and USD 10 trillion according to our calculations. The lower end of this forecast, around USD 3 trillion, is based on a financial market model, calculated by looking at the implied semiconductor fab capacity needed, based on the valuations of GenAI hardware players; it is our base case prediction as it focuses on company development. At the higher end, the figure of USD 10 trillion is based on an optimistic demand forecast that looks at the hardware and semiconductor capacity needed to meet optimistic revenues for GenAI service scenarios, taking into account overall demand expectations for AI use cases.

Massive growth of GenAI will drive the underlying hardware market

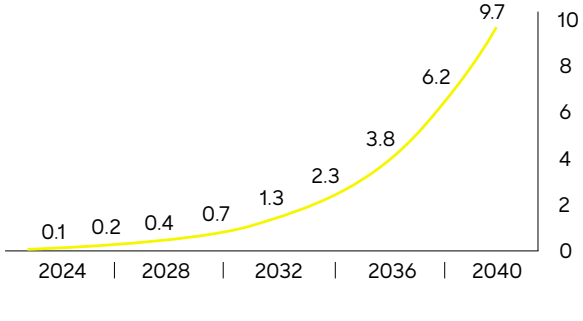
Implied GenAI datacenter hardware markets (GPUs), 2023-40 [USD trillion]

1. Financial market based
Financial market-based model assumes more conservative growth as it focuses on company development



Implies AI service revenue of approx. USD 3 trillion in 2040, i.e., 2 % of GDP

2. GenAI demand based (optimistic scenario)
Demand-based models consider the overall societal demand perspective and expectations for AI use cases



Implies AI service revenue of approx. USD 10 trillion in 2040, i.e., 6 % of GDP

Source Roland Berger

Roland Berger

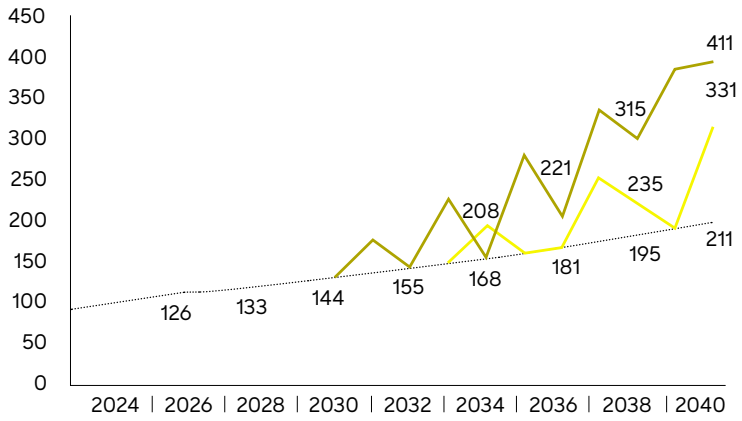
The forecast expansion of the market for GPUs for GenAI datacenters translates into additional demand for semiconductor capacity and equipment of between USD 300 billion and USD 1 trillion, depending on the scenario. Between 2024 and 2040, global investment in semiconductor capacity would amount to USD 3.3-4.0 trillion. Thus, the additional demand for semiconductor capacity for GenAI would lead to a ten to 30 percent upside compared to previous market projections for semiconductor capacity.

Even in our base scenario, production capacities for semiconductors will need to expand significantly. The need for more production capacity will be accompanied by increasing technological complexity. Alongside organic growth, semiconductor manufacturers will need to consider inorganic options, that is, mergers and acquisitions.

Semiconductor equipment demand is expected to double by 2040, mainly due to GenAI

Semiconductor equipment and EPC spend, 2024-40F [USD bn]

Market scenarios



Cumulative spend, 2024-40

- GenAI demand **3,020 + 960**
- Financial markets **3,020 + 280**
- Baseline **3,020**

Source Roland Berger

Roland Berger

A shortage of hardware?

In recent quarters, GPUs for GenAI datacenters have been in short supply. The unfortunate consequence has been that GenAI services, such as image and video generation, have rolled out more slowly than was hoped for.

Of course, shortages of semiconductors are nothing new. But this time, the shortage was driven not by a lack of front-end capacity but rather by a lack of high-end packaging capacity – GenAI computing is very memory intensive, so the memory semiconductors and logic semiconductors need to be very tightly connected via high-bandwidth solutions, for which most companies use a chip-on-wafer-on-substrate (CoWoS) packaging technology that employs multiple dies side by side on a silicon interposer.

For the moment, however, it would appear that this substrate capacity shortage has been resolved. On the other hand, high-end memory semiconductors (e.g., HBM) are already sold out and allocated for 2024 and 2025, while leading-edge logic ICs are also growing strongly. We therefore foresee a robust growth trajectory for GenAI services and their underlying hardware in the current year.

The sustainability challenge

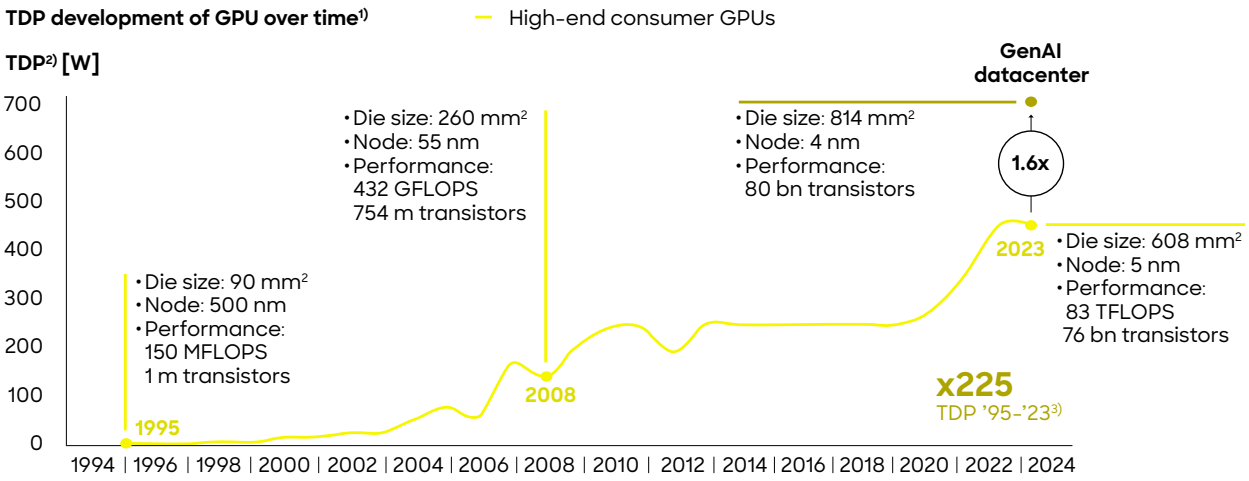
In the longer term, the GenAI hardware market is facing a significant sustainability challenge. This is due to the high levels of energy consumption by the hardware and semiconductors, and the resulting carbon footprint of the industry.

GenAI applications require a lot of computing power, which in turn depends on the performance and efficiency of the hardware and semiconductors. One GenAI accelerator card currently uses approximately 700 W of power, which under full utilization is equivalent to 6 MWh per year. For comparison, a typical four-person household in Europe has an energy consumption of 4 MWh per year. A typical GenAI datacenter will employ 10,000-100,000 of these GenAI accelerators. In addition, it will use substantial amounts of energy for cooling and energy conversion.

To make things worse, since 1995 the die size of desktop GPU chips has increased by a factor of seven, and power consumption of the GPU by a factor of 225. The higher the performance and efficiency, the lower the energy consumption and carbon footprint. In other words, despite improvements in energy efficiency, the energy use and hence carbon footprint of GenAI hardware and semiconductors have increased massively – and show no signs of shrinking any time soon.

Energy consumption is becoming a major issue

Die size of desktop GPUs has increased sevenfold, power consumption by a factor of 225



[1] Representative high-end GPUs for each year [2] Thermal design power [3] TDP grew by a factor of 225 from 1995-2023

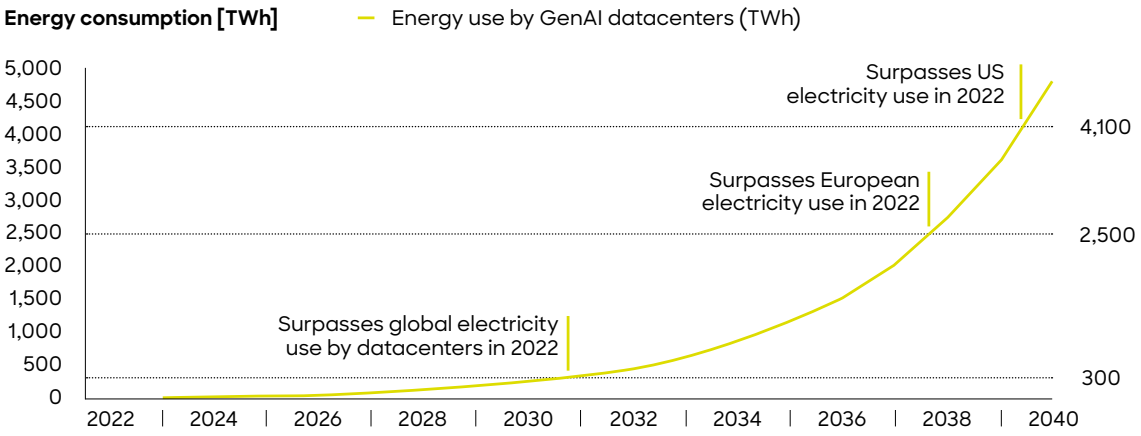
Source Roland Berger; company information

Roland Berger

According to our estimates, the annual energy consumption of GenAI hardware in datacenters could reach up to 4,800 TWh by 2040 – more than the current electricity consumption of the whole of the United States. That is in the lower of our two market scenarios. Clearly, the environmental impact would be enormous, as would the financial costs for the GenAI hardware and semiconductor industry and the wider costs for society. We expect this negative development to continue, as the required performance improvements cannot be achieved with architecture and semiconductor manufacturing technologies alone. What is more, the cost of energy only accounts for around 20 percent of the total cost of operation of GenAI hardware, so there is an incentive to make a trade-off between hardware cost and energy cost. As such, future GenAI hardware accelerators will likely use more than 1,000 W of power.

As demand for GenAI hardware grows, so does demand for energy

Estimated total energy consumption of AI hardware over time (financial market model)



Source Company information, Roland Berger

Roland Berger

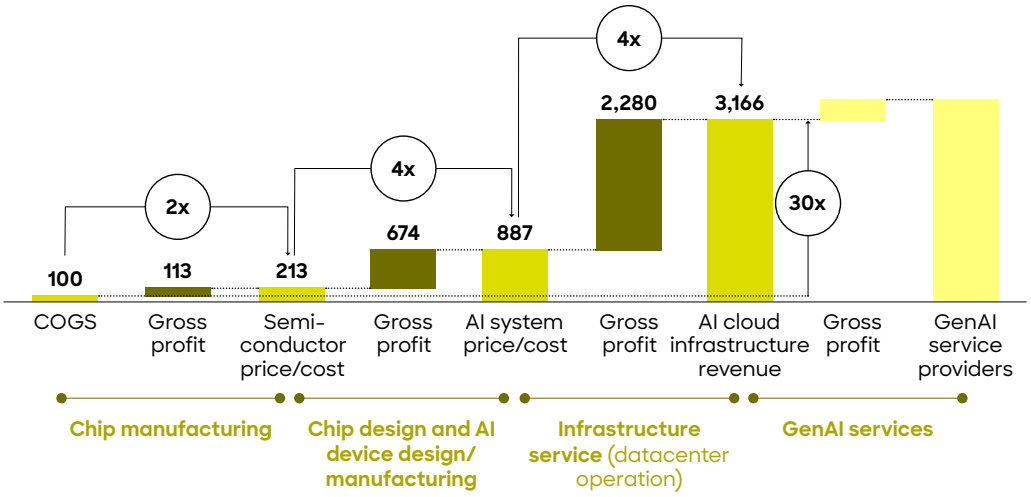
It would appear that the only way to ensure a sustainable future for GenAI will be to quickly expand large-scale renewable energy generation. This will make it possible to continue developing the GenAI industry at full speed, while covering energy demand with limited CO_2 emissions.

- ### The cost of ownership challenge

A further challenge for the industry relates to the cost of ownership for end users. Prices for GenAI cloud infrastructure are driven by 'margin stacking' across the value chain (each member of the supply chain enjoys a profit margin that contributes to the final cost of the service). Currently, it is mainly the GenAI hardware companies that are enjoying larger profits than their peers in traditional computing hardware. The cost of goods sold at the semiconductor raw materials level and the sales price at the data infrastructure level differ by a factor of 30. One AI accelerator card can cost up to USD 40,000. In other words, GenAI infrastructure companies realize a revenue of USD 30,000 on an original cost of production at the semiconductor companies of just USD 1,000.

Consumers of GenAI pay for significant value pools in GenAI hardware and infrastructure

Indicative value pools, Q4 2023



Source Roland Berger; company press releases

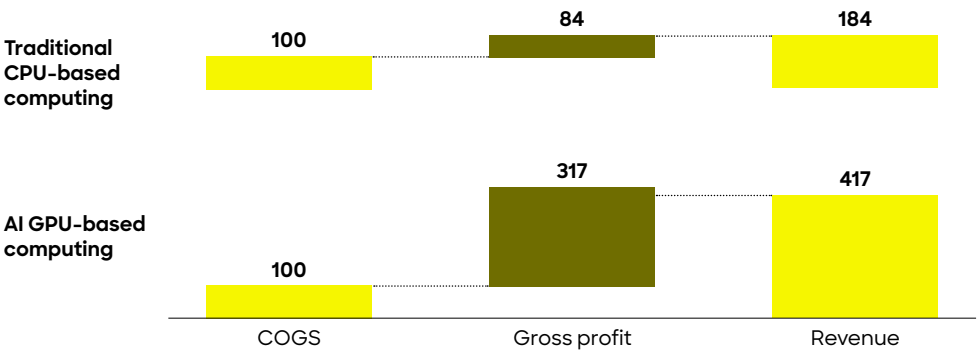
Roland Berger

This ratio of 30:1 is almost three times higher than in traditional CPU-based computing and datacenters. The only option for providers of GenAI services is to actively address their infrastructure costs early on with clever architectural choices. Cloud infrastructure providers venturing into chip and system design should aim to capture a large share of the value.

GenAI hardware is much more costly than traditional datacenter hardware

End users need to carefully manage their cost of operations

Ratio of gross profit to revenue for datacenter computing hardware [indexed, Q4, 2023]



GenAI hardware companies add 2.8x more profit than in traditional CPU-based datacenter computing – leading to the very high cost of GenAI systems

Source Roland Berger

Roland Berger

Opportunities for investment and innovation

Despite the challenges, the GenAI hardware market also offers attractive new opportunities for investment, especially for players along the semiconductor value chain. These players include semiconductor equipment and materials suppliers, chip design and manufacturing companies, hardware and system integrators, and AI service and solution providers.

Semiconductor equipment and material suppliers can benefit from growing demand for semiconductor capacity and equipment, as well as from the increasing complexity of the technologies and processes involved. These suppliers can differentiate themselves by offering more advanced, reliable and efficient solutions, and by expanding their product portfolio and customer base.

Opportunities exist for established and new chip design companies to leverage their expertise and create more customized chips for different AI applications and workloads. Although entry barriers are very high, several startups are trying to enter the market by focusing on energy efficiency.

For their part, EMS and system integrators can take advantage of rising demand for GenAI hardware and systems, and the increasing availability of chips and components. System integrators need strong ties to the chip design companies but at the same time face the threat of downstream integration by chip design companies.

AI infrastructure, service and solution providers can use GenAI systems to enhance their AI offerings, creating greater value for their customers. They can also contribute to the sustainability (or 'circularity') of the GenAI hardware and semiconductor industry by optimizing their data and computing needs and by adopting more energy-efficient, carbon-neutral practices.

Finally, industrial and technology companies need to seriously consider expanding into GenAI semiconductor design and hardware. This is particularly relevant for companies with high computing demands, for example, those working in the area of autonomous driving or humanoid robots.

Roland Berger experts in semiconductors and electronics - The Roland Berger Advanced Technology Center

At Roland Berger, we have extensive experience in the GenAI hardware and semiconductor industry. We also offer established expertise in the automotive, industrial and consumer sectors. We can help you understand the latest market trends, identify the opportunities and implement the best solutions for your business. Please contact us for more information.

Further reading

➔ *QUANTUM TECHNOLOGIES – READY FOR TAKE-OFF AT LAST?*

CONTACT:

THOMAS KIRSCHSTEIN
Partner
Hamburg Office +49 30 3992-3557
thomas.kirschstein@rolandberger.com

ALEXANDER FEY
Principal
Frankfurt Office +49 30 39927-3603
alexander.fey@rolandberger.com

FALK MEISSNER
Senior Partner
Chicago Office +1 510798-7550
falk.meissner@rolandberger.com

CO-AUTHORS:

MAURUS WÜTHRICH

KOUROSH RAJAB SEMNANI