AQ COMPUTE
AI-READY DATA CENTERS

CPUs vs GPUs?

In the world of computer hardware, there are two main components: the CPU (Central Processing Unit) and the GPU (Graphics Processing Unit). While the CPU has been traditionally used for general-purpose processing, the GPU has gained in importance with its specialization in graphics and parallel tasks processing[1]. In this article, we will take a closer look at the differences, advantages and disadvantages as well as the respective areas of application of CPUs and GPUs.

Architectures and designs

A CPU is a microprocessor that is responsible for executing instructions and processing data in a computer. It acts as the "brain" of the system and interprets and executes instructions provided by software. To do this, it fetches and decodes data from the memory of the computer. It then executes the commands received from the software.

In contrast, a GPU is specifically designed to process graphics and visualizations, allowing it to handle complex image processing, rendering and other graphics-intensive tasks with high efficiency[2].

As schematically shown in Figure 1, the architecture of CPUs and GPUs are fundamentally different: CPUs are optimized for sequential processing, which means that they can perform a small number of consecutive large tasks quickly. They are very versatile, good at handling a variety of tasks, even tasks requiring complex logic and control. GPUs, on the other hand, are designed for parallel processing and can simultaneously process many small tasks at high speed[3]. CPUs enjoy an internal low latency to switch between tasks, very important for general-purpose computing. GPUs have usually higher memory compared to CPUs, allowing

---

[1] https://www.cancom.info/2023/11/deshalb-sind-grafikkarten-fuer-ki-loesungen-unverzichtbar/
[2] https://www.heavy.ai/technical-glossary/cpu-vs-gpu
[3] https://www.run.ai/guides/multi-gpu/cpu-vs-gpu; https://kariera.future-processing.pl/blog/when-processor-is-not-enough/

them to handle large amounts of data more efficiently. Since a huge amount of data is required for AI training and it is processed in batches, parallel computation and larger memory from GPUs lead to more efficient training processes.

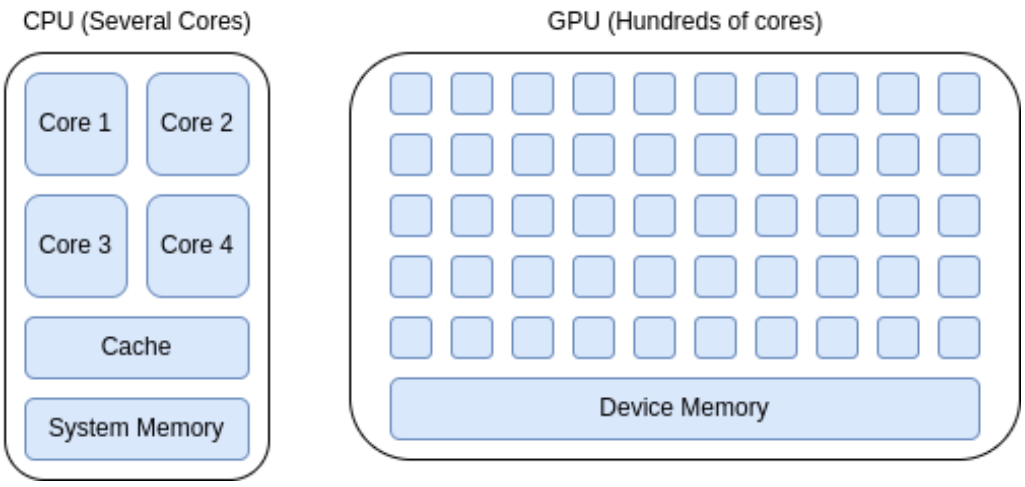

Figure 1- Comparison CPU and GPU architecture, 2024[4]

Table 1 shows a comparison of the features between CPUs and GPUs.

Table 1 – Feature comparison: CPU - GPU

CPU	**GPU**
Performs the main functions of a computer or server	Performs graphic and video rendering and AI applications
2-64 cores	Hundreds of cores
Runs a few parallel processes	Runs many parallel processes
Faster with one big task at a time	Faster with processing several small tasks[5]
Long service life	Quicker to become obsolete[6]

The concept of parallel and serial processing can be scaled up from individual components to an entire server network. Depending on the application, there are different ways in which servers can be interconnected in a

[4] Source: own elaboration
[5] https://www.cdw.com/content/cdw/en/articles/hardware/cpu-vs-gpu.html
[6] https://unicomplatform.com/blog/how-long-should-a-gpu-actually-last-expect-3-5-years/

data center. Star or ring networks can be used for large tasks with little data exchange between servers and with external networks. These are rather outdated and are only used for special use cases such as crypto mining. For HPC or AI applications, it is necessary to combine several servers into a cluster to enlarge the parallel processing capacities needed. These servers must be able to communicate with each other as quickly as possible, which is why mesh networks or fully connected networks are preferred[7].

Applications

The CPU, as a general purpose processor, is good at processing a small number of complex tasks. The basic tasks of the CPU range from simple mathematical operations such as subtracting, multiplying and adding, to coordinating the flow of data within a computer and running the operating system. More specialized tasks include video and audio decoding, encryption and decryption algorithms and SIMD (single instruction, multiple data) tasks. For example, a typical multi-threaded benchmark (i.e. a test in which all CPU cores are utilized) is OpenSSL[8], in which an RSA key[9] with a bit length of 4096 is calculated[10].

The GPU was originally created for graphically displaying the commands processed by the CPU on a screen[11]. However, as GPUs can compute with high degrees of parallelization thanks to the large number of computing cores, most AI algorithms and deep learning models consisting of "neural networks"[12] use them. Neural networks consist of millions of nodes and connections, which can best be simulated using many cores. Parallelization also has the advantage when training such models, since multiple scenarios can be calculated simultaneously thus reducing training time[13]. Another major advantage of the GPU is the "Video Random Access Memory" (VRAM). This is a fast buffer memory into which the large training data and results can be loaded. The VRAM can comprise up to 24 GB of memory, and is therefore considerably larger than the

[7] https://www.hpcwire.com/2019/07/15/super-connecting-the-supercomputers-innovations-through-network-topologies/
[8] https://en.wikipedia.org/wiki/OpenSSL
[9] https://en.wikipedia.org/wiki/RSA_(cryptosystem)
[10] https://aws.amazon.com/what-is/cpu/?nc1=h_ls
[11] https://www.intel.com/content/www/us/en/products/docs/processors/what-is-a-gpu.html
[12] https://news.mit.edu/2017/explained-neural-networks-deep-learning-0414
[13] https://www.run.ai/guides/gpu-deep-learning

cache memory of the CPU which is often in the two-digit MB range[14].

HPC and AI

High-performance computing (HPC) and artificial intelligence (AI) are both separate concepts yet closely linked at the same time. In the past, HPC was mainly focused on scientific research. Supercomputers were used in universities and research institutes to carry out complex calculations such as simulations. These supercomputers were traditionally mostly based on CPUs. AI is a subset of HPC and is becoming more developed in the commercial field, employing mostly GPUs. AI is changing the requirements for computing clusters in commercial data centers, since the hardware they use and the cooling requirements they have are different from classic cloud computing. AI, HPC, their similarities and differences, will be remarked in a separate article.

Power and efficiency

In addition to performance, energy efficiency and thermal aspects play an important role for CPUs and GPUs. In this context, there are two criteria, which are explained below: thermal design power (TDP) and general efficiency.

TDP is measured in watts (W) and represents the maximum theoretical heat output a component can produce. It is important to remember that TDP is not the average power consumption, but equals the maximum power consumption under the maximum load that the processor can support. Components which can handle intense workloads often have a higher TDP and generate more heat. CPUs and GPUs will use less power during low intensity tasks like browsing the web or checking emails. Generally, due to the demanding processing nature of graphics processing, GPUs tend to have higher TDPs compared to CPUs.

TDP plays a key role in choosing the right cooling solution for a system. Components with higher TDPs might require more efficient cooling systems such as liquid cooling, while those with lower TDPs might function well with air coolers.

---

[14]https://www.pcworld.com/article/2066872/how-does-cpu-memory-cache-work.html#:~:text=A%20good%20base%20for%20the,perfectly%20fine%20for%20most%20purposes.

As Figure 2 shows, there is an upward trend in TDP values over time. While this trend tends to be lower for CPUs, the V100 and H100 GPUs, for example, have seen their TDPs double in a short timeframe. TDPs are expected to rise to up to 4000 W by 2030[15]. Such high TDPs will place special demands on cooling and make liquid-based cooling systems necessary.

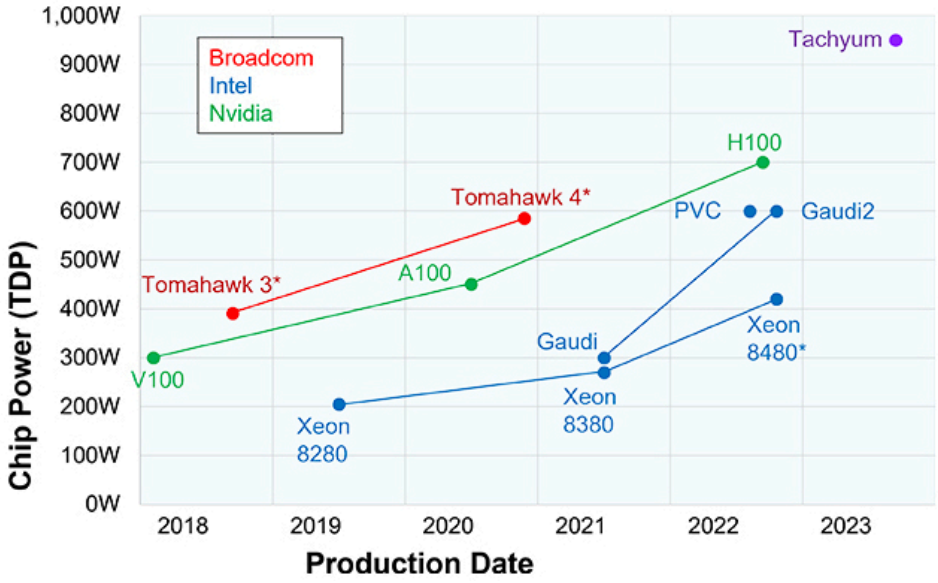

Figure 2 - TDP evolution per chip manufacturer and model[16]

Computing efficiency

While TDP helps us understand a component's thermal output, it does not tell the whole story when it comes to performance. That is where efficiency comes in. Here, efficiency refers to a component's ability to get the most work done while consuming the least amount of power, ideally delivering high performance with low power consumption.

Unfortunately, there is no single, universally accepted metric for measuring CPU and GPU efficiency. For CPUs, metrics like instructions per watt (IPS/W) or floating-point operations per watt (FLOPS/W) are

[15] https://www.ri.se/en/news/blog/generative-ai-must-run-using-liquid-cooling
[16] Source: Techinsights. https://www.techinsights.com/blog/editorial-its-getting-hot-here

commonly used. These metrics indicate how many instructions or calculations a component can perform per watt of power consumed. A higher number signifies better efficiency[17].

For GPUs, frames per second per watt (FPS/W or FPW) can be considered for measuring efficiency. This metric reflects how many frames a GPU can render per watt of power, giving you an idea of its computing efficiency. A higher value equals better efficiency, but not always higher performance.[18].

One way to measure the performance of a server, considering its different components and tasks, is the SERT™ (Server Efficiency Rating Tool) which was developed by SPEC® in collaboration with the U.S. Environmental Protection Agency.

Innovations and trends

The IT hardware sector continues to evolve rapidly. One of the most exciting innovations is the so-called "multi-chiplet design". Instead of a single large monolithic chip, several chips that are easier to produce are combined to provide the same functionality (see Figure 3). These chiplets are more flexible than conventional chips because individual functional groups can be replaced depending on requirements[19, 20].

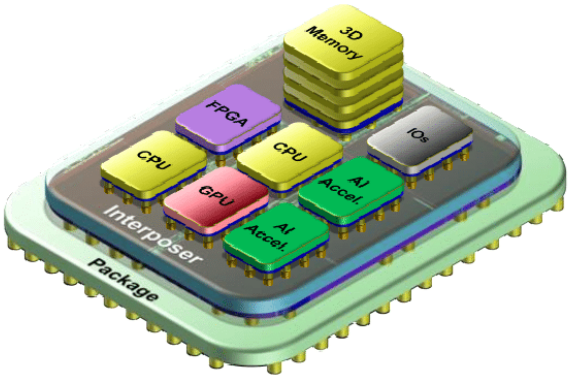

Figure 3 - Example of the chiplet partitioning concept[21]

[17] Source: https://www.intel.com/pressroom/kits/core2duo/pdf/epi-trends-final2.pdf
[18] https://lazer3d.com/2023/06/27/rtx-4060-the-new-power-efficiency-king/
[19] https://www.all-electronics.de/automotive-transportation/diese-vorteile-bieten-chiplets-gegenueber-socs-auch-im-auto-571.html
[20] https://resources.pcb.cadence.com/blog/2023-all-about-chiplet-technology
[21] IntAct: A 96-Core Processor With Six Chiplets 3D-Stacked on an Active Interposer With Distributed Interconnects and Integrated Power Management, Vivet et al., 2020, IEEE Journal of Solid-State Circuits

In semiconductors, an important factor is the structure width, which indicates how many transistors can be arranged on a certain area. Put simply, more transistors means more computing power[22]. While Intel has been producing structure widths of 7nm since 2022, TSMC made the leap to 3nm structures in 2022. TSMC uses this structure to manufacture the M3 CPU for Apple, for example[23]. It is expected that TSMC will be able to realize 2nm structure widths from 2025[24].

The GPU market is dominated by NVIDIA in both the data center and consumer markets. NVIDIA has now introduced the new generation of "Blackwell" GPUs. According to Ian Buck (Nvidia's vice president), the new model enables AI models to be trained four times faster and with up to 25 times better energy efficiency than the previous Hopper architecture[25].

AMD is also making an impact on this market with its MI300 series chipsets. However, this product is not a pure GPU. For reference, the AMD Instinct MI300A combines 228 GPU compute units and 24 "Zen 4" x86 CPU cores[26]. For Q1 2024, sales of the MI300 totaled USD 2.3 billion out of USD 5.45 billion in sales, making it AMD's best-selling product[27, 28].

Intel is also developing in this area and is offering an AI accelerator (not a GPU) with the Intel Gaudi 3 chipset line, which allegedly delivers 50% faster AI performance with 40% better energy efficiency than the H100 from NVIDIA. Interestingly, the Gaudi has a structure width of 5nm[29].

In some instances, AI computing uses general-purpose hardware such as CPUs and GPUs [30]. However, one other possible approach is to customize the hardware for its specific application. One subfield of computer science that has been developed for a specific application is neuromorphic hardware, designed to mimic the structure and function of natural brains. For example, the Intel Loihi 2 can simulate 1 million neurons

[22] https://ourworldindata.org/moores-law
[23] https://www.golem.de/news/apple-m3-ein-wunder-war-nicht-zu-erwarten-2311-178998.html
[24] https://www.heise.de/news/TSMC-Diese-Verbesserungen-bringen-2-Nanometer-Strukturen-9054463.html
[25] https://www.datacenterknowledge.com/hardware/nvidia-launches-next-generation-blackwell-gpus-amid-ai-arms-race
[26] https://www.amd.com/en/products/accelerators/instinct/mi300.html
[27] https://www.fierceelectronics.com/ai/amd-mi300-ai-shipments-helped-drive-q1-revenue-growth
[28] https://www.theverge.com/2024/4/30/24145856/with-1b-in-sales-amds-mi300-ai-chip-is-its-fastest-selling-product-ever
[29] https://wccftech.com/intel-gaudi-3-ai-accelerator-5nm-128-gb-hbm2e-900w-50-percent-faster-nvidia-h100/
[30] https://www.embedded.com/optimizing-embedded-edge-ai-with-neuromorphic-computing/

on each 31 mm^2 chip[31]. For scale, a mouse brain has 71 million of neurons (14 million of which are in the cortex)[32].

On the other hand, XPUs, a term encompassing a variety of specialized processing units such as FPGAs (Field-Programmable Gate Arrays) and TPUs (Tensor Processing Units), are also emerging to handle specific computational tasks more efficiently. These units offer customizable hardware solutions, accelerating applications in machine learning, data centers, and high-performance computing.

Conclusions

CPUs and GPUs are two of the most important components in data centers. CPUs are all-rounders that perform general tasks such as running the operating system and programs as well as performing calculations. The strength of CPUs lies in the sequential processing of a small number of complex tasks. They have a small number of cores (1-64), offering high performance. Graphics processors, on the other hand, specialize in visual tasks, including graphics processing and rendering, and are currently becoming increasingly important for artificial intelligence. Their strength lies in the parallel processing of many small tasks. This is made possible by the high number of cores (hundreds) and high graphics memory.

The newest GPUs, CPUs, and emerging hardware like XPUs and neuromorphic systems are revolutionizing the computing arena. CPUs will keep increasing their efficiency and performance through advanced architectures, staying as the most versatile option for general-purpose tasks. GPUs, essential in AI, big data analytics and simulation applications, will likely see further improvements on components such as memory bandwidth, and on core designs specifically conceived for machine learning.

As software and workloads become more heterogeneous, the demand for XPUs will grow, developing specific hardware to respond to each specific application needs. Neuromorphic hardware will transform how machines interpret their environment and will likely drive innovations in robotics, autonomous vehicles, and cognitive computing. Specific hardware architectures should allow more energy-efficient computing paradigms.

---

[31] https://www.intel.com/content/www/us/en/research/neuromorphic-computing-loihi-2-technology-brief.html
[32] https://www.intel.com/content/www/us/en/research/neuromorphic-computing-loihi-2-technology-brief.html

The future of hardware for computing will facilitate specialization and diversification. The advancements in this field will enable current applications to be faster and better, enabling new applications and technologies. The boundaries of physical limitations are being pushed, generating new architectures allowing incredible processing capacities, narrowing the distance between the performance of human brains and the newest computers being manufactured today.