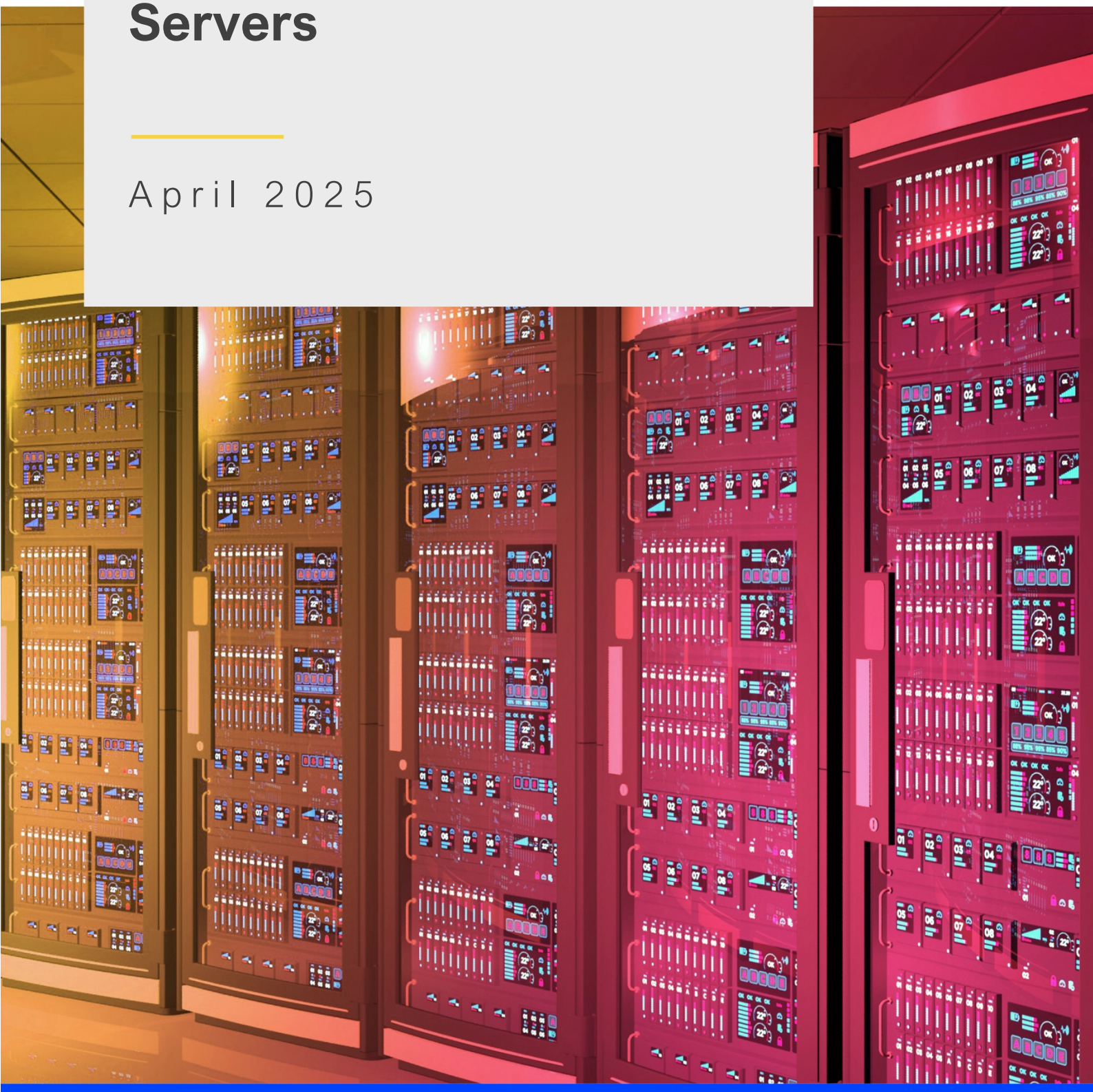
4E Efficient, Demand Flexible Networked Appliances EDNA

Energy Efficiency of Servers

April 2025



Technology Collaboration Programme
by iea

The Technology Collaboration Programme on Energy Efficient End-Use Equipment (4E TCP), has been supporting governments to co-ordinate effective energy efficiency policies since 2008.

Fourteen countries and one region have joined together under the 4E TCP platform to exchange technical and policy information focused on increasing the production and trade in efficient end-use equipment. However, the 4E TCP is more than a forum for sharing information: it pools resources and expertise on a wide a range of projects designed to meet the policy needs of participating governments. Members of 4E find this an efficient use of scarce funds, which results in outcomes that are far more comprehensive and authoritative than can be achieved by individual jurisdictions.

The 4E TCP is established under the auspices of the International Energy Agency (IEA) as a functionally and legally autonomous body.

Current members of 4E TCP are: Australia, Austria, Canada, China, Denmark, the European Commission, France, Japan, Korea, Netherlands, New Zealand, Switzerland, Sweden, UK and USA.

Further information on the 4E TCP is available from: **www.iea-4e.org**

The Efficient, Demand Flexible Networked Appliances Platform of 4E (EDNA) provides analysis and policy guidance to members and other governments aimed at improving the energy efficiency and demand flexibility of connected devices and networks.

Further information on EDNA is available from: **www.iea-4e.org/edna**

This report was commissioned by the EDNA Platform of the 4E TCP and authored by Vlad Coroamă and Simon Hinterholzer. The views, conclusions and recommendations are solely those of the authors and do not state or reflect those of EDNA, the 4E TCP or its member countries.

EDNA Study DC1d – March 2025

Energy Efficiency of Servers

Past and Possible Future Trends

Vlad C. Coroamă* and Oana Dumbravă, Roegen Centre for Sustainability
Simon Hinterholzer*, Kejsi Progni and Ralph Hintemann, Borderstep Institute

***The main authors, Vlad C. Coroamă and Simon Hinterholzer, contributed equally to this work.**

image by Brett-Sayles via pexels.com

Content

Suggested citation

Coroamă, V. C., Hinterholzer, S., Progni, K., Dumbravă, O., & Hintemann, R. (2025). *Energy Efficiency of Servers—Past and Possible Future Trends*. *EDNA – IEA 4E TCP*.

Glossary

AI	artificial intelligence
ARM	advanced RISC machines
ASIC	application specific integrated circuit
AVX-512	advanced vector extensions instructions for x86 instruction set architecture
CAGR	compound annual growth rate
CDNA GPU	compute centric GPU (abbreviation for Compute DNA)
CISC	complex instruction set computing
CISC	complex instruction set computer
CoWoS	chip on wafer on substrate
CPU	central processing unit
CRAC	computer room air conditioner
CRAH	computer room air handler
DC	data center
DDR4	double data rate 4 synchronous dynamic random-access memory
DDR5	double data rate 5 synchronous dynamic random-access memory
DIMM	dual in-line memory module
DRAM	dynamic random-access memory
DVFS	dynamic voltage and frequency scaling
EE	energy efficiency
EI	energy intensity
EUV	extreme ultraviolet (refers to a special lithography system in this document)
FLOP	floating point operation (GFLOPs = GigaFLOPs; one billion FLOPs)
FP64, FP32, FP16, BF16	different precision types; described in chapter 3.1.2
GaN	Gallium-Nitride
GDDR	graphics double data rate
GDP	gross domestic product
GHG	greenhouse gas
GPGPU	general-purpose computing on graphics processing units
GPU	graphics processing unit
HBM	high bandwidth memory
HDD	hard disc drive
high-NA	high numeric aperture
HPC	high performance computing
IEA	international energy agency
IMC	in-memory computing

IO	input/output
ISA	instruction set architecture
IT	information technology
JVM	Java Virtual Machine
LAN	local-area network
LPU	logical processing units
MIG	multi-instance GPU
ML	machine learning
MXUs	matrix-multiply unit
NUMA	non-uniform memory access
NVMe	non-volatile memory express
OPS	operations
PCIe	peripheral component interconnect express
PoW	proof-of-work
PSU	power supply unit
RAM	random-access memory
RISC	reduced instruction set computer
SERT	Server Efficiency Rating Tool
SHA-256	secure hash algorithm (in this case with 256 bit key)
SIMD	single instruction, multiple data
SMT	simultaneous multi-threading
SSD	solid state drive
SSJ	server-side java
TCP	4E Technology Collaboration Programme
TDP	thermal design power
TH/s	tera hashes per second
TPU	tensor processing unit
TSV	through silicon vias
TWh	Terawatt hours
UPS	uninterruptable power supply

Executive summary

Due to current trends in computing, the data stored and processed in data centers (DCs) will continue to grow substantially. These trends include cloud computing, video streaming, the Internet-of-Things (IoT), and in particular artificial intelligence (AI) and machine learning (ML). Mitigating the growing demand, efficiency gains are able to partly offset it and limit the growth of data center energy consumption.

As servers are the main devices responsible for electricity demand in data centers, their energy efficiency is particularly relevant for both the current status and future developments. For several reasons, however, it is notoriously difficult to assess, e.g.: miscellaneous types of hardware and computing workloads, lack of standardized methodologies and assumptions, consistent system boundaries, technological dependencies with other physical and software components, and the dynamic nature of the data center domain and rapid innovation cycles.

The main objective of this study is thus to provide insights into the development of server energy efficiency over the past decade. In order to do so, it first distinguishes three types of devices: general purpose (sometimes also called "volume servers") servers, GPU- and TPU-based accelerated computing servers deployed in AI and high-performance computing (HPC), and application-specific integrated circuit (ASIC) servers typically deployed in cryptocurrency mining.

For each of these main categories of servers, one or two efficiency metrics are deployed as follows: server-side Java operations per second per Watt [SSJ_OP/s/W] and the widely used SERT 2 efficiency metric for general purpose servers, gigaflops per second per Watt [GFLOPs/W] for accelerated computing, and Terahashes per second per Watt [TH/s/W] for ASICs. Due to their specific characteristics, general purpose servers were further divided into one-chip, dual-chip and multiple-chip servers.

The analysis examines the energy efficiency development of servers (or of the chips themselves, where no server data is available) over the last decade in each of these categories and using the corresponding metric(s). To this end, data was gathered on both performance and the power consumption for all categories. For general purpose servers, the data sources were two large databases with hundreds of servers each, which benchmark the SSJ_OP and the SERT 2 metrics, respectively. The first one is a performance metric that needs to be related to power consumption, while the second is an efficiency metric. For GPUs/TPUs and ASICs no such databases exist, and the data was gathered from producer datasheets and third-party sources. The analysis was complemented by literature research and, for the future trends in particular, by interviews with key experts.

The results show a compound annual growth rate for efficiency of 26% for general purpose servers, 49% for accelerated computing chips (for FP16/BF16 data representations), and 47% for ASICs. Representative for all the categories and metrics analyzed within the study, Figure ES 1 shows the efficiency development of the (unitless) SERT 2 efficiency for general purpose servers with more than two CPUs (left) and of the FP16/BF16 efficiency [GFLOPs/Watt] in GPUs and TPUs (right).

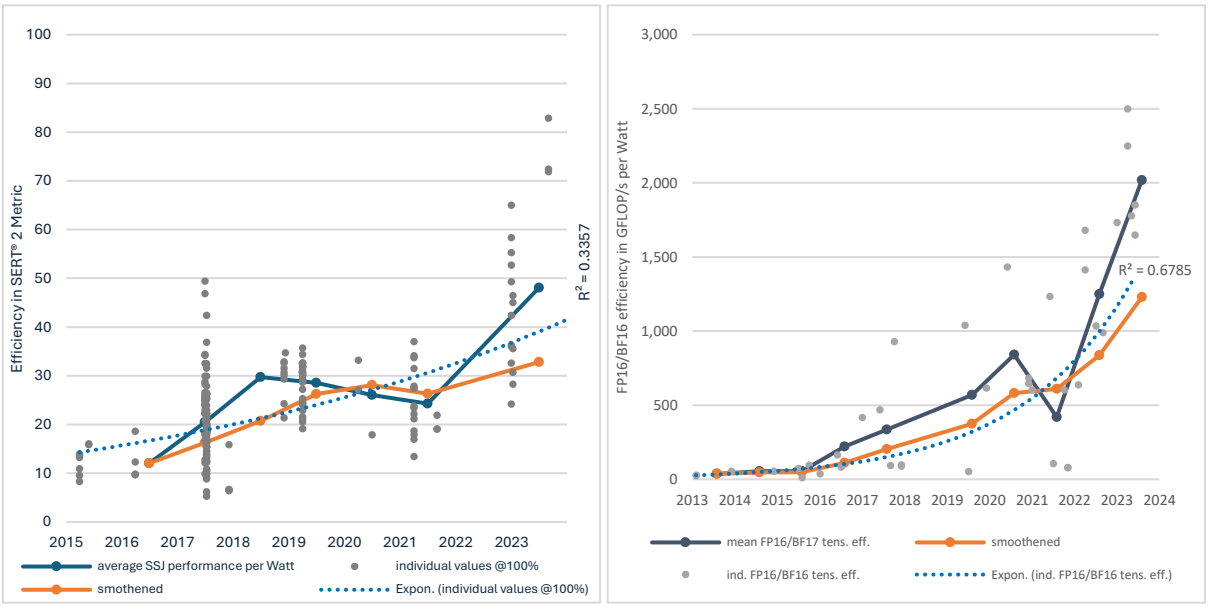


Figure ES 1 Development of the SERT 2 efficiency for general purpose servers with more than two CPUs (left) and of the FP16/BF16 efficiency [GFLOPs/Watt] in GPUs and TPUs (right). Both graphs show the development of the yearly mean raw (blue) and smoothened (orange) as well as the individual values (gray dots).

For each category and metric, the study also presents the development of performance and power consumption. This is not applicable to the SERT 2 metric for general purpose servers, which is directly an (unitless) efficiency metric. For the other categories, efficiency is derived in a third step by dividing performance through power consumption. Performance has generally evolved more rapidly than efficiency, partly explained by the also growing average power consumption in all categories.

Across all hardware types, the main efficiency driver was the ongoing miniaturization of chips circuits, followed by faster, low-latency and task-focused memory. While also important in terms of total power consumption, other efficiency factors such as more efficient power supply units and voltage conversion or the beginning uptake of liquid cooling played a less significant part. For accelerated computing, however, a key role was also played by novel architectures tailored to the needs of machine learning.

For the near future (i.e., by 2030), these trends are expected to continue. There is still improvement potential for all the major as well as the minor efficiency drivers, and no new radically different paradigms could be identified. Correspondingly, and assuming the improvements in efficiency will proceed with the same average growth rate (26% per year in SERT 2 over all systems), the SERT 2 efficiency could reach a value of around 200 in 2030. Beyond five years into the future, the uncertainties become too high to allow for rigorous projections.

1 Introduction

Published estimates of the energy consumption of data centers (DCs) for 2020 diverge by a factor of six, from around 200 TWh/year (Masanet et al. 2020; Malmodin et al. 2024) to nearly 1,200 TWh (Belkhir and Elmeligi 2018; The Shift Project 2019), with more balanced views placing this consumption around 375 TWh/year (Hintemann and Hinterholzer 2019) in 2018, when using a wider definition of data centers (e.g. including small server rooms and crypto mining). For near-term projections, the uncertainty grows quickly, with 2030 scenarios diverging by a factor of 40 (Bremer et al. 2023; Kamiya and Coroamă 2025). These wide-ranging and inconsistent estimates are a source of misunderstanding for the public and pose major challenges for thoughtful policymaking.

Given trends such as cloud computing, artificial intelligence (AI) and machine learning (ML), video streaming and the Internet-of-Things (IoT), the data stored and processed in DCs will keep growing at a substantial pace. Until a few years ago, various types of efficiency gains could largely mitigate this growth. Consequently, global DC energy consumption grew only moderately. Since about 2017-2018, however, efficiency gains could no longer keep up with the increasing demand, and the global DC energy consumption is again on the rise. The estimated total data center electricity use of 60 of the largest operators has doubled between 2018 and 2023, while the combined DC electricity use for the four largest operators has more than tripled (Kamiya and Coroamă 2025). In this context, improving the energy efficiency of data centers will continue being of great importance (Hintemann and Hinterholzer 2019); and understanding its main drivers is a necessary prerequisite.

Servers are responsible for around 70 to 80 % of the electricity demand of IT devices in DCs (Shehabi et al. 2016; Masanet et al. 2020; Hintemann et al. 2020; 2022). Their energy efficiency prospects are thus crucial for the future developments. Due to miscellaneous computing workloads, configurable hardware and software, technological dependencies with other physical/software elements or between server and software, the energy efficiency of servers is difficult to assess. It remains a complex topic and the processes behind it are poorly understood in many aspects.

1.1 Purpose and significance

Solid assessments of current and (to the extent possible) future server efficiency trends are crucial for an accurate understanding of current and future energy demands of data centers. However, this task is hindered by several challenges. The lack of standardized methodologies and assumptions makes comparisons between different studies difficult. Defining system boundaries and considering technological dependencies with other physical and software components presents another obstacle. Furthermore, the dynamic nature of the data center domain and rapid innovation cycles make long-term predictions particularly challenging.

In this context, the main objective of this study is to provide updated insights into the development of the energy efficiency of servers in their operation phase. A thorough breakdown of the historical developments and their drivers is followed by an analysis of the current and expected future state of these drivers, and of the expected further technological developments. For the different server types, various metrics and their significance will also be analyzed in this context. Finally, based on this analysis, a rough estimate for further development of server efficiency by 2030 is provided.

1.2 Main objectives and research questions

This study examines the historical development of energy efficiency in servers during their operation phase (other life cycle phases, in particular raw material extraction and production, are not considered). Next to further insights from literature review and expert interviews, this analysis also informs the assessment of possible future developments.

The research aims to answer several key questions:

- How has server energy efficiency evolved over the past 10 years?
- What factors have had the greatest impact on this efficiency?
- Based on these trends and influencing factors, what is a plausible projection for server energy efficiency up to 2030?
- Finally, how reliable are these estimates given the available server measurement data and what are the key uncertainties that may affect future projections?

The results will further inform EDNA's upcoming Total Economic Model (TEM) version 4.0. EDNA is the Efficient, Demand Flexible Networked Appliances platform of the 4E Technology Collaboration Programme (TCP) of the International Energy Agency (IEA). The platform provides analysis and policy guidance aimed at improving the energy efficiency and demand flexibility of connected devices and networks, reflected, for example, in previous TEM versions v1.0 (Ryan, Smith, and Wu 2019) and v2.0 (Ryan, Smith, and Wu 2021).

2 Scope of the study

Together with networking and storage systems, servers are fundamental IT elements in data centers. To discuss their energy efficiency in Section 4, we first define servers more formally and how they relate to the rest of the data center ecosystem (Section 2.1). Section 2.2 then explores the energy efficiency of servers, deriving it from more general definitions of energy efficiency. After Section 2.3 addresses the main types of servers covered by this study, Section 2.4 finally discusses the scope of the study more broadly, showing what lies within its boundaries, and what outside of them.

2.1 Physical servers and the data center ecosystem

A server is a dedicated, high-performance computer specifically designed to run server software and deliver various services to other computers or devices over a network. It is built with powerful processors (such as CPUs, GPUs, TPUs, chiplets, or others), extensive RAM, and substantial storage, along with multiple network interfaces and redundant power supplies for maximum reliability.

Servers are designed for minimal downtime and can be scaled to meet increasing demands. They also offer advanced remote management capabilities and robust security to protect critical data. They are used in a wide variety of applications, from hosting websites and databases to supporting cloud computing, machine learning, or cryptocurrency mining. They are essential for businesses and organizations that require high performance, reliability, and control over their IT infrastructure.

Most aspects of server energy efficiency can be discussed in this narrow sense of server of physical device. Some of the main factors influencing such a narrow perception of servers as physical devices are shown in Figure 1.

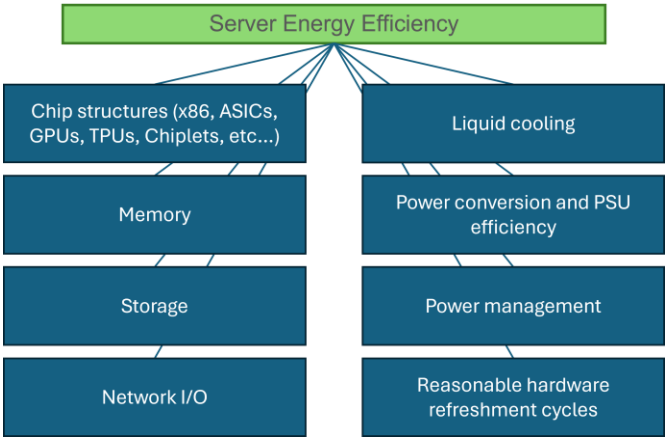


Figure 1 Main factors influencing the energy efficiency of servers as physical devices in a narrower sense.

Servers, however, are part of the larger data center ecosystem. The overall system efficiency (defined as computational output per energy input, see below) is influenced by numerous factors of this wider ecosystem. As shown in Figure 2, these factors include software aspects such as the operating system and the types of applications running on the server, and in particular numerous features of the DC power delivery and cooling infrastructure, such as power generation and distribution, the design of the Uninterruptable Power Supply (UPS) units as well as deployed chillers, ventilators, pumps, heat exchangers, etc. In the widest of senses, even the climate the DC resides in has a strong say in overall system efficiency.

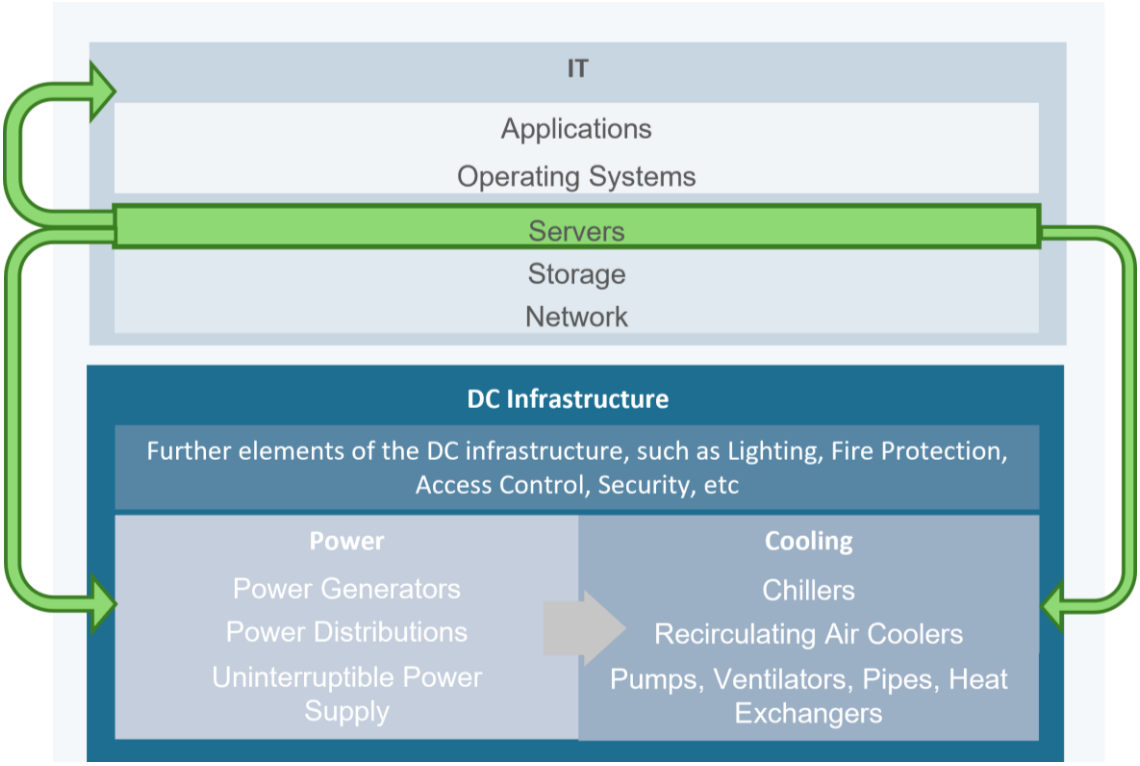


Figure 2 Schematic representation of a server's system interdependencies with software and other physical domains in the data center. They include efficiency interdependencies towards software (e.g., software efficiency, virtualization, CPU power management such as P-/C-states, dynamic voltage and frequency scaling) and efficiency interdependencies towards the data center infrastructure (such as air/liquid cooling, heat reuse, and UPS).

2.2 Energy efficiency of servers

In physics, energy efficiency (EE) typically describes the efficiency of energy conversion[1] and is defined as

$$EE = \frac{useful\ energy\ output}{total\ energy\ input} \tag{1}$$

More generally, however, energy efficiency of a system is the ratio of useful output, which can be in any form (e.g., work, products, services), to the total energy input required to generate that output. So, while the input is always energy, the output can take any quantifiable form:

$$EE_O = \frac{useful\ output}{energy} \tag{2}$$

In transportation, for example, energy efficiency is defined as the distance travelled by passengers or goods divided by the energy input needed to acquire it.[2] For buildings, it often describes the provision of a comfort temperature for a given surface in a building divided by the required energy input.[3] And the energy efficiency of an entire economy is measured as the amount of GDP generated for the energy input.[4]

[1] See https://www.studysmarter.co.uk/explanations/physics/energy-physics/efficiency-in-physics/.
[2] See https://en.wikipedia.org/wiki/Energy_efficiency_in_transport.
[3] See https://www.geze.com/en/discover/topics/energy-efficiency-of-buildings.
[4] See https://ec.europa.eu/eurostat/cache/digpub/energy/2019/bloc-4b.html.

In the computing domain in general and for servers in particular, the useful output are computations: "Energy efficiency refers to maximizing the amount of computational work completed for the amount of energy consumed".[5] This is represented in Equation 3:

$$EE_{compute} = \frac{computations}{energy} \tag{3}$$

According to Equation 3, computational efficiency is the computational performance of an individual server (*output*) relative to its energy consumption (*input*). As with all efficiencies, higher values are better.

The inverted efficiency yields energy intensity (EI) of computation (or "computational intensity" in short), which represents the amount of energy required for a given number of computations. For the computational intensity, lower values are better:

$$EI_{compute} = \left(EE_{compute}\right)^{-1} = \frac{energy}{computations} \tag{4}$$

While the input (i.e., energy consumption) is conceptually straightforward, the output (i.e., computational performance) is difficult to define, as various "compute products" are possible such as computations, storage, database analyses, encryption or cryptography, IO/network bandwidth, etc. Accordingly, the computational performance of servers can be measured and evaluated very differently depending on the application it serves. Hence, Section 2.3 below discusses the types of servers under scrutiny in this study. As part of the deployed methodology, Section 3.1 then defines and discusses the energy efficiency metrics used for each of these server types.

2.3 Main server types

The evolution of servers is closely intertwined with the evolution of computers and networks. The advent of minicomputers in the 1960s and 1970s, which were much more affordable than mainframes,[6] allowed for the first time a more decentralized computing. The rise of file servers in the late 1980s then enabled efficient sharing of files across the local-area network (LAN).[7] Arguably one of the most notable developments regarding computers in data centers were the advent of rack-mounted servers in the 1990s and of virtualization in the late 2000s.[8] They brought about new possibilities for growth and efficiency through enhanced scalability and simplified management and paved the way for today's era of cloud computing. Currently, it can be assumed that 2+ socket servers make the largest share of server types, but 1-socket servers and GPU accelerated servers are expected to grow faster (Shehabi et al. 2024).

2.3.1 CPU-based general-purpose servers

Over the past few decades, general-purpose servers, traditionally CPU-based, have undergone significant evolution driven by advancements in microprocessor architecture, fabrication technologies, and system integration. In the 1990s, server performance was primarily limited by single-core processors following the von Neumann architecture, which processed instructions sequentially. The introduction of pipelining and superscalar designs allowed processors to execute multiple instructions per cycle, enhancing efficiency. However, the growing demand for

[5] See https://www.nvidia.com/en-us/glossary/energy-efficiency/.
[6] See https://en.wikipedia.org/wiki/Minicomputer.
[7] See https://www.gartner.com/en/information-technology/glossary/file-server.
[8] See https://www.thomastechllc.com/articles/server-evolution.

higher performance outpaced the gains from increased clock speeds due to thermal and power constraints. This led to the advent of multi-core processors in the early 2000s, enabling parallelism at the chip level by integrating multiple processing units on a single die, thus improving throughput without proportionally increasing power consumption.

A key technological shift occurred with the development of virtualization technologies, which allowed multiple operating systems and applications to run on a single physical server, improving hardware utilization and scalability. Furthermore, the transition from conventional memory hierarchies to non-uniform memory access (NUMA) architectures enabled better handling of memory-intensive workloads by optimizing data locality. The evolution of instruction set architectures (ISAs) such as x86_64 and the increasing use of specialized instruction sets like SIMD (Single Instruction, Multiple Data) also contributed to improved performance for specific workloads. In recent years, innovations in energy efficiency, the rise of heterogeneous computing, and advancements in interconnect technologies like PCIe and NVMe have driven the performance of modern CPU-based servers, enabling them to handle increasingly diverse and compute-intensive tasks, including AI workloads, cloud computing, and high-performance computing (HPC) environments.

2.3.2 GPUs and TPUs: Accelerated computing for machine learning

Roughly at the same time while volume servers and cloud computing were becoming widespread, another crucial technology was evolving out of its infancy: machine learning (ML). While a relatively young field, ML has roots stretching back to the mid-20th century. Early pioneers like Arthur Samuel in the 1950s developed first programs that could learn from data, laying the foundation for concepts such as neural networks and reinforcement learning.[9] Progress continued through the decades, with key advancements in algorithms and techniques for pattern recognition, decision-making, and artificial intelligence.[10] Early machine learning, however, was limited by the available computing power.

A significant turning point came in the late 2000s with the rise of Graphics Processing Units (GPUs) for general-purpose computing. Originally designed for rendering graphics in video games, GPUs excel at parallel processing, performing many calculations simultaneously. About 2 decades ago, in 2003, two research groups have independently shown (Krüger and Westermann 2003; Bolz et al. 2003) how this capability of GPUs can be used for general-purpose computing in a paradigm known as “General-purpose computing on graphics processing units” (GPGPU).

This capability was also ideally suited for the matrix operations at the heart of most machine learning algorithms (Steinkraus, Buck, and Simard 2005; Chellapilla, Puri, and Simard 2006). By harnessing the parallelism of GPUs, training complex models like deep neural networks became significantly faster and more efficient (Raina, Madhavan, and Ng 2009). This potential was quickly recognized by GPUs producers, who started to first advertise and then increasingly tailor GPU architecture towards the needs of ML algorithms. To some extent this applies to workstation and enterprise-oriented GPUs, but mainly to the increasingly important data-center-grade server GPUs.[11] While most of the quickly expanding ML industry relies on GPUs, Google decided back in 2014 to develop its own custom-built chips for machine learning: the Tensor Processing Units (TPUs)[12]. TPUs are application-specific integrated circuits (ASICs)

[9] See https://www.tableau.com/data-insights/ai/history.
[10] See https://www.sparkfun.com/news/7896.
[11] See https://www.nvidia.com/en-us/data-center/products/.
[12] See https://cloud.google.com/transform/ai-specialized-chips-tpu-history-gen-ai.

designed to accelerate machine learning workloads deployed across Google's data centers.[13] For accelerated computing, our study covers both GPU- and TPU-based servers.

2.3.3 Hash-performing ASICs for proof-of-work in cryptocurrency mining

The TPUs addressed above are one flavor of ASICs. Another type that is widespread in servers are ASICs used in cryptocurrency mining, especially in crypto currencies, which use the proof-of-work consensus mechanism. While in the very beginning of Bitcoin from 2008-2009 on, CPUs were used for mining, due to the advantages of the GPGPU paradigm outlined in Section 2.3.2 above, they were replaced around 2011 by the faster and more energy-efficient GPUs. These in turn were quickly outpaced by field-programmable gate arrays, which were finally replaced by ASICs.

Since about 2013, ASICs have become the dominant force in cryptocurrency mining, particularly for Bitcoin (Bedford Taylor 2017). Unlike general-purpose CPUs or GPUs, these ASICs are specifically designed to perform a single task: compute the cryptographic hash functions required to validate transactions and add blocks to the blockchain in the energy-intensive paradigm (Coroamă 2021; 2022) known as proof-of-work (PoW). Such hash-performing ASICs deployed in cryptocurrency mining are part of this study as well.

2.4 Dimensions within the scope of the study and outside its scope

After having addressed the energy efficiency of servers and discussed the types of relevant logical processing units (LPUs), we can now proceed to define the system boundaries of this study across several dimensions. Given the complexity of the relation between servers and the DCs the typically operate in addressed in Section 2.1, drawing the system boundary is not trivial in all of the dimensions. As our interest lies in the energy efficiency of servers themselves, we exclude the factors that are entirely exogeneous to servers like the physical infrastructure of DCs or software aspects such as operating system or application type.

Other software aspects, such as the server power management or virtualization, however, relate directly to the functioning of servers and thus to their efficiency – they are thus very much within the system boundaries. From the physical aspects related to cooling and power supply, one is tightly interconnected with the way servers' function: the type of cooling, whether air- or liquid-based. Unlike more distant characteristics such as the design of chillers (for air cooling) or the exact refrigerant deployed (for liquid cooling), the cooling paradigm (air or liquid) is considered within system boundaries. Table 1 summarizes the boundaries of this study along several dimensions.

[13] See https://cloud.google.com/tpu/docs/system-architecture-tpu-vm.

Table 1 An overview of this study 's system boundaries across several relevant dimensions.

Dimension	***Within the scope***	***Outside the scope***	***Comments***
Type of equipment	servers, GPU accelerators, crypto mining equipment	end-user computing equipment, Telco equipment (routers, switches), pure storage systems	
Type of logical processing unit (LPU)	CPU, GPU & TPU, dedicated computing hardware (ASICs)		All LPU types are relevant, as long as they are in servers (and not other devices). For non-x86 CPUs, data might be sparse.
Localization of energy consumption	computing, memory, storage, network I/O, PSU, server integrated fan	air cooling, lighting, UPS (i.e., all non-IT PUE contributors)	
External impact on energy efficiency of servers	liquid cooling, environment temperature in IT room	waste heat recovery, more distant technologies (e.g., type of refrigerant)	
Software	server power management and virtualization	operating system, applications	
Time horizon	historic, present, future projections		Projections have inherent uncertainties, which grow over time.
Current server efficiency	state-of-the-art of servers on the market	overall stock	State-of-the-art is more relevant than overall stock for modelling, but not always distinguishable in the data.

3 Methodology

3.1 Efficiency metrics for the individual server types

3.1.1 General purpose servers

As briefly addressed in Section 2.1 above, general purpose servers perform a large variety of tasks. Servers combine processors, memory and networking to a varying degree, and all these hardware types come in different flavors from numerous producers, each of them additionally evolving over time. Some server configurations might excel at a certain type of tasks, while other configurations will excel for different types of tasks. Comparing the energy efficiency of volume servers is thus no straightforward task.

To reflect both this complexity and the variety of server tasks, for about a decade and a half now, a synthetic benchmark which performs various types of operations has been established. It is called "SPECpower_ssj2008" (SPEC 2018). The acronym SPEC stands for "Standard Performance Evaluation Corporation", a non-profit organization that develops standardized benchmarks to measure and compare the performance of computer systems.[14] The acronym "SSJ" stands for "Server-Side Java" and represents the benchmark itself, a synthetic workload of typical (and varied) server-side Java business applications.

The benchmark proves a standardized way to evaluate and compare the efficiency of different servers and has thus become a widely recognized benchmark in the volume server industry. It also puts forward two metrics that are relevant in our context:

- The main *performance* metric it uses are *SSJ_ops* (Server-Side Java operations per second), which measure the throughput (in terms of SSJ operations per second) at different server load levels (from idle to 100% utilization). These results are then aggregated to produce the overall "SSJ_ops" score (SPEC 2018).
- The main *efficiency* metric relates this SSJ performance to the server's power consumption:

$$EE_{x86-SSJ} = \frac{SSJ_{performance}}{P} \; in \left[\frac{SSJ_{OP}/s}{Watt}\right] \tag{5}$$

- As secondary efficiency metric, the SERT 2 metric is used. The SERT 2 metric provides a comprehensive energy efficiency assessment by considering multiple workload types that represent real-world server usage patterns, including CPU, memory, and storage workloads, rather than focusing solely on peak performance scenarios. It also measures power consumption across different load levels, from idle to full utilization, giving a good picture of how the server performs under varying conditions throughout its operational lifecycle. The standardized test methodology and reporting format of SERT 2 enables fair comparisons between different server configurations and vendors, while its recognition by regulatory bodies like the US EPA's ENERGY STAR program adds to its credibility as a reliable efficiency metric. SERT 2 is a weighted average of various efficiency metrics for worklets (including different utilization levels):

[14] See https://www.spec.org/.

$$EE_{load} = \frac{Normalized\ Performance}{P} \left[\frac{1}{Watt}\right] \quad (6)$$

These in turn are summarized in CPU, memory and storage metrics, which in turn are weighted with weighting factors 65% for CPU, 30% for memory and 5% for storage to form the SERT 2 metric.

$$SERT2 = \exp(0.65 * ln(EffCPU) + 0.3 * ln(EffMemory) + 0.05 * ln(EffStorage)) \quad (7)$$

The analysis distinguishes between single-CPU servers, 2-CPU servers and large systems with more than 2 CPUs. Only the number of CPUs that is actually occupied is taken into account, so if a server has two sockets but only one CPU is installed, it is treated as a single-CPU server.

3.1.2 Accelerated computing for ML

For accelerated computing, such an established de-facto industry standard does not exist. It is probably also not needed: Their operations are more homogeneous, focusing almost exclusively on the matrix multiplications required by most machine learning models, especially by deep learning.

As proxies for server performance and efficiency, and in line with the literature (Hobbhahn, Heim, and Aydos 2023), we thus use the performance and efficiency of the hardware accelerators themselves:

- The typical *performance* metric for accelerators (in this study, GPUs and TPUs) is the number of floating-point operations (FLOPs) performed in one second, measured nowadays either in GFLOP/s (gigaflops per second) or TFLOPs (teraflops per second). We will use [TFLOP/s].
- Correspondingly, and symmetrically to volume servers, the *efficiency* metric relates this performance to the power consumption of the accelerators. As the performance often measures dozens or hundreds of TFLOPs and is divided by typically hundreds of Watts, the result would often be sub-unitary (i.e., less than 1) if we were to use [TFLOPs/Watt]. The preferred efficiency metrics is thus [GFLOPs/Watt] instead:

$$EE_{Acc} = \frac{GPU_{performance}}{P} \ in \left[\frac{GFLOP/s}{Watt}\right] \quad (8)$$

In practice, however, various ways to represent real numbers as floating points exist (on 8, 4 or 2 bytes, for example, or by using different coding on the same representation length), and the operations on these representations consequently have differing computing requirements. For the same GPU, the performances (in terms of operations per second) will thus naturally differ for the various floating-point representations. Additionally, ML algorithms can also take advantage of special circumstances, for example that the matrices that are multiplied in ML algorithms are often sparsely populated – special algorithms can thus be devised to speed up the computation. The performance for each of these cases generally needs to be devised separately. Some of the most common representations and algorithmic conditions include:

- **FP64 (Double-precision floating-point)**: Represents numbers using 64 bits, providing high precision and a wide dynamic range. Often used in scientific computing and simulations where accuracy is crucial; it requires more memory and processing power.

- **FP32 (Single-precision floating-point)**: Represents numbers as 32 bits, offering a balance of precision and efficiency. Until some years ago, it used to be the most common format for general-purpose machine learning training and inference.
- **FP16 (Half-precision floating-point)**: 16-bit-representation, requiring less memory and enabling faster computations. Often used for deep learning training and inference, especially on GPUs and TPUs with specialized hardware support. However, it has a smaller dynamic range and lower precision than FP32.
- **FP16 with sparsity**: This refers to techniques that leverage the sparsity (presence of many zeros) in neural network weights or activations to further reduce memory usage and computational costs. It can improve efficiency for specific models and tasks.
- **BF16 (Brain Floating Point)**: Also uses 16 bits, but with a different distribution. As FP16, it has 1 bit for the sign; however, while FP16 uses 10 bits for the fraction (mantissa) and 5 for the exponent, BF16 uses only 7 bits for the mantissa and 8 bits for the exponent. This format sacrifices some precision in the fractional part to accommodate a wider range of exponents. BF16 is designed for deep learning and often preferred over FP16 for certain hardware, particularly for TPUs. While it does not differ from FP16 in terms of the performance considered in this study (operations per second), it uses those operations for larger numbers, thus achieving more high-level computations.
- **Mixed precision** is a technique used in deep learning that combines different numerical precisions (e.g., FP32, FP16, BF16) during model training and inference. Using lower precision (such as FP16 or BF16) for certain operations can significantly speed up computations, especially on hardware with specialized support for these formats (like Tensor Cores in Nvidia GPUs or matrix units in TPUs). Lower precision data types also require less memory, allowing for larger models or larger batch sizes during training. By strategically using higher precision (FP32) for critical operations or parts of the model, however, mixed precision can maintain the accuracy of the final results while still enjoying many performance benefits of lower precision.

Section 3.3 discusses which of these floating-point representations and computation paradigms are relevant to our analysis, and why.

3.1.3 ASICs

For the ASICs used in hashing, the picture is much simpler. As the only relevant operation are hashes, they define both performance and efficiency:

- The *performance* metric is the number of hashes per times, measured in Terahashes (TH) per second.
- As for CPUs and GPUs earlier, the *efficiency* metric relates to the performance to power:

$$EE_{ASIC} = \frac{Hash_{performance}}{P} \ \ in \left[\frac{TH/_S}{Watt}\right] \tag{9}$$

3.2 Data collection

For the analysis of historic trends and current data, a combination of benchmarking datasets for numerous servers and technical data from manufacturers were used. To a lesser extent and only when direct measurements or manufacturer data were not available, third-party measurements or estimates were also considered. The time horizon for historic data was 10 years, i.e., 2015 – 2024.

3.2.1 General purpose servers

For volume servers, SSJ benchmarks were used from two large datasets, with hundreds of data points each: the already mentioned SPECpower dataset (SPEC 2018) and the SPEC "server efficiency rating tool" (SERT) dataset provided by ITI – the green grid.[15] Those benchmarks require to record the respective power consumption at all load stages, which makes it possible to create a dynamic load profile. The data was evaluated for possible categorization, as briefly discussed in Section 3.3 below, and then analyzed in more detail in Section 4.1.

The benchmark data included launch dates of servers. A validation check revealed that these dates were wrong in some cases, as the servers were supposed to have been launched before their chips existed. Therefore, the authors looked up all the chip launch dates and used them to correct the server launch dates. The time axis in chapter 4 refers to chip manufacturing date.

3.2.2 Accelerated computing (GPUs and TPUs)

For GPU- and TPU-accelerated servers, no such comprehensive dataset exists. We thus went to the main GPU manufacturers (i.e., Nvidia, AMD and to a lesser extent Intel) as well as Google as TPU manufacturer and gathered data on their chips designed specifically for data centers and high-performance computing (HPC). The following data was collected:

- their FP32, FP16, BF16 performances (not all were always devised),
- to highlight the importance of the advent of tensor cores, we separately collected FP16 and FP16 with tensor cores performance,
- the Thermal Design Power (TDP), which is the power consumption under maximum load,[16] and was used as standardized proxy – similarly to (Hobbhahn, Heim, and Aydos 2023) – for power consumption whilst computing the EE_{Acc} efficiency according to Equation 8,
- launch date (month and year),
- where available, also the type and size of memory as well as the type and number of cores.

Because the design and complexity for both memory and cores (and especially for the cores) have large variations, the size of memory or number of cores were not used in a quantitative analysis. The important influence of memory evolution for GPU and TPU efficiency is nevertheless qualitatively addressed in Section 5.1. Table 2 shows the individual DC and HPC-grade GPUs and TPUs analyzed for this study.

Table 2 The three GPU manufacturers as together with their DC/HPC-grade GPUs analyzed in this study, as well as Google's TPUs under scrutiny.

Producer	***GPUs (TPUs for Google)***
AMD	FirePro D300, Instinct MI 6, Instinct MI 25, Instinct MI 50, Instinct MI 60, Radeon Pro V340, Radeon Pro VII, Instinct MI 100, Instinct MI 210, Instinct MI 250, Instinct MI 250X, Instinct MI 300, Instinct MI 300X
Intel	Xeon Phi 7295, Arc A380, Arc A750, Arc A770, Habana Gaudi 2, Habana Gaudi 3
Nvidia	Kepler K80, Maxwell M40, Pascal P80, Pascal P100, Volta V100, Turing T4, Ampere A100, Ada L40, Ada L40S, Asa L4, Hopper H100, Hopper H200, Blackwell B100, Blackwell B200 Superchips: Grace-Hopper GH200 (1 Grace CPU, 1 Hopper GPU), Grace-Blackwell GB200 (1 Grace CPU, 2 Blackwell GPUs)
Google	TPU v2, TPU v3, TPU v4, TPU v5e, TPU v5p (for TPU v5e and v5p only performance data, but no consumption data available).

[15] Based on data from The Green Grid®'s TGG Server Energy Efficiency Database v02_00 of SPEC SERT® benchmark results. More details to the SERT benchmark can be found at https://www.spec.org/sert/.
[16] See https://www.intel.com/content/www/us/en/support/articles/000055611/processors.html.

3.2.3 Cryptocurrency mining equipment

Cryptocurrency mining equipment refers to specialized hardware used to solve complex cryptographic puzzles required for validating transactions and securing blockchain networks. This process, known as mining, requires substantial computational power to solve mathematical algorithms, typically based on proof-of-work (PoW) consensus mechanisms. In the early days of cryptocurrencies like Bitcoin, mining was performed on general-purpose servers or standard CPU-based systems. However, as the difficulty of mining increased, more powerful hardware became necessary, leading to the adoption of GPUs and, later, application-specific integrated circuits (ASICs) designed explicitly for high-efficiency mining tasks.

The connection between crypto mining and servers lies in their shared reliance on high computational throughput and efficiency. Initially, CPU-based servers, like those used for conventional tasks in data centers, were adapted for mining. However, as mining became more competitive, the development of dedicated mining rigs—hardware specifically optimized for solving cryptographic functions—outpaced general-purpose servers. Mining equipment has evolved rapidly, with ASICs now dominating the industry due to their superior performance-per-watt ratio. As a result, modern mining operations often require large-scale, server-like infrastructures, equipped with optimized cooling and power management, mirroring data centers in design but purpose-built for mining cryptocurrencies.

As main sources for performance, power consumption and efficiency technical specifications of manufacturers are used, as well as websites that collect such data like (hashrateindex.com 2024).

3.2.4 Future trends

“Predictions are difficult, particularly about the future”; so the adage goes. In the field of digital technologies, with its high dynamics of development and the short innovation cycles, it is particularly challenging to make long-term predictions on technological advancements. Wild cards of revolutionary technological development can always appear, in particular for highly innovative topics such as chip structures; their appearance is impossible to predict.

Conservatively, our analysis of future trends is limited to the trends that can be foreseen to a certain degree of certainty. It thus focuses mainly on the analysis of historical trends deemed as important for past energy efficiency developments (together with the question on whether these trends might continue and can thus be extrapolated for a few more years, or whether they are already showing signs of change), as well as newer trends that are foreseeable and perhaps even technologically required (such as liquid cooling for continuingly more powerful and dense servers).

The main data sources for future trends and the impact of technological developments on server energy efficiency were the insights generated during the analysis of past trends and the relevant past technological leaps, academic and industry literature research for future trends, and a few interviews with chosen experts from industry and research.

3.3 Data analysis

During data analysis, the data was first evaluated for possible categorization. For example, standards volume servers were broken down into 1-socket, 2- socket, and multi-socket servers, and the analyses were performed separately on each of these categories. Further performance

indicators and benchmark results were selected to reflect the output in different terms (such as computations, memory, etc.).

The exact performance and efficiency metrics, as well as the assessment conditions were chosen. In actual usage, servers are typically operated in a certain partial load range. Efficiency in the partial load range is thus also very relevant; for volume servers, for which this data exists, performance and efficiency were thus devised for 0% load (i.e., idle), 12,5%, 25%, 37,5%, 50%, 62,5%, 75%, 87,5% and 100% load.

For GPUs, the exact performance and efficiency metrics were chosen among those discussed in Section **Error! Reference source not found.**. As will be discussed further in the Results, the trend in ML algorithms is clearly (and has been for some time now) to trade precision for speed and large data volumes. For ML, FP64 is thus essentially irrelevant and will not be discussed.

Although FP32 is also becoming less and less prevalent and for many of the more recent GPUs, manufacturers often no longer even bother to mention their FP32 performance, we do consider it in our analysis for the historic developments. FP16 and BF16 are central to our analysis; because of their equal size and our metric (i.e., operations per second), we treat them jointly. Sparsity and mixed precision are not considered in our analysis, because they can be implemented (and interpreted) differently and are thus not well-suited for benchmarking performance or efficiency.

The changes over time in each of the thus-determined categories was then documented. The numeric results are presented in Section 4 below. For each category, we present 3 types of results: performance, power (i.e., TDP), and efficiency as the ratio of the two. For trends, we devise for each year the mean value across all servers / chips launched in that year. Due to noise in the data (in particular due to sparse data for some of the years), however, we also devise a 3-year smoothened curve.

Past trends were identified from the data analysis itself (corroborated with important shifts in performance or efficiency), literature review, and interviews. The individual historic trends were then also evaluated on possible continuation for the future analysis. The future analysis was also complemented by literature research and with insights gathered from the interviews. In total, four experts were interviewed for the study. They represent perspectives from science (building energy and cooling technologies), data center operations (hyperscale respectively colocation data centers) and a microchip and power converter producer.

4 Results: Server efficiency developments over the last decade (Quantitative analysis)

4.1 General purpose servers

4.1.1 Performance

The two diagrams provided in Figure 3 illustrate the performance trends of one-CPU (left) and two-CPU (right) servers in terms of maximum SSJ operations at 100% utilization from 2015 to 2023. Each graph shows individual performance values (for each SERT benchmarked system) as scattered dots, along with an average performance curve (blue) and a smoothed trend line (orange), allowing for an interpretation of overall developments in processing power across the observed period.

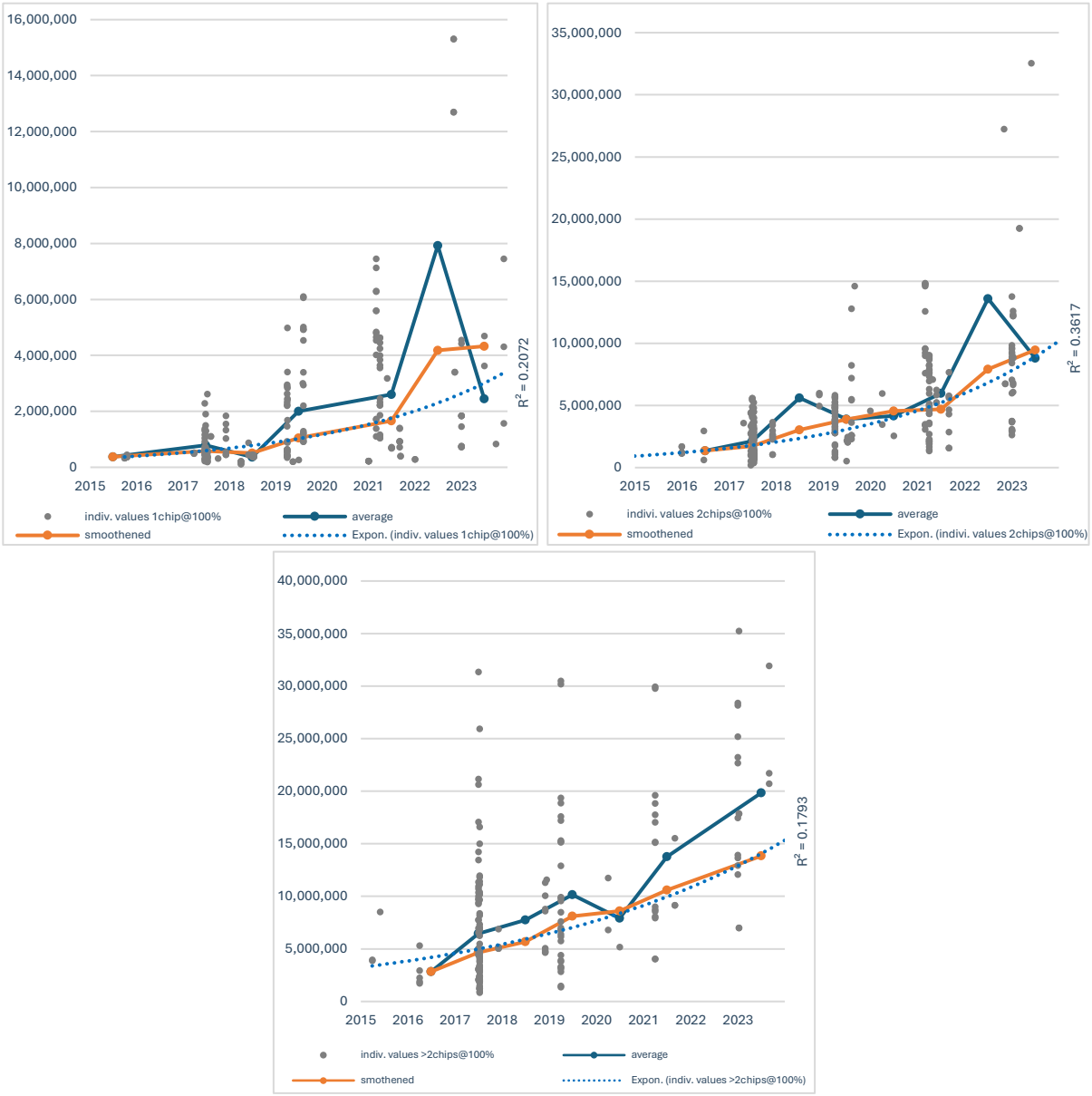


Figure 3: The growth in performance of CPU and Memory of single-CPU servers (left) and dual-CPU servers (right)) and servers with more than two CPUs (bottom) based on the SSJ benchmark.

As the market launch date (often only available as a quarter) of the CPUs installed in the servers was used to arrange the servers on the timeline, there are always a relatively large number of points horizontally above each other. Similarly, the chip generations, some of which were introduced alternately with high-performance cores (P-cores) and efficiency cores (E-cores), show an upward and downward movement in the trend, which overlays the long-term upward trend. The performance of single-CPU servers as well as dual-CPU servers has increased significantly over the last years. For single-CPU servers, the average annual growth rate was about 55% in the years between 2015 and 2022. For dual CPU servers, the growth rate in this period is slightly lower at 45% per year.

The improvements in SSJ benchmark results over recent years can be attributed to a range of advancements in server hardware and software. One of the key drivers has been processor architecture improvements, with modern CPUs offering higher core counts, increased clock speeds, and advanced instruction sets like AVX-512, all of which enhance the ability to handle parallel workloads and compute-intensive tasks. Larger, faster CPU caches have also reduced memory latency, further boosting performance.

In tandem, the memory subsystem has seen significant upgrades. Faster memory technologies such as DDR4 and DDR5 have increased bandwidth and reduced latency, allowing for quicker data access by the CPU, while more efficient memory management technologies have optimized multi-channel configurations to handle large datasets common in Java workloads which is especially relevant for the SSJ benchmark.

Multi-threading and virtualization improvements, such as enhanced Simultaneous Multi-Threading (SMT) and hyper-threading, allow servers to run multiple threads per core more efficiently. Additionally, modern Java Virtual Machines (JVMs) have introduced optimizations like better garbage collection and thread management, allowing applications to use resources more effectively. Virtualization technologies have also matured, improving the performance of containerized and virtualized workloads.

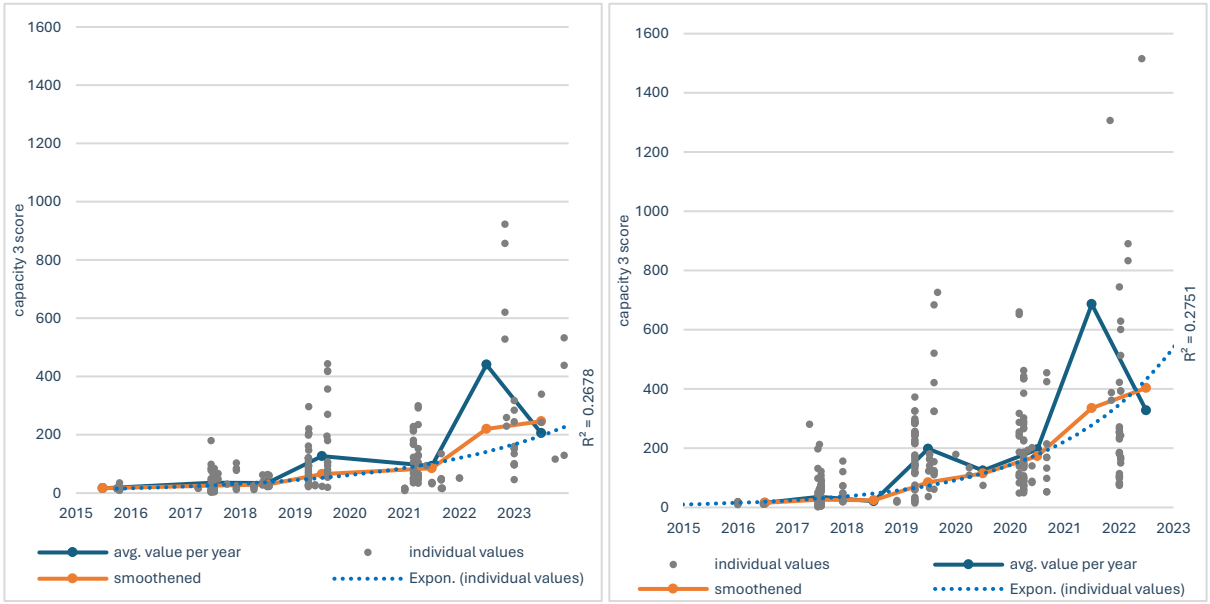


Figure 4: The results of memory capacity benchmarks for single-CPU servers (left) and dual-CPU servers (right). Not sufficient data points for servers with >2 CPUs.

I/O and storage technologies have advanced with faster interconnects like PCIe 4.0 and NVMe SSDs, greatly reducing bottlenecks in data transfer and access times, which is crucial for the data-heavy operations in Java workloads. At the same time, the sizes of typical memory have increased massively in servers. The effect of this can also be seen in the memory capacity benchmark (see Figure 4).

Improved network speeds and storage access times also contribute to higher throughput and lower latency. Software advancements, such as optimizations in operating systems, Java compilers, and modern garbage collection algorithms, have improved how effectively hardware resources are used, further contributing to better benchmark results. Power management and thermal designs have improved as well, allowing servers to run at higher speeds for longer durations without overheating.

Lastly, innovations in server architecture, including chiplet designs and hybrid CPU architectures—have enabled more scalable and efficient resource sharing. Cloud-native applications and microservices, increasingly common in enterprise environments, are also driving optimizations in hardware usage, making modern workloads more efficient.

These collective improvements in hardware, software, and architecture have driven significant gains in SSJ performance, reflecting the overall trend of more efficient and powerful server systems.

4.1.2 Power consumption

Max computation load

Albeit not by far as the performance, the maximum power consumption per device (server) has also been rising in the previous years (see Figure 5). For single-CPU servers it has almost doubled (factor 1,95) between 2017 and 2022 at an average increase of 14% per year. With a factor of 1.8, the increase of dual CPU servers has been slightly lower in that period. Larger systems, though, have increased their average power consumption even by a factor of 2.85 between 2016 and 2021, but this could also be explained by the fact that at this time multiple very large single CPU systems (with heavy memory and storage) have been benchmarked.

The rise in the maximum power consumption of general-purpose servers, particularly under conditions of full CPU and RAM utilization like in SSJ 100% benchmark, can be attributed to several architectural and technological advancements aimed at increasing computational performance and system throughput. These trends, though driven by the need for higher processing power and memory capacity, have inevitably resulted in greater energy demands during peak operational loads and can also be seen in technical specifications of servers and server chips (Sun et al. 2019).

One primary factor is the evolution of processor design over the past decade. Modern server CPUs have experienced a dramatic increase in the number of cores per chip, as manufacturers have shifted toward multi-core architectures to meet the demands of multi-threaded applications, virtualization, and parallel processing. Each additional core increases the processing capacity of the CPU but also contributes to higher power draw when all cores are fully engaged. Moreover, server CPUs have been designed with higher clock speeds and enhanced instruction sets, further boosting their performance capabilities. However, higher clock frequencies and the associated dynamic power scaling (which increases power consumption quadratically with clock speed) result in a significant rise in energy consumption under full load. As a result, CPUs at maximum utilization today consume more power than their

predecessors, despite advances in power management techniques such as dynamic voltage and frequency scaling (DVFS).

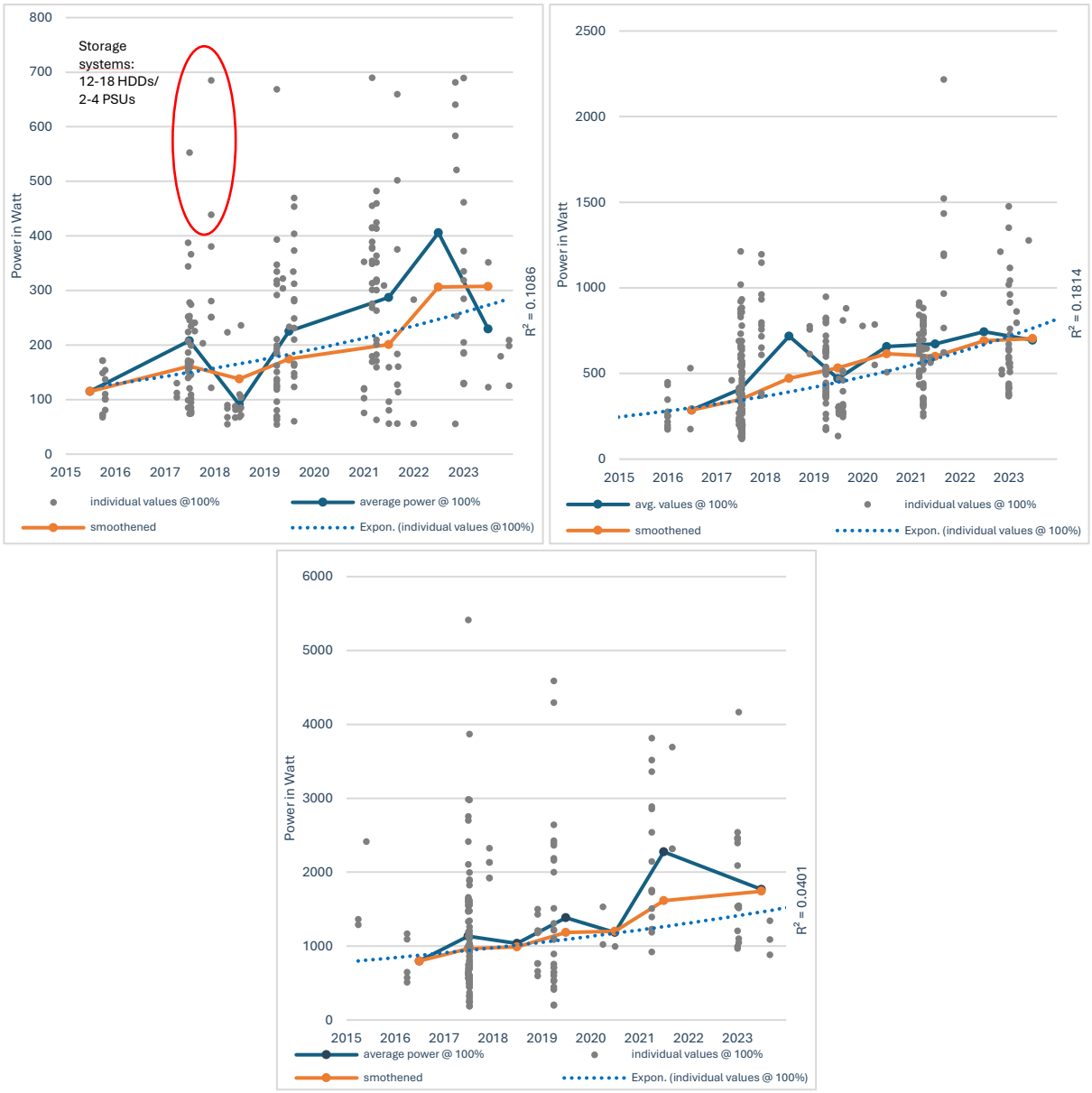


Figure 5: Maximum power use of single-CPU servers (left) and dual-CPU servers (right) and servers with more than two CPUs (bottom)

Another critical factor is the increased memory bandwidth and capacity in modern servers. As data-intensive applications have become more prevalent, the demand for faster and larger memory has grown. Newer generations of DRAM, such as DDR4 and DDR5, offer significantly higher bandwidth and lower latencies compared to earlier memory technologies. However, these improvements come at the cost of higher energy consumption, particularly when memory is fully utilized. Larger memory modules and higher data transfer rates inherently increase power requirements due to the greater number of memory cells that must be powered and the increased energy needed to drive data transfers between the CPU and RAM. Furthermore, as servers are often configured with more memory channels and higher memory densities, the total power consumption of the memory subsystem has risen substantially during peak utilization.

Thermal management also plays a role in the rising power consumption of servers under maximum load. The more powerful processors and memory systems found in modern servers generate significantly more heat than earlier generations, requiring advanced cooling mechanisms. While cooling systems themselves are a secondary load on overall power consumption, the increased heat dissipation requirements influence the design of server components. For instance, servers are designed to maintain operational stability at high temperatures, which often requires additional energy overhead in the form of internal fans and cooling solutions that operate more aggressively as thermal output increases under heavy computational loads. This additional high-power consumption for the fans is also reflected in the disproportionately high increase in power consumption at high performance levels of modern servers (see the high values of the line diagram at 80% and 90% in Figure 6).

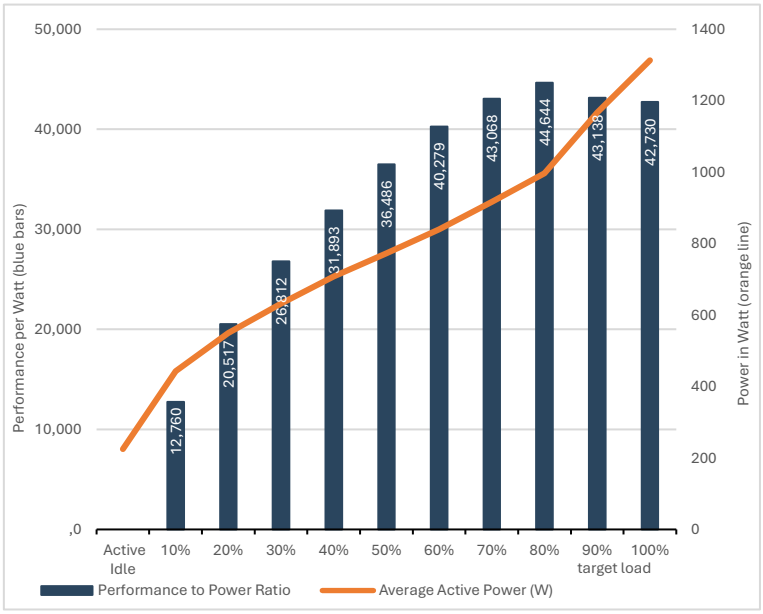


Figure 6: Power consumption and Performance to Power Ratio of a modern server (random example from SPEC Power results[17])

Advances in server interconnects and communication pathways, such as PCIe and memory controllers, also contribute to the higher power draw under maximum utilization. Modern general-purpose servers often support high-bandwidth interconnects to enable rapid communication between CPUs, memory, and storage devices. The shift toward PCIe 4.0 and PCIe 5.0 standards, which enables increased data transfer rates, has introduced new possibilities in peak power demand through new power connectors which supply up to 600 W in PCIe 5.0. Similarly, integrated memory controllers and other auxiliary processing components operate at higher frequencies, leading to higher power consumption when handling large volumes of data.

Finally, improvements in server architecture have paradoxically increased power requirements despite efforts to enhance energy efficiency. While new fabrication processes (such as 7nm and 5nm technology) have reduced the power consumption of individual transistors, the overall energy savings have been offset by the sheer increase in the number of transistors integrated into each processor (Sun et al. 2019). These transistors are employed not only to increase raw

[17] The data from this result can be found under https://www.spec.org/power_ssj2008/results/res2024q2/power_ssj2008-20240327-01387.html

computational power but also to support features like larger caches, more complex pipeline architectures, and enhanced parallelism, all of which draw additional power during peak utilization.

Idle power consumption

The analysis of server CPU idle power consumption trends from 2015 to 2023 (see Figure 7) reveals significant patterns across different server configurations.

A notable observation is the increasing divergence between maximum power capabilities and idle power consumption. This divergence is particularly pronounced in multi-CPU configurations, where maximum power thresholds have increased substantially while idle power consumption has maintained relatively moderate growth. This phenomenon suggests significant advancements in power management technologies and CPU idle state optimizations.

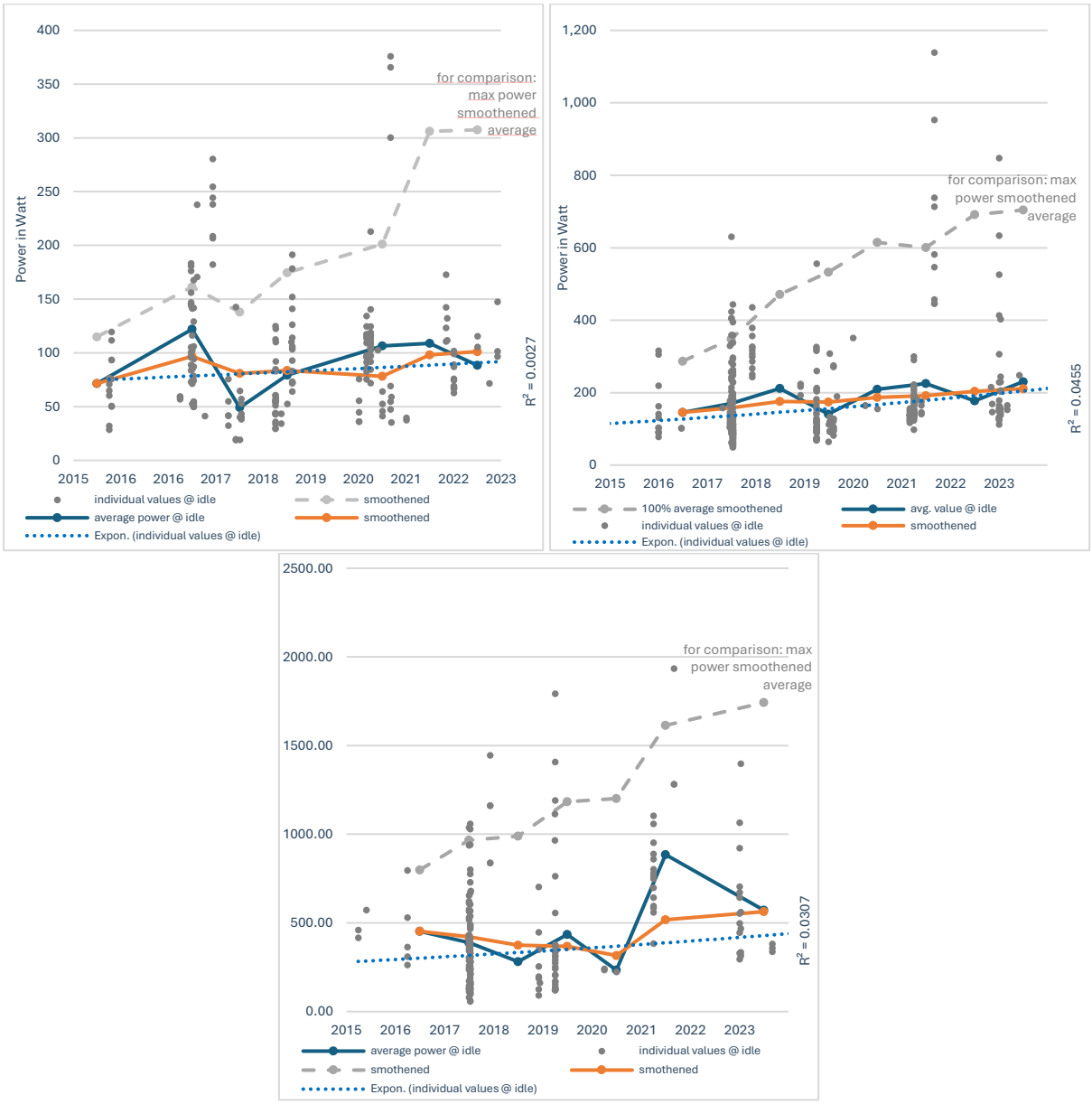


Figure 7: IDLE power use of single-CPU servers (left) and dual-CPU servers (right) and servers with more than two CPUs (bottom), the smoothened maximum power consumption from above is shown as dashed grey line

More recent server models have a wide variety of idle power consumption, mostly in multi-CPU configurations, indicating diverse architectural approaches and system design methodologies. The period from 2020 to 2023 demonstrates a relative stabilization or marginal reduction in idle power consumption, with only a minimal upward trajectory.

Perhaps most significantly, the data indicates that despite substantial increases in computational capabilities, as evidenced by rising maximum power thresholds, idle power consumption has remained comparatively stable.

The smoothed averages underscores both the achievements in power efficiency and the persistent challenges in minimizing idle state energy consumption in high-performance server systems.

4.1.3 Efficiency

SSJ_op/s efficiency

In Figure 8, the previous data is used to describe the efficiency of servers, as described in Equation 5 as function of computations (expressed as SSJ_op/s) per power consumption (Watt). Again, the first diagram shows single-CPU servers, the second (top right) shows double-CPU servers and the one on the bottom shows multi-CPU (>2) servers.

For one-CPU systems (top left graph in Figure 8), there is a gradual upward trend in performance per watt, particularly after 2019. The performance appears relatively steady in the early years (2015-2019), with noticeable improvements after 2020. The scatter of individual points suggests that while some one-CPU systems achieve significantly higher efficiency, there is a broad range of performance levels, possibly indicating variability in design or technology among different systems. The smoothed line highlights a continuous improvement in average efficiency, particularly noticeable from 2020 onward.

The two CPU systems (top right graph in Figure 8) show a more pronounced increase in efficiency from 2020, with average SSJ operations per watt accelerating more sharply than for one-CPU systems. A notable trend is the increasing clustering of individual values around higher performance levels after 2021, suggesting greater consistency in efficiency gains across systems. The exponential trend line shows this rapid improvement, particularly in recent years, indicating that advancements in two-CPU architecture are contributing significantly to better energy efficiency. The drop in 2023 in performance might be due to a low total number of CPUs in this period, which included many servers with the rather low-end Intel XEON Gold 5415+ CPUs, which had more modest SSJ/Watt results (below 8000 SSJop/s per Watt) as compared to other contemporary chips.

In the case of multi-CPU systems (bottom graph in Figure 8), the scatter of individual values is larger, indicating significant variability in performance. However, the average performance per Watt also follows an upward trajectory, especially after 2020. This suggests that while there is substantial room for improvement, certain systems are achieving high efficiency, driving the overall trend upwards. The exponential curve reflects a steady improvement but with more pronounced variance compared to one- and two-CPU systems, likely due to the complexity of multi-CPU architectures.

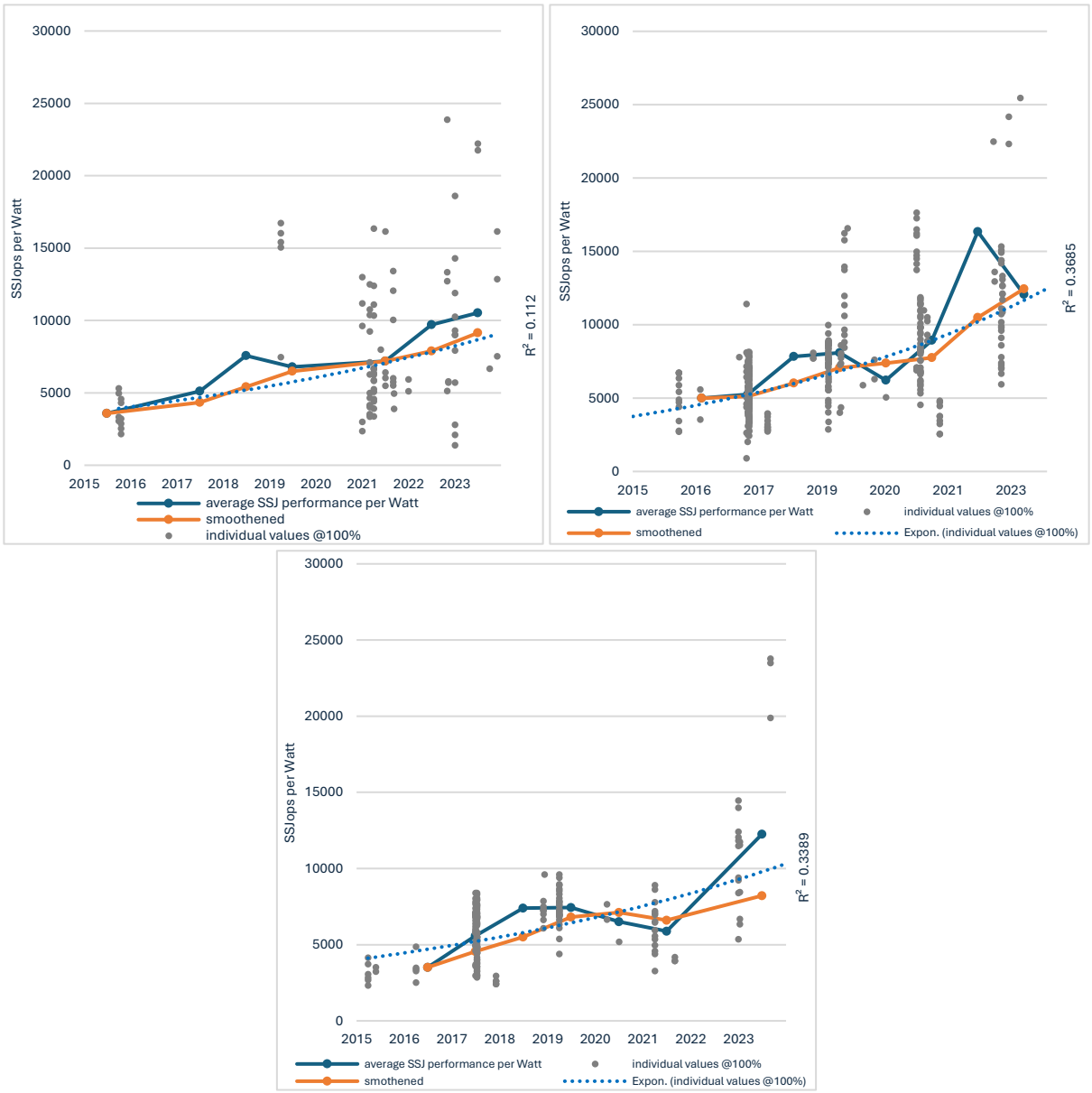


Figure 8: Efficiency [SSJops per Watt] of single-CPU servers (left) and dual-CPU servers (right) and servers with more than two CPUs (bottom), the smoothened maximum power consumption from above is shown as dashed grey line

SERT 2 Efficiency

The diagrams in Figure 9 illustrate the evolution of server energy efficiency measured using the SERT 2 metric across different CPU configurations from 2015 to 2023.

In the single CPU segment, there is a notable upward trend in efficiency scores, starting from approximately 10 points in 2016 and reaching peaks of around 55 points in 2021, followed by a moderate decline to approximately 40 points by 2023 which is comparable to the trends in the previous sections. As in previous graphs, the individual servers (grey dots) show considerable variance, particularly during the 2020-2023 period, ranging from about 10 to 80 points.

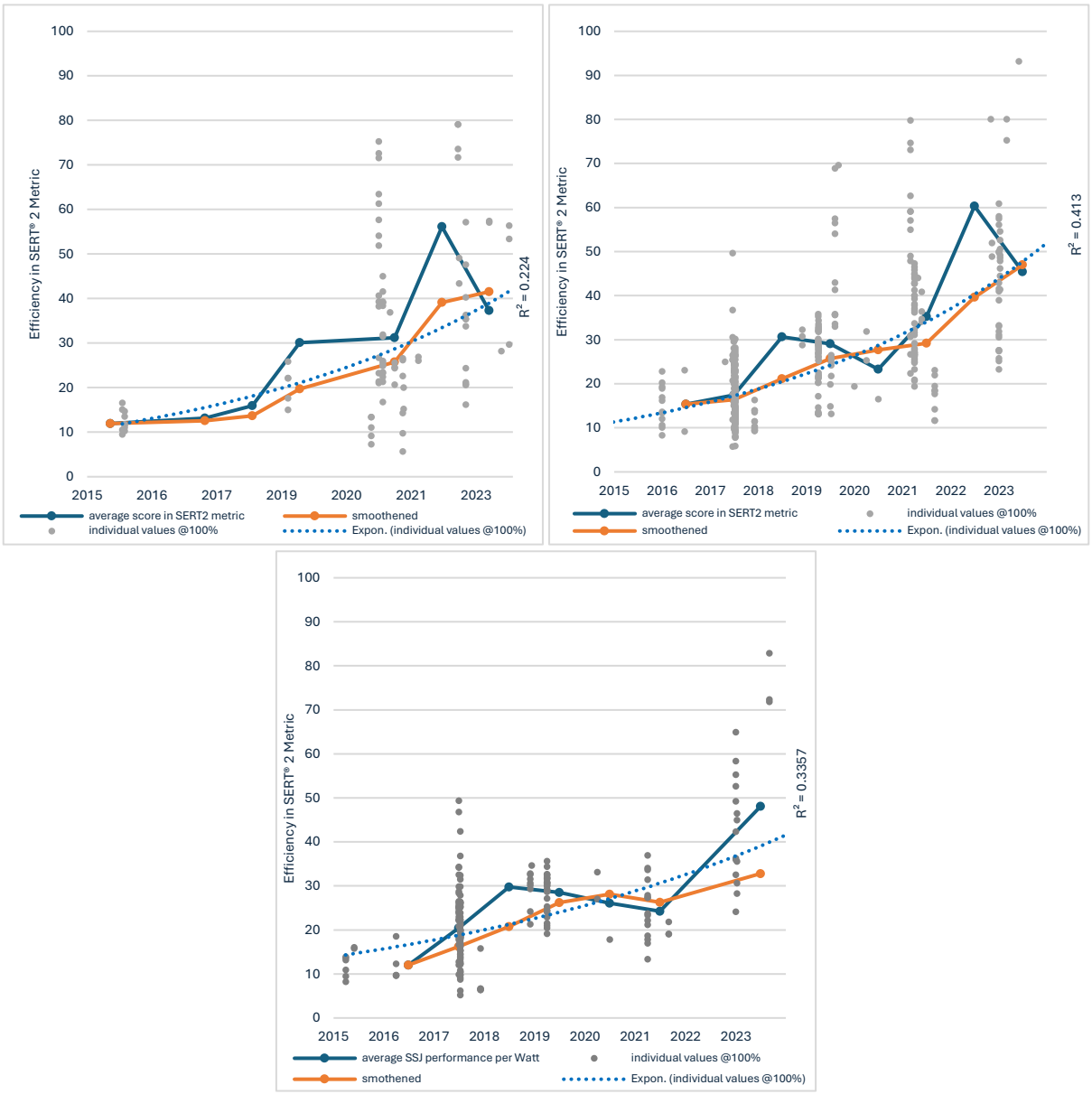


Figure 9: SERT 2 efficiency for single-CPU servers (left), dual-CPU servers (right) and servers with more than two CPUs (bottom), the smoothened maximum power consumption from above is shown as dashed grey line.

The dual CPU server category exhibits a similar but more pronounced pattern of improvement, with efficiency scores initially around 15 points in 2016, rising to approximately 60 points at their peak in 2022. This category shows the highest individual measurements, with some servers achieving scores above 90 points in 2023. The data points display significant dispersion, particularly from 2020 onwards, indicating substantial variability in efficiency among different server models. The exponential trend line suggests a consistent improvement rate, albeit with more pronounced fluctuations in the average scores as compared to single CPU systems.

The multi-CPU server segment, depicted in the second image, has a more moderate efficiency progression. Starting from around 10-15 points in 2015-2016, these systems show a gradual increase to approximately 47 points by 2023. Individual measurements in this category exhibit less extreme variation compared to single and dual CPU systems, though there are notable outliers, particularly in 2023, with some systems reaching efficiency scores above 80 points. It is remarkable that the top systems all contain a CPU from AMD (see Figure 10).

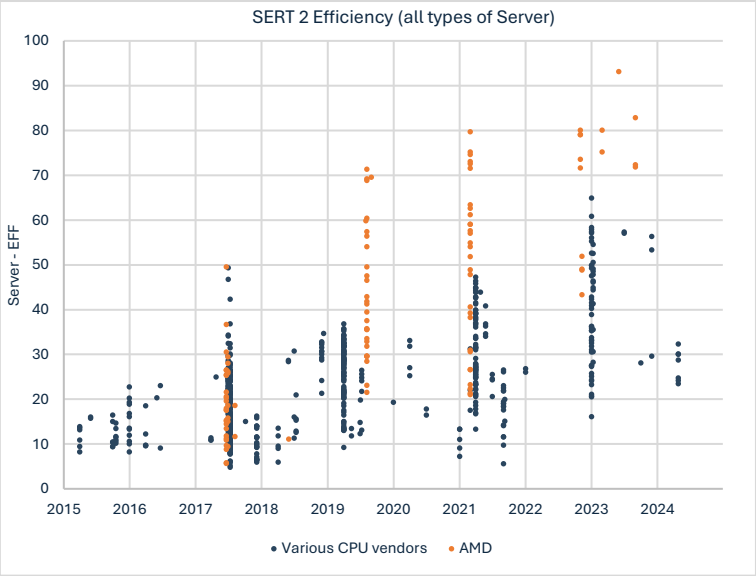


Figure 10: SERT 2 efficiency for all types (single-, dual- and multiple CPU), that shows that the top performing Servers in terms of energy efficiency all have AMD CPUs

Across all three categories, the exponential trend lines indicate a consistent improvement in energy efficiency over the observed period, though the rate and pattern of improvement vary significantly between CPU configurations. As shown in Figure 9, the current SERT 2 results of single-CPU, dual-CPU, and multi-CPU servers are relatively similar. The relative gains between 2017 and 2022, however, seem to be highest for the dual-CPU systems, as also shown by the individual trendlines. However, due to the relatively small number of individual measured values in the segments, there is a high level of uncertainty here.

4.2 Accelerated computing (GPUs and TPUs)

As discussed in Section 3.3, we analyze throughout this section mainly two performance indicators (and subsequently the correspondingly derived efficiency indicators) for GPUs: FP32 and FP16/BF16.

We additionally distinguish between FP16 performance with tensor cores and without tensor cores. As will be discussed further down, the advent of tensor cores with the Nvidia Volta GPU in December 2017 was one of the pivotal moments in GPU performance for deep learning. Tensor cores are specialized processing units designed specifically for the matrix operations used in deep learning. They dramatically accelerated training and inference for neural networks.

Because deep learning are major drivers of data center demand, there is hardly any GPU without tensor cores (or its equivalent for other companies, e.g. “matrix cores” for AMD) today. It nevertheless makes sense to still follow the development of FP16 performance and efficiency without (or with disabled) tensor cores not only for historic reasons, but also to isolate the performance contribution of tensor cores from all the other contributing factors. For some GPUs, either manufacturers or third-party sources[18] still devise the performance abstracting from the tensor cores. All in all, three performance and efficiency indicators will be devised throughout this section: FP32, FP16/BF16 abstracting from the tensor cores, and FP16/BF16 considering the contribution of tensor cores.

[18] See e.g. this source for Nvidia’s H100 GPU: https://www.techpowerup.com/gpu-specs/h100-sxm5-96-gb.c3974.

4.2.1 Performance

Figure 11 shows the mean annual FP32 performance together with the individual FP32 performance values. While from 2014 to about 2019, there was incremental growth, from about mid-2020 on, accelerated performance growth can be seen, although FP32 performance is no longer a priority for most GPUs which are optimized for lower or mixed precisions.

This growth is partially explained through general technological advancements, partially by the fact that from about 2020 on, there was also a substantial growth in power consumption / TDP (see Figure 14). FP32 thus did not grow as fast as FP32 performance; see Figure 15.

The outlier in Figure 11 is Nvidia's GB200 superchip, which contains two Blackwell B200 GPUs and one Grace ARM CPU (a "superchip" being defined as the combination of GPUs and CPUs on a single chip, such as the Grace-Hopper superchip comprising one Grace CPU and one Hopper GPU or the more recent Grace-Blackwell superchip comprising one Grace CPU and two Blackwell GPUs).

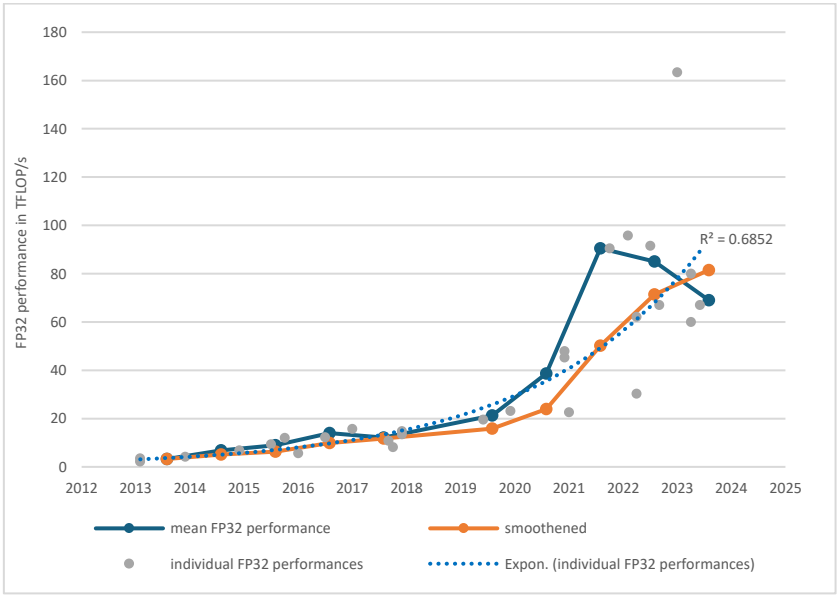


Figure 11 Mean annual FP32 performance [TFLOP/s], raw and smoothened as well as the individual FP32 performance values for all GPUs and TPUs from Table 2 that indicate it.

The trends of the FP16 performance, indicated in Figure 12, look much noisier. The main reason is that for 2022, only data for two GPUs were available: Intel's Habana Gaudi 2, launched in May 2022, and Nvidia's Ada Lovelace L40 GPU, launched September 2022; both relatively low-power, low-performance GPUs.

This sparse data for 2022 skewed the results, causing the downwards dent in the blue line on the left part of Figure 12. The smoothened orange line reflects the trend better; however, both trendlines hide a bit the partially spectacular performance improvements from mid-2020 on. The right side of the figure was therefore included, which only shows the individual FP16 performance values. As highlighted in the figures, from mid-2020 on, substantial FP16 performance growth took place for some of the GPUs. As for FP32 before, part of this growth is due to an increase in TDP; hence, the efficiency gains were not quite as high.

Another reason, however, is that GPU design was more and more geared towards half-precision (i.e., FP16) and even lower precisions and away from single-precision (FP32) and higher. The FP16 performance, which used to be double the FP32 performance, decoupled from it, growing faster. Accounting for the influence of tensor cores strengthens the effect, as argued below.

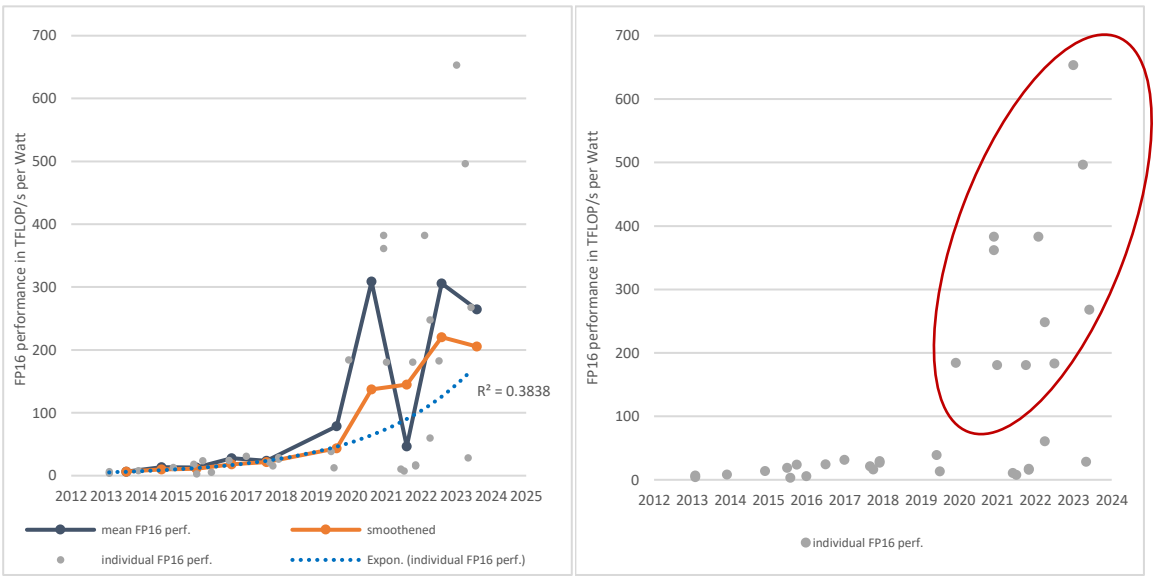


Figure 12 Left: Mean annual FP16 performance [TFLOP/s], raw and smoothened, as well as the individual FP16 performance [TFLOP/s] values for all GPUs and TPUs from Table 2 that publish it, whilst abstracting from the influence of tensor cores. Right: Only individual FP16 performance values. While the existence of low-performance (and typically low-power) chips persists, from mid-2020 on, many of the GPUs show substantial FP16 performance gains (highlighted).

The left side of Figure 13 shows the mean annual FP16 performance, as well as the individual performance values, with the effect of tensor cores included. Since 2023, there has been a GPU performance explosion for all the main manufacturers and their GPUs Habana Gaudi 3 (Intel), Instinct MI300 and MI300x (AMD) as well as the Hopper and Blackwell generations from Nvidia. When considering the performance of the Grace-Hopper and Grace-Blackwell superchips (also considered in Figure 13, left), this effect becomes even stronger (the outlier at 5000 TFLOP/s, in fact, is the GB200 superchip). As compared to Figure 12, the downwards dent in 2022 (due to sparse data for that year) is now barely noticeable; so fast does the FP16 (with tensor cores) performance grow in 2023 and 2024. For the decade 2014-2024, the average annual growth rate of the FP16 performance was 80%, with an accelerating trend lately.

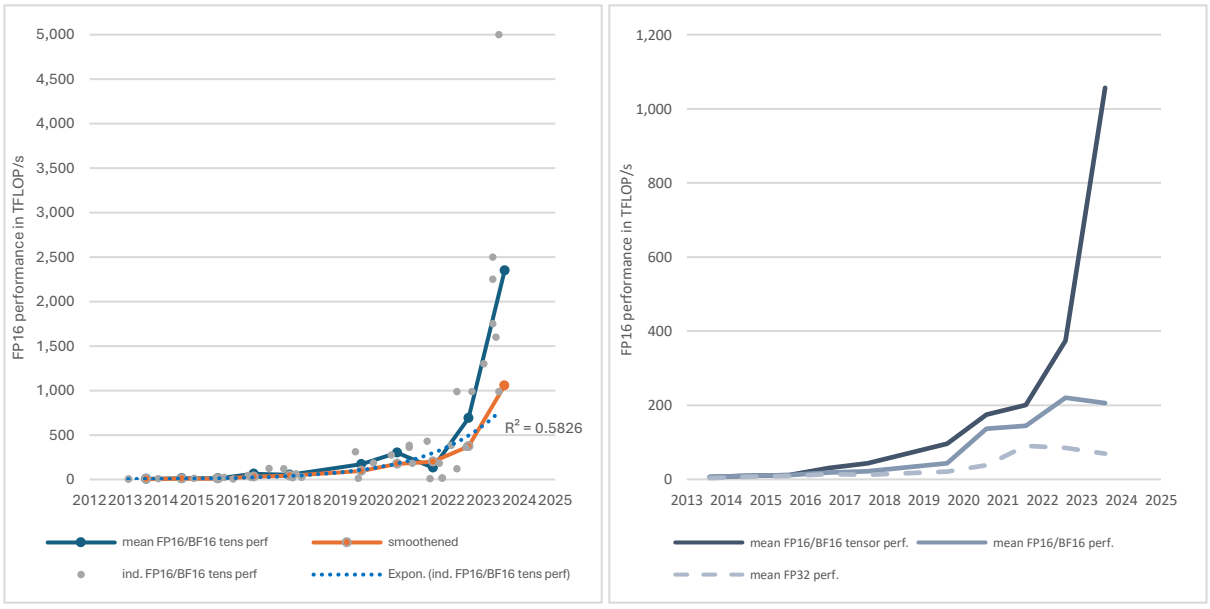


Figure 13 Mean annual FP16 performance [TFLOP/s] (raw and smoothened) as well as the individual FP16 performance values for all GPUs and TPUs from Table 2; tensor cores included (left), and comparative performance development for FP16 w/ tensor cores, FP16 w/o tensor, and FP32 (right, all TFLOPs, smoothened).

Finally, the right side of Figure 13 compares the mean performance growth of FP32, FP16 without the influence of tensor cores, and FP16 with tensor cores considered. It is evident how the advent of tensor cores decouples the growth in performance from the other, much more moderate developments.

4.2.2 Power consumption

Figure 14 shows the evolution of the Thermal Design Power (TDP) of GPUs and TPUs over the last decade. As mentioned earlier, the TDP is used as proxy for the power consumption of chips at maximum load, although the actual value might differ (Govind et al. 2023). As can be seen from the upper left part of Figure 14, the TDP was fairly constant for the period 2013 – 2020. There were mainly two categories of GPUs, most with a TDP of 200-300W, and some lower-power, lower-performance ones (but typically more efficient) with a TDP of 72 – 150W.

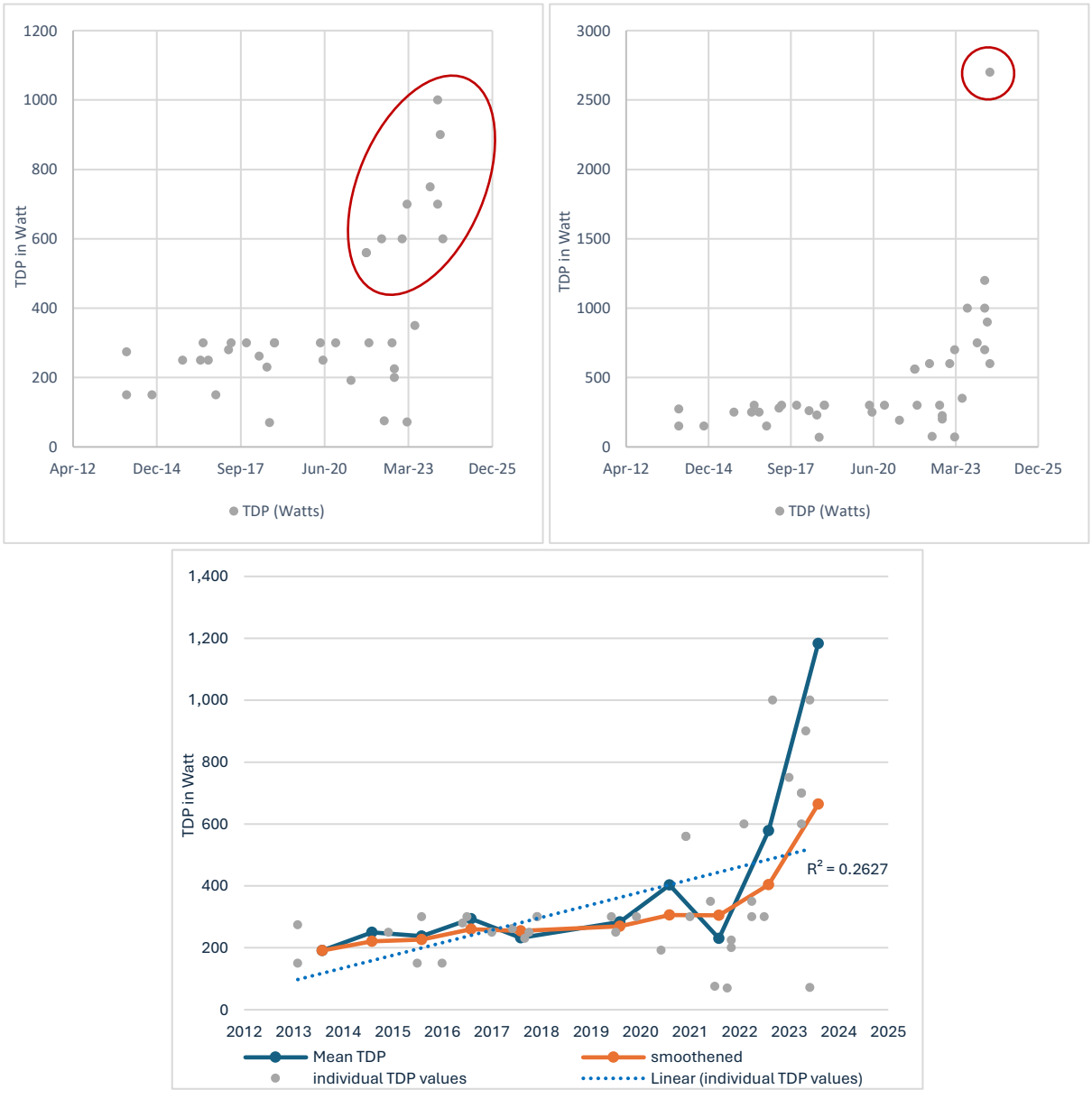


Figure 14 Individual TDPs [Watt] for GPUs and TPUs (upper left), the same but including superchips (upper right), and mean TDP for each year, raw and smoothened (lower part).

In late 2020, however, chips with dual the TDP (i.e., 550 – 600W) emerged. From that moment onward, the TDP kept growing towards 1kW, and the new norm seems to be 600 – 1,000W, as highlighted in the upper left corner of Figure 14, there's hardly any new GPUs now with a lower TDP. Accounting for superchips as well (right upper part of Figure 14), this range extends to the 1200W of the Grace-Hopper superchip (comprising one Hopper GPU and one Grace CPU with TDPs of 900W and 300W, respectively) and in particular the 2.7kW of the recent Grace-Blackwell GPU, which comprises two Blackwell GPUs and one Grace CPU, with TDPs of 2*1200W and 300W, respectively).

The lower part of Figure 14 presents the development of the mean TDP over this last decade. As expected from the individual values presented before, the curve is first quite flat, in 2020 it starts growing rapidly, and shoots up for the last two years. Due to these two quite distinct phases, one single trendline can only do a poor job in approximating this development.

4.2.3 Efficiency

As defined in Equation 7, relating performance and power consumption / TDP yields the efficiency for each GPU and TPU. While the TDP is a invariable of the chip, the performance depends on the indicator chosen, as thoroughly discussed in Section 4.2.1. Corresponding to the three performance indicators chosen, there are thus also three different efficiency indicators, that all need to be addressed: FP32 efficiency, FP16/BF16 without tensor cores, and FP16/BF16 with tensor cores efficiencies.

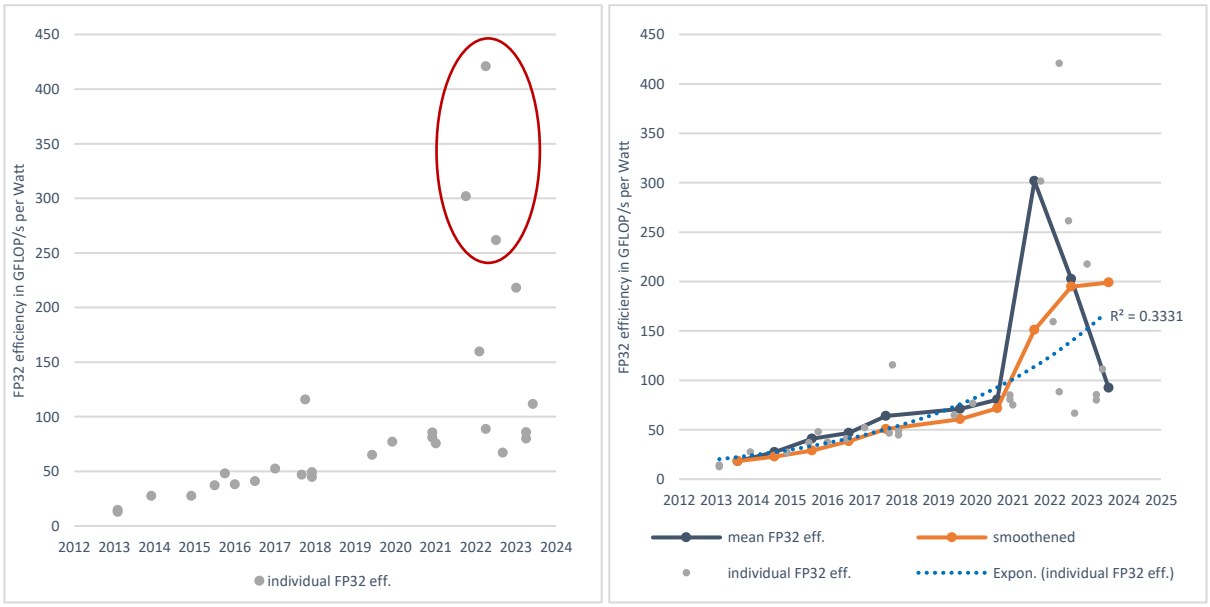


Figure 15 Left: FP32 efficiency development [GFLOPs/Watt] over time; all three outliers are the relatively low-power, low-performance, but very FP32-efficient Nvidia Ada Lovelace GPU generation. Right: Development of the mean FP32 efficiency, raw and smoothened.

Figure 15 presents the development of FP32 efficiency. Except for a few outliers, its left part shows constant, incremental efficiency gains at a steady, good pace. The highlighted outliers are all from the 2021-2022 Ada Lovelace generation of Nvidia GPUs, which are low-power and low-performance, but very FP32-efficient. The right side of the figure shows the development of the mean FP32 efficiency, both raw and smoothened. The 2024 dip is certainly partly explained by sparse data available for the latest chips (Google, for example, released only performance data on versions 5e and 5p of its TPUs, but no power consumption or TDP data; the efficiency can thus not be determined. But it is probably also explained by the fact that focus in ML is shifting away from single precision towards half precision, mixed precision, or even lower

precisions. The newest Nvidia GPU generations Hopper and Blackwell, for example, have less FP32 efficiency than the previous Ada Lovelace generation.

As can be seen in Figure 16, for the period 2013 – 2020, the FP16/BF16-efficiency (without the contribution of tensor cores) experienced incremental increases along the tendency of the FP32-efficiency. Given its half precision as compared to FP32 (16 bits instead of 32), the FP16-performance (and thus efficiency as well) had traditionally been twice the FP32-efficiency. This is highlighted in the lower part of the left side of Figure 16. From late-2020 on, the increasing design for half precision and mixed precision ML training saw the FP16 representation increasingly being displaced by the BF16 representation, and in parallel a decoupling of the FP16/BF16 efficiency from the FP32 one. Compared to its predecessor, for example, the FP32 performance of the AMD MI100 GPU grew by 75%, while the FP16 performance grew by 300%. The same 4-fold growth difference reflected in the respective efficiency gains, showing the decoupling of FP16/BF16 from FP32.

Given this decoupling and substantial growth of the FP16/BF16 efficiency, the graph on the right side of Figure 16 might at first glance seem paradoxical. With ups and downs due to sparse data in 2022, the FP16/BF16 efficiency seems to have stalled since 2021, and perhaps even regressing now. Even the smoothened curve seems to have flattened out and now perhaps going down. This of course did not happen, and the explanation behind this seeming paradox is more pragmatic: As a dominating part of computing loads are for deep learning, the non-tensor FP16/BF16 performance starts being so irrelevant in the DC / HPC domain that manufacturers stopped devising it. It is only sporadically devised for lower-end chips, which of course skews the results. As Figure 17 below shows, the development of the full FP16/BF16 efficiency (including the effect of tensor cores) is on the contrary quite impressive.

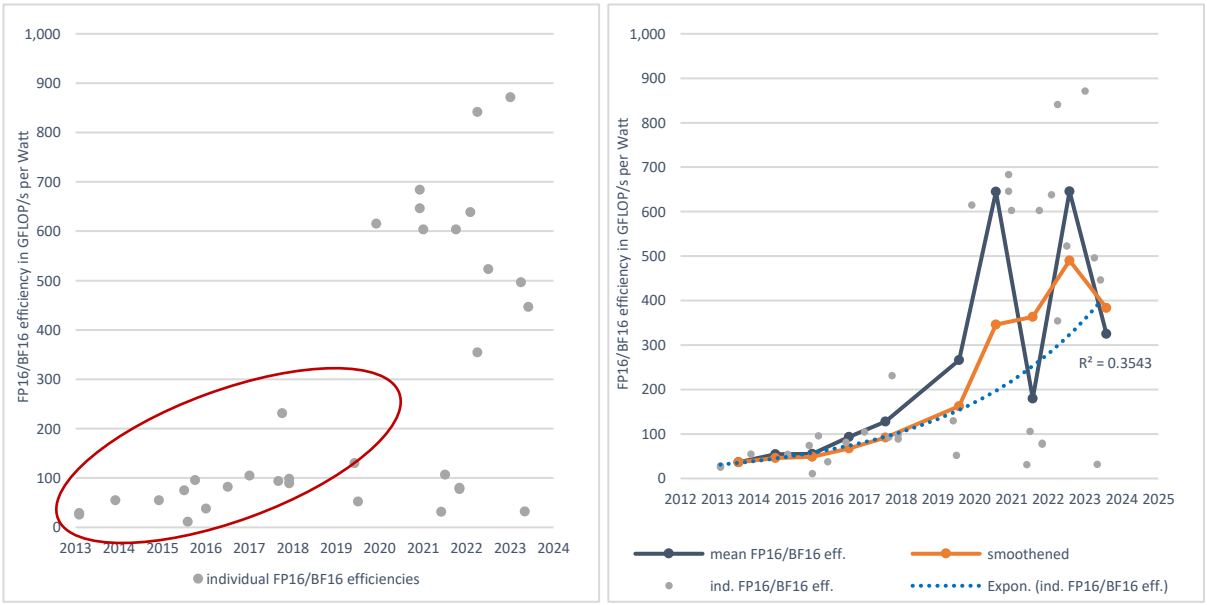


Figure 16 FP16/BF16 efficiency [GFLOPs/Watt] without the contribution of tensor cores. Individual values (left) and evolution of the yearly mean, raw and smoothened (right).

Accounting for the ever more important contribution of tensor cores substantially changes the picture. Ever since the first GPU with tensor cores appeared, it became evident that the impact of tensor on matrix operations (which are typical for numerous ML algorithms, and in particular for deep learning) would be vast. The first GPU to include tensor cores was Nvidia's Volta V100 in December 2017. It had a FP16/BF16 efficiency of 416.7 GFLOPs/Watt (125 TFLOPs performance at 300 W). Compared to its predecessor Pascal P100 from 1.5 years earlier, which

had an efficiency of 74.8 GFLOPs/Watt (18.7 TFLOPs performance at 250 W), it had an efficiency jump by a factor of 5.5 (or 450% increase).

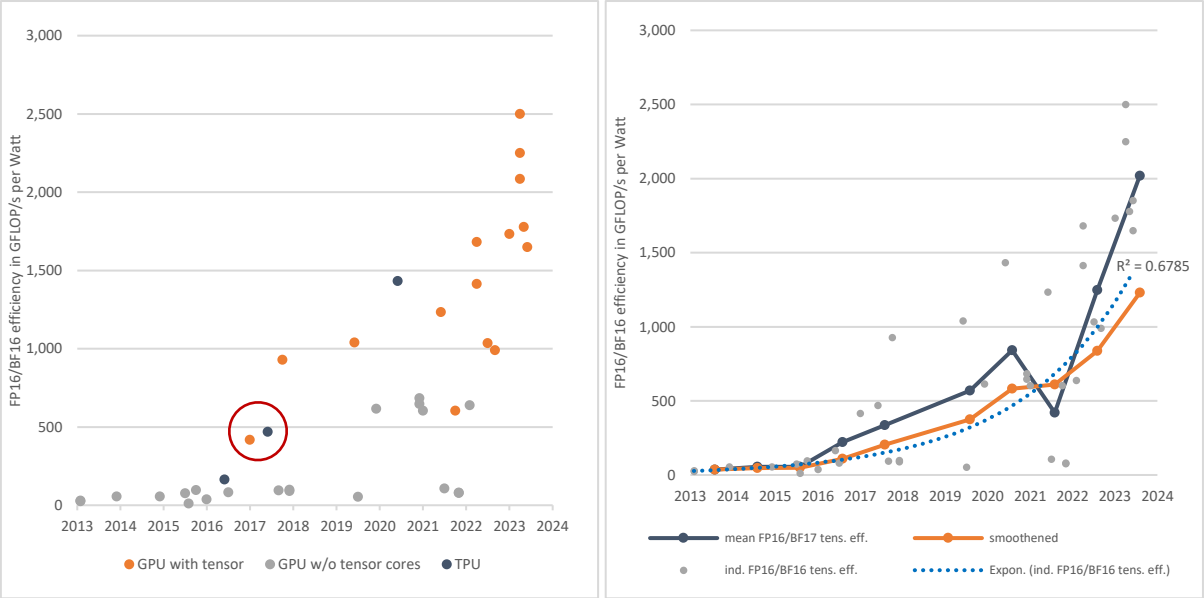


Figure 17 Full FP16/BF16 efficiency [GFLOPs/Watt], including the contribution of tensor cores. Individual values (left), highlighting the contribution of TPUs (blue dots) and tensor cores (orange dots, as opposed to GPUs without tensor cores in grey), and evolution of the yearly mean, raw and smoothened (right).

On the left side of Figure 17, this jump is highlighted. The small circle encloses the first GPU and first TPU, respectively, of this new wave of efficiency: Nvidia Volta V100 (December 2017, 416.7 GFLOPs/Watt) and Google's v3 TPU (May 2018, 469.5 GFLOPs/Watt). Today's top value is around 2,500 TFLOPs/Watt (Nvidia Blackwell). Overall, in the 7 years between 2017 and 2024, the top GPU/TPU FP16/BF16 efficiency grew by a factor of about 25, from around 100 to 2,500 GFLOPs/Watt.

This also reflects in the evolution of the yearly mean values of the full FP16/BF16 efficiency shown on right of Figure 17: With the exception of the 2022 dent due to sparse data, the mean grows strongly; a growth that has even been accelerated over the past two years. The advent of superchips is less important for efficiency than for performance: While putting two Blackwell GPUs together duals the performance of the superchip, it also duals its TDP, with a neutral effect on the efficiency. By contrast, as also shown in Figure 18, tensor cores and algorithms that are being optimized for them, have an important influence. Overall, the mean FP16/BF16 efficiency grew from 36.5 GFLOPs/W in 2014 to 2018.7 GFLOPs/W in 2024, yielding a compound annual growth rate (CAGR) of 49% over this decade.

In a similar analysis, which also considered individual GPU and TPU performance values and related them to the TDP as proxy for the peak power consumption, (Hobbhahn, Heim, and Aydos 2023) reach a different result. The report concludes that the performance and efficiency of accelerators doubled every 2.3 and 2.7 – 3 years, respectively, corresponding to performance and efficiency CAGRs of 35% and 26-29%, respectively. By contrast, our results indicate 80% average yearly performance increases (see Section 4.2.1) and 49% yearly efficiency growth, as discussed above. These differences are explained through three main differences in scope:

i) the period under scrutiny, 2010-2023 vs. 2014-2024, with a relatively flat curve before 2015 and the most substantial jumps taking place over the past two years,

ii) the different GPU set under consideration, a larger set of GPUs which includes e.g. Nvidia's GeForce RTX series in (Hobbhahn, Heim, and Aydos 2023), while we only considered GPUs for DCs and HPC, as shown in Table 2), and

iii) the metrics defined for FP32 vs. FP16/BF16 precisions, respectively, with diverging trends, as shown in Figure 18.

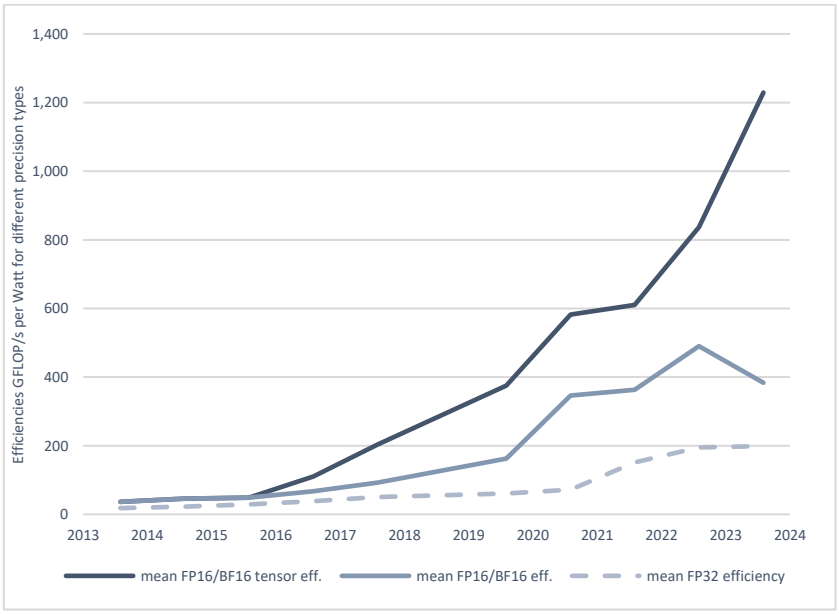


Figure 18 Compared development of all three efficiencies [GFLOPs/Watt]: FP32, FP16/BF16 without tensor cores, and full FP16/BF16 with sensor cores.

4.3 ASICs

4.3.1 Performance

As described in section 2 already briefly, ASICs (Application-Specific Integrated Circuits) are highly specialized hardware designed for specific tasks, like the cryptographic computations in Proof of Work (PoW) blockchains such as Bitcoin. In cryptocurrencies, ASICs are optimized for the hashing algorithms used to solve complex mathematical puzzles (e.g., Bitcoin's SHA-256). Their efficiency makes them highly relevant because they dramatically outperform general-purpose hardware, like CPUs and GPUs, in mining.

However, ASICs contribute significantly to the high energy consumption in PoW blockchains. Mining, which involves solving cryptographic puzzles, is the most energy-intensive part of PoW, far outweighing the energy used for network communication or transaction validation (Coroamă 2021). Since miners compete to solve these puzzles faster, the combined computational power (and energy consumption) of the network grows, which in turn leads automatically (every two weeks) to an increase of the difficulty to solve a block (ensuring, that on average one block is solved every 10 minutes).

The diagram in Figure 19 illustrates the exponential growth in mining hardware performance measured in TH/s (Terahashes per second) from 2014 to 2024. The data points (individual values, shown as gray dots) demonstrate a slight variance around the average trend line (shown in orange), with performance values escalating from approximately 1 TH/s in 2014-2015 to over 400 TH/s by 2024.

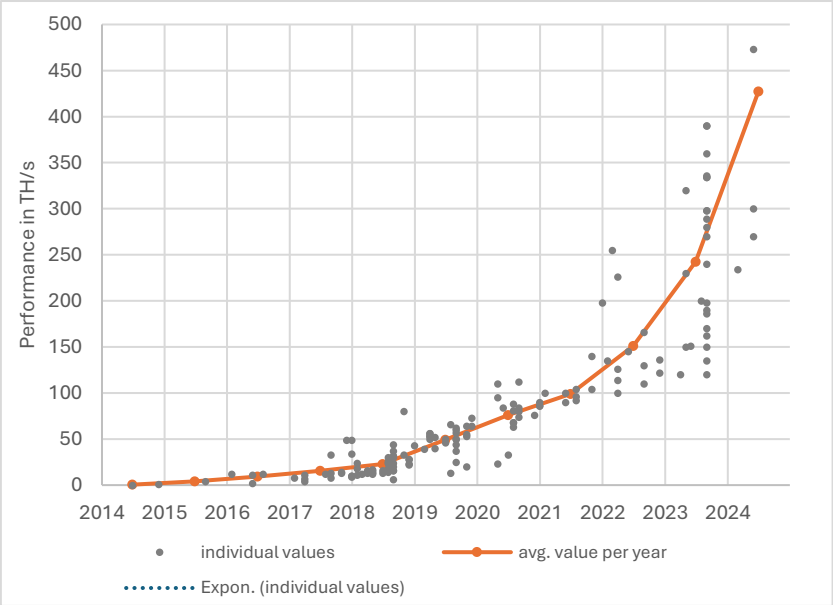


Figure 19: The growth in performance of ASCI miner hardware in Tera-Hash per second (TH/s), based on data from (hashrateindex.com 2024)

This remarkable performance increase can be attributed to several technological advancements in ASIC design and manufacturing processes like extreme ultra-violet (EUV) laser lithography systems. The primary drivers are similar to those of CPUs and GPUs, i.e. the transition to smaller semiconductor process nodes (from 28nm down to 5nm and below), enhanced chip architecture efficiency, improved power delivery systems, and more sophisticated cooling solutions. The exponential growth pattern aligns with Moore's Law-like scaling in the semiconductor industry, though specifically optimized for the SHA-256 hashing algorithm used in cryptocurrency mining.

4.3.2 Power

This power consumption trend in Figure 20 reveals several interesting patterns. The data shows a general upward trend in power consumption from approximately 500W in 2014 to around 5500-6000W by 2024, though this increase is notably more linear compared to the exponential performance growth observed in the previous graph.

The graph illustrates the power consumption of a range of devices, from home miners to professional hardware. It is evident that multiple devices operate at a level just below the 3680 W limit, which is a typical design limit for one electrical socket in many regions, such as central Europe. This limit is determined by the fuse capacity of the socket, which typically is 230 V x 16 A = 3680 W.

Furthermore, the growth in power consumption demonstrates that the aforementioned gains in performance are not solely attributable to miniaturization (Moore's and Dennard's laws). They also result from parallelization, which entails the integration of multiple chips (or larger chips) into a single device, albeit at the cost of higher power consumption.

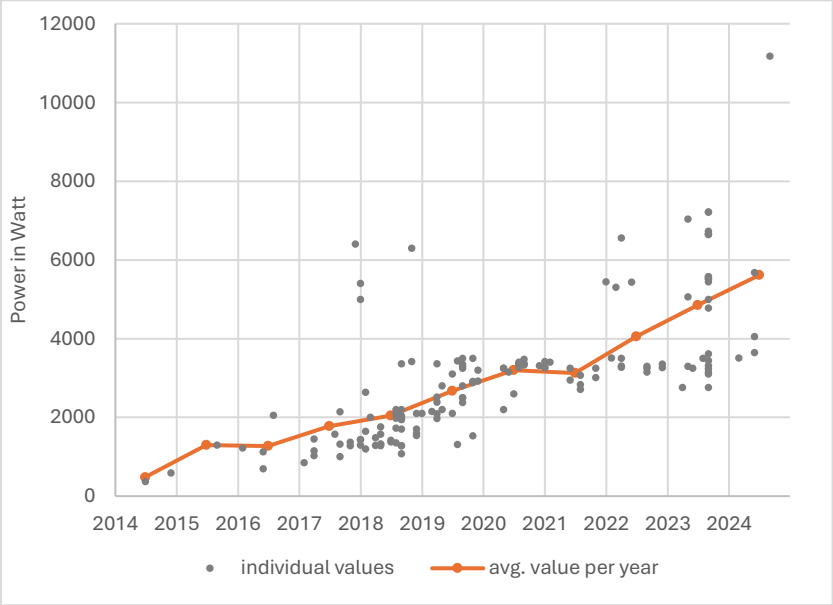


Figure 20: The power consumption of ASCI miner hardware in Watt, based on data from (hashrateindex.com 2024)

4.3.3 Efficiency

The trend in Figure 21 demonstrates a clear exponential improvement in energy efficiency, increasing from approximately 0.002 TH/W in 2014 to about 0.075 TH/W by 2024 - representing roughly a 37-fold improvement over this decade.

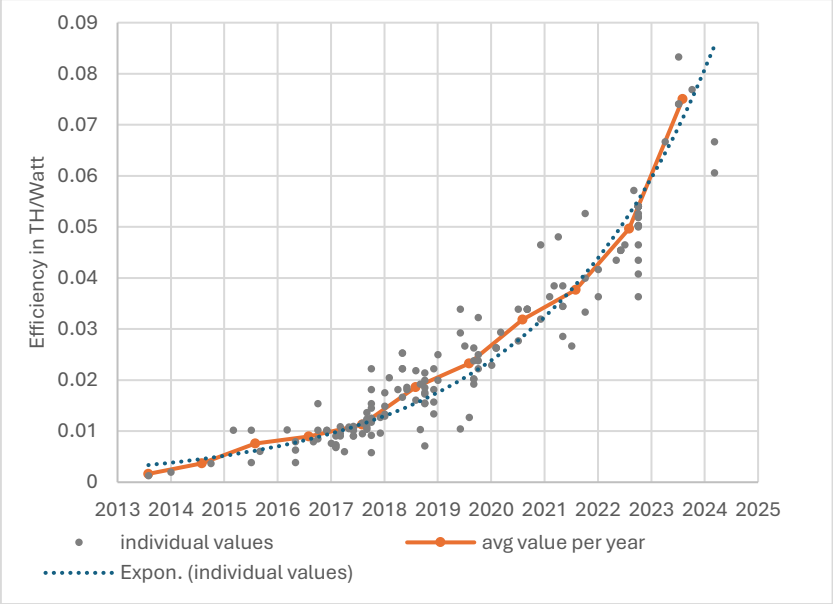


Figure 21: The efficiency of various mining hardware (mining rigs), based on data from (hashrateindex.com 2024)

The exponential nature of the efficiency gains (as indicated by both the orange average line and blue exponential trend line) can clearly be seen in the case of mining hardware, which might be also due to their very specific purpose. This efficiency improvement suggests that manufacturers have successfully leveraged advances in semiconductor technology, achieving better performance per Watt through:

- Process node shrinks (progression to smaller fabrication technologies)

- New chip architectures and system-of-chips/chiplets to better adapt chip structures to specific workloads
- Improved thermal management

The scatter of individual data points shows increasing variance in later years, indicating a wider range of design approaches and technical solutions in the market. This diversity likely reflects different optimization strategies among manufacturers, with some achieving breakthrough efficiency levels while others maintain more conservative designs.

The sustained exponential growth in efficiency also suggests that, despite the maturity of SHA-256 mining hardware, significant technological headroom still exists for further improvements, though perhaps with diminishing returns as fundamental physical limits are approached.

This development also leads to an ever-decreasing energy consumption per computation overall in the network. The Cambridge Centre for Alternative Finance estimates the total energy use by bitcoin and Ethereum (CBECI 2024), but also the average energy used to provide one Terahash of computations in the bitcoin network (Figure 22).

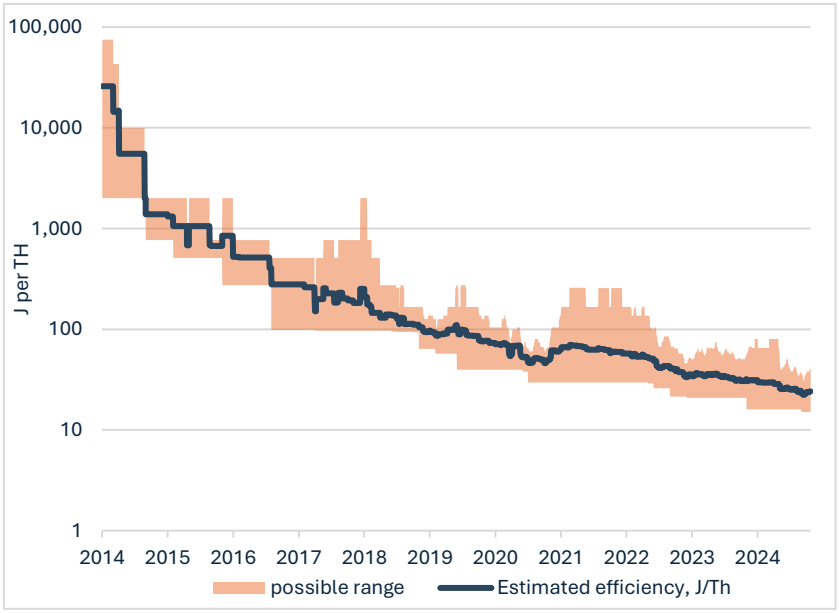


Figure 22: The average amount of energy used to provide 1 TH of computations, based on the data of Cambridge Bitcoin Electricity Consumption Index (CBECI 2024), vertical axis is logarithmic

5 Discussion: Technological developments leading to the past efficiency gains

After the quantitative results presented throughout Section 4, the current section discusses the factors identified as main contributors to these performance and efficiency gains. Unlike Section 4, however, this section is not structured along server types (i.e., general-purpose, AI, and crypto servers), but aims at uniting these different server types and finding the common trends that led in the past to their respective performance and efficiency gains (while also highlighting the difference and retaining some aspects thar are specific to specific server types). Instead, this section is structured after the main factors influencing the energy efficiency of servers (as physical devices in a narrower sense) as shown in the beginning of this document in Figure 1. It thus addresses logical processing units / chips (Section 5.1), memory (Section 5.2), storage (Section 5.3), a specific deep dive into AI-specific architecture (Section 5.4), power management (Section 5.5), cooling (Section 5.6) as well as a discussion on the relation between hardware refreshment cycles and the pace of efficiency gains (Section 5.7).

5.1 Logical processing unit

Smaller Lithography Nodes + Advanced packaging / 3D packaging/3D ICs

Probably the most critical factor in improving energy efficiency regarding computing technologies are the advances in semiconductor manufacturing processes. As chips are produced using smaller process nodes – such as the shift from 14nm to 7nm or 5nm technologies – multiple effects affect the power consumption. For example, a lower operating voltage is needed to switch transistors, which leads to lower power consumption. Also, the size of the transistors affects the leakage current, which further reduces the power consumption. Finally, the miniaturization of the transistors has led to a significantly reduced power use per computation.

Such advances in miniaturization are primarily achieved through improved production processes. Major progress has been made in the last decade by switching to light with ever shorter wavelengths in photolithography processes, known as extreme ultraviolet lithography. In recent years, further challenges have arisen in miniaturization and manufacturers have achieved this through improved optical instruments, in particular higher numerical aperture. Whilst the aperture can in principle be achieved by increasing the size of mirrors etc., problems arise at the same time due to increased angles and the loss of reflection from the mask. This problem can be countered by using anamorphic optics, which reduce the size of the pattern to be printed more in one direction than in the other (Christine Middleton 2024).

3D stacked ICs (3d packaging)

For a few years now, new chip designs have been available that stack the cache directly above or below the computing unit (CPU/GPU). In principle, independent chips are placed on top of each other and connected to each other by electrically conductive contacts. Due to the shortened paths, both the latency between memory and CPU can be greatly reduced as well as the electrical power dissipation due to the shorter paths and thus resistors and capacities (AMD 2024; KJ Jacoby 2023).

ARM-based chip architectures in servers

ARM processors were initially designed for mobile devices (smartphones, tablets) where battery life is paramount. This design heritage has ingrained a focus on power efficiency from the ground up. Every aspect of ARM processors is optimized to minimise power consumption. By contrast, x86 processors, while also improving in efficiency, historically prioritized raw performance. The core of ARM's efficiency lies in its RISC design philosophy. RISC processors use a smaller, simpler set of instructions. Each instruction typically performs a single, well-defined operation. This contrasts with x86's CISC (Complex Instruction Set Computing) architecture, which has a much larger and more complex instruction set. It can't be stated in general that every software on ARM processors has a lower energy consumption and is therefore more energy-efficient to operate. However, some research results in recent years (Simili et al. 2022; Tahmid Noor Rahman, Nusaiba Khan, and Zarif Ishmam Zaman 2024) as well as the relatively cheap and huge offer of some hyperscalers (Patrizio 2025) suggest that ARM processors have the potential to increase energy efficiency.

5.2 Memory

Memory has seen significant improvements over the past decade, both for CPUs and GPUs. While significant leaps positively affected the compute performance, some novelties were beneficial for the power consumption as well. The joint effects of substantial performance gains and decreasing (or stagnating, or at the very least not proportionally increasing) power consumption combined into quite a spectacular effect on compute efficiency.

As already mentioned in Section 4.1.1, with the rise of data-intensive applications, there's a growing need for faster, larger memory. Newer memory such as DDR4 and DDR5 boost speed and cut lag for quicker CPU data access. Better memory management optimizes how this works with large datasets, which is key for Java and the SSJ benchmark.

This performance boost, however, also comes with increased energy consumption, especially when memory is fully utilized. To counter this undesired effect at least to some extent, each generation of DDR memory works at a lower voltage. While the first-generation DDR required 2.5V, DDR4 works at 1.2V and DDR5 at 1.1V.[19] Additionally, both DDR4 and DDR5 memory technologies incorporate sophisticated deep power-down modes specifically designed to minimize energy consumption during periods when the memory is not actively engaged in data access operations. DDR 5 Modules for servers now use voltage converters on the DIMM, reducing the resistance losses for the low voltage power transport (a similar but less extreme approach as the vertical voltage conversion of 48 V within the chip package; described in Section 6.3 below) (Lenovo 2024).

DDR prioritizes low latency and all-around performance and is the memory of choice for general purpose computing. For GPUs, however, the main priority is the high bandwidth demands of graphics processing (initially) and ML algorithms (today). Hence, the DDR version optimized for the high bandwidth required by GPU applications is called GDDR (“graphics dual data rate synchronous dynamic random-access memory”). It has a wider memory bus and higher clock speeds than DDR, allowing it to transfer data to the GPU very quickly.[20] However, it typically has higher latency (delay) than DDR RAM.

[19] See https://www.corsair.com/us/en/explorer/diy-builder/memory/is-ddr5-better-than-ddr4/.
[20] See https://www.exxactcorp.com/blog/hpc/gddr6-vs-hbm-gpu-memory.

GDDR, which is currently in its 6th iteration GDDR6, has been the standard GPU memory since GPUs were used for their initial graphics purpose, before being deployed for AI applications. The matrices that are being multiplied in ML algorithms, however, and especially for the huge models that are being trained today (for generative AI and beyond) require even more bandwidth than traditional graphics applications. For the GPUs deployed in high-performance machine learning, GDDR is thus increasingly replaced by a new type of GPU-specific RAM, called "high-bandwidth memory" (HBM).

By contrast to traditional DDR or GDDR, HBM stacks several layers of RAM vertically (Patel, Xie, and Wong 2023). The stacked DRAM dies are linked by tiny electrical connections called Through Silicon Vias (TSVs) and bonded together. At the base of this stack sits a logic die that acts as a controller, managing the flow of data. This stashing is based on a fairly novel "2.5D" layering technique called CoWoS, "Chip on Wafer on Substrate" (Patel, Xie, and Wong 2023).

The HBM memory is also typically placed all around the GPU. Both measures aim at maximizing bandwidth, while at the same time minimizing trace lengths. The latter "measure in millimeters for HBM vs cm for GDDR and DDR" (Patel, Xie, and Wong 2023). Higher bandwidth and short transmission paths each contribute to a vastly improved performance. A desired side-effect of the former is also the massively improved energy efficiency of the data transfer, which next to the increased performance further contributes to efficiency gains.

Lower power consumption (or at least less rapid power consumption growth) through closeness to the GPU has also been the reason for developing HBM in the first place. It was originally developed in 2013 for AMD, and not for ML algorithms but for computer graphics. In 2008, AMD had predicted that constantly increasing memory bandwidth to keep up with gaming GPU performance would require so much power that it would hinder the GPU's processing capabilities, ultimately hurting overall performance. It hence approached its suppliers to develop a higher-bandwidth, more energy efficient memory, which ultimately resulted in the development of HBM in 2013 (Patel, Xie, and Wong 2023).

The first HBM was deployed in AMD's Fiji gaming GPUs in 2015, followed by the Vega series with HBM2 in 2017. However, HBM didn't significantly boost gaming performance and, due to its higher cost, AMD reverted to GDDR. Even today, the top gaming GPUs still use the more affordable GDDR6. While AMD was right that scaling memory bandwidth is challenging, this issue primarily affects data center GPUs. For consumer gaming, Nvidia and AMD now typically use large caches, allowing them to stick with lower-bandwidth GDDR memory.

GPUs in DCs, however, increasingly deploy HBM. All recent and current high-performance GPUs and TPUs use HBM memory: AMD's Radeon Pro V340 (August 2018) and Radeon Pro VII (June 2020), AMD's Instinct series (MI25 from June 2017, MI50 and MI60 from November 2018, MI100 from November 2020, MI210, MI250 and MI250x from November 2021, MI300 from January 2023, and MI300x from December 2023), all of Google's TPUs since v3 (May 2018) and all of Nvidia's Tesla (June 2016), Volta (December 2017), Pascal (May 2020), Hopper (March 2023), and Blackwell (March 2024) GPU series.

HBM is clearly necessary for the best possible GPU performance nowadays. While its short traces are indeed beneficial for energy efficient communication within the GPU, the sheer amount of memory (e.g., 192 GB in both AMD's Instinct MI300x and Nvidia's Blackwell B200 GPUs) also affects of course the overall power consumption, and thus also the efficiency. Hence, some of the less performant but also much more energy-frugal DC GPUs, such as the Nvidia's Ada Lovelace series (September 2022 – March 2023) continue using GDDR6 memory.

5.3 Storage

The most remarkable technology shift in data storage within the last decade was surely the shift towards non-moving SSD storage. Although in end-user devices SSDs are dominating the market, in DCs HDDs are still widely used according to a large storage provider (Herreria 2024). Both technologies usually consume roughly 5-15 Watt per device. SSDs have a great advantage in latency-sensitive applications, at random data access. Also, applications with very high I/O requirements, for which traditionally multiple (parallel) HDDs were used can be handled by SSDs usually much more energy efficient. These requirements are often needed in cloud environments with multiple virtual machines on one physical server, where the much faster access to random data brings a huge advantage to the various applications. Other, less frequent and more sequently written/read data can be handled still very efficiently on HDDs.

5.4 Deep dive: Microelectronic architectures tailored for machine learning

A substantial part of the DC growth today, both in terms of compute loads and resulting energy consumption, is due to ML training and inference (Google 2024; Microsoft 2024). As discussed in Section 2.3.2, over the past 20 years and taking advantage of the GPGPU paradigm, GPUs have seen a radical change of main usage from computer graphics to machine learning, and in particular deep learning algorithms.

Taking Nvidia as case study, this started in 2007 with the manufacturer's Tesla line of DC- and HPC-grade GPUs. Tesla GPUs were built until 2020, with numerous sub-generations such as Tesla Fermi, Tesla Kepler, Tesla Maxwell, Tesla Pascal, Tesla Volta, and Tesla Turing.[21] The Tesla line gradually became more focused on high-performance computing (HPC) and machine learning. Around 2012, with the Kepler architecture, Tesla GPUs started incorporating features specifically designed for deep learning, like improved dual-precision performance and optimized memory access. The Pascal architecture, launched in 2016, marked a significant turning point. With features like NVLink for faster GPU-to-GPU communication and improved deep learning performance, Pascal GPUs (such as the Tesla P100) were widely adopted for AI research and development.

Volta, launched in 2017, introduced tensor cores, specialized processing units designed specifically for the matrix operations used in deep learning. The substantial efficiency benefits of tensor cores for the matrix operations typical for ML algorithms have been addressed in Section 4.2. While this study calls them "tensor cores" using the Nvidia's terminology (the company that first introduced them), all other major manufacturers use similar concepts. AMD, for example, uses the term "matrix cores", and introduced them November 2020 with the Instinct MI100 GPU and along the novel CDNA architecture.[22]

The Ampere series (2020) and the more recent Hopper and Blackwell series (2023 and 2024, respectively) further enhanced deep learning capabilities with features like multi-instance GPU (MIG) for improved resource utilization and transformer engine for accelerating transformer models. MIG is a technology developed by Nvidia that allows a single physical GPU to be partitioned into multiple smaller, isolated virtual GPUs, each with its own dedicated resources

[21] See https://de.wikipedia.org/wiki/Nvidia_Tesla.
[22] See https://en.wikipedia.org/wiki/CDNA_(microarchitecture) and https://www.amd.com/en/technologies/cdna.html.

such as memory, compute cores, and cache.[23] This allows multiple workloads to run concurrently on a single physical GPU, improving resource utilization and efficiency.

Google's TPUs, meanwhile, use the same terminology of "tensor cores" as Nvidia, but referring to a different (albeit increasingly converging) concept. While this study addressed TPUs jointly with GPUs, they are in fact application-specific integrated circuits (ASICs), albeit specific not to performing hashes such as the ASICs discussed in Section 4.3, but for machine learning.[24]

Driven by the increasing computational demands of their AI models for various services, Google started exploring specialized hardware for machine learning around 2013 (Gartenberg 2024). Realizing that CPUs and even GPUs were becoming a bottleneck for the rapidly growing AI workloads, Google designed and deployed its first-generation TPUs internally in 2015. They were specifically tailored for the TensorFlow framework and optimized for inference. Publicly revealing their existence in 2016, the second-generation TPUs was launched in 2017, with enhanced performance for both training and inference. Via Cloud TPU, the technology also became available for external users via Google Cloud.

Although using the terminology "tensor cores", being a purpose-built ASIC, the entire TPU is essentially a tensor core in itself; a dedicated processor specifically designed for tensor operations, the fundamental building blocks of deep learning. A TPU does not have separate tensor cores within it, as a GPU does; the whole chip is optimized for that purpose. The key components of a TPU are:

- **Matrix Multiply Unit (**MXU): The heart of a TPU, responsible for performing high-speed matrix multiplications. It is analogous to Nvidia's tensor cores, but represents a more central part of the TPU's design.
- **Vector Unit**: Handles other vector computations and mathematical operations.
- **Scalar Unit**: Manages control flow and other scalar operations.

Intel follows a similar paradigm to Google: Its Habana ML processors (from the acquired company Habana Labs) are dedicated processors specifically developed for machine learning. The main product lines (Gaudi processors for training, and Greco processors for inference, respectively) are complete processors, not just specialized cores within a larger GPU. Within the Gaudi and Greco processors, there are dedicated hardware blocks for matrix multiplication and other tensor operations, like Google's MXUs.

As can be seen from this discussion, addressing GPUs and TPUs together stood to reason: They not only converge in their aim (i.e., machine learning loads), but also the architectures grow increasingly similar, with their focus on fast matrix operations.

At the same time, as discussed in Section 4.2, there was also substantial algorithmic and data representation progress over the past decade. Lower precisions substituted higher ones, trading precision for speed and amount of ML computing. Their analysis is not directly part of this study, which focuses on the technological efficiency, excluding software enhancements from its scope (See Table 1). Had they been considered (e.g., the shift from FP32 to FP16/BF16 with tensor cores and sparsity), then the overall performance gains would have been another order of magnitude or so larger. But even focusing solely on hardware-based efficiency, GPUs

[23] See https://www.nvidia.com/en-us/technologies/multi-instance-gpu/.
[24] See https://cloud.google.com/tpu/docs/intro-to-tpu.

and TPUs had performance and efficiency CAGRs over the past decade of 80% and 49%, respectively; a growth compatible with (and even superior to) Moore's law.

5.5 Power management

Many servers are often poorly utilized when looking at a 24/7 average values. This is a great motivation to get closer to a so-called energy-proportionality in computing (Barroso and Hölzle 2007). In servers, this is implemented through multiple approaches under the term "Power management" which recognize a low utilization and automatically reduce the capacity (e.g. frequency, voltage, active cores etc.) of a server, to save energy. The crucial part here is to be able to bring the capacities back within milliseconds, when required. Shehabi et al. (2024) assume (p.27) that the utilization of non-AI servers is still around 50% in hyperscale DCs, below 30% in Colocation DCs and below 20% in all other data centers, which underlines the importance of power management.

One of the most significant contributors to lowering the idle power consumption is the introduction of advanced power management techniques. Dynamic Voltage and Frequency Scaling (DVFS) plays a crucial role by allowing processors to adjust both their voltage and operating frequency according to the workload. When a server is idle or lightly loaded, DVFS reduces the power supplied to the processor, lowering energy consumption. Additionally, power gating is another essential technique that reduces power consumption by completely shutting off power to unused sections of the chip, such as inactive cores or memory controllers. This complements clock gating, which minimizes energy usage by halting the clock signal to inactive parts of the processor, reducing unnecessary switching activity within the circuits.

Most servers nowadays reach very low levels of power, when in idle state. But at many servers the power consumption rises more than proportional, when there is just a very small task (this can also be seen in Figure 6 at 10%). It can be assumed that many efficiency mechanisms are activated in active idle and switched off at even low utilizations. This at least partially counteracts the goal of energy-proportional computing somewhat, as many servers often operate in the 5% to 50% utilization range. Nevertheless, it is a great achievement of such regulations, that idle power consumption has remained almost constant for many servers even with much higher rated power.

The integration of smarter cache management systems has also contributed to energy savings. By optimizing how data is fetched and stored, modern processors reduce the need for costly memory access, which consumes far more power than on-chip cache operations.

Power efficiency improvements in servers are also driven by energy-efficient peripheral technologies. For instance, modern memory and I/O systems have been designed with power-saving modes that reduce energy consumption during idle periods. DDR4 and DDR5 memory, for example, include deep power-down modes that minimize power use when the memory is not actively being accessed. Similarly, PCIe interfaces support power management features like Active State Power Management (ASPM), which allows the I/O system to enter low-power states when idle. Additionally, the widespread adoption of solid-state drives (SSDs) over traditional hard disk drives (HDDs) has contributed to significant power savings. SSDs not only consume less power overall, but they also offer more advanced power management options.

Moreover, software optimizations at the operating system and hypervisor levels contribute to lower idle power usage. Modern operating systems make use of advanced power states (such as the C-states in x86 processors), where processors can enter progressively deeper sleep

states as their workloads diminish. Software-driven workload balancing also helps minimize the number of active servers, consolidating tasks so that idle servers can be powered down or placed in deep power-saving modes. In larger data center environments, AI and machine learning algorithms are being employed to analyze workload patterns and optimize resource allocation, ensuring that servers consume only the power required for their workloads.

5.6 Cooling of servers: The rise of liquid cooling

As server processing power and especially density continue to grow, traditional air-cooling methods are reaching their limits in effectively removing the generated heat, calling for different methods of heat removal. *Liquid cooling* is such a paradigm, which leverages the superior heat transfer properties of fluids as compared to air. As name suggests, a liquid is used to capture and remove the heat from servers, liquid which can be either a dielectric (such as oil) in direct contact with the components or any refrigerant (including water) pumped through cold plates that are attached to the heat-generating components (Schneider Electric 2024).

Liquid cooling is not a new idea. Its roots can be traced back to the mainframe era of the 1960s, with IBM's enterprise-grade computer line System/360 deploying a hybrid cooling using both air and liquid cooling (Anghel 2023). And while individual prototypes and demonstrators kept popping up ever since – such as IBM's Aquasar around 2010 (Zimmermann et al. 2012) – they remained until recently a niche product for specialized applications. As well-established, reliable and comparatively inexpensive technology that could scale well and has also become increasingly efficient, air cooling has dominated server cooling in data centers for decades.

Air cooling, however, reaches its physical limits at a certain power density of servers, and thus of cabinets in DCs. While these limits were believed to be around 20-30 kW per cabinet until a decade ago, thanks to advances in air moving technologies within both servers and DCs, they were later pushed towards 40-50 kW per cabinet. This newer limit, however, could not be pushed further so far, as stated by the "American Society of Heating, Refrigerating and Air-Conditioning Engineers" (ASHRAE TC9.9 2019) and confirmed by our interviewees.

For accelerated computing in particular, however, this limit has long been surpassed. As discussed in Section 4.2.2, the TDP of GPUs has increased over the last decade from around 300 W to 1-1.2 kW, while "superchips" consisting of both GPUs and CPUs can reach up to 2.7 kW. Correspondingly, GPU clusters such as Nvidia's NVL72 (which consists of 36 GB200 superchips, i.e. unites a total of 72 Blackwell GPUs and 36 Grace CPUs)[25] already surpass 100 kW of power.

As shown in Section 4.1.2, however, server CPUs also tend to require ever higher performance. Correspondingly, ASHRAE distinguishes three phases of CPUs in servers (ASHRAE TC9.9 2021):

- In the single-core phase (2000-2010), the performance of CPUs increased slowly but steadily.
- In the second phase of multi-core processors (approx. 2010-2018), performance increases were made possible by increasing the number of cores while distributing the same total power across more sources within the chip; this meant that power consumption was relatively constant.
- Since around 2018, however, the third phase has begun, which has been suggestively named the "power wars era" (Lawrence 2024): now performance increases have to be

[25] See https://www.nvidia.com/en-us/data-center/gb200-nvl72/.

bought with more power. This leads to increased power density, inducing what ASHRAE calls a higher "*degree of cooling difficulty*" (ASHRAE TC9.9 2021), yielding in turn liquid cooling ever more likely.

As liquid cooling is still a marginal phenomenon (our interviews indicated that it is well below 5% of current orders), it is not an important factor in past efficiency gains. As Section 6.3 discusses, however, it is likely an important future technology and thus a relevant factor for future efficiency trends.

5.7 The pace of efficiency gains and hardware refreshment cycles

The energy and environmental impact of hardware production is not within the system boundaries of this assessment, which focuses on the operational energy efficiency. However, the frequency of hardware refreshment cycles is directly relevant to the average server energy efficiency. Hence, this subsection discusses the relation between energy efficiency developments and hardware refreshment cycles.

In past decades, rapid gains in processing power and energy efficiency supported frequent hardware upgrades, with organizations benefiting from lower operational costs and increased performance. Today, however, as advancements in CPU efficiency face diminishing returns, while at the same time the production of devices receives more intense economic and environmental scrutiny, the traditional economic and environmental advantages of frequent upgrades become less attractive. This shift prompts a reconsideration of refresh cycles, with longer intervals potentially yielding greater overall benefits by reducing manufacturing impacts and electronic waste. For general-purpose servers, which commonly power data centers, efficiency improvements continue, but at a reduced pace as compared to efficiency gains in the 1990s or the early 2000s. The energy savings from newer chips are now smaller on a per-generation basis, and when balanced against both financial and environmental costs of manufacturing, transporting, and installing new hardware, extended refresh cycles may prove more economically and/or environmentally favorable.

The decreasing greenhouse gas (GHG) intensity of electricity in many regions further strengthens this argument. As electricity generation shifts increasingly toward low carbon power sources, the operational GHG emissions associated with data center power consumption decline, reducing the pressure to adopt more efficient hardware solely for emissions reduction. This trend suggests that existing hardware may remain in service longer, especially as incremental efficiency gains deliver smaller returns in terms of both cost and carbon savings. Consequently, in our interviews, the industry indicated that server lifespans have been lately expanded to 5-6 years, while just a few years ago they were in the range of 3-4 years (Moss 2023).

ASICs, being highly specialized for tasks like cryptocurrency mining or AI inference, represent a slightly different case. While these chips can quickly become obsolete due to rapid advancements in specific application areas (resulting e.g. in a decline in mining revenues), the principle of extending refresh cycles holds if energy efficiency gains per generation are modest. Furthermore, since the environmental impact of ASIC production can be considerable due to specialized manufacturing processes, longer refresh cycles may help mitigate the lifecycle GHG emissions associated with production. In scenarios where GHG emissions from electricity are low, extending the useful life of ASICs could prove environmentally beneficial, particularly if newer models offer limited efficiency improvements.

6 Outlook on future efficiency developments

Estimating possible further developments of the efficiency drivers presented in Section 5 is crucial for understanding feasible overall developments of server energy efficiency. Based on literature review and expert interviews, this section discusses the developments which seem to some extent foreseeable. These developments, however, are of course uncertain. After some likely developments for the next years, some of the drivers might change substantially and entirely new ones might appear.

6.1 Further potential in miniaturization

The production of modern microchips is already an enormous technological feat. Nevertheless, EUV chip manufacturing processes using lasers with a wavelength of 13.5 nm still offer room for improvement, although there is still only one manufacturer, ASML, that is capable of producing such devices. China is in the process of investing heavily in research regarding EUV technology; it remains unclear, however, whether this will succeed, as ASML holds many patents and utilizes large number of international (mainly Western) suppliers and their innovations (Lovati 2025).

Within the EUV lithography systems, the focus is currently on improvement of the optical systems to increase the numeric aperture towards 0,55 (high-NA) and 0,75 (hyper-NA). The first EUV high-NA machines are in test operations, e.g. at Intel (Morescalchi 2024)and TSMC (Trueman 2024) and according to (van Monsjou 2025; via Paul van Gerven 2025) ASML is just about to ship manufacturing-capable high-NA devices. Whether efficiency gains are possible at the same pace for the future depends primarily on how quickly the new chip manufacturing processes are technically capable of producing large quantities of chips at marketable prices. As described above, further progress in miniaturization is likely in the near future, particularly through high-NA EUV (and later hyper-NA EUV).

6.2 Further memory developments

The number of stacked memory layers grew with the HBM generations. While the first version – HBM1, first deployed in 2015 – had up to 4 such layers, the second generation of HBM2 doubled the maximum number of layers to 8.[26] The following HBM3 generation then had up to 12 layers[27] and the currently newest generation HMB 3e can stack up to 16 RAM layers.[28] This trend towards more layers, however, is slowing down: The next-generation HBM 4 HBM4 is anticipated to feature the same 16 layers in its release[29] most likely in 2026, while only from 2027 forward it might also feature 24 layers. Correspondingly, the memory increase is also moderate, from a maximum of 192 GB in HBM 3e to a maximum of 256 GB for HBM 4.[30] To achieve this, HBM 4 will deploy 5nm technology.

[26] See https://www.jedec.org/news/pressreleases/jedec-updates-groundbreaking-high-bandwidth-memory-hbm-standard.
[27] See https://news.skhynix.com/sk-hynix-develops-industrys-first-12-layer-hbm3/.
[28] See https://www.digitimes.com/news/a20240222PD215/sk-hynix-hbm-dram-2024-production.html.
[29] See https://www.jedec.org/news/pressreleases/jedec-approaches-finalization-hbm4-standard-eyes-future-innovations.
[30] See https://www.tomshardware.com/tech-industry/preliminary-hbm4-specs-point-to-major-performance-uplift-for-gpus.

Nevertheless, the current computer architecture is processor centric, a paradigm that can be traced back to John von Neumann's seminal contributions after WW2. Data processing (i.e., computing) only takes place in the processor. Despite the computation and efficiency gains brought about by ever-faster memory (both DDR and HBM), processors nowadays can thus waste up to 60% of their capacity simply waiting for data from storage or memory, and substantial energy on moving the data between storage/memory and processor (Mutlu 2023). This performance and energy bottleneck is sometimes referred to as "von Neumann bottleneck" (Wolters et al. 2024).

In-memory computing (IMC), also referred to as "compute-in-memory", "memory-centric computing", or "in-memory processing", is a potentially promising solution to this problem. In this paradigm, data storage is removed altogether, all data being stored in memory. It is then either processed using logic attached to the memory arrays ("Processing-near-Memory", PnM), or even without any logical processing unit ("Processing-using-Memory, PuM), which exploits the inherent analogue properties of memory for computation (Mutlu et al. 2024).

IMC optimizes communication times no longer in the microsecond range, but in the range of nanoseconds. By shortening communication distances even more or eliminating them altogether, it also has the potential to be more energy efficient. It is particularly well-suited for AI inference (Wolters et al. 2024). While numerous prototypes exist, IMC is not yet deployed on a large scale due to challenges related to programming, system support, costs, and analogue computation challenges such as tolerating circuit variation and noise specific to PuM (Mutlu 2023).

6.3 Shift in power conversion efficiencies

The power supply unit (PSU) converts the input voltage (often 115 or 230 V AC) into DC voltage of, for example, 12V, 5V or 3.3 V to supply the various components on a server board.

Since this conversion leads to thermal losses, which in the form of hot air also have to be transported out of the server and data center, there have been efforts to make PSUs energy-efficient for a very long time. Back in the 80s and 90s, PSUs usually had poor efficiency at low loads, especially due to their static losses, which is why it has always been relevant not to massively oversize them. Servers often have a redundant power supply to meet the N+1 redundancy criterion, which is on standby in case the regular power path fails. This reinforces the motivation to minimize static losses / improve efficiency at low utilization. Over the 2000s and 2010s, the 80plus standard evolved with it's bronze, silver, gold, platinum and titanium labels for energy efficient power supplies (from bronze to titanium ascending efficiency) (Mpitziopoulos 2018). A titanium PSU guarantees a 90% power conversion efficiency even at 10% load.

In recent years, more and more chip designs for GPUs and now also for CPUs are designed to allow much more than 300 Watt; even into the range of 1 kW. Such power can simply not be transported with very low voltage; therefor already in 2016, the 48 V server mainboard design was introduced in the Open Compute Project (OCP). This concept is quite promising in computing environments with high power density and power-hungry chips. Especially in combination with highly efficient power conversion directly at the Point of Load (POL) with wide band-gap semiconductors like Gallium Nitride (GaN) as DC-DC power conversion module. Using these high efficient semiconductors, the power conversion towards the very low voltage power for the chip usually does not happen laterally (besides the chip) anymore, but directly below or above it, by integrating the power conversion right below the chip, which is called

vertical power delivery (Krishnakumar and Partin-Vaisband 2023). Together with a more efficient AC-DC conversion (e.g. with SiC semiconductor-based converters), the total efficiency chain could be improved from 96% * 96% * 90% to 98% * 98% * 95%; reducing total losses to roughly half of previous power conversion architectures.

6.4 Wider adoption of liquid cooling

Section 5.6 discussed the rise of liquid cooling. So far, liquid cooling has been more of a niche technology, only marginally influencing overall server efficiency. With increasing power densities in both general purpose servers and – even more so – in AI-dedicated servers, however, liquid cooling becomes more and more of a necessity. As discussed in Section 5.6 above, some GPU racks already require more than 100 kW of power. During our interviews, we found out that this trend is likely to continue, and some next-generation racks will be designed for up to 300 kW.

While the expansion of liquid cooling seems unavoidable (as it has become a necessity), several types of liquid cooling exist, and it is hard to predict which will prevail or whether several will coexist. The basic design uses *cold plates* which circulate the coolant and are directly in contact with the heat-generating components, placed in contact with the cold plate. Cold plates evolved into *direct-to-chip (D2C) cooling*, where specially designed cold plates are placed directly onto the CPU or GPU die, the coolant thus being channeled through capillaries as close as possible to the heat-generating components. A quite different paradigm is called *immersion cooling*, where entire servers are submerged in a thermally conductive, dielectric fluid (Schneider Electric 2024).

Finally, *two-phase liquid cooling* leverages the latent heat of vaporization for heat removal. In these systems, the coolant is designed to boil at a relatively low temperature when it absorbs heat from server components. As the liquid coolant vaporizes, it absorbs a significant amount of heat (much more than just raising its temperature in one-phase cooling) and carries it away as vapor. This vapor is then condensed back into a liquid, typically in a heat exchanger, releasing the heat to a secondary cooling loop or the ambient environment. Two-phase systems often employ specialized dielectric fluids with low boiling points. Two-phase liquid cooling can be combined with both D2C cooling (where boiling happens at processor level) and immersion cooling (where servers are submerged in a bath of two-phase coolant). The interviews revealed that two-phase systems are still few and far between, but that customer interest in them is steadily growing.

As liquids are denser, have a higher specific heat capacity and a lower thermal resistance than air, heat can be removed with substantially less volumetric fluid flow as compared to air cooling (Lawrence 2024). And because the movement of fluid (whether gas or liquid) requires energy, liquid cooling thus has the potential to be more efficient than air cooling.

These efficiency gains are hard to quantify, though, and not only because the exact type of liquid cooling that will prevail is hard to predict. It is also impossible to tell now to which extent liquid cooling will substitute air cooling: in our interviews, the experts agreed that hybrid cooling is the likely future of data centers, in which the share of liquid cooling could be anywhere between 30-80%. Furthermore, it is also difficult to say how fast this substitution will happen: Retrofitting current DCs for liquid cooling is more challenging and costlier than foreseeing hybrid cooling from the outset for newly built data centers; how fast new DCs will be built depends however on various socio-economic and technological factors.

Finally, the energy savings induced by liquid cooling will occur both in the data center infrastructure (CRAC units etc.) and in servers themselves. In servers, they occur for two reasons: due to the diminished (or often entirely disappearing) server fans, but also because of less electric leakage at lower temperatures in the chip, for which up to 15% was recorded in some measurements (Thomas 2024). For the purpose of this study (i.e., the energy efficiency of servers), the infrastructure efficiency gains lie outside the boundaries and the server gains within, making an allocation of the efficiency gains necessary (if these were reasonably predictable).

As with the trend towards 3D packaging and 3D monolithic chips, the distance that heat needs to travel through the semiconductor is rising, as even liquid cooling usually happens on the surface. Even though there are some approaches to better conduct the heat to the surface (e.g. Deng et al. 2024), chip manufacturers are discussing for some years if there should be an intra-chip / or microfluid cooling be implemented directly into the chip (Judge 2024; Tyson 2021; Ao and Ramiere 2024).

6.5 Future hardware refreshment cycles

Section 5.7 discussed the lately longer hardware refreshment cycles of general-purpose servers, which is supported by both an economic and an environmental argument in the context of diminishing efficiency gains per generation and decarbonization of electricity. As energy grids worldwide transition toward lower-carbon power, the embodied carbon of new hardware – which is harder to decarbonize than operational electricity – becomes more relevant. From an environmental sustainability perspective, longer refresh cycles also align with circular economy principles, minimizing resource extraction, e-waste, and emissions from manufacturing.

However, this approach must be balanced against potential performance needs and the cumulative energy savings that could arise from strategic upgrades. This is also not to argue that operational electricity (and thus energy efficiency) would not remain crucial. Not only for computing, but also environmentally: the smaller the global overall electricity consumption, the easier it is to decarbonize.

But as the embodied carbon gains importance, organizations are increasingly adopting lifecycle assessments when planning refresh cycles, weighing the embodied carbon and resource costs of new equipment against the marginal efficiency gains of newer chips. Optimal refreshment strategies in this context aim not only to reduce operational costs but to balance environmental impacts across the hardware lifecycle, supporting both economic and environmental sustainability goals.

7 Conclusions

7.1 Summary of key conclusions

The analysis of energy efficiency across servers, GPUs, and ASICs in this study revealed substantial variation in measured data points, reflecting the inherent challenges in deriving clear indicators for technology trends. Across all categories, however, it can be relatively clearly deduced that the hardware is becoming much more performant (rather exponentially), the power consumption per device is increasing (rather linearly, but also strongly), and thus the efficiency in terms of computations per watt is improving significantly. Specialized hardware (GPUs, TPUs and ASICs) has established itself in the application areas of extremely high workloads and complex calculations such as crypto mining and machine learning, where it creates massive performance and efficiency leaps that cannot be achieved with general purpose CPUs. Currently it can be assumed, that the servers with two CPU sockets have still the biggest impact in the overall market, but the market for GPU based servers is currently growing much faster and could be more important within the next years (Shehabi et al. 2024).

7.1.1 Continuing efficiency gains, but at differing paces

With the results of this study, it can be quantitatively shown that general purpose servers as well as graphics cards, special AI chips such as TPUs and ASICS are still developing very dynamically in terms of their performance per energy input. For general purpose servers, this development is not quite as fast as in earlier times of Moore's Law. GPUs, TPUs and ASICs, on the other hand, have shown efficiency gains compatible with (or even higher than) Moore's law over the last years. As they are younger technologies, it is perhaps not surprising that there was (and most likely still is) more untapped potential to exploit.

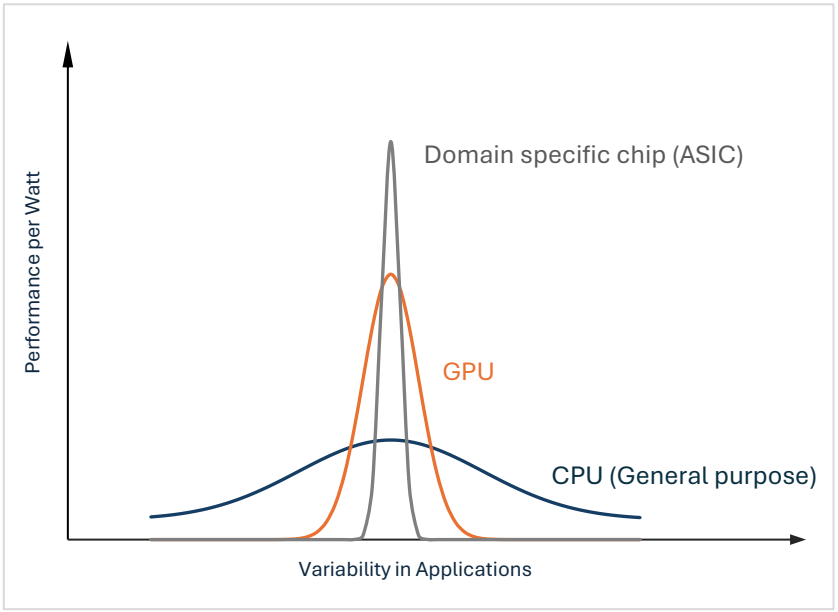


Figure 23: Workload specific chips vs. general purpose CPUs qualitative illustration, derived from (Naffziger 2023)

A significant challenge in assessing the energy efficiency of different hardware types lies in their diverse architectures and workload specializations. ASICs, for example, are designed for peak efficiency in single-purpose tasks, unlike general-purpose servers and GPUs, which are adaptable to a range of operations but often at a higher energy cost per task (see Figure 23). This specialization is a critical factor to consider when interpreting energy metrics, as ASICs will consistently outperform general-purpose hardware in narrowly defined tasks but cannot

compete on flexibility. In contrast, GPUs, while adaptable to multiple tasks, require parallel workloads to achieve energy efficiency, meaning that specific workload matching is crucial for accurate energy assessments. This is why very different metrics like SSJ-OP/s, Teraflop/s and Terahashes/s had to be considered when discussing the performance – and thus the energy efficiency – of compute chips and servers. It becomes most difficult within the general-purpose server segment, as their circuits might be optimized for different workloads, which is probably also a reason for the very widely varying results shown by the benchmarks, especially for very recent models.

7.1.2 Main drivers for efficiency and possible developments

For general purpose servers, efficiency improvements stem from various factors such as faster CPUs with more cores and larger caches coupled with faster memory technologies, advances in multi-threading and virtualization which enable more efficient resource utilization, faster interconnects, but also faster networks and storage and optimized server-side software. Chiplet designs and hybrid CPUs enable greater scalability and resource sharing. Improved idle power management also contributed to overall general purpose server efficiency.

For GPUs, the most important driver was arguably the trend towards the inclusion of tensor cores, which are specialized processing units within GPUs that accelerate the matrix multiplications crucial to deep learning. The emergence of tensor cores accelerated the technological convergence of GPUs towards TPUs, which are ASICs designed specifically for this task of matrix operations. Despite their name, DC-grade GPUs thus become less and less suited for graphics, and increasingly fit for their main usage nowadays, ML training and inference.

Faster memory also played a crucial role for both CPUs and GPUs. Not only faster, in fact, but also memory that is better suited for its task. CPUs saw DDR4 and DDR5 memory emerge, while for GPUs, GDDR memory has been increasingly replaced by HBM. Its multi-layer stack of RAM enables high bandwidth, low-latency communication to the processor, which for the large volumes of matrices involved in ML is of crucial importance. To further increase memory bandwidth and decrease the communication distance between processing place and data, in-memory computing might grow in importance in future.

The greatest increases in performance and efficiency continued to be achieved through the miniaturization of circuits in chips, and similar increases are expected to continue. At the same time, however, for both CPUs and GPUs, the performance gains have partly been achieved through increased per-chip or per-server power consumption. Over the last decade, the mean TDP of data center general purpose servers increased by a factor of 2-3, while for data center GPUs, it increased by a factor of 3-4. These trends show that performance gains were not only achieved by miniaturization but also by increasing the size (more parallel transistors of the same node size) – and, correspondingly, increased power consumption. As efficiency is performance divided by power, the efficiency gains are naturally reduced by these factors as compared to the performance gains. This tendency towards more power density also has implications for server cooling, as well be addressed below.

Traditional air-cooling methods are reaching their limits in removing heat from increasingly powerful and dense servers, requiring alternative solutions such as liquid cooling. With historic roots dating back to the 1960s, liquid cooling remained a niche technology so far due to the effectiveness, reliability, and scalability of air cooling. However, air cooling reaches its limit at around 40-50 kW per cabinet, a limit now surpassed by high-performance computing components such as GPU racks, which can exceed 100 kW and are projected to further

increase their power density. The share of liquid cooling of servers is thus bound to grow, most likely leading to more frequent hybrid cooling of data centers. This will also increase server efficiency as it reduces or eliminates the need for fans.

Efficiency in the area of voltage conversion within the server is already very advanced and only losses in the single-digit percentage range remain. Nevertheless, optimization is still relevant, especially with regard to the distance from voltage conversion to consumption, as the power consumption per chip continues to increase and the very high-power consumption of modern CPUs and GPUs at extremely low voltage (<1.5V) leads to very high currents and thus losses. With the new paradigm of vertical voltage conversion, it is expected that losses can be significantly reduced once again compared to conventional conversion.

As the old adage goes, "predictions are difficult, particularly about the future". Quantifying the future development of server energy efficiency is not only subject to many drivers (only some of which are technological, others being socio-economic), but also to various ontological uncertainties.

This study therefore focused on identifying the most influential technological drivers of server efficiency. It now wraps up by assessing for each of these drivers their potential for i) inducing efficiency gains, and ii) the expected adoption by 2030. Given the high uncertainties, both attributes are presented on an ordinal scale, as presented in Table 3. They reflect the subjective assessment of the authors based on the analysis throughout this study, which is itself rooted in the literature review and interviews performed for this study.

While this approach cannot be directly deployed in e.g. future server energy modeling, it can inform such assessments and models on the developments that are likely to be more important (and thus deserve more focus, careful methods, or uncertainty analyses). It can further serve to validate already existing models.

Table 3 Drivers for energy efficiency in servers, together with the author's assessment of their potential for energy gains and likely adoption by 2030, both on an ordinal scale reaching from 1 to 5.

Efficiency driver	***Potential efficiency gains (1 – 5)***	***Expected adoption by 2030 (1 – 5)***	***Brief explanation***
Further miniaturization	4	3 – 4	Technology mature to continue at 26% / year; immediate large-scale adoption guaranteed.
Memory improvements & in-memory computing	3	2	In-memory computing has solid potential but confronts adoption challenges and is not universally deployable.
Power management	1	4	Further efficiency improvements are almost certain, but from a high level; remaining potential about 20% overall.
Liquid cooling	1	2 – 3	Liquid cooling an increasing necessity, but with slow progress and limited efficiency implications.
HW refreshment cycles	-1	3	Slower refreshment cycles slow down efficiency gains; adoption high for general-purpose servers due to diminishing returns.

As shown in Table 3, it is expected that in the near future, the main efficiency gains will continue to be delivered by the miniaturization of circuits in chips. These come with both strong efficiency

gains, and also with a fairly rapid diffusion rate. Further memory improvements, including mew paradigms, such as in-memory computing, are also expected to continue and will have an important influence; their adoption, however, is likely to proceed at a more moderate pace. By contrast, improved power management and liquid cooling can only bring modest efficiency gains, and adoption rate of the latter is also expected to be slow. Finally, the length of hardware refreshment cycles is inversely correlated with efficiency gains; however, after it has grown substantially over the past years, changes from now on are likely to be only incremental; hence, the overall effect is rather small as well.

Overall, given that the important trends are expected to continue for some time and no obvious efficiency revolution is on the radar, the best guess for the near future is that hardware efficiency gains[31] will continue along similar rates. Assuming thus that the improvements in efficiency will proceed with the same average growth rate (26% p.a. in SERT 2 over all systems), the SERT 2 efficiency would reach values above 250 in 2030 (see Figure 24). Tweaking the values from Table 3 due to different expectations towards individual efficiency gains or adoption rates, would correspondingly shift the curve from Figure 24 up- or downwards.

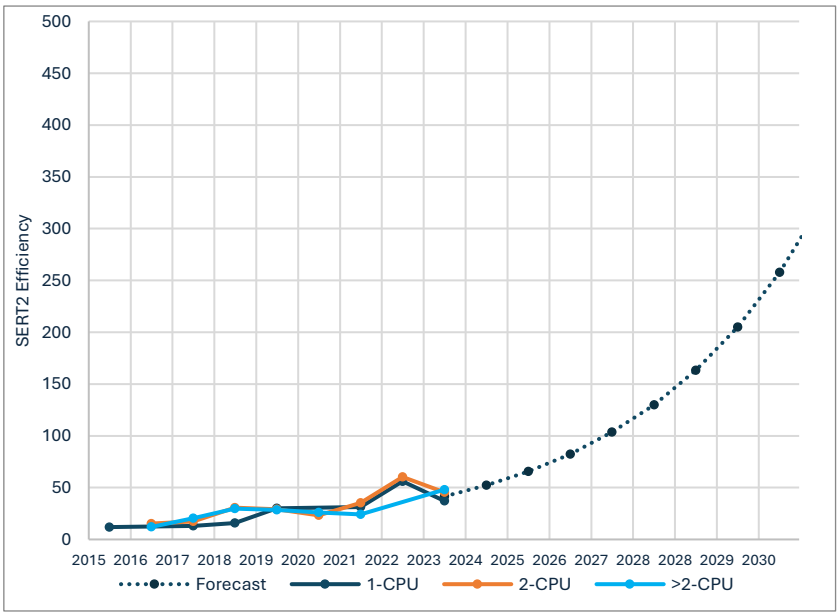


Figure 24: Extrapolation of the average efficiency improvements to 2030

Assuming the improvements in efficiency will proceed with the same average growth rate (26% p.a. in SERT 2 over all systems), the SERT 2 efficiency would reach 200 in 2030 (see Figure 24).

7.2 Recommendations

7.2.1 Miniaturization and energy efficient chip design is the key factor for energy efficiency

Modern processors with a mix of high-performance (P-cores) and energy-efficient cores (E-cores) allow servers to dynamically allocate workloads based on task requirements. Efficient processors also benefit from advanced manufacturing nodes (e.g., 7nm or below), which reduce power consumption by decreasing the energy needed per operation. Additionally, features like Dynamic Voltage and Frequency Scaling (DVFS) allow the processor to adapt its

[31] Next to hardware efficiency gains, the efficiency of computing is also subject to algorithmic efficiency gains. These are particularly relevant for AI, with an estimated doubling time of just 8.4 months on average (Ho 2024), but they lie outside the scope of this study.

power use in real time, based on workload demands, saving energy when full performance isn't needed.

7.2.2 Automatic hardware allocation based on more detailed performance indicators

The diversification of CPU architectures - with E-cores (Efficient Cores), P-cores (Performance Cores) and additional chiplets - reflects a shift towards hardware tailored for heterogeneous workloads. These innovations allow CPUs to balance energy efficiency and performance by matching core types to task requirements. However, this adds complexity to hardware selection in multi-tenant, virtualized environments such as cloud Infrastructure-as-a-Service (IaaS), where a wide range of workloads such as machine learning, database management and microservices coexist. Static hardware allocation can result in underutilized resources or inefficient power consumption. A dynamic allocation system that matches VMs to optimal hardware configurations would achieve better efficiency by using hardware tailored to specific tasks.

In this adaptive approach, each hardware configuration would undergo standardized benchmarking to generate 'capability profiles' that reflect performance in relevant areas (e.g. matrix operations, I/O throughput). At the same time, each VM request would be assigned a 'task demand profile' based on expected requirements (e.g. floating-point computation for ML or low-latency memory access for web hosting). A resource orchestration system would then match these profiles and dynamically allocate VMs to the most appropriate hardware.

To continuously improve this process, a feedback loop would monitor performance and refine both task and hardware profiles over time using machine learning. This automated, adaptive allocation could optimize energy efficiency and performance by ensuring workloads receive the most appropriate hardware, reducing over-provisioning and maximizing performance per watt, contributing to more sustainable cloud operations.

7.2.3 Server virtualization and utilization

In the study analysis it was found that servers aren't perfectly energy proportional. Especially the fact that in low utilization levels, the efficiency in terms of Performance/Watt is very bad, this is a relevant recommendation towards those who operate servers (even if data centers itself are not in the scope of the study). High server utilization through consolidation is therefore critical for overall energy efficiency in a system of servers like a compute cluster or a data center. Low utilization leads to "stranded" energy, where servers draw power without contributing proportionately to computational tasks, effectively wasting energy, even if the idle consumption has been reduced as shown in section 4.1.2. Modern servers are often underutilized, with utilization rates in some data centers averaging only 10–30%, leaving substantial room for improvement through workload consolidation and virtualization. By increasing server utilization, data centers can reduce the number of active servers needed, cutting both energy consumption and cooling requirements, thus improving operational efficiency and sustainability. A next step would be to switch at least some of the servers that are currently not in use (e.g. during the night) to an active standby mode (<5 watts), from which they can be reactivated via the network within milliseconds when needed, just like a modern smartphone.

7.2.4 Energy efficient power-supply-units in servers

Energy-efficient power supply units (PSUs) are crucial for overall energy efficiency in data centers because they minimize energy loss during power conversion, reducing the amount of electricity drawn from the grid. By decreasing wasted energy, efficient PSUs lower the heat

output of servers, which in turn reduces the cooling demand and saves on additional power consumption. Standard PSUs in modern data centers already achieve efficiencies around 90–94% under typical load conditions (Van Geet and Sickinger 2024), but further gains—especially approaching 96% or higher—are still possible with optimized designs and advanced power conversion technologies like 48V at rack level. These incremental improvements could still reduce both operational costs and the environmental footprint of large-scale server facilities.

7.2.5 Energy efficient cooling methods

The cooling of the data center and the server have a significant impact on the energy efficiency of the overall system. While the power usage effectiveness (PUE) is an effective measure of the data center's infrastructure losses (cooling, UPS, lighting, etc.), the cooler in the server can also result in a notable increase in energy consumption. In specific instances, this can result in an increase of over 20% in the power consumption of the server. Experimental evidence indicates that environmental conditions, including the supply and exhaust air temperature and airflow of a server, must be carefully calibrated to achieve optimal energy efficiency (Sarkinen et al. 2020). Insufficient attention to these factors can result in significantly higher total power consumption than necessary, even when PUE values are excellent.

For certain contemporary high-performance chips (CPUs, GPUs and TPUs), manufacturers have already advised lowering the supply air temperatures below the ranges recommended by ASHRAE. This results in elevated energy requirements for cooling. In such instances, it would be prudent to consider the use of increasingly liquid-cooled systems (immersion and on-chip), as higher power densities are not an issue here due to enhanced heat capacity and transfer. Server manufacturers are also offering heat sinks for liquid cooling as standard for an ever-increasing range of products, and there is no risk of losing the warranty or incurring significant costs for the cooling system.

7.2.6 Regulation and standardized metrics

It is very important that regulators from different economic areas, such as the EU, US and China, harmonize which parameters they regulate as efficiency criteria and how these are defined. Good coordination with industry (e.g. ITI The Green Grid, ITU) and science and research is essential. Ongoing monitoring and adaptation of regulations is particularly necessary in the area of new capacities being built for AI. Here, policymakers need to be very careful to walk the fine line between efficiency targets and openness to innovation, in order to avoid being left behind technologically or, at the other extreme, operating very inefficient hardware.

Literature

AMD. 2024. ‘AMD 3D V-Cache™ Technologie’. *AMD* (blog). 2024. https://www.amd.com/de/products/processors/technologies/3d-v-cache.html.

Anghel, Vlad-Gabriel. 2023. ‘An Introduction to Liquid Cooling in the Data Center’. 28 March 2023. https://www.datacenterdynamics.com/en/analysis/an-introduction-to-liquid-cooling-in-the-data-center/.

Ao, Lihong, and Aymeric Ramiere. 2024. ‘Through-Chip Microchannels for Three-Dimensional Integrated Circuits Cooling’. *Thermal Science and Engineering Progress* 47 (January):102333. https://doi.org/10.1016/j.tsep.2023.102333.

ASHRAE TC9.9. 2019. ‘Water-Cooled Servers. Common Designs, Components, and Processes’. American Society of Heating, Refrigerating and Air-Conditioning Engineers. https://www.ashrae.org/File%20Library/Technical%20Resources/Bookstore/WhitePaper_TC099-WaterCooledServers.pdf.

———. 2021. ‘Emergence and Expansion of Liquid Cooling in Mainstream Data Centers’. American Society of Heating, Refrigerating and Air-Conditioning Engineers. https://www.ashrae.org/file%20library/technical%20resources/bookstore/emergence-and-expansion-of-liquid-cooling-in-mainstream-data-centers_wp.pdf.

Barroso, Luiz André, and Urs Hölzle. 2007. ‘The Case for Energy-Proportional Computing’. *IEEE Computer* 40. http://www.computer.org/portal/site/computer/index.jsp?pageID=computer_level1&path=computer/homepage/Dec07&file=feature.xml&xsl=article.xsl.

Bedford Taylor, Michael. 2017. ‘The Evolution of Bitcoin Hardware’. *Computer* 50 (9): 58–66. https://doi.org/10.1109/MC.2017.3571056.

Belkhir, Lotfi, and Ahmed Elmeligi. 2018. ‘Assessing ICT Global Emissions Footprint: Trends to 2040 & Recommendations’. *Journal of Cleaner Production* 177 (March):448–63. https://doi.org/10.1016/j.jclepro.2017.12.239.

Bolz, Jeff, Ian Farmer, Eitan Grinspun, and Peter Schröder. 2003. ‘Sparse Matrix Solvers on the GPU: Conjugate Gradients and Multigrid’. *ACM Trans. Graph.* 22 (3): 917–24. https://doi.org/10.1145/882262.882364.

Bremer, Christina, George Kamiya, Pernilla Bergmark, Vlad C. Coroamă, Eric R. Masanet, and Reid Lifset. 2023. ‘Assessing Energy and Climate Effects of Digitalization: Methodological Challenges and Key Recommendations’. nDEE Framing Paper Series. Research Coordination Network on the Digital Economy and the Environment. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4459526.

CBECI. 2024. ‘Cambridge Blockchain Network Sustainability Index’. 2024. https://ccaf.io/cbnsi/cbeci/methodology.

Chellapilla, Kumar, Sidd Puri, and Patrice Simard. 2006. ‘High Performance Convolutional Neural Networks for Document Processing’. In . Suvisoft. https://inria.hal.science/inria-00112631.

Christine Middleton. 2024. ‘5 Things You Should Know about High NA in EUV’. ASML. 2024. https://www.asml.com/en/news/stories/2024/5-things-high-na-euv.

Coroamă, Vlad C. 2021. ‘Blockchain Energy Consumption. An Exploratory Study’. 68053. Bern, Switzerland: Swiss Federal Office of Energy SFOE. https://www.aramis.admin.ch/Default?DocumentID=68053.

———. 2022. ‘Exploring the Energy Consumption of Blockchains through an Economic Threshold Approach’. In *2021 Joint Conference - 11th International Conference on Energy Efficiency in Domestic Appliances and Lighting & 17th International Symposium on the Science and Technology of Lighting (EEDAL/LS:17)*, 1–10. https://ieeexplore.ieee.org/abstract/document/9932250.

Deng, Yaohui, Peisheng Liu, Zhao Zhang, Jiajie Jin, Pengpeng Xu, and Lei Yan. 2024. ‘3D Package Thermal Analysis and Thermal Optimization’. *Case Studies in Thermal Engineering* 64 (December):105465. https://doi.org/10.1016/j.csite.2024.105465.

Gartenberg, Chaim. 2024. ‘TPU Transformation: A Look Back at 10 Years of Our AI-Specialized Chips’. 31 July 2024. https://cloud.google.com/transform/ai-specialized-chips-tpu-history-gen-ai.

Google. 2024. ‘2024 Environmental Report’. Google. https://www.gstatic.com/gumdrop/sustainability/google-2024-environmental-report.pdf.

Govind, Anish, Sridutt Bhalachandra, Zhengji Zhao, Ermal Rrapaj, Brian Austin, and Hai Ah Nam. 2023. ‘Comparing Power Signatures of HPC Workloads: Machine Learning vs Simulation’. In *Proceedings of the SC ’23 Workshops of the International Conference on High Performance Computing, Network, Storage, and Analysis*, 1890–93. SC-W ’23. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3624062.3624274.

hashrateindex.com. 2024. ‘Bitcoin And Crypto ASIC Miner Profitability’. 2024. https://hashrateindex.com/rigs.

Herreria, Anne. 2024. ‘A Balancing Act: HDDs and SSDs in Modern Data Centers’. Western Digital Corporate Blog. 25 July 2024. https://blog.westerndigital.com/a-balancing-act-hdds-and-ssds-in-modern-data-centers/.

Hintemann, Ralph, Monika Graß, Simon Hinterholzer, and Tim Grothey. 2022. ‘Rechenzentren in Deutschland: Aktuelle Marktentwicklungen, Stand 2022’. Berlin: Bitkom. https://www.bitkom.org/Bitkom/Publikationen/Rechenzentren-in-Deutschland-2022.

Hintemann, Ralph, and Simon Hinterholzer. 2019. ‘Energy Consumption of Data Centers Worldwide - How Will the Internet Become Green?’ In *Proc. of the 6th Int. Conf. on ICT for Sustainability (ICT4S)*. Lappeenranta, Finland. http://ceur-ws.org/Vol-2382/ICT4S2019_paper_16.pdf.

Hintemann, Ralph, Simon Hinterholzer, F Montevecchi, and T Stickler. 2020. ‘Energy-Efficient Cloud Computing Technologies and Policies for an Eco-Friendly Cloud Market’. Berlin, Vienna: Borderstep Institute & Environment Agency Austria. https://op.europa.eu/en/publication-detail/-/publication/bf276684-32bd-11eb-b27b-01aa75ed71a1/language-en/format-PDF/source-183168542.

Ho, Anson. 2024. ‘Algorithmic Progress in Language Models’. *Epoch AI* (blog). 12 March 2024. https://epoch.ai/blog/algorithmic-progress-in-language-models.

Hobbhahn, Marius, Lennart Heim, and Gökçe Aydos. 2023. ‘Trends in Machine Learning Hardware’. Epoch AI. https://epoch.ai/blog/trends-in-machine-learning-hardware.

Judge, Peter. 2024. ‘Microfluidics: Cooling inside the Chip’. 2024. https://www.datacenterdynamics.com/en/analysis/microfluidics-cooling-inside-the-chip/.

Kamiya, George, and Vlad C. Coroamă. 2025. ‘Data Centre Energy Use: Critical Review of Models and Results’. IEA 4E TCP Efficient, Demand Flexible Networked Appliances (EDNA).

KJ Jacoby. 2023. ‘Tech Explainer: What’s the Deal with AMD’s 3D V-Cache Technology?’ Performance Intensive Computing. 14 November 2023. https://www.performance-intensive-computing.com/objectives/tech-explainer-what-s-the-deal-with-amd-s-3d-v-cache-technology.

Krishnakumar, Sriharini, and Dr Inna Partin-Vaisband. 2023. ‘Vertical Power Delivery for Emerging Packaging and Integration Platforms -- Power Conversion and Distribution’. arXiv. https://doi.org/10.48550/arXiv.2309.10141.

Krüger, Jens, and Rüdiger Westermann. 2003. ‘Linear Algebra Operators for GPU Implementation of Numerical Algorithms’. *ACM Trans. Graph.* 22 (3): 908–16. https://doi.org/10.1145/882262.882363.

Lawrence, Stuart. 2024. ‘Liquid Cooling Is More Sustainable. And It’s Coming to a Data Center near You.’ Stream Data Centers.

Lenovo. 2024. ‘Introduction to DDR5 Memory’. https://lenovopress.lenovo.com/lp1618.pdf.

Lovati, Stefano. 2025. ‘China Invests €37 Billion to Develop Domestic EUV Lithography Systems’. *Power Electronics News* (blog). 11 February 2025. https://www.powerelectronicsnews.com/china-invests-e37-billion-to-develop-domestic-euv-lithography-systems/.

Malmodin, Jens, Nina Lövehagen, Pernilla Bergmark, and Dag Lundén. 2024. ‘ICT Sector Electricity Consumption and Greenhouse Gas Emissions – 2020 Outcome’. *Telecommunications Policy*, January, 102701. https://doi.org/10.1016/j.telpol.2023.102701.

Masanet, Eric, Arman Shehabi, Nuoa Lei, Sarah Smith, and Jonathan Koomey. 2020. ‘Recalibrating Global Data Center Energy-Use Estimates’. *Science* 367 (6481): 984–86. https://doi.org/10.1126/science.aba3758.

Microsoft. 2024. ‘2024 Environmental Sustainability Report’. Microsoft. https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RW1lMjE.

Monsjou, Daan van. 2025. ‘ASML levert “in de komende maanden” eerste high-NA-machine voor massaproductie’. Tweakers. 2025. https://tweakers.net/nieuws/231220/asml-levert-in-de-komende-maanden-eerste-high-na-machine-voor-massaproductie.html.

Morescalchi, Daniela. 2024. ‘With High NA EUV, Intel Foundry Opens New Frontier in Chipmaking’. Newsroom. 18 April 2024. https://newsroom.intel.com/de/intel-foundry/intel-foundry-opens-new-frontier-chipmaking.

Moss, Sebastian. 2023. ‘Google Increases Server Life to Six Years, Will Save Billions of Dollars - DCD’. 2023. https://www.datacenterdynamics.com/en/news/google-increases-server-life-to-six-years-will-save-billions-of-dollars/.

Mpitziopoulos, Aris. 2018. ‘Tom’s Explains: What Do 80 PLUS Bronze, Silver, Gold & Titanium Signify?’ Tom’s Hardware. 2018. https://www.tomshardware.com/news/what-80-plus-levels-mean,36721.html.

Mutlu, Onur. 2023. ‘Memory-Centric Computing’. arXiv. https://doi.org/10.48550/arXiv.2305.20000.

Mutlu, Onur, Ataberk Olgun, Geraldo F. Oliveira, and Ismail E. Yuksel. 2024. ‘Memory-Centric Computing: Recent Advances in Processing-in-DRAM (Invited)’. In *2024 IEEE*

International Electron Devices Meeting (IEDM), 1–4. https://doi.org/10.1109/IEDM50854.2024.10873410.

Naffziger, Sam. 2023. 'How AMD Is Advancing the 30x25 Energy Efficiency Goal in High-Performance Computing and AI'. AMD.Com. 2023. https://community.amd.com/t5/corporate/how-amd-is-advancing-the-30x25-energy-efficiency-goal-in-high/ba-p/650223.

Patel, Dylan, Myron Xie, and Gerald Wong. 2023. 'AI Capacity Constraints - CoWoS and HBM Supply Chain', 5 July 2023. https://www.semianalysis.com/p/ai-capacity-constraints-cowos-and.

Patrizio, Andy. 2025. 'Graviton Progress: 50% of New AWS Instances Run on Amazon Custom Silicon | Network World'. 2025. https://www.networkworld.com/article/3631134/graviton-progress-50-of-new-aws-instances-run-on-amazon-custom-silicon.html.

Paul van Gerven. 2025. 'Shipment of ASML's High-NA Tools to Start "within a Few Months" - Bits&Chips'. Https://Bits-Chips.Com/. 30 January 2025. https://bits-chips.com/article/shipment-of-asmls-high-na-tools-to-start-within-a-few-months/.

Raina, Rajat, Anand Madhavan, and Andrew Y. Ng. 2009. 'Large-Scale Deep Unsupervised Learning Using Graphics Processors'. In *Proceedings of the 26th Annual International Conference on Machine Learning*, 873–80. ICML '09. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/1553374.1553486.

Ryan, Paul, Terence Smith, and Anson Wu. 2019. 'Total Energy Model for Connected Devices'. International Energy Agency. https://www.iea-4e.org/wp-content/uploads/2020/11/A2b_-_EDNA_TEM_Report_V1.0.pdf.

———. 2021. 'Total Energy Model V2.0 for Connected Devices'. International Energy Agency. https://www.iea-4e.org/wp-content/uploads/publications/2021/02/EDNA-TEM2.0-Report-V1.0-Final.pdf.

Sarkinen, Jeffrey, Rickard Brännvall, Jonas Gustafsson, and Jon Summers. 2020. *Experimental Analysis of Server Fan Control Strategies for Improved Data Center Air-Based Thermal Management* *. https://doi.org/10.1109/ITherm45881.2020.9190337.

Schneider Electric. 2024. 'Navigating Liquid Cooling Architectures for Data Centers with AI Workloads'. https://www.se.com/uk/en/download/document/SPD_WP133_EN/.

Shehabi, Arman, Sarah Smith, Alex Hubbard, Alex Newkirk, Nuoa Lei, Md Abu Bakar Siddik, Billie Holecek, Jonathan Koomey, Eric Masanet, and Dale Sartor. 2024. '2024 United States Data Center Energy Usage Report'. LBNL-2001637. Berkeley, CA: Ernest Orlando Lawrence Berkeley National Laboratory. https://eta-publications.lbl.gov/sites/default/files/2024-12/us_data_center_energy_usage_report_lbnl-2001637_0.pdf.

Shehabi, Arman, Sarah Josephine Smith, Dale A. Sartor, Richard E. Brown, Magnus Herrlin, Jonathan G. Koomey, Eric R. Masanet, Nathaniel Horner, Inês Lima Azevedo, and William Lintner. 2016. 'United States Data Center Energy Usage Report'. LBNL-1005775. Lawrence Berkeley National Laboratory. https://eta-publications.lbl.gov/sites/default/files/lbnl-1005775_v2.pdf.

Simili, Emanuele, Gordon Stewart, Samuel Skipsey, Dwayne Spiteri, and David Britton. 2022. 'Power Efficiency in HEP (X86 vs. Arm)'. https://indico.cern.ch/event/1106990/contributions/4991256/.

SPEC. 2018. ‘SPECpower_ssj2008 Result File Fields’. 2018. https://www.spec.org/power/docs/SPECpower_ssj2008-Result_File_Fields.html.

Steinkraus, D., I. Buck, and P.Y. Simard. 2005. ‘Using GPUs for Machine Learning Algorithms’. In *Eighth International Conference on Document Analysis and Recognition (ICDAR’05)*, 1115-1120 Vol. 2. https://doi.org/10.1109/ICDAR.2005.251.

Sun, Yifan, Nicolas Agostini, Shi Dong, and David Kaeli. 2019. *Summarizing CPU and GPU Design Trends with Product Data*.

Tahmid Noor Rahman, Nusaiba Khan, and Zarif Ishmam Zaman. 2024. ‘Redefining Computing: Rise of ARM from Consumer to Cloud for Energy Efficiency’. *World Journal of Advanced Research and Reviews* 21 (1): 817–35. https://doi.org/10.30574/wjarr.2024.21.1.0017.

The Shift Project. 2019. ‘Lean ICT: Towards Digital Sobriety’. https://theshiftproject.org/en/article/lean-ict-our-new-report/.

Thomas, Albert. 2024. ‘Testing the Impact of Liquid Cooling on Intel’s Core Ultra 7 265K Using the Tryx Panorama 240’. Tom’s Hardware. 2024. https://www.tomshardware.com/pc-components/liquid-cooling/testing-the-impact-of-liquid-cooling-on-intels-core-ultra-7-265k-using-the-tryx-panorama-240.

Trueman, Charlotte. 2024. ‘TSMC to Receive First High NA EUV Lithography Machine from ASML in Q4’. 2024. https://www.datacenterdynamics.com/en/news/tsmc-to-receive-first-high-na-euv-lithography-machine-from-asml-in-q4/.

Tyson, Mark. 2021. ‘TSMC Reckons Intrachip Cooling Might Become Necessary Soon - Cooling - News - HEXUS.Net’. 2021. https://m.hexus.net/tech/news/cooling/148094-tsmc-reckons-intrachip-cooling-might-become-necessary-soon/.

Van Geet, Otto, and David Sickinger. 2024. ‘Best Practices Guide for Energy-Efficient Data Center Design’. Colorado: National Renewable Energy Laboratory. https://www.nrel.gov/docs/fy24osti/89843.pdf.

Wolters, Christopher, Xiaoxuan Yang, Ulf Schlichtmann, and Toyotaro Suzumura. 2024. ‘Memory Is All You Need: An Overview of Compute-in-Memory Architectures for Accelerating Large Language Model Inference’. arXiv. https://arxiv.org/html/2406.08413v1.

Zimmermann, Severin, Ingmar Meijer, Manish K. Tiwari, Stephan Paredes, Bruno Michel, and Dimos Poulikakos. 2012. ‘Aquasar: A Hot Water Cooled Data Center with Direct Energy Reuse’. *Energy*, 2nd International Meeting on Cleaner Combustion (CM0901-Detailed Chemical Models for Cleaner Combustion), 43 (1): 237–45. https://doi.org/10.1016/j.energy.2012.04.037.