STUDY

From Chips to Grids

Why Energy Constraints Influence AI Compute Expansion in Europe

Maria Nowicka

May 05, 2026

Tech analysis and policy ideas for Europe

interface

Table of Contents

Key takeaways

Over the next decade, Europe's ambition to become an 'AI continent' will hinge not only on access to compute capacity but also on its ability to host very large AI compute clusters. As AI uptake accelerates, the latter is set to become a central constraint within an already strained power system transitioning away from fossil fuels, shaping what is realistically achievable. Large AI compute clusters, including prospective AI gigafactories (AIGFs), generate concentrated energy demands of hundreds of megawatts at a single site, effectively making them power system assets in their own right. The recent global energy crisis triggered by the war in Iran and the prolonged closure of the Strait of Hormuz underscores the importance of factoring energy constraints into infrastructure planning, including in the context of the European Commission's upcoming call for proposals for AIGFs. Large AI labs have also begun to reconsider planned infrastructure in Europe – with OpenAI pausing its UK data centre project in April 2026 citing high electricity costs and regulatory uncertainty and pulling back from its Norway investment shortly after.

Whether these clusters strengthen or strain Europe's energy system will depend on how and where they are built and under which conditions they operate. If executed well, they can anchor high-utilisation, low-carbon infrastructure that supports both AI development and grid stability. If executed poorly, they risk becoming power-intensive prestige projects that crowd out electrification incentives in other sectors and undermine the security of supply goals. The key question, therefore, is how to integrate large AI clusters into the existing partly fossil-based, constrained power systems. For this integration to succeed, it is essential to consider the key characteristics and particularities of large AI clusters.

Large AI clusters behave like industrial loads, not like traditional data centres

Clusters built for AI workloads differ from traditional data centres in three ways that matter for energy usage– they are built around specialised AI accelerators rather than general-purpose central processing units (CPUs); they push rack power densities to levels several times higher than legacy designs; and they are optimised for long, high-intensity workloads rather than variable, transactional traffic. In operational terms, a 100–300 MW AI cluster behaves less like the traditional IT infrastructure and more like an electro-intensive industrial plant connected to constrained grids.

Utilisation is a missing piece in current AI energy debates

Current monitoring frameworks in Europe focus on power usage effectiveness (PUE) and total electricity use but do not address how intensively AI hardware is used. In large AI

clusters, a substantial portion of the total power consumed goes to compute servers and accelerators. Regardless of whether they sit idle or underused, cooling, networking and power conversion systems still run in the background and consume energy. Evidence from both training and inference deployments suggests that typical utilisation in multiclient clusters is often in the 30–40% range, with contracted grid capacity often underused. Without utilisation and compute-per-energy metrics, a large AI cluster can seem efficient on paper, demonstrating good PUE and a high renewable share, while still being a poor steward of scarce grid capacity.

Operating models determine whether AI clusters become energy assets or liabilities

How and to whom AI compute capacity is sold and how workloads are scheduled largely determine whether large AI clusters become energy assets or liabilities. When one or a few large users commit to using most of the capacity of one cluster over many years, keeping utilisation high and producing relatively predictable load profiles are easier to align with low-carbon generation. In a multiclient scenario, where many small tenants share the cluster, their diverse and often latency-sensitive workloads are much harder to aggregate and schedule efficiently. Given Europe's current AI ecosystem, the most likely future is predominantly multiclient clusters, which are structurally more exposed to low utilisation, spiky demands and complex scheduling constraints.

Location will drive system impacts

Where and how large AI clusters are built will shape their interactions with the power system for decades. Clusters of 100–300 MW typically require new or reinforced high-voltage grid connections, triggering the need for upstream transmission upgrades. Training-heavy clusters can, in principle, be located where renewable energy is available and affordable, and where demand can be easily shifted. Inference-heavy clusters, on the other hand, must be closer to users and networks, often in already congested regions with high prices and limited grid capacity. Placing very large, inflexible loads into stressed grid nodes without clear demand guarantees and integration plans risks hindering other electrification projects and raising system costs.

Energy is only one part of the picture, but it is a binding constraint

Large AI clusters raise environmental concerns that go beyond electricity use, including water consumption, local air pollution and persistent noise. Even against this broader backdrop, energy and grid compatibility remain a hard boundary condition in Europe. Without credible plans to secure low-carbon supply, reinforce grids and engage in demand response, very large AI clusters cannot be built or operated at scale in a socially and politically sustainable manner.

Policy recommendations for large AI compute clusters in Europe

Building on these findings, we put forward the following recommendations for the design and governance of large AI clusters, including the EU's AIGFs:

Differentiate AI clusters in policy frameworks

Treat large AI clusters as a distinct category from traditional centres in the EU and member-state instruments. The European data centre rating scheme stemming from the Energy Efficiency Directive (EED) and other related policies should reflect the structural differences of AI clusters – higher power densities, accelerator-centric architectures, liquid cooling and denser interconnects – through separate reporting tracks, thresholds and design obligations. Tighter utilisation, flexibility and siting requirements should apply specifically to facilities whose primary purpose is AI training and high-volume inference, while mixed workload centres continue to follow the baseline EED framework.

Use minimum energy performance standards as a baseline and add flexibility requirements

Consider the EED's minimum energy performance standards – including targets such as PUE <1.3 from 2027 and a fully renewable electricity supply by 2030 – as the baseline for all large AI clusters, rather than creating parallel metrics. Integrate AIGF and similar initiatives into the EED-based rating framework and explore flexibility obligations, defined as a minimum level of demand response or interruptible capacity made available each year, to reflect the grid value of shifting some AI workloads over time.

Mandate utilisation disclosure and minimum utilisation targets

Require operators of large AI clusters to disclose AI compute utilisation and compute-per-energy metrics from the start of commercial operations. Regular reporting of quarterly metrics (such as TFLOPS per MWh) and average AI accelerator utilisation would link compute performance to energy intensity and reveal how episodic AI workloads actually are. Over time, regulators could use these disclosures to converge on minimum utilisation bands that are above the current typical range of 30–40% but remain achievable, with third-party verification to ensure integrity and comparability across clusters.

Integrate grid-compatibility assessments into decisions on AI infrastructure expansion

Make grid compatibility assessments a standard part of designating, permitting and supporting large AI clusters. These assessments should examine peak and baseload

profiles, connection points, proximity to existing infrastructure and the potential for demand response participation. National examples – such as emerging preferred zone approaches in France and the UK's linkage of data centre siting to AI Growth Zones – illustrate how spatial planning and grid planning can be aligned; similar principles should guide decisions on large AI clusters across the EU.

Enhance operational transparency through disclosure of scheduling methodologies

Require high-level disclosure of how capacity in large AI clusters is allocated between workload categories (for example, real-time inference versus batch training, anchor customers versus external tenants versus public-sector users). The aim is to ensure that utilisation metrics reflect genuine accelerator activity rather than accounting artefacts, without exposing proprietary scheduling algorithms. In parallel, the European Commission should encourage or contract work on interoperability standards for AI-cluster scheduling to reduce the risk of European infrastructure becoming locked into opaque, non-European systems.

Create strong institutional coordination on AI infrastructure

Treat large AI clusters as a shared responsibility across energy, digital and grid-operator communities. A standing joint working group on AI infrastructure, bringing together energy and digital policymakers from the European Commission (e.g. from DG ENER and DG CNECT) as well as European transmission and distribution system operators, could help align siting, grid connection and decarbonisation objectives. At a minimum, designation of an AIGF-type project should automatically trigger enhanced EED reporting and grid integration obligations, so that special AI-infrastructure initiatives and the general data centre framework at the EU level function as complementary layers of a single system rather than separate regimes.

Together, these recommendations aim to ensure that Europe's AI compute clusters – including the AIGFs – are made for purpose and supports strategic AI capabilities while remaining compatible with the physical constraints and overarching decarbonisation goals of Europe's power system.

Introduction

Over the past five years, AI has triggered a global race to build ever larger compute infrastructure, bringing attention to the physical footprint of this expansion. A central constraint in this race is energy. Improvements in model performance are now tightly coupled to the ability to deliver very large, continuous power flows to a single large AI

compute cluster. This shift is reflected in the rapid growth of leasing facilities, with the power capacity of large AI clusters increasing from approximately 13 MW in 2019 to an estimated 280–300 MW for xAI's Colossus in 2025, comparable with the demand of roughly 250,000 European households.[1] In this global contest, Europe emerges both as a competitor and a bottleneck, as it seeks to host world-class AI infrastructure but does so within some of the most grid-constrained electricity systems among advanced economies, marked multi-year grid connection queues and limited capacity for new large loads.

Consequently, Europe is now entering a new phase of AI industrial policy in which questions of compute capacity, energy security and climate policy are becoming tightly intertwined. Recent initiatives by the European Commission and EuroHPC Joint Undertaking point to an ambition to scale up AI infrastructure. The AI Continent Action Plan envisions up to five publicly co-funded AI Gigafactory (AIGF) clusters, each built around roughly 100,000 advanced AI processors[2] and designed to cover the full life cycle of very large models.[3] These plans build on an already evolving landscape of AI infrastructure, with hyperscalers, neoclouds and AI labs announcing multi-hundred–megawatt clusters across Europe and beyond. Earlier work by *interface* that analysed the AIGF initiative has already warned that Europe's debate risks focusing too much on overcoming past shortages of compute supply and too little on whether there will be sufficient, well-defined demand to justify gigafactory-scale projects.[4] At the same time, complementary *interface* analysis on AI Factories has already highlighted uncertainties about demand and gaps in governance in Europe's first wave of projects.[5]

On top of these demand- and governance-side challenges, Europe's room to manoeuvre is also highly constrained by grids. Across EU member states, grid connection capacity, connection lead times, local congestion and, most recently, energy prices have already become binding constraints, delaying or redirecting large deployments despite initial investment interest.[6] Decisions about where to locate large AI compute clusters and how

1 Konstantin F. Pilz et al. (2025). Power requirements of leading AI supercomputers have doubled every 13 months. https://epoch.ai/data-insights/ai-supercomputers-power-trend.

2 In the AI Continent Action Plan and other policy documents and throughout this study, advanced AI processors are expressed in H100 equivalents, a normalised unit that converts different accelerator types into the approximate compute capacity of a single NVIDIA H100-class GPU. This provides a common reference for comparing AIGFs and for relating their compute capacity to power demand and energy metrics.

3 The scope attributed to AIGFs has broadened progressively across policy documents. The AI Continent Action Plan published in April 2025 described AIGFs as facilities designed to 'develop and train complex AI models at an unprecedented scale', making no explicit reference to inference or deployment. The Call for Expression of Interest in AIGFs, published alongside the action plan, extended this to clusters 'designed to develop, train, and deploy very large AI models and applications at an unprecedented scale', therefore adding deployment and inference to the mandate for the first time. The most expansive definition appeared in the updated EuroHPC Regulation in January 2026, which defines an AIGF as a 'state-of-the-art large-scale facility with sufficient capacity to handle the complete lifecycle, from development to large-scale inference, of very large AI models and applications'. This study uses the definition established in the EuroHPC Regulation as the operative definition while noting that the shift from training only to full–life cycle framing has significant implications for the scale, purpose and energy profile of the envisioned AIGFs.

4 Hess and Sieker conclude that AIGFs are only feasible if sufficient demand is factored in from the beginning and recommend that consortia selection explicitly considers projected workloads and concrete user commitments, not just the supply side.

5 AI Factories are national or regional flagship facilities that bundle access to high-performance compute, relevant datasets and technical support to develop and deploy advanced AI models, building on EuroHPC-class supercomputing capacity and related initiatives. They are meant to act as user-facing ecosystems that connect researchers, startups and industry with shared infrastructure. By contrast, AIGFs are conceived as very large AI clusters focused on supplying raw compute at scale, with dedicated industrial-grade power and cooling, while assuming that data, talent and support services are organised around them by separate initiatives.

large they can realistically be are increasingly determined by grid availability and electricity prices.[7] This siting logic is already visible in some of the largest announced AI compute investments in Europe. For example, Mistral committed to investing 1.2 billion EUR in an AI cluster in Sweden, attracted by the country's relatively low-cost, largely low-carbon electricity mix.[8] More broadly, AI infrastructure bids and investments are now distributed across Europe—from France and Germany to Spain and the Nordics—reflecting a combination of electricity costs and the availability of renewable energy, as well as political and industrial considerations.[9]

These developments unfold against a demanding macropolitical energy backdrop. The International Energy Agency (IEA) expects global data centre electricity use to more than double by 2030,[10] largely owing to AI workloads, while NVIDIA alone projects around 1 trillion USD in orders for its AI systems through 2027, signalling an unprecedented demand for AI hardware.[11] In Europe, this surge is now visible even in system-adequacy studies, where ENTSO-E (European Network of Transmission System Operators for Electricity), the association of Europe's high-voltage grid operators, models large digital loads, including data centres and AI clusters, when they exceed roughly 50–100 MW at a single connection point, treating them as system-relevant assets in their own right.[12] At the same time, Europe confronts this spike in demand while its energy-security position remains fragile: Fossil fuels still account for nearly one-third of its electricity generation, and recent conflicts in the Middle East and Russia's war on Ukraine have underlined the risks of tight gas and power markets.[13]

Against this backdrop, this study argues that not all digital infrastructures look the same from the perspective of the power system. It investigated how AI clusters differ from traditional data centres in not only how they consume electricity but also how tightly they are coupled to the grid and how easily (or not) their demand can be adjusted.

To capture the broader class of AI-focused infrastructure, this study uses the umbrella term 'large AI compute cluster', referring to facilities hosting 100,000 or more advanced

6 Mark Bergen (2026). OpenAI Pauses Stargate UK Data Center Citing Energy Costs. https://www.bloomberg.com/news/articles/2026-04-09/openai-pauses-stargate-uk-data-center-effort-citing-energy-costs?embedded-checkout=true.

7 In the FLAP-D markets—Frankfurt, London, Amsterdam, Paris and Dublin—new facilities wait on average 7–10 years for a grid connection, rising to 13 years in the most congested primary markets. Ireland has imposed a de facto moratorium on new data centres in Dublin until 2028, while the Netherlands and Frankfurt have effectively banned new connections until at least 2030.

8 Data Center Dynamics (2026). French AI firm Mistral Signs $1.2bn Deal to Build EcoDataCenter Facility in Sweden. https://www.datacenterdynamics.com/en/news/french-ai-firm-mistral-signs-12bn-deal-to-lease-ecodatacenter-facility-in-sweden/.

9 EuroHPC JU (2024). Selection of the First Seven AI Factories to Drive Europe's Leadership in AI. https://www.eurohpc-ju.europa.eu/selection-first-seven-ai-factories-drive-europes-leadership-ai-2024-12-10_en.

10 IEA (2026). Energy Demand from AI. https://www.iea.org/reports/energy-and-ai/energy-demand-from-ai.

11 Ian King (2026). Nvidia Makes Trillion-Dollar Forecast at Annual Product Expo. https://www.bloomberg.com/news/articles/2026-03-16/nvidia-expects-to-make-1-trillion-from-ai-chips-through-2027.

12 ENTSO-E (2025). European Resource Adequacy Assessment. 2025 Edition. https://eepublicdownloads.blob.core.windows.net/public-cdn-container/clean-documents/sdc-documents/ERAA/ERAA_2025_ExecutiveReport_ENTSOEProposal_Dec2025.pdf.

13 Eurostat (2026). Production of Electricity and Derived Heat by Type of Fuel. https://ec.europa.eu/eurostat/databrowser/view/nrg_bal_peh__custom_20322651/bookmark/table?lang=en&bookmarkId=7fd37d6c-0606-4c34-8b31-63fa5283e674&c=1772444282549.

AI accelerators dedicated primarily to AI workloads. This included EU-supported AIGFs and private projects developed or operated by hyperscalers or neocloud providers or owned by AI labs. Therefore, the AIGF concept is one prominent expression of a wider trend rather than the sole focus of the analysis.

Therefore, the AIGF concept is one prominent expression of a wider trend rather than the sole focus of the analysis.

Building on this distinction, the study examined how large AI compute clusters interact with the energy system, why they pose different challenges compared with traditional data centres, and what this implies for Europe's plans to expand AI compute capacity and for EU-level policy frameworks in a constrained decarbonising power system.

Chapter 1 maps the energy architecture of large AI clusters from facility to rack to chip. It distinguishes five tightly coupled subsystems: compute, memory and storage, cooling and thermal management, networking and power delivery. It also shows how each shapes total power draw and load patterns. This provides a technical baseline for the rest of the report.

Chapter 2 builds on this technical foundation to argue that large AI clusters constitute a distinct class of energy-intensive infrastructures. It analyses how different workload regimes, particularly training and inference, affect load profiles, siting choices and flexibility options, and develops the notion of 'hybrid workloads as an energy variable' to link utilisation and customer base to system impacts. In doing so, it complements earlier *interface* work on frontier compute demand by linking questions around frontier compute demand directly to Europe's grid and decarbonisation constraints.[14]

Chapter 3 turns to policy design, focusing on the AIGF model. It examines what would be required for AIGFs to function as 'responsible energy citizens' in a European power system facing increasing demand, rapid data centre expansion and renewed concerns around security of supply. The chapter identifies structural bottlenecks around utilisation, flexibility and scheduling that shape how AI clusters interact with grids and markets.

In combination, this study advances the claim that investments in large AI compute clusters, such as AIGFs, can underpin Europe's AI ambitions only under specific energy-system conditions. The analysis examines ways to secure the long-term value of AIGF projects while keeping in mind the regional energy landscape. It aims to demonstrate that planning the EU's AI compute expansion within the physical limits of a still fossil-dependent power system aligned with climate and security objectives and subject to utilisation and grid-based accountability will therefore be key to truly deliver

14 Julia Christina Hess & Felix Sieker (2025). Id.

on the promise of becoming an AI continent.

Chapter 1 - How AI hardware uses energy: Inside large AI compute clusters

This chapter unpacks what exactly drives the energy characteristics of a large AI compute cluster by breaking down the cluster into the subsystems that shape its power draw and the power demand profile over time.

The first section singles out the structural drivers of the energy characteristics of the AI cluster, pointing towards three main elements that make it distinct from traditional data centre infrastructure. We then map the main subsystems at the AI cluster level to show how they shape the power draw and where energy bottlenecks arise the most. This provides the technical basis that the following chapters use to analyse AI clusters as a distinct kind of energy-intensive infrastructure.

Why AI clusters prompt higher energy demands than traditional data centres

Large AI compute clusters differ from conventional data centre infrastructure, and these differences explain why energy use has emerged as a central constraint in scaling modern AI.[15] While traditional data centres are designed for stable, mixed non-AI workloads, such as web hosting, databases and enterprise software, AI infrastructure is built to sustain much higher computational throughput at scale. Three structural differences drive this pattern and underpin the technical analysis in this chapter.

From CPUs to specialised accelerators

The first difference is architectural. Conventional data centres rely primarily on general-purpose central processing units (CPUs) for their computation power, whereas AI clusters are centred on specialised AI accelerators. AI accelerators are optimised for the matrix mathematics that underpins modern AI models. Unlike general-purpose CPUs, they execute thousands of operations simultaneously, delivering up to 150 times faster end-to-end training and inference than CPU-only systems.[16]

15 IEA (2025). Energy and AI. https://www.iea.org/reports/energy-and-ai.

16 Piotr Bigaj, Dawid Majchrowski, Kyle Kranen & Pawel Morkisz (2021). Accelerating the Wide and Deep Model Workflow from 25 Hours to 10 Minutes Using NVIDIA GPUs. https://developer.nvidia.com/blog/accelerating-the-wide-deep-model-workflow-from-25-hours-to-10-minutes-using-nvidia-gpus/.

The main accelerator families in current large AI clusters include the following:

- GPUs (graphic processing units, e.g. NVIDIA's H100 or GB300): general-purpose accelerators that support a broad range of workloads, including AI training and inference. Current high-end GPUs provide 3–8 TB/s high-bandwidth memory (HBM) per chip, which is crucial for large-scale training and inference.[17]
- ASICs (application-specific integrated circuits): customised chips hard-wired for specific AI workloads, trading flexibility for higher performance and efficiency. This includes the following:
  - Tensor processing units (TPUs): accelerators optimised for dense tensor operations used for large-scale training and high-volume inference.
  - Inference-optimised ASICs (e.g. AWS Inferentia, Meta MTIA and Groq language processing units [LPUs]): tailored to specific parts of the AI inference pipeline or particular AI model families.[18]

These AI accelerators allow the execution of large numbers of operations simultaneously and are optimised specifically for AI workloads, such as training and inference. On the other hand, CPU-based architectures use a small number of powerful general-purpose cores that execute instructions sequentially and scale less well across thousands of servers.[19] [20] In practice, a finance or e-commerce company running web front ends, databases and small recommendation models will typically consume CPU-centred capacity, while AI labs training frontier large language models or serving generative AI queries cluster around GPU-rich AI compute clusters.

The second, closely related feature is that GPU-based systems rely on much deeper parallelism—a process where multiple AI accelerators are used to execute many tasks at once, increasing throughput and efficiency compared with sequential processing, which is used in traditional CPU-reliant data centres.[21] CPU-centred architectures typically run independent workloads where the need for parallel computing is limited. On the other hand, AI systems rely on tightly coordinated execution across large accelerator pools, which requires the rapid exchange of intermediate results between the processing units and their memory.[22] To minimise this communication bottleneck, in a modern AI accelerator, HBM and GPU are on the same interposer, shortening data paths and allowing terabytes per second of memory bandwidth to feed parallel computations efficiently.[23]

17 Clarifai (2025). NVIDIA H100: Price, Specs, Benchmarks and Decision Guide. https://www.clarifai.com/blog/nvidia-h100.

18 How AI Works (2025). TPUs vs GPUs vs ASICs: AI Hardware Guide 2025. https://howaiworks.ai/blog/tpu-gpu-asic-ai-hardware-market-2025.

19 Tim Lu (2024). CPU vs GPU: How They Work and When to Use Them. https://www.datacamp.com/blog/cpu-vs-gpu.

20 With the rise of agentic AI systems, the CPU re-emerges as a bottleneck and control plane. Studies of agentic workloads have shown that tool processing and orchestration on CPUs can account for 50–90% of the total latency, with CPU core count being one of the determinants of how well GPUs are utilised. Thus, recent analyses argue that next-generation AI clusters will need to scale CPU capacity alongside accelerators to keep agentic AI clusters efficient.

21 Xinyao Yi (2024). A Study of Performance Programming of CPU, GPU Accelerated Computers and SIMD Architecture. https://arxiv.org/abs/2409.10661.

22 This phenomenon is also referred to as the 'Von Neumann bottleneck'—the communication delay that occurs when data moves slower than computation because of separation between the compute and memory.

Finally, these design choices concentrate compute into far fewer, denser racks. Traditional data centres were built around roughly 5–15 kW per rack, whereas AI-oriented clusters now routinely deploy racks in the 30–120 kW range, with some reference designs going higher, requiring fundamentally different thermal and power delivery designs.[24] [25] From an energy perspective, this means that the compute architecture in an AI cluster is dominated by high-power racks rather than a fleet of moderate power racks. This shifts both the magnitude and spatial distribution of power demand inside the facility, and it sets the stage for the rack-level bottlenecks described later in this chapter.

Taken together, the three features (GPU-centered architecture, deep parallel computing and rack density) mean that AI clusters are best understood as tightly coupled, high-density compute systems where energy consumption becomes a structural constraint. The following sections unpack how these drivers manifest in concrete subsystems, where energy demand concentrates inside the cluster and why the rack becomes the central chokepoint for power and cooling.

Subsystems of a large AI compute cluster

From an energy perspective, five subsystems are particularly important for the total power draw and the unique patterns of energy use in a large AI compute cluster: compute, memory and storage, cooling and thermal management, networking and power delivery. Figure 1 below illustrates how these subsystems sit within the physical structure of a cluster. It distinguishes two levels of organisation.

- At the cluster level, the subsystems are distributed across the facility: dedicated areas for memory and storage, cooling, core networking infrastructure and power delivery all feed rows of AI compute racks. This level makes it clear how electricity, cooling and data have to be routed across the building to reach the compute floor.
- At the rack level, the illustration zooms into a single rack and then into one server. In the rack view, it shows servers containing AI accelerators mounted in vertical slots, linked with interconnects, and connected to power distribution units. Inside the server, it highlights how AI accelerators with high-bandwidth memory (HBM) are linked via a networking fabric, and how direct-to-chip cooling is arranged.

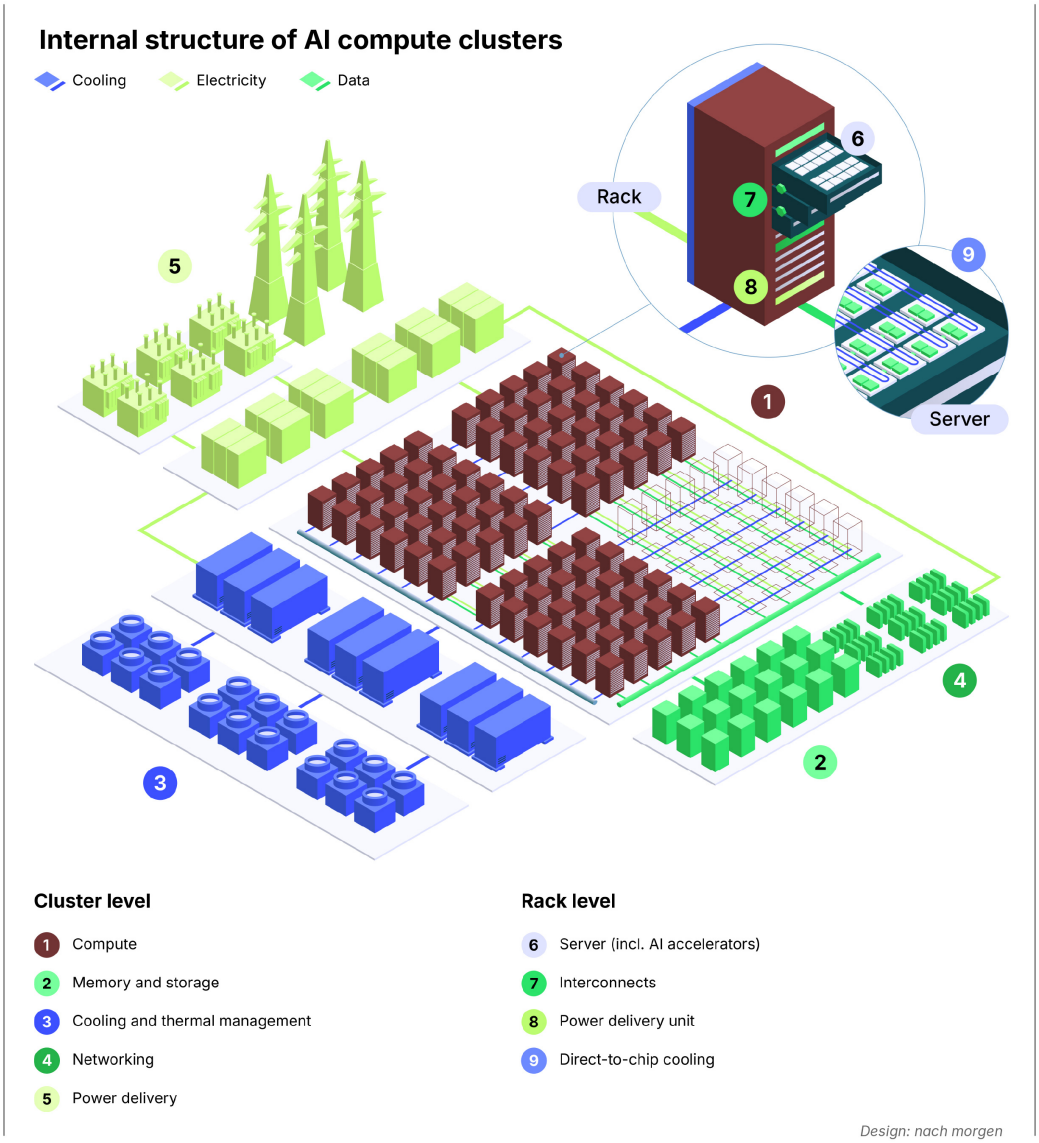
The colour coding traces how all three resource streams run through both levels simultaneously: blue for cooling flows, neon green for electricity and dark green for data. A transparent rendering of a part of the compute subsystem makes this integration visible at the rack level, where the same electricity, cooling and data connections that span the

23 For the H100, the HBM3 memory subsystem delivers 3–3.35 terabytes per second of bandwidth per GPU, depending on the model.
24 Michael Wilson (2025). Data Center Rack Power Costs: A Condensed Analysis. https://www.nlyte.com/blog/data-center-rack-power-costs-a-condensed-analysis/.
25 Marcus Chen (2026). H100 GPU Server Speed vs CPU Comparison Guide. https://ventusserver.com/speed-vs-cpu-comparison/.

whole cluster converge into individual racks.

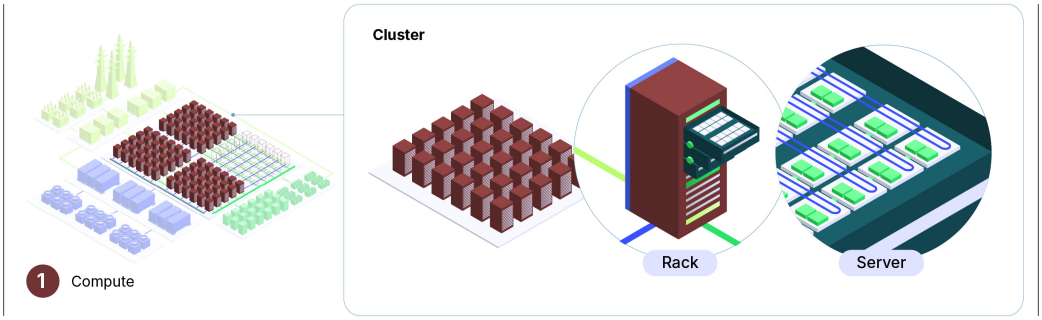
This section introduces subsystems across the mapped levels and explains their most important technical components. The following section returns to these subsystems at the rack level, where their interactions become constrained by local power density and cooling limits and where the practical bottlenecks for AI workloads emerge.

Figure 1: Internal structure of AI compute clusters



Compute

Figure 2: The compute subsystem



At the heart of every large AI cluster, the compute subsystem is located within one or more dedicated compute rooms, or ‘AI halls’, via so-called racks. A compute rack in this context is a tall cabinet that holds multiple servers, which are self-contained computers that each include AI accelerators (GPU or other ASICs customised for AI workloads) and host central processing units (CPUs), memory, storage and networking components mounted on circuit boards. In traditional data centres, most computation is handled by CPUs, which are designed to run many small tasks at moderate power per server. By contrast, in AI clusters, GPUs take over the bulk of computation, as they are optimised for the large matrix operations used in model training and inference.[26] Each server concentrates far more compute and thus more electrical power into the same physical space.

Unlike general-purpose enterprise racks, AI compute racks are typically deployed in large homogenous blocks, as they execute and coordinate thousands of matrix operations simultaneously to handle the development and deployment of AI models. As a result, the compute layer dominates electricity at the cluster level, with servers accounting for roughly 55–65% of total facility power in a frontier AI cluster at peak operation.[27]

In this sense, compute equipment not only is the largest energy consumer inside the cluster but also sets the spatial and electrical design planning that every other subsystem in the facility must be built around. This power concentration in compute equipment is also what makes large AI clusters behave like continuous industrial processes rather than traditional IT facilities.[28] In Section 3.3, we dive deeper into the internal arrangement of the compute subsystem.

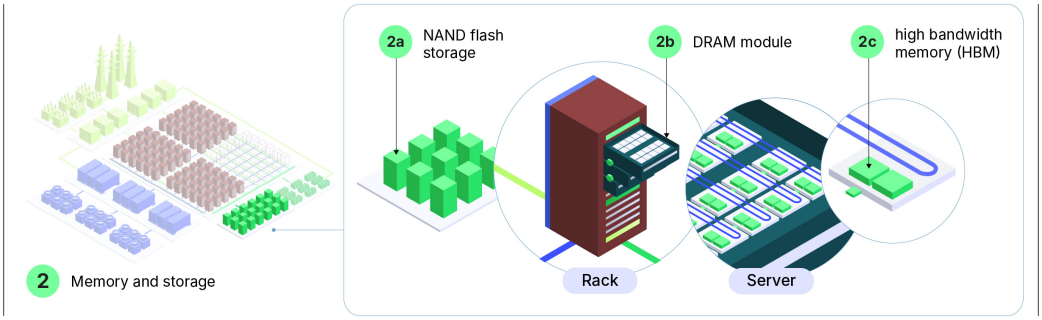
26 This and the following descriptions of GPU-centric architectures equally apply to other application-specific accelerators for AI workloads (ASICs), which are likewise designed to execute large matrix-style operations efficiently.

27 Luke Emberson and Ben Cottier (2025). GPUs account for approximately 40% of power usage in AI data centers. https://epoch.ai/data-insights/gpus-power-usage-in-ai-data-centers.

28 Throughout this report, ‘industrial loads’ refers to large, typically high-voltage electricity consumers engaged in production or processing activities, with high and relatively continuous demand profiles. This usage follows how EU law and system operators treat electro-intensive users and large industrial consumers in state-aid rules, energy-taxation schemes and resource-adequacy assessments, where sectoral criteria and energy intensity determine eligibility rather than a single capacity threshold.

Memory and storage

Figure 3: The memory and storage subsystem



In an AI cluster, memory spans three distinct tiers, each serving a different function and located at a different level of the hardware stack.

The first and fastest tier sits inside the AI accelerators within the server (no. 2c on the chart). Each accelerator integrates HBM directly onto the chip package, providing the very high memory bandwidth that GPU workloads require.[29] HBM is short-term volatile memory. It is set apart from other volatile memory by its speed. Its vertically stacked 3D architecture allows it to deliver 3–8 TB/s of memory bandwidth to each GPU compared with less than 0.5 TB/s for the DDR5 memory attached to a typical server-class CPU.[30] [31]

The second tier is system dynamic random access memory (DRAM - no. 2b on the chart), which serves the same short-term volatile function as HBM but at the server level rather than the accelerator level. DRAM modules are mounted on the server motherboard and serve as the general (short-term) system memory for CPU and GPUs.[32] It stores the data and intermediate results that the systems are actively using but is wiped when the power is turned off.[33] Therefore, HBM and DRAM are volatile working memory systems that operate at different points in the memory hierarchy, with HBM as the dedicated memory of the accelerator and DRAM as the server's general-purpose working memory, serving both the CPU and GPUs.

29 For example, in NVIDIA's H100 and B200 architectures, HBM stacks are mounted directly on the GPU interposer alongside the compute dies, forming a single integrated package. This design maximises memory bandwidth (up to 3.35 TB/s on the H100 SXM) while keeping the memory physically inseparable from the accelerator itself.

30 Clarifai (2025). NVIDIA H100: Price, Specs, Benchmarks and Decision Guide. https://www.clarifai.com/blog/nvidia-h100.

31 Micron (2024). Boost HPC Workloads with Micron DDR5 and AMD EPYC Processors. https://www.micron.com/about/blog/applications/data-center/boost-hpc-workloads-with-micron-ddr5-and-4th-gen-amd-epyc-processors.

32 Simmtester (2026). HBM, DRAM, and NAND: How AI is Reshaping the Memory Market. https://www.simmtester.com/News/IndustryArticle/27782.

33 GeeksForGeeks (2025). What Is Volatile Memory? https://www.geeksforgeeks.org/computer-organization-architecture/what-is-volatile-memory/.

The third tier is non-volatile storage (no. 2a on the chart). NAND flash acts as a long-term storage, keeping data even when powered down, and is used in hard drives and larger storage systems.[34] In an AI cluster, NAND-based storage holds the largest datasets, model checkpoints and outputs, and streams data towards system DRAM and HBM as the data moves through the system.[35] In AI clusters, storage is therefore organised around keeping the GPUs busy. The most frequently used data are kept as close as possible to the accelerators, while older data sit on slower object storage further away.

As GPUs themselves already account for approximately 40% of the total power and total server power increases to 55–65%, storage adds a further layer of power demand at the rack level, which makes the combined compute-and-storage footprint a crucial energy bottleneck for large clusters.[36] The precise share of storage systems in cluster power demand is difficult to isolate.[37] Most available breakdowns cover traditional data centres rather than dedicated AI clusters, and operators rarely publish figures for storage. An additional complication is that HBM draws power already counted within the GPU's energy budget, which means that any estimate risks double-counting.[38]

In traditional data centres, most data sit on large shared hard-disk (HDD) systems, with a thinner layer of faster solid-state drives (SSDs) on top.[39] Because traditional data centres mainly run business applications, websites and databases that send many small requests, HDD-plus-SSD layering is designed to keep most data on large, inexpensive disks and to put only a small fraction of very frequently used data on fast SSDs.[40] That way, operators optimise storage costs while still serving the most frequent queries at low latency.

Cooling and thermal management

Figure 4: The cooling subsystem

34 Vikram Sekar (2025). Role of Storage in AI, Primer on NAND Flash, and Deep-Dive into QLC SSDs. https://www.viksnewsletter.com/p/role-of-storage-in-ai-primer-on-nand.

35 Vikram Sekar (2025). Role of Storage in AI, Primer on NAND Flash, and Deep-Dive into QLC SSDs. https://www.viksnewsletter.com/p/role-of-storage-in-ai-primer-on-nand.

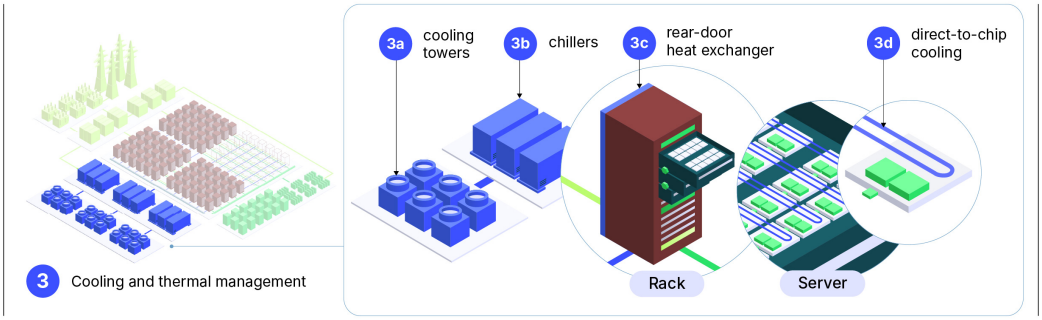
36 IEA (2025). Data Centre Energy Use: Critical Review of Models and Results. https://www.iea-4e.org/edna/publications/data-centre-energy-use-critical-review-of-models-and-results/.

37 George Kamiya & Vlad C. Coroamă (2025). Data Centre Energy Use: Critical Review of Models and Results. https://www.iea-4e.org/wp-content/uploads/2025/05/Data-Centre-Energy-Use-Critical-Review-of-Models-and-Results.pdf.

38 Luke Emberson & Ben Cottier (2025). GPUs Account for About 40% of Power Usage in AI Data Centers. https://epoch.ai/data-insights/gpus-power-usage-in-ai-data-centers/.

39 Hanwha Data Centers (2025). AI Data Centers vs. Traditional Data Centers. https://www.hanwhadatacenters.com/blog/ai-data-centers-vs-traditional-data-centers/.

40 Scality (2025). Why Tiered Storage Is the Secret to Scalable AI. https://www.youtube.com/watch?v=VQf6VRCxZf4&t=2s.



Cooling equipment regulates the thermal environment for the compute hardware to ensure it operates efficiently, preventing overheating and maintaining the lifespan of components.

Traditional data centres built around 5–15 kW CPU racks rely mainly on refined room-level air cooling, in which chilled air is blown through racks, across heat sinks and then removed at room level by computer room air conditioning units. On the other hand, GPU-heavy AI clusters increasingly require liquid-based systems, as rack densities higher than 30 kW result in highly concentrated heat that air can no longer remove fast enough.[41]

This shift has driven two main approaches. The first and currently most widespread in AI clusters is direct-to-chip cooling, in which cold plates are mounted directly onto GPU and CPU packages inside each server (no. 3d on the chart), and coolant (typically water) is circulated through them via rack-level distribution loops that run along the back of each rack (no. 3c on the chart). These loops connect to a facility-level primary cooling loop, which carries the warmed water to mechanical chillers (no. 3b on the chart) and to cooling towers (no. 3a on the chart), where heat is rejected to the outside air or atmosphere. The second approach, which has been increasing in popularity for highest-density AI clusters, is full immersion cooling (not pictured on the chart). Here, entire servers or rack assemblies are submerged in an electrically non-conductive liquid that absorbs heat directly from all components. The heated fluid is then circulated to heat exchangers that transfer the thermal load to a cluster-level water loop, again ultimately rejected via chillers or cooling towers.

In AI clusters, cooling typically accounts for a significant share of total facility power, varying substantially by cooling technology, and facility design.[42] The rapid shift

41 Global Data Center Hub (2025). Why AI Is Forcing a Complete Rethink of Data Center Design. https://www.globaldatacenterhub.com/p/why-ai-is-forcing-a-complete-rethink.

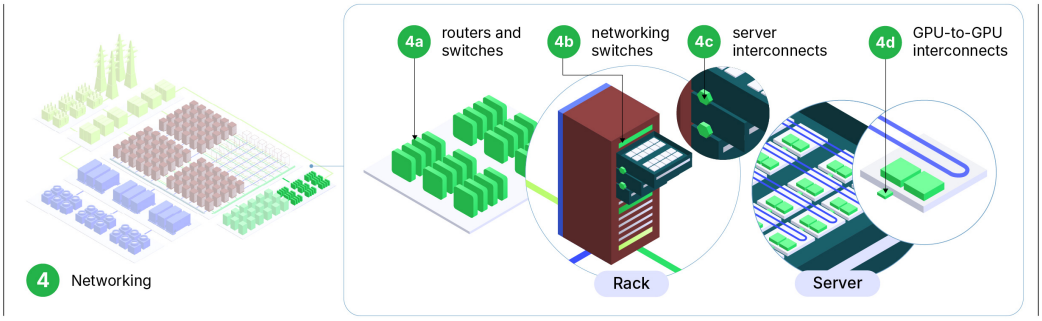
42 Estimates of cooling's share of total data centre facility power vary considerably depending on cooling technology, rack density, climate, and facility age, and no single authoritative figure exists for AI clusters specifically. The European Parliament Research Service reports that cooling and environmental control accounts for '7% to over 30% of electricity use [in data centres], depending on energy efficiency'; the floor of 7% is associated with small or edge facilities benefiting from passive or free-air cooling strategies and is not representative of hyperscale AI clusters operating at high power densities. Industry benchmarks for traditional air-cooled data centres converge around 30–40% of total facility power.

towards liquid-based solutions is pushing the best-performing facilities towards the lower end. The most efficient hyperscale operators, whose newer facilities increasingly host AI workloads, report cooling shares as low as 7% of the total facility power.[43] Older facilities relying partly on air cooling or those in warmer climates sit closer to the higher end.[44] Most of these figures are reported for data centres that mix CPU- and GPU-heavy workloads rather than isolating large AI compute clusters, which reflects the broader measurement gaps in AI energy reporting discussed further in Section 3.4.

Because cooling demand scales with both compute workloads and ambient conditions, it is also one of the more variable components of a cluster's total power draw—a point that matters for how AI clusters interact with the grid (see discussion in Chapter 2).

Networking

Figure 5: The networking subsystem



In a traditional data centre, the network is built mainly to serve user-facing applications, such as websites, databases and file services. Most traffic flows from users or other systems, hits a front-end server and then goes back out again.[45] The core components here are Ethernet switches and routers arranged in a three-tier or leaf-and-spine topology, using standards such as 10, 25 and 100 gigabit-per-second Ethernet.[46]

However, in an AI cluster, the networking system consists of two intertwined layers: a conventional vertical network for user and control traffic (sometimes also referred to as 'north–south' traffic) and a dense internal fabric that effectively treats thousands of GPUs as one large parallel computer (also called the 'east-west' traffic).

43 IEA (2025). Energy and AI. https://www.iea.org/reports/energy-and-ai.

44 Xin Chen et al. (2025). Electricity Demand and Grid Impacts of AI Data Centers: Challenges and Prospects. https://arxiv.org/html/2509.07218v3.

45 Terakraft (2024). Data Center Design Requirements for AI Workloads. A Comprehensive Guide. https://www.terakraft.no/post/datacenter-design-requirements-for-ai-workloads-a-comprenshive-guide.

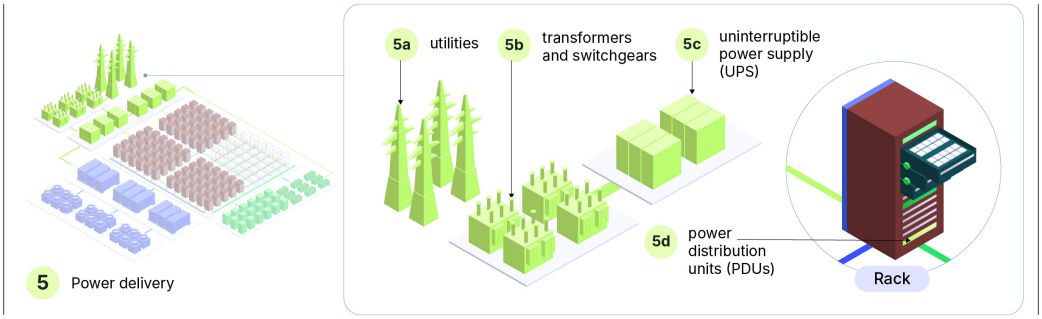
46 Asterfuison (2026). AI Data Centres vs. Traditional Data Centres: Key Differences Explained. https://cloudswit.ch/blogs/ai-data-centers-and-traditional-data-centers/.

AI clusters are wired very differently from traditional data centres because they see heavier east–west flows inside the facility, as GPUs continuously exchange model parameters and activations within and across racks.[47] To support this, operators add a dedicated compute fabric alongside the conventional Ethernet network. Inside servers, high-bandwidth links, such as NVLink, or similar GPU-to-GPU interconnects (no. 4d on the chart) handle fast short-reach communication. Across servers, low latency fabrics (no. 4b and 4c on the chart), such as InfiniBand or high-density Ethernet at 400 or 800 gigabits per second per port, carry traffic between AI accelerators and racks.[48] At the edge of the facility, both traditional and AI-oriented data centres still rely on Ethernet routers and switches (no. 4a on the chart) to connect to storage systems, other data centres and the public internet. However, AI clusters require far more ports per rack and several times more fibre connections than those in conventional designs to support this internal east–west load.

Recent analyses of frontier-scale AI facilities suggest that IT equipment outside the GPU servers, dominated by inter-server switches, fabric links and management nodes, adds roughly 14% on top of server power, implying that networking accounts for 8–10% of the total IT power in large AI clusters.[49] Studies of mixed-workload, traditional data centres typically place networking in a similar ballpark or slightly lower, often 5–10% of the total facility electricity use.[50] Section 3.3 returns to this networking stack at the rack level and examines how the power and heat from dense fabrics and optical links interact locally with accelerator density and cooling limits.

Power delivery

Figure 6: The power delivery subsystem



47 Gruve (2026). Id.

48 AscentOptics (2025). Comprehensive Guide to 400G/800G QSFP-DD Optical Modules. https://ascentoptics.com/blog/comprehensive-guide-to-400g-800g-qsfp-dd-optical-modules/.

49 Luke Emberson and Ben Cottier (2025). GPUs Account for About 40% of Power Usage in AI Data Centers. https://epoch.ai/data-insights/gpus-power-usage-in-ai-data-centers/.

50 SolarTech (2026). How Much Electricity Does a Data Center Use? Complete 2025 Analysis. https://iaeimagazine.org/electrical-fundamentals/how-much-electricity-does-a-data-center-use-complete-2025-analysis/.

From the grid to each GPU, power passes through a short chain of different facilities as follows:

- utilities providing high-voltage current at the site boundary (no. 5a on the chart),
- transformers and switchgears (no. 5b on the chart) that take in medium-voltage current and lead circuits into the cluster,
- uninterruptible power supply (UPS) and backup units (no. 5c on the chart) that smooth out disturbances and keep critical loads running during failures,
- rack-level distribution (i.e. 'power distribution units' [PDUs]) and voltage-regulation stages (no. 5d on the chart) that fan this power out to servers and finally convert it into the low-voltage current used on the chips.

Grid electricity arrives as an alternating current (AC), where the direction of flow reverses many times per second, but servers and chips run internally on direct current (DC), where current flows in one direction only. In conventional data centres, power typically flows through several AC-to-DC and DC-to-DC conversion stages: from medium-voltage AC at the site boundary to low-voltage AC in the main IT hall, intermediate DC power rails and, finally, server power supplies. Cumulative conversion and distribution losses typically account for 8–11% of the total facility power in conventional data centres.

In GPU-dense AI clusters, similar chains are used today, but the much higher rack power and more volatile load profiles mean that these losses are likely higher in practice, prompting operators and vendors to experiment with alternative topologies, such as high voltage direct current (HVDC) distribution and 800V DC power rail architectures. The reason is structural. As previously established, AI-oriented racks draw 3–10 times more power than conventional CPU-based systems. They also continuously push far more current through PDUs, power rails and voltage regulation stages than through conventional racks operating at 10–15 kW, and GPU servers often operate close to their rated power for long periods. Conventional AC-centric power supplies were designed to be the most efficient when running at moderate, relatively smooth loads. At AI rack densities, losses in conductors and conversion inefficiencies accumulate quickly, so even a few percentage points of additional loss translate into megawatts at the cluster scale.

Thus, newer architectures aim to cut losses at each step. They reduce how many times electricity is converted from AC to DC and between different DC voltages by, for example, using HVDC distribution that does most of the conversion in central units near the grid connection, which can cut distribution losses.[51] At the same time, improvements in power semiconductors, such as gallium-nitride–based converters and very efficient DC-to-DC direct current transformer modules, make each remaining conversion stage

51 Vikram Sekar (2025). Understanding High-Voltage DC Power Architectures for AI Megafactories. https://www.viksnewsletter.com/p/understanding-hvdc-architectures-ai-megafactories.

waste less energy and heat.[52]

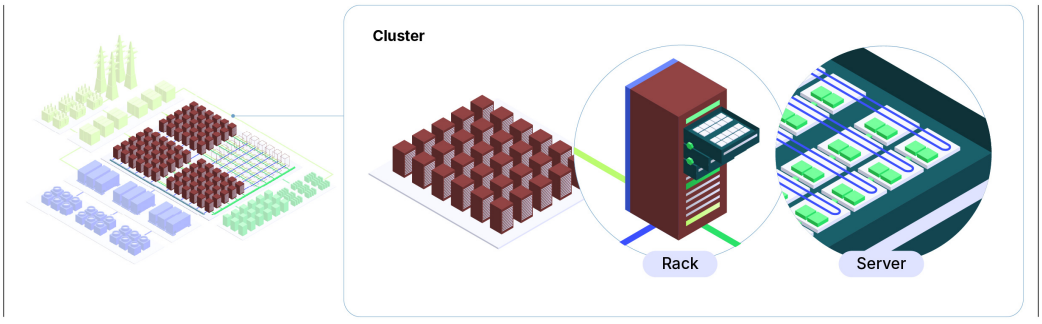
In effect, innovations in power delivery design and in the underlying power management integrated circuits are becoming as important to the energy performance of large AI clusters as improvements in the GPUs themselves.

Together, these subsystems form an integrated energy system where AI accelerators and CPUs set the primary demand, while memory, storage, networking, cooling and power delivery shape how each watt is transported and dissipated through the facility. The next section examines how these same subsystems are co-located and constrained within a single rack.

Zooming in on rack-level bottlenecks

The subsystems described in Section 3.2 are concentrated again at the rack level where power delivery, thermal limits and physical space converge. This is also the scale at which the structural drivers in Section 3.1 become physically binding, with accelerator power density, dense interconnects and advanced cooling systems all sharing the same confined volume. The remainder of this section focuses on three rack-internal elements where these constraints are currently most acute: AI accelerators, high-speed interconnects and cooling.

Figure 7: Compute infrastructure from cluster to server



Energy constraints at the AI accelerator level

Across AI accelerators, typical power draw ranges from 300–1,200 W per chip,[53] and most are operated at high utilisation for long periods.[54] Utilisation in this context refers

52 Tom Truman (2026). When It Comes to Powering AI Data Centres, GaN Is Taking Center Stage. https://www.renesas.com/en/blogs/when-it-comes-powering-ai-data-centers-gan-taking-center-stage?srsltid=AfmBOooSjKJ6IBM5t_NF2UfdzgBhDqjJ8ecON5i3XDNWU653v7uBdEfi.

53 The wide 300–1.200 W range reflects differences between accelerator generations and designs: Mid-generation or inference-optimised GPUs and ASICs typically sit around 300–400 W, while the newest frontier-training parts, which combine multi-die packages, large HBM stacks and liquid-cooled power modes, are rated at 700–1.200 W per chip.

54 The term utilisation is also used at the facility level to describe how much of a data centre's total installed capacity—servers, racks and

to the share of a chip's theoretical peak compute capacity that is actively engaged at a given moment, usually expressed as a percentage. What makes AI accelerators distinct from other components is that they routinely run close to their peak utilisation ceiling. Empirical assessments of NVIDIA H100 AI accelerators during large language model training show GPU-level utilisation at approximately 93%. Each H100 GPU carries a rated power of approximately 700 W. At the node level, where 8 GPUs are combined with CPUs, memory, interconnect and cooling, this compounds to an actual power draw of 7.9 kW, close to the 10.2 kW manufacturer-rated maximum.[55] The combination of high power per chip and sustained near-peak loads at the node level is what makes accelerators the primary energy chokepoint in the rack.

The way the racks are configured also reshapes the physical fabric of the data centre. Traditional data centres were designed around average rack densities of 5–10 kW. AI clusters typically pack four of these 8-GPU nodes into a single rack, pushing densities to 30–100 kW per rack, with some vendor reference designs in the case of, for example, NVIDIA's GB200 going beyond 135 kW.[56] To keep high-density compute nodes close enough together for low-latency interconnects, operators pack many of these racks side by side, turning individual racks into multi-megawatt 'blocks' that anchor the layout of an AI compute room (sometimes referred to as 'AI hall') and drive requirements for cooling distribution, power delivery and even how much capacity can be connected to a single grid substation.

Comparison of power density and total power

The high-energy draw at the rack level ultimately originates in the way individual accelerators use power. Rather than just asking how many watts a chip or server draws in total, designers must consider how tightly that power is packed, that is, the power density on each unit area and within each rack. Two similar chips may draw the same total power, but the one that concentrates that power into smaller physical space, that is, the one with higher power density, is much more difficult to cool and operate reliably. This is why advances in chip manufacturing that shrink transistors and pack more compute onto smaller dies do not necessarily reduce cooling demands. Designers typically use efficiency gains to deliver more compute at similar or higher total power, increasing power density rather than reducing it.[57]

---

floor space—is actively in use at a given time. A facility can be physically full of hardware (high asset utilisation) while its chips are sitting largely idle (low compute utilisation) or vice versa. This chapter uses utilisation in the compute sense, that is, how hard individual chips and nodes are working, unless otherwise specified. The distinction matters for policy. Facility utilisation shapes investment and planning decisions, while compute utilisation determines actual energy draw.

55 Imran Latif et al. (2024). Empirical Measurements of AI Training Power Demand on a GPU-Accelerated Node. https://arxiv.org/html/2412.08602v2.

56 Marc Hamilton (2025). Chill Factor: NVIDIA Blackwell Platform Boosts Water Efficiency by Over 300x. https://blogs.nvidia.com/blog/blackwell-platform-water-efficiency-liquid-cooling-data-centers-ai-factories/.

57 This reflects the end of so-called 'Dennard scaling', a principle which states that as transistors shrink, their power density remains constant because voltage and current scale downward proportionally with size, keeping total power in check even as transistor counts rise. This made successive chip generations faster without consuming more power per unit area. The principle broke down around

Modern AI accelerators exemplify this distinction. The latest GPUs and specialised accelerators are designed to routinely exceed several hundred watts per chip (e.g. NVIDIA's Blackwell GB200 generate up to 1,000 W per chip[58]). At the rack level, this translates into power densities that now reach anywhere between 30 and 140 kW per rack. These densities are 3–10 times higher than those for which many legacy data centres were designed, and they fundamentally reshape the energy profile of AI infrastructure. As density rises, removing heat becomes the dominant design constraint, and the energy required for cooling and auxiliary systems rises accordingly.

Thermal design limits

AI accelerators have strict operating temperature ranges for reliable operation, set by manufacturers to optimise performance, reliability and lifetime. Training large models requires high throughput via parallel processing, while economics incentivise maximising the use of expensive hardware (with state-of-the-art accelerators costing up to 40,000 EUR per unit).[59] In practice, this means that accelerators often run close to their thermal design limits for extended periods.

This operating regime has implications for hardware ageing. For example, Black's equation, a standard model for chip wear from heat-driven metal atom migration, shows that every 10°C temperature increase typically doubles failure rates by speeding up microscopic damage in the chip's wiring, effectively halving its lifetime.[60] In large AI clusters, where individual GPUs may already have expected service lives of only a few years, even modest thermal overshoots can translate into more frequent replacements.

Added complexity of interconnects

High-speed interconnects (no. 4c and no. 4d from Figure 5) are the second major source of rack-level energy stress. To keep accelerators fed with data and synchronise model updates, GPUs must communicate continuously within servers and across the rack so that a training job behaves as if it were running on one very large processor. Section 3.2 describes how this required a dense fabric of NVLink inside servers and InfiniBand or high-performance Ethernet between servers and racks. At the rack scale, the same fabric concentrates additional power and heat into the same confined volume as the accelerators themselves.

2005–2007. At less than 90 nm, further shrinking of transistor dimensions stopped increasing the speed of the transistor. While individual transistors have continued to become more energy efficient with each process node, chip designers have consistently applied those gains to increase compute throughput rather than reduce total power draw, resulting in increasing power density per generation.

58 NVIDIA. NVIDIA GB200 NVL72. https://www.nvidia.com/en-us/data-center/gb200-nvl72/.

59 Epoch AI (2025). NVIDIA's B200 Costs Around $6,400 to Produce, with Memory Accounting for Half. https://epoch.ai/data-insights/b200-cost-breakdown.

60 Wikipedia (2026). Black's equation. https://en.wikipedia.org/wiki/Black%27s_equation.

Within each AI server, GPUs are linked by ultra-high-bandwidth, short-reach interconnects such as NVLink. Instead of sending data over the standard connection that normally links a GPU to the rest of the server, NVLink creates a direct ‘fast lane’ between GPUs so that they can exchange model data without going through the CPU, as it is the case for the standard connection.[61] This high-speed fabric effectively pools the memory and compute of multiple GPUs inside a rack but also raises local power density, adding extra heat that needs to be removed at rack level.[62]

Across servers in an AI rack, traffic is carried by a separate scale-out fabric, typically based on a combination of InfiniBand and very high-speed Ethernet running at 400–800 gigabits per second per port, fast enough that a single link can move in a second what a typical home internet connection would need several minutes to transfer.[63] Each of these links is driven by energy-intensive switch ASICs and optical transceivers, with per-port power now reaching tens of watts and aggregate switch power in a single high-density rack often measured in the low kilowatt range. Therefore, compared with traditional CPU racks, AI racks host many more fibre cables per cabinet, and the resulting bundle of switches, optics and cables occupies valuable rack space and obstructs airflow paths that could otherwise be used for air cooling.[64]

These interconnect layers remain moderate contributors to total energy use when viewed at the facility scale. As discussed in Section 3.2, networking and fabric equipment together account for 8–10% of IT power in large AI clusters. However, at the rack scale, their impact is amplified because their watts and waste heat are co-located with the accelerators that already dominate the power budget, tightening the same electrical and thermal limits that GPUs themselves create. In practice, this means that raising available rack power to host more accelerators usually requires a parallel upgrade of switch capacity, optics, cabling and cooling, making interconnects a key part of the rack-level bottleneck.

Rack cooling systems and the shift beyond air

From an architectural perspective, the cooling requirements of CPU- and GPU-centred racks are increasingly distinct.[65] CPU racks in traditional data centres typically draw 5–15 kW and can usually be cooled with room-level or in-row air systems that have been

61 Vishnu Subramanian. What Is the Difference Between NVLink and InfiniBand? https://jarvislabs.ai/ai-faqs/what-is-the-difference-between-nvlink-and-infiniband.

62 Blake Crosley (2026). NVLink and Scale-Up Networking: When 800G Ethernet Isn‘t Enough. https://introl.com/blog/nvlink-scale-up-networking-gpu-interconnect-infrastructure-2025.

63 Nijole Simaitiene (2024). What Is a Good Broadband Speed? https://cybernews.com/uk/best-broadband-deals/what-is-a-good-broadband-speed/.

64 Compared with traditional data centre racks, AI racks are typically equipped with 4 or 5 times more fibre interconnects to link all the accelerators at high speed.

65 Global Data Center Hub (2025). Why AI Is Forcing a Complete Rethink of Data Center Design. www.globaldatacenterhub.com/p/why-ai-is-forcing-a-complete-rethink.

refined over decades. By contrast, GPU racks designed for large AI clusters start around 30 kW and can exceed 100 kW per rack, pushing air cooling to its physical limits as air volume, velocity and temperature deltas become insufficient to carry away the heat.[66] Industry guidance now treats roughly 30 kW per rack as the point beyond which liquid cooling—usually direct-to-chip or increasingly full immersion—becomes either necessary or strongly preferred for reliable operation.[67] Operators focused on dense AI workloads use these liquid-based approaches not only to keep GPUs within their thermal envelope but also to place more accelerators in each rack and to reduce fan and chiller power compared with stretching legacy air systems.

Modern AI GPUs can operate at junction temperatures in the 80–95°C range, and packaging trends (e.g. multi-die designs, stacked HBM and tightly packed chiplets) have increased local heat flux dramatically. Tests show air cooling failures at densities exceeding approximately 41 kW per rack. Above this point, the required air volume and velocity create thermal recirculation and acoustic issues that cannot be abated easily.[68] Therefore, for AI racks that are now routinely designed to exceed 50 kW per rack, air systems do not seem to be suited to remove heat efficiently.

Section 3.2 describes how direct-to-chip and immersion cooling address this at the facility level. At the rack level, the distinction between the two approaches comes down to how completely they displace air. In direct-to-chip systems, liquid typically removes typically 60–80% of the server heat (and up to 90% in the highest-density configurations), while air still cools the remaining 20–40% from components such as storage, power delivery and other peripherals that cold plates do not cover.[69] By submerging the entire server or rack, immersion cooling eliminates the need for additional air cooling at the rack level.[70]

Cooling as an integral energy cost

At high power densities, cooling becomes an integral component of an AI cluster's energy and capacity planning. As highlighted earlier, traditional air-cooled architectures typically show power usage effectiveness (PUE) in the 1.4–1.8 range, with cooling often accounting for up to 40% of the total electricity use at scale. Well-implemented liquid systems can reduce PUE to 1.05–1.15.[71] This gap matters directly for AI clusters that are designed to run at high utilisation, where every percentage point of overhead power multiplies across hundreds of megawatts.[72]

66 Ibid.

67 Tess Sohngen (2025). Why You Need Liquid Cooling for AI Performance at Scale. https://www.coreweave.com/blog/why-you-need-liquid-cooling-for-ai-performance-at-scale.

68 Blake Crosley (2026). Liquid Cooling vs Air Cooling for AI Data Centers: 2025 Analysis. https://introl.com/blog/liquid-vs-air-cooling-ai-data-centers.

69 Blake Crosley (2026). id.

70 Kawsar Haghshenas et al. (2022). Comparing of Data Center Cooling Solutions. https://ar5iv.labs.arxiv.org/html/2205.04257.

71 Blake Crosley (2026). Water Usage Efficiency: AI Data Center Cooling Without Crisis. https://introl.com/blog/water-usage-efficiency-wue-ai-data-center-cooling-guide-2025.

While liquid cooling systems are more effective at heat removal, they do not eliminate cooling-related energy bottlenecks. Pumps, heat exchangers, chillers and control systems all consume electricity, and their energy use scales with both compute load and ambient conditions characteristic to the region where the cluster is based. In high-density AI facilities, cooling can still account for a substantial share of total electricity use, especially during periods of peak utilisation or in warmer climates.

Moreover, cooling demand does not necessarily track compute demand perfectly. The physical mass of cold plates, coolant volumes and facility chillers retain heat even after GPUs ramp down from training peaks.[73] Pumps and chillers must continue operating for minutes to hours to gradually stabilise temperatures, which then impacts the power curve at the facility level—a topic we will address in the following sections.

What these characteristics mean at scale

Taken together, this chapter shows that large AI compute clusters differ from traditional data centres in both their internal architecture and in how they draw power. Traditional facilities are built around general-purpose CPUs, moderate rack densities of roughly 3–15 kW and room-level air cooling, so the total demand is spread across many relatively low-power servers and a comparatively thin networking fabric. By contrast, large AI clusters centre on specialised accelerators that consume 300–1.200 W per chip, are run at high utilisation and are packed into racks that routinely operate in the 30–100 kW range, supported by dense interconnects, liquid-based cooling and upgraded power delivery chains. The result is a much more concentrated and less forgiving energy footprint.

PUE: A helpful but incomplete metric

PUE is the standard ratio used to describe traditional data centre infrastructure efficiency:

PUE = total facility power ÷ IT equipment power

A PUE of 1.2 means that for every 1 kW used by servers, storage and networking, the building consumes another 0.2 kW for cooling, power conversion, lighting and other overheads. Lower PUE values indicate that a larger share of electricity reaches the IT equipment rather than being lost in supporting systems.[74]

PUE has been central to efficiency efforts in conventional data centres and underpins the EU's Energy Efficiency Directive reporting framework and upcoming European data centre rating scheme.[75] However, for AI-specific infrastructures, it gives an incomplete picture:

72 Seamus Nayduch (2025). CoreWeave Data Center Operations: Built for AI. https://www.coreweave.com/blog/coreweave-data-center-operations-built-for-ai.
73 Training peaks are further explained in Section 4.1.
74 EDP Europe. Understanding PUE and Its Impact on Data Centre Sustainability. https://www.edpeurope.com/power-environmental/understanding-pue-and-its-impact-on-data-centre-sustainability/.
75 https://energy.ec.europa.eu/topics/energy-efficiency/energy-efficiency-targets-directive-and-rules/energy-efficiency-directive_en.

- It does not measure utilisation: A facility can have an excellent PUE and still run its GPUs at 30–40% utilisation, which means that much of the ‘IT power’ counted in the denominator is idle.
- It says nothing about the carbon intensity or timing of electricity use. The Commission’s technical assessment notes that ‘a low PUE does not provide any indication of whether the electricity is derived from fossil fuels, the duration of server utilisation, or the extent to which waste heat is reused’.
- In AI clusters, aggressive liquid cooling and high rack densities can lower PUE (e.g. from 1.4–1.8 to 1.05–1.15), while total electricity demand still increases because more accelerators are installed.

For this reason, the rest of the report treats PUE as a useful indicator of building-level efficiency but not as a proxy for the energy productivity of AI compute itself, a gap later chapters address through utilisation-based metrics.

Within the AI cluster’s footprint, the rack emerges as the key unit of energy concentration. All five subsystems described in Section 3.2 are co-located within each cabinet. At this scale, three structural drivers from Section 3.1 become physically binding: accelerator power density, the heat and space taken up by high-bandwidth interconnects and the limits of rack-level cooling and power conversion hardware. These interactions explain why AI racks hit electrical and thermal ceilings long before floor space is exhausted, why liquid cooling shifts from an efficiency option to a necessity and why incremental improvements in chip-level efficiency do not automatically translate into lower facility-level power if they are used to add more accelerators per rack.

Viewed from the above-mentioned findings, these rack-scale properties give large AI clusters a distinct profile even when their total megawatt demand overlaps with that of traditional sites. Their load is dominated by a relatively small number of high-density compute blocks, connected by equally specialised networking and cooling infrastructures, and designed to run near capacity for extended periods. In operational terms, they behave less like generic co-location centres with diversified, partially idle workloads and more like firm baseload industrial actors that run at sustained high intensity. This is because their operation is tightly coupled to hardware amortisation and model training cycles or AI deployment. At the same time, large AI compute clusters in the 100–300 MW range typically require new or reinforced high-voltage grid connections in the 110–220 kV band and, in many systems, are likely to trigger upstream 380 kV reinforcement needs.

Therefore, it is useful to treat large AI compute clusters as a separate category of infrastructure rather than as a simple continuation of existing data centre trends. The structural differences mapped in Chapter 1 and summarised in Table 1 below shape how these facilities interact with the grid, how flexible their demand can be and what kinds of risks and opportunities they create for energy and climate goals.

Table 1: Comparison of architectures – AI compute cluster and traditional data centre

	AI compute cluster	Traditional data centre
Compute	**Hardware:** Specialised AI accelerators (GPUs/TPUs) with HBM **Parallelism:** Deep, large homogenous tasks **Workload:** Massive parallel tasks	**Hardware:** General-purpose CPUs and minimal accelerators **Parallelism:** Many small independent tasks **Workload:** Mixed, latency-sensitive applications
Memory and storage	**Hierarchy:** HBM-local SSD-central HDD/ object storage **Priority:** Keep accelerators fed; data proximity is key **Access pattern:** Sequential, large-batch reads	**Hierarchy:** Server DRAM-thin SSD layer-central HDD **Priority:** Cost optimisation and request throughput **Access pattern:** Random, small, frequent requests
Networking	**Topology:** Dense east–west compute fabric **Interconnects:** NVLink, InfiniBand, and high-speed Ethernet **Scale:** Synchronises thousands of accelerators	**Topology:** North–south user to server flows **Interconnects:** Conventional Ethernet **Scale:** Fewer high-speed ports per rack
Power delivery	**Density:** 3–10 times the conventional rack power **Distribution:** HVDC + advanced power electronics **Design driver:** Managing conversion losses at scale	**Density:** Standard rack power levels **Distribution:** Conventional AC-centric chain **Design driver:** Smooth predictable loads
Cooling	**Method:** Liquid (direct-to-chip or immersion) **Threshold**: Design assumption higher than 30 kW/rack	**Method:** Air cooling (room level or in-row) **Threshold:** Adequate up to 15–30 kW/ rack

Analysing the energy characteristics of AI infrastructure, this chapter demonstrates that this AI-versus-traditional distinction should be made explicit in policymaking strategies, that is, in how new projects are planned and sited, in how performance standards and reporting schemes are designed, and in how Europe's broader AI-industrial ambitions are reconciled with the physical limits of its power system.

Chapter 2 - How AI workloads use energy: Comparison of training and inference

Chapter 1 showed that large AI compute clusters differ from traditional data centres because they concentrate 30–100 kW per rack around high-utilisation accelerators, dense interconnects and liquid cooling. This chapter takes that physical baseline as given and asks how different workload regimes—training versus inference—turn the same racks into very different load profiles at the system level.

AI workload regimes as separate energy profiles

Building on the structural and subsystem characteristics outlined in Chapter 1, this chapter examines how AI workloads actually use that hardware over time. The way power demand unfolds over hours, days and weeks is generally captured in what electrical engineers call a load profile—the time-varying pattern of electricity draw. AI clusters produce load profiles with distinct features that may challenge power system planning, cooling infrastructure and integration with intermittent renewables. To understand the latter, we examined the two main AI workload categories: training and inference.

Most AI accelerators (including NVIDIA H100 and its successors) are designed to handle both the training (development) and inference (deployment) of AI models. At the same time, an increasing number of chips are now being optimised for narrower roles, especially cost-efficient inference, such as AWS Inferentia, Meta's MTIA family and Groq's LPUs, and this specialisation is likely to continue as operators search for cheaper alternatives to general-purpose GPUs. To make the best use of the hardware available, large-scale AI clusters use unified scheduling systems to co-schedule training and inference jobs on the same GPU pool.[76] Thus, training and inference are best understood as two different ways of using the same chips, producing different load profiles and energy characteristics, as the sections below set out.[77] The workload patterns build directly on the rack-level characteristics described in Chapter 1: GPUs already operating at 300–1.200 W per chip, near-constant utilisation during training phases and per-rack densities of tens of kilowatts. When these utilisation patterns change, the entire 30–100 kW rack follows.

Dimension 1 – Load profile and intensity

Training typically consists of three stages. Pre-training uses massive, mostly unlabelled datasets to teach the model general patterns in language, images or other data. Mid-training fthen mixes in more specialised or higher-quality data to steer capabilities towards particular domains, and post-training (often called fine-tuning) aligns the model for specific tasks using smaller, curated datasets.

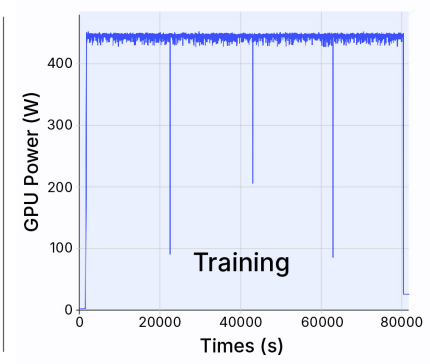
Training runs are time concentrated, using power sharply as thousands of accelerators synchronise at near-peak utilisation for days or weeks.[78] GPT-4 training reportedly

76 Lingling Zeng et al. (2025). Kant: An Efficient Unified Scheduling System for Large-Scale AI Clusters. https://arxiv.org/abs/2510.01256.

77 While this chapter focuses on flexible accelerators that can serve both training and inference (e.g. NVIDIA H100/H200-class GPUs), major cloud providers also deploy specialised chips that are optimised for one regime only, such as AWS Trainium for model training and AWS Inferentia or Google's Ironwood TPU for large-scale inference.

consumed approximately 46 GWh in total energy, which is equivalent to a sustained 20 MW draw over three months and enough to power the entire Brussels Capital Region for more than four days.[79] At the server level, training draws 7–8 kW per server and saturates hardware and cooling for extended periods.

Figure 8: Training load profile[80]



What you see: The horizontal axis shows time in seconds, and the vertical axis shows power drawn by a single GPU in watts. Over the whole training run, the curve stays almost flat at approximately 430 W, with only very short dips roughly every 20,000 seconds (around 5.5 hours) as the workload or job stage changes.

What it means: This pattern illustrates that once a training job is underway, each GPU operates at close to its maximum power for many hours at a time. As a result, utilisation and thus electricity demand, remains high and steady rather than fluctuating like typical web or enterprise workloads.

Inference serves user queries with a trained model using several times less energy per operation but easily scales to massive volumes of requests, especially across popular services.[81] At the node level, inference draws 4–6 kW per node but runs nearly continuously and is tied to user activity patterns.

Like training, inference can also be divided into three phases. Pre-deployment optimisation compresses and restructures models before launch. Active serving is the steady, high-volume baseline for processing live user requests, where accelerators run continuously to keep latency low. Finally, background adaptation covers periodic model

78 Roberto Vercellino et al. (2026). Measurement of Generative AI Workload Power Profiles for Whole-Facility Data Center Infrastructure Planning. https://arxiv.org/html/2604.07345v1.

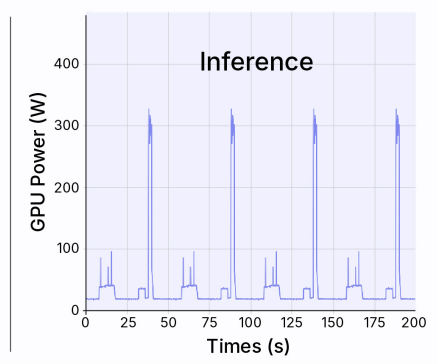
79 Several publications suggest that the training of OpenAI's latest model GPT-5.2 released in December 2025 entailed a power draw many times higher than that of GPT-4. Recent frontier models, such as GPT-5.2 (released December 2025), are estimated to have required 200–500 GWh, with cluster-level peaks reaching 100–300 MW during ramp-up.

80 *From: Sheng, Y., Zhang, C., Zhu, Z., Xu, H., Wen, J., Wang, R., Yang, J., Wang, Q., & Bu, S. (2026). Power for AI Data Centers: Energy Demand, Grid Impacts, Challenges and Perspectives.* Energies, 19*(3), 722.* https://doi.org/10.3390/en19030722

81 Felipe Oviedo et al. (2025). Energy Use of AI Inference: Efficiency Pathways and Test-Time Compute. https://arxiv.org/abs/2509.20241.

refreshes or evaluation workloads.[82] Each mode contributes distinct load patterns at the cluster scale. Optimisation phases briefly saturate accelerators similar to training. Active serving sustains a near-constant draw across the AI accelerators, while adaptation adds unpredictable peaks as AI models are partially retrained based on new data.

Figure 9: Inference load profile[83]



What you see: The horizontal axis shows time in seconds, and the vertical axis shows GPU power in watts. Power rises in a repeating pattern roughly every 25 seconds. First, a smaller ramp-up to approximately 100 W occurs, and then a narrow spike exceeding 300 W can be observed before another decrease and the pattern is repeated.

What it means: This shows that inference drives GPUs in short, frequent bursts rather than at a flat maximum. Each wave of user requests or background updates produces a brief, high-power peak on top of a lower baseline, so the cluster's load is spiky but nearly continuous over time.

Cooling adds another dimension to these load patterns. Owing to thermal inertia, cooling demand does not track compute demand perfectly. The physical mass of cold plates, coolant volumes and facility chillers retains heat even after GPUs ramp down from training peaks. Pumps and chillers must continue operating for minutes to hours to gradually stabilise temperatures. It is widely understood that liquid cooling systems are much better placed to handle this efficiently. For example, direct-to-chip H100 systems run a 10–20°C cooler and deliver higher performance per watt than air-cooled systems under the same AI workloads.[84] Studies also show that inertia increases the need for AI clusters to be equipped with dynamic cooling response controls. With AI compute power changing quickly and cooling power adjusting more slowly, the systems risk experiencing

82 Within each request during active serving, inference technically splits into prefill (the compute-intensive step where the model processes the full prompt) and decode (the memory-related step of generating tokens one by one). However, at the cluster level, these micro-phases blend into a flatter profile than the synchronised ramps of training.

83 *From: Sheng, Y., Zhang, C., Zhu, Z., Xu, H., Wen, J., Wang, R., Yang, J., Wang, Q., & Bu, S. (2026). Power for AI Data Centers: Energy Demand, Grid Impacts, Challenges and Perspectives.* Energies, 19*(3), 722.* https://doi.org/10.3390/en19030722

84 Imran Latif et al. (2025). Cooling Matters: Benchmarking Large Language Models and Vision-Language Models on Liquid-Cooled Versus Air-Cooled H100 GPU Systems. https://arxiv.org/abs/2507.16781.

periods of over-cooling and higher cooling energy.[85]

Dimension 2 – Latency sensitivity as a siting constraint with energy consequences

Given the rack-scale power densities and continuous operating regimes described in Chapter 1, siting decisions for AI clusters now resemble siting decisions for other large industrial loads, with latency adding an extra constraint for inference-heavy deployments. Training and inference differ fundamentally in where they can be located, and siting emerges as one of the most consequential energy variables for large AI clusters.[86]

Training workloads prioritise throughput and are not latency sensitive, which means that clusters can be located wherever energy conditions are most favourable; that is, areas where renewables drive lower electricity prices or far from population centres are entirely viable. In contrast to inference, training is more flexible in terms of scheduling and can theoretically be shifted to off-peak periods. Several AI labs are also testing to connect a training run via multiple locations in the same region. SemiAnalysis reports that racks belonging to one cluster are typically kept within roughly 30 meters of the network core and that synchronous multi-datacentre training works best when facilities are only kilometres to a few tens of kilometres apart (up to 100 km) so that additional round-trip latency stays in the sub- to low-millisecond range.[87] Beyond regional distances, even the speed of light in fibre imposes tens of milliseconds of delay. At that point, it is no longer possible to update all GPUs everywhere at the same moment, so very large training runs start to use multistage or partly asynchronous update schemes instead of treating distant sites as one synchronous super-computer.

On the other hand, inference workloads benefit from reaching the lowest possible latency. Active serving of live user queries requires response times usually less than 200 milliseconds, roughly the time it takes to blink. Every step in that chain consumes time. The request travels from the user's device to the data centre, the model processes it, and the response travels back. The further the AI cluster is from the user, the more the 200-millisecond budget is consumed just by the physical journey of data across the network, leaving less time for the actual computation.[88] This means inference workloads pull towards sites closer to population centres, typically where grid connection is more constrained, energy is more expensive, renewable availability is lower and natural cooling

85 Jiaqiang Wang et al. (2025). MPC-based joint optimization for rack-based cooling data centers: Modeling, performance evaluation, and time-delay characteristic analysis. *Building and Environment*, 286, 113695. https://doi.org/10.1016/j.buildenv.2025.113695.

86 Nicoleta Kyosovska & Andrea Renda (2025). EU Plans for AI (Giga)Factories: Sanctuaries of Innovation or Cathedrals in the Desert? https://www.ceps.eu/ceps-publications/eu-plans-for-ai-gigafactories-sanctuaries-of-innovation-or-cathedrals-in-the-desert/.

87 Dylan Patel, Daniel Nishball, & Jeremie Eliahou Ontiveros (2024). Multi-Datacenter Training: OpenAI's Ambitious Plan To Beat Google's Infrastructure. https://newsletter.semianalysis.com/p/multi-datacenter-training-openais.

88 Prodia (2026). Inference Speed Benchmarks Explained: Compare LLM Performance for Developers. https://blog.prodia.com/post/inference-speed-benchmarks-explained-compare-llm-performance-for-developers.

is less accessible.[89]

The practical implication is that training and inference clusters built on identical hardware operate under fundamentally different energy conditions.

Dimension 3 – Demand response and grid flexibility

In the context of this study, we use ‘demand response’ and ‘demand-side flexibility’ in the same broad sense as recent IEA and EU works, defining it as the ability to adjust the timing or level of electricity use in response to power system conditions, typically by reducing demand during peaks, increasing it when low-carbon generation is abundant, or shifting it across hours or days based on price or grid signals.[90] For large AI clusters, this means modulating training or non–time-critical inference workloads in response to system needs rather than treating their load as fully inflexible.

As training workloads have a better tolerance for higher latency, they have more flexibility potential than inference. Users do not wait for a training run to complete within seconds; thus, training loads can, in principle, be scheduled around periods of grid stress, shifted to off-peak hours, or curtailed during emergency events, with the run resuming from a checkpoint once conditions normalise. Research on the demand response participation of AI data centres confirms that training workloads offer greater scheduling flexibility than inference workloads, with modelled cost savings of 17–54% in reserve service programmes.[91] Google has formalised this in agreements with utilities in the United States, committing to reschedule non-urgent AI workloads during grid stress events.[92] Recent research in the European context also shows that even relatively modest flexibility can materially change how many sites can be connected on constrained grids. One modelling exercise for Belgium’s transmission system operator finds that allowing large data centres to operate flexibly for roughly 5% of their total operation time (or to shed a similar share of their annual demand) substantially increases the number of locations where such loads can connect to the grid without triggering major modernisation or reinforcement work.[93]

In practice, this flexibility is most available in dedicated frontier training clusters, where large continuous pre-training jobs run without competing for the same GPU pool. However, most AI clusters run mixed workloads, where training and inference share the

89 Aaron Delp & Terry Takouridis (2025). 3 Reasons AI Inference Belongs at the Metro Edge. https://blog.equinix.com/blog/2025/05/27/3-reasons-ai-inference-belongs-at-the-metro-edge/.

90 Florence School of Regulation (2025). Flexibility in Power Systems. https://fsr.eui.eu/flexibility-in-power-systems/.

91 Fatih Acun et al. (2026). Investigating Power Consumption Flexibility in AI Data Centers for Demand Response Participation. https://www.bu.edu/peaclab/files/2026/03/FlexDC_Sim_ACM_E_Energy26.pdf.

92 Tobias Mann (2025). Google Agrees to Pause AI Workloads to Protect the Grid When Power Demand Spikes. https://www.theregister.com/2025/08/04/google_ai_datacenter_grid/.

93 Pawel Czyzak (2026). Data Center Flexibility (Intro). https://paczyzak.substack.com/p/data-center-flexibility-intro?publication_id=7847607&post_id=194047574&isFreemail=true&r=7avnm.

same hardware, and the inference component must remain continuously available, limiting how freely the cluster can respond to grid signals. At the same time, most AI operators, even those with the technical capability to shift training loads, still show a strong preference for approaches with zero workload impact. They largely choose onsite generation and battery storage over curtailment, pointing towards the high operational costs of restart cycles on large training jobs.[94] Ultimately, the flexibility potential of training workloads remains theoretical unless operators expose it through contractual mechanisms (e.g. interruptible load schemes). As the number of AI workloads shifts from training to inference, the window for load flexibility might narrow further.

The first empirical trials already point to the magnitude of flexibility that AI clusters can deliver without compromising core services. In the UK, a demonstration ran by Nebius, Emerald AI, EPRI and the National Grid showed an AI cluster sustaining 10–40% load reductions for a period of 10 hours while delivering 99% performance on the highest priority workloads.[95] Coupled with the Belgian system-level analysis, these findings support the argument that including flexibility should be treated as part of the core design parameters for AI clusters, on par with peak demand and utilisation, when evaluating their compatibility with constrained European grids.

Inference clusters face more challenging constraints. Real-time serving of user interactions cannot simply be delayed because curtailment directly shows up as slower or delayed responses. By contrast, offline batch inference—large jobs where results are not time critical, such as generating recommendations—can often be scheduled into off-peak hours, and agentic AI tasks that run over minutes rather than milliseconds also somewhat allow for more flexibility than traditional query-response interactions.[96] However, even though online providers internally ‘batch’ many live requests to improve efficiency, the continuously available portion of inference demand cannot participate in standard curtailment programmes. This makes inference-dominated facilities structurally less useful as demand response assets from a grid operator’s perspective.

These flexibility asymmetries have direct consequences for how large AI clusters interact with power systems. The inference baseline requires a firm dispatchable capacity or storage to cover 24/7 demand. Training peaks, which can surge 15–25% higher than the nominal load during synchronised ramp-up, stress local distribution and transmission infrastructure in ways that require advance planning and, in some cases, dedicated grid upgrades.

Therefore, a mixed workload AI cluster creates two distinct challenges for grid planners.

94 ChonLam Lao et al. (2024). TrainMover: An Interruption-Resilient and Reliable ML Training Runtime. https://arxiv.org/abs/2412.12636.
95 National Grid (2026). UK-First Trial of AI Grid Technology Successfully Demonstrates the Ability of Data Centres to Adjust Power Needs. https://www.nationalgrid.com/uk-first-trial-ai-grid-technology-successfully-demonstrates-ability-data-centres-adjust-power-needs.
96 Yi Wang et al. (2025). Providing Load Flexibility by Reshaping Power Profiles of Large Language Model Workloads. https://www.sciencedirect.com/science/article/pii/S2666792425000265.

On the one hand, the continuously available inference load behaves much like a conventional, firm industrial demand that must be backed by reliable capacity or long-duration storage. On the other hand, large shared-cluster training runs behave more like episodic, high-power industrial processes, as they can, in principle, be shifted or slowed in response to system conditions but only if operators are willing to accept higher training times and if appropriate demand-response contracts and interconnection terms are in place. This is why recent work on AI training load fluctuations and first-of-a-kind gigawatt-scale AI clusters argues for differentiated grid responses: firm capacity and resilience planning for inference baseline and flexible interconnection capacity plus explicit demand management agreements for training-driven peaks.[97]

Hybrid workloads as an energy variable

This chapter has argued that the AI cluster's infrastructure creates energy characteristics that are structurally different from those of a traditional data centre. Specialised accelerators, dense interconnects and liquid cooling form a fundamentally different computational regime, one that concentrates far more power into smaller physical footprints and runs at sustained near-peak utilisation for extended periods. These design choices are not neutral; they push cooling and power delivery equipment into new regimes and produce load profiles that look more like industrial process plants than traditional server farms.

The training-versus-inference analysis sharpened this point. The two primary AI workload types differ not only in their load profiles but also in the energy conditions they require, the locations they can occupy and the flexibility they can offer to grid operators. Training produces sustained high-intensity plateaus, whereas inference produces a lower but highly variable and largely inflexible baseline tied to user demands. Training is power hungry but geographically free and at least theoretically deferrable. Inference must be close to users, must remain continuously available and locks in a demand baseline that cannot easily be curtailed. These structural features are set at the design stage and are difficult to adjust once construction is underway.

Crucially, apart from a few frontier AI labs dedicated to training facilities, largest AI clusters run both workload types on the same hardware. Therefore, the energy profile is often a hybrid, shaped by the proportions of its workload mix in ways that neither peak power figures nor infrastructure efficiency metrics can easily reveal. A facility's nameplate capacity informs about the maximum it can draw but nothing about how much of that capacity is doing useful work, for whom and under what conditions.

97 Jeremie Eliahou Ontiveros, Dylan Patel, & Ajey Pandey (2025). AI Training Load Fluctuations at Gigawatt-Scale—Risk of Power Grid Blackout? https://newsletter.semianalysis.com/p/ai-training-load-fluctuations-at-gigawatt-scale-risk-of-power-grid-blackout.

The energy characteristics established here (load profile, siting constraints, demand response potential and fixed infrastructure overhead) produce a range of possible propositions, depending on what the cluster runs and at what utilisation levels. Chapter 3 applies that reasoning to the demand scenarios most plausible for Europe and asks what they imply for how AIGFs could be integrated into the continent's constrained energy system.

Chapter 3 - How to make AIGFs work: Demand, utilisation and system-level implications

Why utilisation emerges as a key metric

The previous chapters established the structural characteristics of large AI compute clusters and why its energy profile differs from conventional data centre infrastructures. However, understanding the energy demand of a single cluster at full load is only part of the picture. The more pressing question, both for the economics of these facilities and for their position in Europe's energy system, is how often they run at that load. This is where the concept of utilisation is central.

At the chip level, utilisation refers to the share of a GPU's theoretical compute capacity that is actively performing useful work at any given moment. In tightly optimised benchmark or hyperscale training runs, cluster-level GPU utilisation can reach approximately 90% for extended periods once data pipelines and storage bottlenecks are carefully engineered. However, this figure represents the best-case scenario. Studies of commercial deployments where workloads are more varied and scheduling policies face limitations find that utilisation is considerably lower at 30–70% on average. At the cluster level, utilisation takes on a different meaning, as it describes how much of the installed compute capacity is actively drawing load over time, as opposed to sitting idle or in standby. [98]

The two concepts are related but distinct. A cluster can be running at high capacity (with all racks being powered and all servers on), while its GPUs perform useful computation only intermittently. The reason for this is largely structural, although the mechanism differs by workload type.

98 Akhmadillo Mamirov (2025). Reducing Fragmentation and Starvation in GPU Clusters Through Multi-objective Scheduling. https://arxiv.org/html/2512.10980v1.

For training, GPU idle time is built into how the workload runs. Research on training workloads at scale finds that preprocessing is often the main bottleneck. In some systems, up to 65% of the time in each pass through the training data can be spent on loading, transforming and shuffling training data before GPUs can perform any calculations.[99] Training large models across thousands of GPUs also requires splitting the workload into sequential stages, where each GPU processes its stage and then waits for the next one to finish before it can begin again. These waiting periods (sometimes referred to as 'pipeline bubbles') can consume 15–30% of GPU time in typical large-scale training runs.[100] Communication overhead adds further waste. In distributed training, GPUs must continuously exchange updates with each other, and studies have found that 14–32% of all GPU hours are spent on this inter-GPU communication, with no useful computation occurring in parallel.[101]

For inference, the problem is different but equally consequential. As established in Chapter 2, inference workloads are inherently bursty, which means that GPUs provisioned to handle peak demand sit partially idle for much of the day. Studies of commercially deployed AI inference models have found that request volumes for conversational AI can be three times higher at peak than at off-peak hours and that for coding assistants, peak demand can reach more than 30 times the overnight minimum.[102] At the same time, the inference cluster must always be optimised for the busiest moment. During quieter periods, the same hardware continues drawing power while performing a fraction of its potentially useful work.

In both cases, GPUs continue drawing power through idle and underutilised periods. The result is that facility-level utilisation tends to fall well below the peak figures that hardware specifications and cluster announcements typically cite. Training produces high utilisation during active campaigns but low utilisation in the gaps between them. Inference produces more consistent but structurally lower utilisation because the cluster must always be provisioned for the peak demand it rarely reaches.

The energy cost of this gap is worth zooming into. GPUs account for only around 40% of the total facility power at peak operation in a large AI cluster, with cooling, networking, power conversion and server overhead accounting for the rest. A large portion of that infrastructure runs regardless of whether the GPUs are computing or waiting. When utilisation is low, this fixed infrastructure power is spread over fewer useful computations, so the energy cost per unit of output rises sharply. This pattern is clearly

99 Keshav Vinayak Jha (2025). HyCache: Hybrid Caching for Accelerating DNN Input Preprocessing Pipelines. https://www.usenix.org/system/files/atc25-jha.pdf.

100 Daiyaan Arfeen et al. (2024). PipeFill: Using GPUs During Bubbles in Pipeline-parallel LLM Training. https://arxiv.org/abs/2410.07192.

101 Samuel Hsia et al. (2024). MAD Max Beyond Single-Node: Enabling Large Machine Learning Model Acceleration on Distributed Systems. https://arxiv.org/abs/2310.02784.

102 Jovan Stojkovic et al. (2024). DynamoLLM: Designing LLM Inference Clusters for Performance and Energy Efficiency. https://arxiv.org/abs/2408.00741.

visible at the inference level, where benchmarking on H100 clusters shows that increasing batch utilisation on the same hardware can reduce energy per token by a factor of two to three because more tokens share the same baseline and overheads.[103]

This has direct implications for how we evaluate the energy productivity of large AI clusters, such as AIGFs. A facility consuming 300 MW continuously is not in the same category, whether its accelerators are mostly engaged in large training runs at 85% GPU utilisation or are fragmented across many smaller enterprise inferences and fine-tuning workloads at 40%. The grid connection required is identical, but the useful output per megawatt-hour differs. In practice, even frontier-scale clusters are rarely 'pure' training facilities, as a GPT-4-class pre-training run may occupy tens of thousands of GPUs for a few months, after which capacity is progressively reassigned to other workloads, such as fine-tuning, evaluation and high-volume inference for deployed models. However, the closer the facility operates to high-utilisation, training-dominated workloads, the more AI work it produces for a given continuous power draw.

Despite this, utilisation is largely absent from the EU's current data centre legislative and policy frameworks. The European sustainability reporting rules for data centres, established by the Energy Efficiency Directive (EED) and operationalised through a delegated act, require operators of facilities with an installed IT power demand above 500 kW to report annually on 18 key performance indicators, including energy consumption, PUE, water use, waste heat and renewable energy share.[104] The four sustainability indications that the Commission calculates from these data (PUE, water usage effectiveness [WUE], energy reuse factor and renewable energy factor) are all measures of infrastructure efficiency, showing how well the building serves its IT load.[105] At the moment, none of them measure how effectively its compute capacities are used.

Within the upcoming Roadmap for digitalisation and AI in the energy sector, the Commission is now developing a European rating scheme for data centres based on the sustainability reporting framework stemming from the EED. However, the logic is still similar to the framework that is already in place, with the scheme designed to evaluate building infrastructure performance. More complex IT performance metrics, including any measure of compute utilisation, have been explicitly deferred by the industry to after 2030, pending standardisation.[106]

---

103 In large language models, a token is the basic text unit that the model consumes or produces (e.g. a sub-word fragment). Reporting energy 'per token' is a standard way to compare inference efficiency across hardware and system configurations.

104 European Commission (2023). Directive (EU) 2023/1791 of the European Parliament and of the Council of 13 September 2023 on Energy Efficiency and Amending Regulation (EU) 2023/955 (Recast) (Text with EEA Relevance). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ%3AJOL_2023_231_R_0001&qid=1695186598766.

105 European Commission (2024). Commission Delegated Regulation (EU) 2024/1364 of 14 March 2024 on the First Phase of the Establishment of a Common Union Rating Scheme for Data Centres. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32024R1364&qid=1775210757146.

106 Industry representatives, such as DigitalEurope, the association representing all major technology companies operating data centres in Europe, opposed including an accelerated compute metric in the upcoming rating scheme, arguing that real-world GPU utilisation is too variable and too poorly standardised to be meaningfully captured in a regulatory label yet.

Should the Commission follow this advice, the upcoming rating scheme would be likely to indicate how efficiently a data centre uses energy relative to its IT load and how green that energy is but would be less likely to show whether the IT load represents useful work.[107] For general-purpose colocation data centres hosting a mix of web servers and databases, this gap matters relatively little. For an AIGF requesting grid capacity comparable to a small city and drawing on public co-financing, it is precisely the question that accountability should turn on.

While the EED framework was not designed for AI-specific infrastructure, it has already made significant political effort to get the current reporting requirements in place. The first reporting cycle of September 2024 already revealed data quality problems, with the technical report commissioned by DG ENER showing that only 36% of all European data centres subject to the reporting requirements participated in the first survey and that industry feedback pointing to unclear guidelines, inconsistent definitions and overly complex obligations were the main barriers to compliance.[108] These early data-quality and participation problems can be seen as the expected growing pains of a regulatory instrument that is being built in real time.

Nevertheless, some implications for AIGFs might be drawn here. Both the EED sustainability reporting scheme and the upcoming European rating scheme are starting points for establishing the structure for EU-level data centre accountability. However, the discussion about AIGFs, especially in the context of the upcoming publication of tender criteria for those facilities, is an opportunity to address the layer that the EED framework is structurally unable to provide, which is a requirement to report and eventually meet minimum targets for compute utilisation. Without this layer, large AI compute clusters, such as AIGFs, could earn strong sustainability labels based on low PUE, high renewable share and efficient cooling while operating at utilisation levels that make its energy consumption (per unit of useful AI output) highly wasteful.[109] To prevent this from happening, we must dive a bit deeper into different demand scenarios and their implications for utilisation for AIGFs.

107 The concept of ‘useful energy output’ is commonly defined as the share of energy that goes towards the desired output of the end-use application rather than being lost in conversion or overheads. In the context of a large AI compute cluster, useful output can be understood as the share of energy that actually serves AI workloads (training or inference), as opposed to the energy used for supporting infrastructures, such as cooling, power distribution and networking systems.

108 Simon Hinterholzer et al. (2025). Assessment of Next Steps to Promote the Energy Performance and Sustainability of Data Centres in EU, Including the Establishment of an EU-Wide Rating Scheme—First Technical Report. https://euagenda.eu/publications/assessment-of-the-energy-performance-and-sustainability-of-data-centres-in-eu.

109 The second technical report commissioned by DG ENER to assess the reporting scheme notes that ‘a low PUE does not provide any indication of whether the electricity is derived from fossil fuels, the duration of server utilisation, or the extent to which waste heat is reused’ and explicitly acknowledges that ‘emerging indicators such as CPU utilisation or workload-specific metrics (e.g. for AI/accelerated computing) are not included at this stage’ in the proposed minimum performance standards.

Energy compatibility of large AI clusters in the European context

Building on the technical baseline from Chapter 1 and the workload regimes analysed in Chapter 2, this section analyses the physical and operational constraints for two potential AIGF use cases.

AIGFs and the underlying market dynamics

AIGFs emerge from a specific market constellation. Understanding this constellation is essential because it shapes who builds AIGFs, which users they serve and how these AI clusters appear to the energy system. On the provider side, a small group of US hyperscalers, specialised neoclouds and a handful of European infrastructure developers are competing to secure high workloads, long-term power contracts and suitable sites.[110] Hyperscalers, such as Microsoft, Google and Amazon, increasingly bundle compute, storage and proprietary scheduling software as vertically integrated dedicated AI compute infrastructure, while neoclouds, such as CoreWeave and Fluidstack, position themselves as GPU-rich alternatives that rent capacity to model developers and enterprises, often emphasising high-density GPU clusters and specialised commercial arrangements for large AI workloads.[111] In turn, European energy companies and data centre operators are beginning to look at large AI clusters as potential priority customers for new grid connections and generation projects. Where utilisation can be kept high and contracts are long enough, a single AIGF can underwrite a significant share of the revenue needed to justify new network and generation investments.[112]

On the demand side, at least three user groups shape how these facilities are likely to operate. First, a very small number of frontier AI labs (currently located almost entirely outside the EU) require tightly coupled, high-utilisation clusters to train and deploy successive generations of large AI models. These users favour long-term, dedicated capacity and are willing to sign multi-billion–euro contracts, as recent deals between OpenAI, Google, xAI and hyperscalers and neoclouds illustrate.[113] Second is a broad base of enterprises and public bodies that want to integrate AI into existing processes through APIs or managed platforms, generating mostly inference and light fine-tuning loads that are more variable and latency sensitive, and that often prefer to source from

110 McKinsey & Company (2025). The AI Infrastructure of the Future. https://www.mckinsey.com/capabilities/tech-and-ai/our-insights/the-ai-infrastructure-of-the-future.

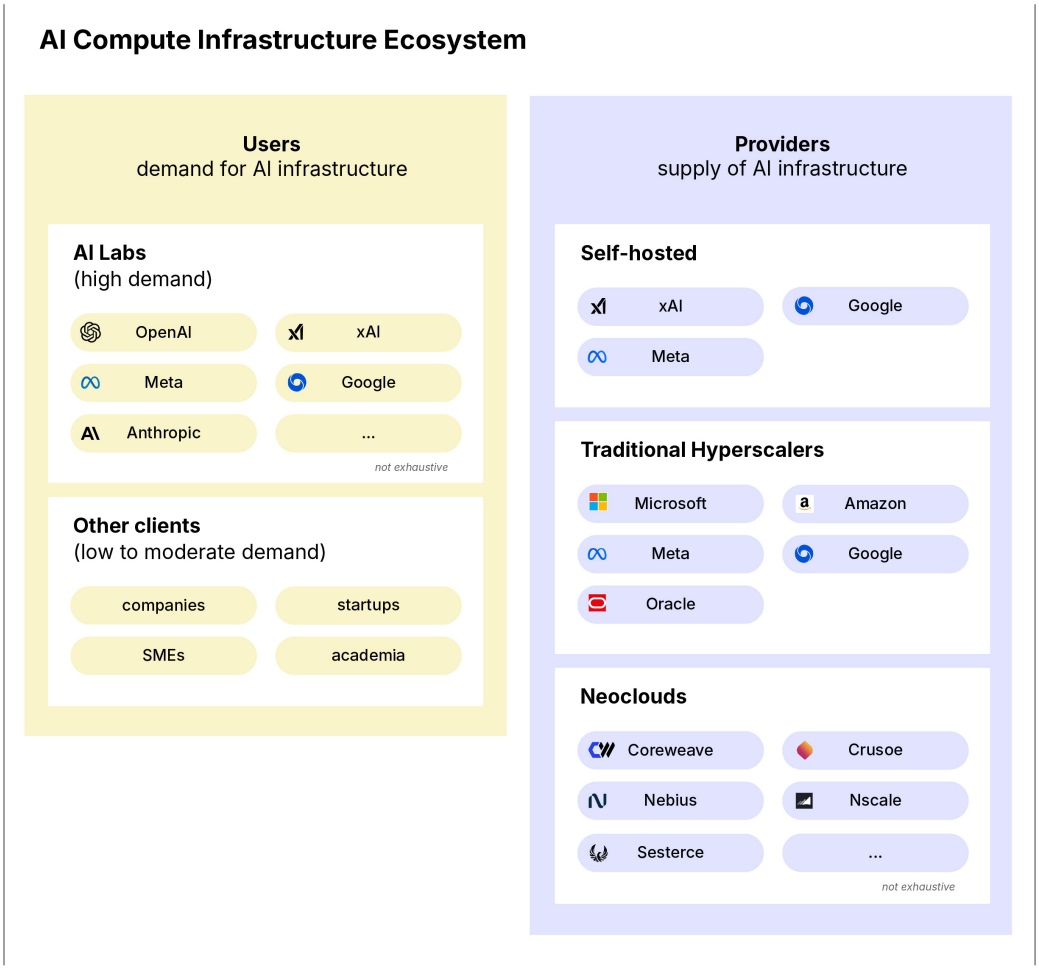
111 Netboxlabs. CoreWeave: Accelerating AI Infrastructure with NetBox Cloud. https://netboxlabs.com/customer-stories/coreweave/.

112 Elisabeth Cremona & Pawel Czyzak (2025). Id.

113 Blake Crosley (2026). xAI Colossus Hits 2 GW: 555,000 GPUs, $18B, Largest AI Site. https://introl.com/blog/xai-colossus-2-gigawatt-expansion-555k-gpus-january-2026.

hyperscalers' global clouds rather than from single sites.[114] Third, a growing startup and research ecosystem in Europe requires AI compute for experimentation and mid-scale training but tends to be price sensitive and intermittent in its demand, making it difficult to underwrite multi-hundred–megawatt facilities on its own.[115]

Figure 10: AI Compute Infrastructure Ecosystem



Source: https://www.interface-eu.org/publications/ai-gigafactories

These provider–user combinations matter for the energy system because they lead to different demand structures for the same physical hardware. An AI cluster built around an AI lab as an anchor customer running continuous inference presents itself to the grid as a relatively steady, high-utilisation industrial load. The same AI compute capacity offered as a shared service to hundreds of smaller tenants—each with spiky, generous workloads

114 Dave Davies (2026). AI Data Centers: Key Differences, Benefits, and Trends. https://wandb.ai/wandb_fc/genai-research/reports/AI-data-centers-Key-differences-benefits-and-trends--VmlldzoxNjMyMjQ0NA.

115 Julia Christina Hess & Felix Sieker (2025). Built for Purpose? Demand-Led Scenarios for Europe's AI Gigafactories. https://www.interface-eu.org/index.php/publications/ai-gigafactories.

and contractual latency guarantees—produces lower average utilisation, sharper short-term swings and less scope for deliberate curtailment.[116] In other words, who books the capacity and how they use it determines whether an AIGF behaves like a cluster of bursty, always-on micro-services.

This is where the AIGF concept and European industrial reality diverge. As defined in the AI Continent Action Plan, AIGFs are intended as large, partly publicly co-funded GPU clusters of at least 100,000 advanced accelerators for training and deploying frontier-scale models.[117] In the updated EuroHPC Regulation, this has been formalised as 'state-of-the-art large-scale facilities' with the capacity to support the complete life cycle of very large AI models—from training through fine-tuning to large-scale inference—and offer priority access for European public and industrial and research users.[118] In practice, however, Europe currently hosts only one AI lab operating at the frontier training scale, Mistral, which has already partnered with the neocloud provider Fluidstack for a gigawatt-scale GPU cluster in France. There is no public record on whether Mistral is also interested in participating in an AIGF, although the company is simultaneously developing its own large-scale compute platform (Mistral Compute).[119] Most of the remaining European demand comes from small and medium enterprises (SMEs) and public institutions that will never fill an AIGF on their own and from research and startup projects with episodic needs.[120] This configuration makes a US-style anchor customer model structurally unlikely and pushes AIGFs toward a multi-client platform role, competing with hyperscalers and neoclouds on services and governance rather than on raw compute alone.[121]

Therefore, the rest of this section builds explicitly on previous analysis from our programme, which formalised these market dynamics into two demand-led scenarios for AIGFs.[122] The anchor customer scenario assumes one or a few large AI labs as tenants with high and continuous AI training workloads. The multi-client scenario assumes a fragmented user base of enterprises, SMEs, public institutions and research organisations with heterogeneous, largely inference-driven workloads.[123] The paper concludes that given Europe's provider landscape and user base, the anchor customer scenario is improbable and the multi-client scenario is far more realistic. This chapter takes that

116 Zhiheng Lin et al. (2025). Data-Driven Load-Forecast-Aided Microgrid for AI Data Center https://www.techrxiv.org/doi/pdf/10.36227/techrxiv.176300486.61825852/v1.

117 European Commission (2025). The AI Continent Action Plan. https://digital-strategy.ec.europa.eu/en/library/ai-continent-action-plan.

118 Paloma Villa Mateos & Elisabeth Prieto Strobl (2025). EU's AI Continent Action Plan: A Turning Point for Digital Sovereignty. https://www.telefonica.com/en/communication-room/blog/eus-ai-continent-action-plan-turning-point-digital-sovereignty/.

119 businesswire (2025). EU's AI Continent Action Plan: A Turning Point for Digital Sovereignty. https://www.businesswire.com/news/home/20250210579531/en/Fluidstack-to-Build-1GW-AI-Supercomputer-in-France.

120 Julia Christina Hess & Felix Sieker (2025). Built for Purpose? Demand-Led Scenarios for Europe's AI Gigafactories. https://www.interface-eu.org/index.php/publications/ai-gigafactories.

121 Maximilian Henning (2026). EU Recalibrates the Case for AI Gigafactories. https://www.euractiv.com/news/eu-recalibrates-the-case-for-ai-gigafactories/?utm_source=euractiv&utm_medium=newsletter&utm_content=Pro+articles+(2+articles)&utm_term=0-0&ut

122 Julia Christina Hess & Felix Sieker (2025). Id.

123 Ibid.

conclusion as its starting point. The following subsections show how a multi-client AIGF structure sharpens three bottlenecks—the utilisation gap and its energy costs, load variability and system-level impacts, and scheduling complexity—and contrast these outcomes with the anchor customer model to illustrate what the same GPU hardware would look like under a different use case.

Bottleneck 1 – Utilisation gap and its energy cost

Chapter 2 established that utilisation is absent from the EU's current monitoring framework. The Commission's own technical report on data centre performance confirms the gap explicitly: Compute utilisation metrics for AI and accelerated computing are 'not included at this stage' in the proposed minimum performance standards.[124] Section 3.1 establishes that this gap exists and why it matters. It leaves the questions of why the multi-client scenario makes it particularly consequential and what it translates to at the scale of infrastructure currently being planned.

The 35–45% GPU utilisation figure that DigitalEurope cites for typical multi-client AI deployments reflects the structural reality of serving many clients with heterogeneous, latency-sensitive workloads on shared infrastructure.[125] Peak provisioning is unavoidable regardless of how well the facility is managed. The hardware provided for those peaks continues drawing power through the quiet periods. Independent grid data corroborates the scale of the gap. Ember's analysis of European TSO data revealed that data centres use, on average, only 44% of their contracted grid capacity, using 34% of it in Ireland and 30% in Norway. These figures cover conventional data centres broadly but establish a baseline. Even before accounting for the additional variability introduced by mixed AI workloads, the gap between nameplate capacity and actual draw is already substantial.

The multi-client scenario makes this even more complex. Research on multi-client inference cluster design shows that allocating separate GPU pools to different workload types is the natural response to serving clients with different latency requirements. However, this approach increases total cluster energy consumption by 20% compared with a consolidated architecture because each pool carries its own idle overhead.[126] An anchor customer scenario running a single large workload avoids this extra energy consumption entirely.

At the scale of a 100k GPU facility, the numbers are striking. Facilities such as AIGFs draw approximately 150 MW at full GPU load, increasing to more than 200 MW with cooling, networking and power conversion overhead.[127] At 40% GPU utilisation, the

124 Simon Hinterholzer et al. (2025). Id.
125 DigitalEurope (2025). Id.
126 Jovan Stojkovic et al. (2024). DynamoLLM: Designing LLM Inference Clusters for Performance and Energy Efficiency. https://arxiv.org/abs/2408.00741.

facility is consuming close to its full power envelope while producing less than half the useful AI output that full utilisation would deliver.

Bottleneck 2 – Load variability and its grid system cost

Section 3.1 introduces the point that inference loads are bursty and periodic. What this means for the grid system is worth developing further because the consequences extend beyond the AIGF itself to costs imposed on other electricity users.

When a large facility connects to the grid, it applies for a maximum import capacity, that is, a contracted level that the grid operator must keep available at all times. Planning the transmission infrastructure around that contracted level means building for the worst case. The grid operator has built and must maintain infrastructure for loads that never materialises. The capital cost of that overbuilt infrastructure is socialised across other electricity users through network tariffs. The Eurelectric analysis of European distribution operators documents that maximum import capacity contracts of data centres have led directly to capacity being blocked for other users, including renewable generators, in already constrained areas.[128]

For AI clusters running inference-heavy workloads, the problem goes beyond average underutilisation. As shown in Chapter 2, real-time inference produces highly bursty load profiles, with short, sharp demand spikes on top of a lower baseline. Research on AI data centre grid impacts shows that large-scale GPU clusters can produce power fluctuations of hundreds of megawatts within seconds, driven by the non-linear nature of real-time inference requests.[129] Grid operators must hold an aFRR reserve, generation capacity kept online but idle and ready to respond within seconds to cover unexpected demand swings. The more variable the load, the more reserve the system must maintain at cost to all users. A multi-client AIGF producing sub-second power swings of this magnitude is therefore actively increasing the system's balancing costs, not merely underutilising its connection.

Short-term load forecasting adds a further layer of difficulty. Inference-dominated clusters are structurally more difficult to forecast than training-dominated ones because the underlying drivers (user behaviour, query complexity and concurrent tenant activity) are more stochastic than the controlled cadence of a training run.[130] In practice, largest AI clusters operate with a shifting mix of training and inference workloads rather than a

127 Chenxu Niu et al. (2025). TokenPowerBench: Benchmarking the Power Consumption of LLM Inference. TokenPowerBench: Benchmarking the Power Consumption of LLM Inference. https://arxiv.org/html/2512.03024v1.

128 Eurelectric (2025). From Backlog to Breakthrough: Managing Connection Queues in Distribution Networks. https://www.eurelectric.org/wp-content/uploads/2025/04/From-Backlog-to-Breakthrough-Managing-Connection-Queues-in-Distribution-Networks.cleaned.pdf.

129 Xin Chen et al. (2025). Electricity Demand and Grid Impacts of AI Data Centers: Challenges and Prospects. https://arxiv.org/html/2509.07218v4.

130 Mariam Mughees et al. (2025). Short-Term Load Forecasting for AI-Data Center. https://arxiv.org/html/2503.07756v1.

clean separation between the two, which further complicates forecasting. Grid operators managing a region with multiple AI facilities would not be able to accurately predict aggregate demand even hours ahead, complicating the scheduling of dispatchable generation and reserve procurement.

The demand response picture only adds to the asymmetry. Training workloads, as the anchor customer benchmark illustrates, can in principle be shifted, paused or curtailed during grid stress events, and research indicates that this flexibility is worth 17–54% in reserve service cost savings.[131] Inference serving real-time applications cannot be deferred in the same way. For example, in a multi-client scenario with inference focus, it would be beneficial if the AIGF remains continuously available to serve live requests and cannot offer the flexibility that would make it a useful grid partner or benefit from the faster grid connections and lower tariffs that flexible connection agreements make possible.[132] IEA analysis revealed that if European data centres offered just 30 hours of flexibility annually, their available grid capacity could more than double.[133] A multi-client AIGF is structurally constrained from providing that flexibility occupies a grid connection that would deliver substantially more system value in other hands, and AIGF designation criteria currently have no mechanism to account for this.

Bottleneck 3 – Scheduling complexity and the ‘self-hosting trap’

The multi-client scenario’s energy complexity is not only due to the diversity of workloads but also how those workloads are allocated across the GPUs. Workload scheduling is the mechanism by which a cluster decides which tenant gets which GPUs, when and for how long. In a single-tenant training cluster, scheduling is relatively straightforward. One workload occupies the cluster for an extended period, and the scheduler’s job is primarily to maximise throughput. In a multi-client AIGF serving heterogeneous tenants simultaneously, scheduling becomes a continuous optimisation problem with significant energy consequences.

The core tension is between latency and throughput. Serving real-time inference queries requires small batch sizes, processing requests quickly as they arrive, even if only a handful of GPUs are active at any moment. This leads to poor GPU utilisation for structural reasons, as modern AI accelerators are designed to deliver peak performance when executing large, dense matrix operations across hundreds of parallel threads. When a GPU processes a batch of one or two requests, the vast majority of its compute units sit idle, including its hardware, which is physically present and powered but has no work to

131 Fatih Acun et al. (2026). Id.

132 Elisabeth Cremona & Pawel Czyzak (2025). Id.

133 Ibid.

execute. The GPU draws near-peak power regardless of how little computation it is performing, so energy is consumed without proportional useful output. Large batch sizes keep GPUs busy processing many requests together but introduce latency that degrades the user experience for interactive applications.[134]

For a multi-client AIGF to serve both real-time inference and batch workloads simultaneously, the scheduler must continuously balance these competing demands. The natural response allocating separate GPU pools to latency-sensitive and latency-insensitive workloads produces systematic underutilisation because each pool is provided for its own peak demand and cannot absorb the other's idle capacity. Research on the inference and deployment of large AI models shows that this siloing leads to 'significant under-utilisation of expensive accelerators due to load mismatch'.[135]

The self-hosting arrangements embedded in the AIGF governance structure also introduce a governance constraint on scheduling flexibility. Under the EuroHPC Regulation, the Union's financial contribution, capped at 17% of the capital expenditure, entitles it to proportional access, with priority for public law entities, industrial users on EU-funded projects and SMEs.[136] Consortium members who contribute capital similarly expect reserved capacity in return. Indeed, it is technically possible to dynamically reallocate idle reserved capacity to other tenants, as modern AI cluster management systems can automatically assign other workloads to GPUs that are temporarily idle, and production deployments have demonstrated that dynamic sharing can increase utilisation from less than 30% to more than 70%.[137] However, the stronger the access guarantees owed to capacity holders, the less freedom the scheduler has to fill their idle periods with other workloads; we call this 'the self-hosting trap'. Public law entities expecting guaranteed on-demand access cannot accept their reserved capacity being occupied by other jobs at the moment they submit their own workloads. Therefore, the self-hosting trap is a governance constraint rather than a technical one. The priority and availability guarantees that justify public co-financing structurally might limit the scheduling optimisation that could otherwise reduce idle time.

The scheduling systems themselves represent further dependency. Large AI cluster schedulers, such as Microsoft's Singularity, Google's and Meta's equivalent internal systems, are developed by a small number of hyperscalers and are not designed with European public interest objectives in mind. A European AIGF that licences or adapts one of these systems inherits its optimisation logic, transparency limitations and vendor

134 Jovan Stojkovic et al. (2025). TAPAS: Thermal- and Power-Aware Scheduling for LLM Inference in Cloud Platforms. https://arxiv.org/abs/2501.02600.

135 Shashwat Jaiswal et al. (2025). SageServe: Optimizing LLM Serving on Cloud Data Centers with Forecast Aware Auto-Scaling. https://arxiv.org/abs/2502.14617.

136 Zuzanna Warso (2026). Who Controls Europe's AI Future. https://openfuture.eu/blog/who-controls-europes-ai-future/.

137 Yihao Zhao et al. (2023). MuxFlow: Efficient and Safe GPU Sharing in Large-Scale Production Deep Learning Clusters. https://arxiv.org/abs/2303.13803.

dependencies. Research on AI workload scheduling specifically in the European context notes that GPU resource pools cannot be efficiently managed across HPC and cloud stacks without integrated scheduling and that the absence of interoperable scheduling infrastructure is a source of systematic resource waste.[138]

Table 2: Comparison of energy characteristics – anchor-customer and multi-client scenario

	Anchor customer	Multi-client
GPU utilisation	Higher: concentrated demand from large AI labs that sustains high utilisation	Lower: fragmented demand across many clients with heterogenous workloads
Load profile	Sustained high, more predictable	Bursty, periodic, high short-term variability
Grid flexibility	Low: Large training workloads can be shifted or deferred (but barriers are high), with an ability to shift between training and inference workloads	Lower: real-time serving across many clients that cannot be curtailed
Grid system costs	Lower: predictable, more flexible load	Higher: variable load that raises balancing costs
Scheduling complexity	Lower: fewer tenants, longer-running workloads	Higher: latency/throughput tension across heterogenous tenants
Self-hosting obligations	Lower: few capacity reservation obligations	Higher: reserved capacity for many stakeholders that cannot be reassigned when idle

The anchor customer model performs better on every energy dimension. A single AI workload or portfolio of workloads that keeps GPUs highly utilised over time (e.g. through sustained training campaigns followed by large-scale inference) produces relatively predictable load profiles and makes it easier to align siting decisions with abundant low-carbon supply.[139] However, Hess and Sieker's analysis revealed that this model is highly improbable for Europe's AIGFs because the continent lacks multiple European frontier-scale labs willing to act as long-term anchor tenants and because the one AI lab operating at that scale, Mistral, has already secured its own large-scale infrastructure through a partnership with neocloud Fluidstack.[140] The multi-client model is what Europe is actually building. It aggregates many smaller tenants and mixes training and latency-sensitive inference. The planned size of the AIGFs present very large,

138 Pedro Garcia Lopez et al. (2025). AI Factories: It's time to rethink the Cloud-HPC Divide. https://arxiv.org/pdf/2509.12849.
139 Fatih Acun et al. (2026). Id.
140 Julia Christina Hess & Felix Sieker (2025). Id.

centralised clusters, concentrating a more difficult energy profile, the most variable and costly grid-integration challenges and the most complex scheduling environment.

An important implication of this comparison is that cluster size and geography are not neutral design choices. Smaller inference-oriented clusters closer to users and sized to local grid capacity could, in principle, ease connection bottlenecks and reduce the need for over-built transmission, especially if paired with dedicated, efficient inference hardware.[141] By contrast, very large multi-client AIGFs that combine heterogeneous workloads in a single site maximise the risk of underutilised capacity, volatile load and stranded grid investments. In practice, the perceived prestige, inter-state competition and the goal of broad geographical representation are likely to weigh heavily on siting decisions, which found its confirmation in the overwhelming response to the initial AIGF call for interest launched by the Commission in 2025, resulting in 76 expressions of interest across 60 proposed sites in 16 member states.[142]

In light of these findings, building large AI compute clusters, such as AIGFs, does not appear to be a sufficient or even obviously legitimate policy objective on its own. Our demand-side work shows that Europe's likely user base will not turn AIGFs into anchor customer facilities hosting frontier LLMs, while the analysis in this chapter shows that multi-client configurations are prone to chronic underutilisation, higher system costs and limited flexibility if they are not tightly governed. Constructing multi-hundred–megawatt facilities that fail to use their contracted capacity effectively would be unsustainable not only economically but also from an energy-and climate-system perspective, especially if they are not matched with substantial additional renewable supply.[143] Therefore, the task for policy is to define much more concrete goals and governance conditions, including utilisation and flexibility requirements, siting rules linked to renewable build-out and clear limits on cluster size and purpose. These factors form the basis for making any AIGF compatible with Europe's energy and climate objectives. The sections that follow examine what such criteria could contain.

Towards concrete criteria for AI infrastructure in Europe

The analysis in the previous sections points towards specific and actionable additions to the ongoing policy discourse around AI compute infrastructure expansion, including the upcoming AIGF projects.

141 Terakraft (2024). Data Center Design Requirements for AI Workloads. A Comprehensive Guide. https://www.terakraft.no/post/datacenter-design-requirements-for-ai-workloads-a-comprenshive-guide.

142 European Commission (2025). Overwhelming response as 76 respondents express interest in the European AI Gigafactories initiative. https://digital-strategy.ec.europa.eu/en/news/overwhelming-response-76-respondents-express-interest-european-ai-gigafactories-initiative.

143 Elisabeth Cremona & Pawel Czyzak (2025). Id.

The following recommendations draw on the conclusions from our analysis and are proposed as input to the Commission's ongoing work on AIGF criteria, which are currently under development in parallel with the EED data centre rating scheme, the plans to triple data centre capacity by 2030, the Cloud and AI Development Act and further policy instruments targeting the nexus between energy infrastructure planning and management and AI compute capacity expansion.

On the basis of the technical analysis in Chapter 1, we recommend that policymakers take the following actions:

Differentiate AI clusters from traditional data centres in policy instruments

At both the EU and member state levels. Given the structural differences documented in our analysis, such as GPU-focused compute, 30–100 kW AI racks versus 3–15 kW traditional racks, liquid-cooling dependence and denser interconnects, the upcoming European data centre rating scheme and related policies, should not treat all facilities as a single category. Instead, they should explicitly distinguish AI clusters from conventional data centres. This distinction could be reflected in separate reporting tracks, thresholds and design obligations by, for example, applying tighter utilisation, flexibility and siting requirements only to facilities whose primary purpose is AI training and high-volume inference, while continuing the use of the existing EED framework for mixed-workload traditional data centres.

On the basis of the findings from Chapter 2, we recommend that policymakers take the following actions:

Affirm the EED baseline and consider flexibility requirements.

AIGF criteria should adopt the Energy Efficiency Directive's Minimum Performance Standards (PUE < 1.3 from 2027; 100% renewable by 2030) as a baseline. These should be treated as minimum common denominator and integrated with the EED rating framework rather than duplicated. Modest flexibility requirements, defined through the demand response capacity to be made available annually, might also be considered given their proven grid value.

Require GPU utilisation disclosure and set minimum utilisation targets.

Stakeholders may wish to examine the disclosure of GPU utilisation from the outset of commercial operations. Regular (e.g. quarterly) reporting based on metrics such as TFLOPS per MWh could help link compute performance with energy intensity and fill current monitoring gaps while reflecting the episodic nature of AI workloads more accurately than annual data. Over time, these disclosures could serve as a point for further alignment of utilisation targets. A threshold in the 50–55% range could be a

pragmatic starting point, as it sits above the 30–50% utilisation commonly observed in GPU clusters but remains below best-case engineered workloads. Third-party verification is essential to ensure integrity and comparability across facilities.

Integrate grid compatibility assessments into AI infrastructure expansion strategies.

Grid compatibility assessments could become an important component of designation processes. Evaluating peak and baseload profiles, proximity to infrastructure and potential for demand response participation may help ensure that facilities are sited and designed with system resilience in mind. Examples from national siting policies in France, where preferred zones and fast-track grid connection procedures for large data centres are being developed, and the UK, where national planning policy now requires local plans to identify suitable data centre locations and links grid capacity to AI Growth Zones, illustrate how spatial planning can support such alignment.[144]

On the basis of the considerations presented in Chapter 3, we suggest the following policy instruments:

Enhance operational transparency through scheduling methodology disclosure.

Transparency in workload scheduling approaches could also enhance accountability. Requiring limited disclosure—for example, how GPUs are allocated between real-time inference and external customers, internal model-development jobs for anchor customers and reserved capacity for public sector or research users—would make it easier to verify that reported utilisation metrics reflect genuine GPU activity rather than accounting artefacts. Such disclosure, focused on high-level rules rather than proprietary algorithms, would reduce the risk of misreporting while improving the understanding of how efficiently different user groups are sharing the same hardware. This requirement closes the most obvious vector for gaming the utilisation metric.

– The Commission could additionally contract further work on interoperability standards for AI cluster scheduling to reduce the risk that European AIGFs become permanently dependent on proprietary scheduling systems from non-European providers.

Finally, we believe that building on the findings of this study, a structural need arises for:

---

144 Simon Cudennec & Sandra Hahn Duraffourg (2025). Building Data Centers in France: Navigating Regulatory Hurdles and Unlocking Growth. https://natlawreview.com/article/building-data-centers-france-navigating-regulatory-hurdles-and-unlocking-growth.

Stronger coordination across policymakers who share responsibility for AI-related infrastructure.

European Commission's DG ENER and DG CNECT provide one salient example: Without closer cooperation between energy and digital policy communities and without the systematic involvements of transmission and distribution system operators, the recommendations cannot be implemented effectively because institutional fragmentation keeps the relevant policy processes separate.

– A standing joint working group on AI infrastructure that brings together DG ENER, DG CNECT and TSOs/DSOs might help solve overlaps, harmonise policy strategies and embed AI compute clusters within a coherent EU-level accountability framework.

– At a minimum, an AIGF designation should trigger enhanced EED reporting obligations automatically, treating the two frameworks as complementary layers of a single system rather than separate instruments with separate logics.

Conclusion

This study has shown that large AI compute clusters, including prospective AIGFs, constitute a distinct class of energy-intensive infrastructure, fundamentally different from traditional data centres. Traditional facilities are built around CPU-based servers in 3–15 kW racks, cooled mainly by room-level air systems and serving diversified, partly idle workloads, whereas large AI clusters concentrate thousands of accelerators, drawing 300–1,200 W per chip into 30–100 kW racks with tightly coupled liquid cooling, networking and power delivery subsystems and load patterns that resemble continuous industrial processes more than conventional IT.

We demonstrated how these architectural differences translate into energy constraints. Power is concentrated first at the rack, where accelerator density, fabric power and cooling capacity jointly set hard limits. Only then is it aggregated to the facility and grid. Our analysis revealed that incremental chip-efficiency gains do not remove these rack-scale bottlenecks and that from an energy perspective, AI clusters must be treated as a separate infrastructure category rather than as a continuation of traditional data centre trends.

We then examined how different workload regimes further map this physical baseline onto the power system. Training-heavy operation drives long, high-utilisation plateaus that can, in principle, provide limited flexibility services, while multi-tenant, inference-heavy operation produces bursty, harder-to-forecast loads, structurally lower utilisation and a wider gap between contracted grid capacity and actual draw. We

introduced the idea of hybrid workloads as an energy variable. The same hardware can behave like either a relatively steady industrial plant or a spiky service platform, depending on how much training versus inference it runs, for whom and at what utilisation levels.

We finally translated these technical and operational insights into governance questions around AIGFs and other large AI clusters. Our analysis results demonstrated that Europe is the likeliest to build multi-client, inference-heavy AIGFs rather than anchor customer training clusters and that without explicit utilisation, siting and scheduling conditions, such facilities risk becoming underutilised, inflexible and heavily dependent on opaque, vendor-controlled workload scheduling and cluster management systems. In response, the chapter argued for differentiating AI clusters from traditional data centres in all relevant instruments, embedding utilisation as a core accountability metric alongside PUE, integrating grid compatibility and flexibility requirements into AIGF design and siting, and tightening governance around scheduling and access.

Taken together, this body of work supports the cross-cutting claim that headline investments in large AI clusters, including AIGFs, are not inherently sufficient for Europe's AI ambitions. Placed in the wider debate, these findings echo and deepen the concerns raised by other analysts. Our programme's previous work on AIGFs argues that Europe's demand structure makes anchor customer AIGFs improbable and that multi-client facilities face serious utilisation and governance risks.[145] This study adds that the same multi-client configurations are also the most challenging from an energy system perspective, combining very high rack-level power densities with volatile loads and limited demand-response potential.

Several other analyses raise concerns that point in a similar direction to our findings. Centre for European Policy Studies (CEPS) argues that the EU's AIGF plans underestimate grid constraints and overstate Europe's ability to host frontier-scale clusters, warning of 'white elephant' risk if utilisation and siting are not addressed.[146] A Centre for European Policy Analysis (CEPA) commentary on 'Europe's troubled bet on AI factories' stresses governance and demand-side uncertainties, noting that the EuroHPC AI factories could lock in expensive infrastructure that fails to attract sufficient workloads or renewable-aligned power contracts.[147] The German AI Association's report on the AIGF initiative also documents scepticism from industry and utilities about grid capacity, timelines and unclear responsibilities between Brussels and national authorities, calling for a rethink of the current model.[148] Environmental groups and energy-focused

145 Julia Hess & Felix Sieker (2025). Id.
146 Nicoleta Kyosovska & Andrea Renda (2025). Id.
147 Anda Bologa (2025). Europe's Troubled Bet on AI Factories. https://cepa.org/article/europes-troubled-bet-on-ai-factories/.
148 KI Bundesverband (2025). Die AI Gigafactories - Richtiger Ansatz, falsche Finanzierung? https://ki-verband.de/wp-content/uploads/2025/08/KI-Gigafactories.pdf.

non-government organisations, including Beyond Fossil Fuels and Ember, similarly caution that uncontrolled growth of large data centre projects could undermine Europe's decarbonisation pathway, especially where grid capacity and low-carbon generation are already tight.[149] However, this notion is not confined to grid-constrained markets. France, which had a surplus grid capacity originally planned for transport and heating, is finding that AI clusters have been faster to claim, displacing the very sectors the headroom was intended to serve.[150]

The analysis in this study is deliberately narrow in one respect, as it focuses solely on electricity use and power system integration and not on the full environmental footprint of AI infrastructure. Many local stresses that already appear and are likely to intensify around the energy supplies in data centre regions (within and beyond FLAP-D) also manifest in other domains, including water consumption, greenhouse gas emissions, air quality and noise pollution.

Recent work on the carbon and water footprints of data centres has begun to quantify how AI-driven growth in electricity demand interacts with power sector emissions and direct and indirect water use, often in already water-stressed regions.[151] Studies from the United States have shown that large AI clusters can consume millions of litres of water per day, rival the water use of small cities and rely on electricity mixes that keep their operational emissions well above national averages.[152] [153] Parallel research has documented heightened air pollution burdens and persistent low-frequency noise for communities living near data centre clusters, driven by backup generators, onsite power plants and cooling equipment, prompting local pushback and new zoning and permitting debates.[154] [155]

These strands of work underscore that energy is only one dimension of the environmental and social footprints of large AI clusters. A full assessment of AIGF projects will need to combine the demand-, utilisation- and grid-focused criteria developed here with a more granular analysis of water availability and competing uses, life cycle emissions, local air-quality impacts, noise exposure and land-use change, as well as distributional

149 Beyond Fossil Fuels (2025). Majority of Europeans Polled Want Rules to Limit New Data Centres' Impacts on Energy, Water and Economy. https://beyondfossilfuels.org/2025/10/27/most-europeans-polled-want-rules-to-limit-new-data-centres-impacts-on-energy-water-and-economy/.

150 Gauthier Roussilhe (2026). AI Data Centres: Between Energy Requisition and Environmental Eclipse. https://gauthierroussilhe.com/en/articles/ai-data-centres-between-energy-requisition-and-environmental-eclipse.

151 Jon Gorey (2025). Data Drain: The Land and Water Impacts of the AI Boom. https://www.lincolninst.edu/publications/land-lines-magazine/articles/land-water-impacts-data-centers/.

152 Charlotte Jennings (2025). The Cloud Is Drying Our Rivers: Water Usage of AI Data Centers. https://ethicalgeo.org/the-cloud-is-drying-our-rivers-water-usage-of-ai-data-centers/.

153 Sentinel Earth (2026). Data Center Water Consumption: AI Uses More Water Than Entire Cities. https://www.sentinelearth.com/post/data-centers-now-drink-more-water-than-entire-cities.

154 Harvard School of Public Health (2026). Analyzing Air Pollution Health, Economic Risks from AI Data Centers. https://hsph.harvard.edu/news/analyzing-air-pollution-health-economic-risks-from-ai-data-centers/.

155 Environmental and Energy Study Institute (2026). Communities Are Raising Noise Pollution Concerns About Data Centers. https://www.eesi.org/articles/view/communities-are-raising-noise-pollution-concernsabout-data-centers.

questions about who bears these burdens and who benefits from the resulting AI capacity. Integrating these dimensions into future work and into the design of EU-level governance frameworks is essential if Europe's AI infrastructure expansion is to remain within planetary and social boundaries.

However, within the narrower focus on electricity and grid integration, the analysis in this study points to a clear conclusion. The long-term value and acceptability of large AI compute clusters will depend on whether they are conceived, regulated and operated as critical energy infrastructure distinct from traditional data centres, aligned with decarbonisation and grid constraints and accountable for how fully and flexibly they use the power they demand. Therefore, decisions made in the AIGF process over the next few months and years will help shape how Europe balances its AI ambitions and energy priorities in a tightening geopolitical and climate context.

Glossary

Automatic frequency restoration reserve (aFRR)	Automatically activated reserve power that grid operators use to correct lasting frequency deviations and rebalance supply and demand within a few minutes.
AI accelerator	Specialised processor optimised for matrix and tensor operations used in modern AI, such as GPUs, TPUs and inference-specific ASICs, delivering far higher throughput and energy efficiency than general-purpose CPUs for training and inference workloads.
AI compute cluster/ large AI compute cluster	A tightly coupled group of interconnected servers with AI accelerators, networking, storage and cooling, operated as a single system to train and serve large AI models. In this paper, 'large AI compute cluster' typically refers to facilities with 100,000 or more advanced accelerators and power demand in the 100–300 MW range or higher.
AI Gigafactory	Large, publicly co-funded AI compute facility proposed under the AI Continent Action Plan and EuroHPC, designed to host roughly 100,000 or more advanced AI accelerators and support the full life cycle of very large AI models, from training to large-scale inference.
AI models	Mathematical and computational systems trained to recognise patterns or make decisions based on input data.
Anchor customer scenario	AIGF operating model in which one or a small number of large users underwrite a substantial share of capacity through long-term binding commitments, keeping GPUs highly utilised and producing relatively predictable load profiles over the life of the cluster.
Batch inference	Mode of running inference where a trained model processes a large set of inputs in bulk on a schedule or as a background job, prioritising throughput and energy efficiency over per-request latency.
Frontier AI models	Very large AI models with billions of parameters that are often trained with massive training compute.
Hybrid workload regime	AIGF configuration in which one or a small number of large users underwrite a substantial share of capacity through long-term binding commitments, keeping

	GPUs highly utilised and producing relatively predictable load profiles over the life of the cluster.
Hyperscalers	The largest cloud providers operating global massive-scale infrastructures (e.g. AWS, Microsoft Azure, and Google Cloud).
H100 equivalent	Normalised measure of installed AI compute capacity that expresses diverse accelerators in terms of the number of NVIDIA H100-class GPUs with comparable performance and memory bandwidth to allow comparisons of cluster size.
Idle load/baseline load	The power a data centre or AI cluster draws even when accelerators are under-utilised or idle, reflecting fixed consumption from servers and cooling, networking and power-delivery systems.
Inference	Phase in which a trained AI model is used to generate outputs from new inputs, often at lower per-request compute cost than training but with more variable, sometimes latency-constrained, load, especially when models are exposed to users or integrated into online services.
Latency-sensitive workload	Workload whose value depends on very fast responses to external requests (e.g. interactive generative AI services) and thus constrain siting and network design because compute must be placed close enough to users and networks to keep end-to-end delay within strict bounds.
Load profile	The pattern of power draw of a facility over time shaped by workload mix, scheduling and contracts.
Multi-client scenario	AIGF operating model that serves a broad set of smaller clients (enterprises, SMEs, startups, public institutions, and research organisations) with diverse, often inference-oriented workloads, requiring aggregation of demand and more complex scheduling to avoid low utilisation.
Neoclouds	A new class of cloud providers specialising in GPU capacity for AI. Many evolved from crypto-mining operators, which repurposed these facilities into GPU clusters. They offer flexible contracts and rapid deployment.
Power Usage Effectiveness (PUE)	Standard data-centre efficiency metric defined as the ratio of total facility energy use to the energy delivered to IT equipment.
Rack power density	Amount of electrical power drawn by the equipment installed in a single rack, usually expressed in kilowatts (kW) per rack. AI-oriented clusters commonly operate at 30–100 kW per rack or more, several times more than the energy required in traditional data-centre designs, with major implications for cooling and power delivery.
Training	Phase in which the parameters of an AI model are optimised on large datasets, typically using many accelerators in parallel at high utilisation for extended periods, resulting in long, near-flat, high-power load profiles.
Useful energy/ useful output	Portion of energy input that produces the desired end-use, which, in this context, directly powers AI computation on accelerators, as opposed to being lost in conversion, cooling, networking overheads or idle operation.

Acknowledgments

I would like to thank Julia Christina Hess, Manuel Sánchez, Felix Sieker, Nicoleta

Kyosovska, Gauthier Roussilhe, Jan Peter Kleinhans, Thibault Pirson, Catherine Schneider, and Nicole Lemke for their constructive feedback and support during the research and writing process; Maximilian Gottwald for his help in research and text edits; Alina Siebert for guiding the design process and publication layout, and for developing subsystem detail graphics; Luisa Seeling for her support in editing the text and Iana Pervazova and Sebastian Rieger for helping me spread the word about this publication.

Concept and Design Infographics: nach morgen

Author

Maria Nowicka
Policy Researcher Global Chip Dynamics
mnowicka@interface-eu.org

Imprint

interface – Tech analysis and policy ideas for Europe
(formerly Stiftung Neue Verantwortung)

W www.interface-eu.org
E info@interface-eu.org
T +49 (0) 30 81 45 03 78 80
F +49 (0) 30 81 45 03 78 97

interface – Tech analysis and policy ideas for Europe e.V.
c/o Publix
Hermannstraße 90
D-12051 Berlin

Design by Make Studio
www.make.studio
Code by Convoy
www.convoyinteractive.com