Trends in AI Supercomputers

Konstantin F. Pilz [*][1] **James Sanders** [2] **Robi Rahman** [2] **Lennart Heim** [2][3]

Abstract

Frontier AI development relies on powerful AI supercomputers, yet analysis of these systems is limited. We create a dataset of 500 AI supercomputers from 2019 to 2025 and analyze key trends in performance, power needs, hardware cost, ownership, and global distribution. We find that the computational performance of AI supercomputers has doubled every nine months, while hardware acquisition cost and power needs both doubled every year. The leading system in March 2025, xAI's Colossus, used 200,000 AI chips, had a hardware cost of $7B, and required 300 MW of power—as much as 250,000 households. As AI supercomputers evolved from tools for science to industrial machines, companies rapidly expanded their share of total AI supercomputer performance, while the share of governments and academia diminished. Globally, the United States accounts for about 75% of total performance in our dataset, with China in second place at 15%. If the observed trends continue, the leading AI supercomputer in 2030 will achieve 2×10^{22} 16-bit FLOP/s, use two million AI chips, have a hardware cost of $200 billion, and require 9 GW of power. Our analysis provides visibility into the AI supercomputer landscape, allowing policymakers to assess key AI trends like resource needs, ownership, and national competitiveness.

---

[1]Georgetown University [2]Epoch AI [3]Centre for the Governance of AI [*]Correspondence to Konstantin at kfp15@georgetown.edu. For inquiries about our dataset and future updates, please contact Robi at robi@epoch.ai.

Executive Summary

AI progress has relied on exponentially larger AI supercomputers. The compute used to train the most notable AI models has grown by $4.1\times$ per year since 2010, enabling breakthroughs like advanced chatbots, image generation, and protein structure prediction. This training compute growth relied primarily on larger AI supercomputers that now consist of more than 100,000 AI chips, have hardware costs of billions of dollars, and consume as much power as a medium-sized city. We compile a dataset of over 500 AI supercomputers worldwide by systematically collecting public data from 2019 to 2025. We define an AI supercomputer as a system using AI chips that achieved at least 1% of the computational performance of the leading AI supercomputer when it first became operational. We estimate our dataset captures 10-20% of all existing AI supercomputer capacity, based on comparing the total performance to public AI chip production and sales estimates.

The computational performance of leading AI supercomputers has doubled every 9 months, driven by deploying more and better AI chips (Figure 1). Two key factors drove this growth: a yearly $1.6\times$ increase in chip quantity and a $1.6\times$ annual improvement in performance per chip. While systems with more than 10,000 chips were rare in 2019, companies deployed AI supercomputers more than ten times that size in 2024, such as xAI's Colossus with 200,000 AI chips.

Computational performance of AI supercomputers

Figure 1: The performance of leading AI supercomputers (in FLOP/s, for 16-bit precision) has doubled every 9 months (a rate of $2.5\times$ per year).

Power requirements and hardware costs of leading AI supercomputers have doubled every year. Hardware cost for AI supercomputers has increased by $1.9\times$ every year, while power needs increased by $2.0\times$ annually. As a consequence, the most performant AI supercomputer as of March 2025, xAI's Colossus, had an estimated hardware cost of $7 billion (Figure 2) and required about 300 MW of power—as much as 250,000 households. Alongside the massive increase in power needs, AI supercomputers also became more energy efficient: computational performance per watt increased by $1.34\times$ annually, which was almost entirely due to the adoption of more energy-efficient chips.

If the observed trends continue, the leading AI supercomputer in June 2030 will need 2 million AI chips, have a hardware cost of $200 billion, and require 9 GW of power. Historical AI chip production growth and major capital commitments like the $500 billion Project Stargate suggest the first two requirements can likely be met. However, 9 GW of power is equivalent to 9 nuclear reactors, a scale beyond any existing industrial facility. To overcome power constraints, companies may increasingly use decentralized training approaches, which would allow them to distribute a training run across AI supercomputers in several locations.
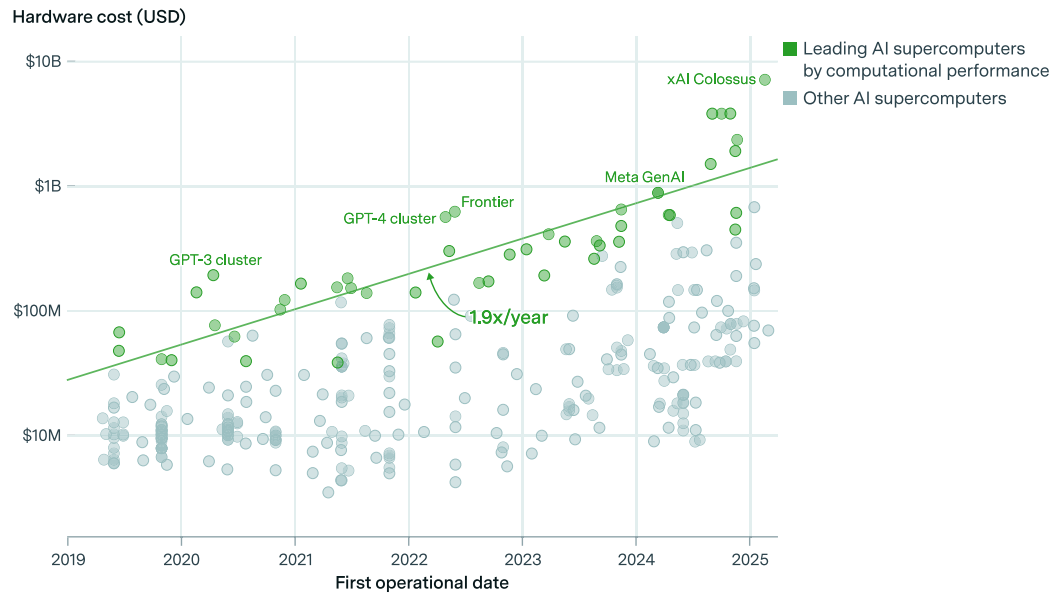
Cost of AI supercomputers

Hardware cost (USD)

Figure 2: The cost of leading AI supercomputers (in 2025 USD) has doubled roughly every year.

Companies now dominate AI supercomputers. As AI development has attracted billions in investment, companies have rapidly scaled their AI supercomputers to conduct larger training runs. This caused leading industry system performance to grow by 2.7× annually, much faster than the 1.9× annual growth of public sector systems. In addition to faster performance growth, companies also rapidly increased the total number of AI supercomputers they deployed to serve a rapidly expanding user base. Consequently, industry's share of total AI compute surged from 40% in 2019 to 80% in 2025, as the public sector's share fell below 20% (Figure 3).

Share of aggregate AI supercomputer performance by public/private sector over time
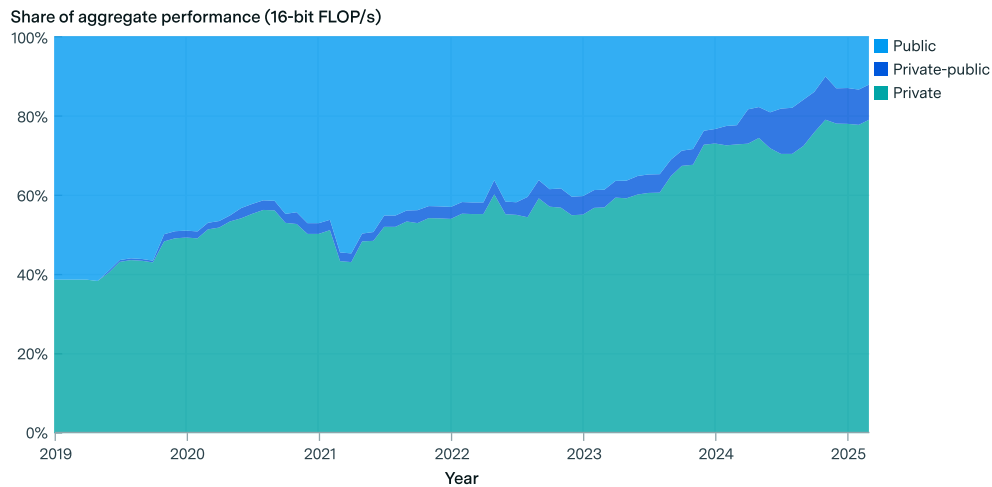
Share of aggregate performance (16-bit FLOP/s)

Figure 3: Share of aggregate AI supercomputer performance over time in the public vs private sector.

The United States hosts 75% of AI supercomputers, followed by China. The United States accounts for about three-quarters of total AI supercomputer performance, with China in second place at 15% (Figure 4). Meanwhile, traditional supercomputing powers like the UK, Germany, and Japan now play marginal roles in AI supercomputers. This shift reflects the dominance of large, U.S.-based companies in AI development and computing. However, AI supercomputer location does not necessarily determine who uses the computational resources, given that many systems in our database are available remotely, such as via cloud services.

Share of aggregate AI supercomputer performance by country over time
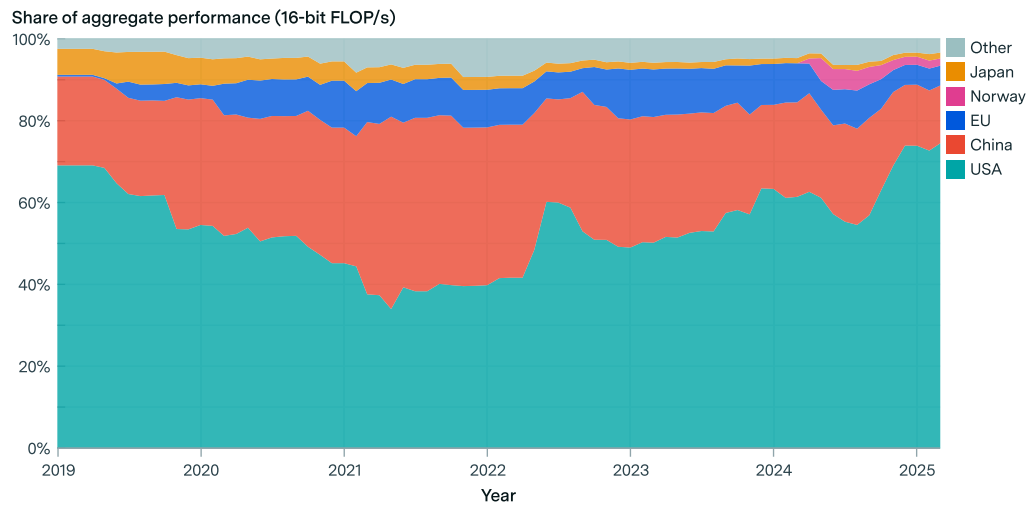
Share of aggregate performance (16-bit FLOP/s)

Our dataset covers an estimated 10–20% of global aggregate AI supercomputer performance as of March 2025. While coverage varies across companies, sectors, and hardware types due to uneven public reporting, we believe the overall distribution remains broadly representative. Future country shares may change dramatically as exponential growth continues in both AI chip performance and production volume.

Figure 4: Share of AI supercomputer computational performance by country over time. We are visualizing all countries that held more than a 3% share at some point in time.

We are releasing our dataset along with documentation soon after this publication. Our data will be part of Epoch AI's Data on AI hub and maintained with regular updates.

Contents

1. Introduction

The computing resources (compute) used to train notable AI models have increased at a rate of 4–5× per year since the beginning of the deep learning era in 2010 (Sevilla & Roldan, 2024). This exponential increase has been a major driver of improvements in AI capabilities across many domains, such as in large language models or image generation (Erdil & Besiroglu, 2022; Ho et al., 2024). Most of this increase in compute has been driven by larger, higher-performance AI supercomputers (Hobbhahn et al., 2023; Frymire, 2024).

Given their importance for AI development, systematically collecting data on AI supercomputers allows us to better understand trends such as their hardware costs, power requirements, and global distribution. This analysis is relevant to policymakers, because compute is both an enabler of AI progress and a potential tool for governance (Sastry et al., 2024; Khan & Mann, 2020). For instance, information about the distribution of AI supercomputers across countries allows governments to assess their national competitiveness in AI, and data on the growth of power requirements can help with electrical grid planning.

However, despite the importance of AI compute, no comprehensive dataset of AI-specific supercomputers exists. Resources like the Top500 list or the ML-Perf benchmark rely on voluntary submissions and thus lack sufficient data to reliably analyze trends (Top500).[1] Meanwhile, databases used for business intelligence, such as SemiAnalysis's data center model, are not available for public analysis and focus on present-day systems rather than historical trends (SemiAnalysis, 2024).

We attempt to close this gap by collecting data from various public sources and establishing a dataset of 500 AI supercomputers between 2019 and 2025. We use this to study several key trends: the growth of AI supercomputer performance, hardware costs, power consumption, and the distribution of AI supercomputing power between countries and sectors.

2. Methods

AI supercomputer definition

We define an AI supercomputer as a computer system that can support training large-scale AI models, deployed on a contiguous campus. We use two criteria to assess whether a given system can support training large-scale AI models:

1. The system contains chips that can accelerate AI workloads, such as NVIDIA's V100, A100, H100, and GB200, Google's TPUs, and other chips commonly used to train frontier AI models. To assess if a given chip is suitable for large-scale AI training, we use a dataset of machine learning hardware created by Hobbhahn et al. (2023). If a chip is not part of that dataset, we consider it an AI chip if it has the following features:

   - Support for precisions commonly used in AI training, such as FP16 or INT8.
   - Compute units dedicated for matrix multiplications, such as tensor cores in NVIDIA GPUs.
   - High-bandwidth memory (HBM) or other memory types enabling a high memory bandwidth.
   - Was used to train a model in Epoch AI (2025)'s notable AI models dataset.

2. The system has a high theoretical computational performance on AI-relevant precisions.[2] Due to the rapid pace of hardware improvements, we use a moving definition and only include systems that have at least 1% of the performance of the most performant existing AI supercomputer at that time.[3]

To balance data collection effort and representativeness, we limit the scope of our data collection to about six years, from the start of 2019 to February 2025. We will maintain the dataset at `https://epoch.ai/data/ai-supercomputers` and integrate it with Epoch AI's Data on AI hub.

Data collection

We use the Google Search API, existing compilations of (AI) supercomputers, and manual searches to collect a dataset of 501 leading AI supercomputers between 2019 and 2025. We also cover an additional 225 systems pre-2019 for a total of

---

[1]For a review of existing data sources, see Appendix A.

[2]We consider 32, 16, and 8-bit number formats as AI-relevant in our study period.

[3]Our inclusion criteria compare the system's highest performance available in 32-, 16-, or 8-bit arithmetic formats to the highest performance rate of the leading AI supercomputer at the time. Note that we exclude systems that do not support 32-bit or lower precision formats from the analysis. See Appendix B.8 for details on our approach.

726 AI supercomputers.[4] Our most significant sources are company announcements, Top500 entries with significant GPU numbers, and the Epoch AI (2025) dataset of notable AI models. For each potential AI supercomputer, we manually search for details such as the number and type of chips used by the system, when it was first operational, its reported performance, owner, and location.

We estimate our dataset covers about 10% of the aggregate performance of all AI chips produced until 2025 and about 15% of the AI chip stocks of the largest companies as of early 2025. Our dataset covers about half of all systems used in the 25 largest training runs in Epoch AI's notable models dataset as of March 2025 (Epoch AI, 2025). For a detailed analysis of our coverage, see Appendix C.1.

Analysis

We combine our collected data with Epoch AI's data on machine learning hardware to estimate the total performance[5], hardware cost, and power requirements of systems in our database (Epoch AI, 2024; Hobbhahn et al., 2023). We filter our dataset to 389 high-certainty, confirmed operational systems between 2019-01-01 and 2025-03-01. We then fit regressions for key metrics of the 57 AI supercomputers in our study period that were in the top-10 worldwide by 16-bit FLOP/s when they first became operational. The metrics we analyze include computational performance, number of chips, power requirements, energy efficiency, and hardware costs. We further assess the distribution across sectors and countries of the aggregate performance of all AI supercomputers in our dataset, including pre-2019 systems, for a total of 470 systems. Appendix B contains detailed information on our data collection, estimations for hardware cost and power, and methods for data analysis.

3. Results

We first assess growth in performance, power, and hardware cost for the leading AI supercomputers in our dataset. We then examine how AI supercomputers in our dataset are distributed across the private vs public sector, and across different countries.

3.1. Computational performance of the leading AI supercomputers has doubled every nine months

The computational performance of leading AI supercomputers increased by $2.5\times$ per year between 2019 and 2025 (Figure 5).[6] Performance increased at an even faster rate when considering only AI supercomputers owned by companies (Section 3.1.3). The rapid increase in performance resulted in the leading system in March 2025, xAI's Colossus achieving over 50 times the performance of Oak Ridge National Laboratory's Summit, the leading AI supercomputer in 2019.[7]

We found several large AI supercomputers in 2017 and 2018, significantly above the trend suggested by our post-2018 results. It is unclear to what extent this reflects a lack of coverage in our dataset or whether these genuinely were the largest deployed systems until 2021. We discuss in Section 4.1 how these early systems were primarily used for scientific research, rather than for conducting large training runs, and may not be directly comparable to later systems.

---

[4]Our dataset includes an additional 99 systems that we exclude because they are below our inclusion threshold or otherwise outside our definition. When including these excluded systems, our total is 825 entries.

[5]We define computational performance for an AI supercomputer as the advertised theoretical maximum non-sparse FLOP/s in a given numerical precision for an AI chip, summed over all AI chips in the system.

[6]In our study period AI training workloads shifted from 32-bit precision, to 16-bit precision and partially to 8-bit precision (see Appendix B.8 for an explanation of different precisions and how we handle them). We provide an overview table of all metrics for each precision in Appendix D.1.

[7]Summit achieved a performance of 3.5×10^{19} FLOP/s (16-bit precision) while Colossus achieved a performance of 2.0×10^{20} FLOP/s.

Computational performance of AI supercomputers

Figure 5: Performance of top-10 leading AI supercomputers, with performance measured in FLOP/s at 16-bit precision, has increased by 2.5× per year (90% Confidence Interval (CI): 2.4–2.7×). We start the regression in 2019, but consider pre-2019 AI supercomputers to identify which systems were in the top-10 at the start of 2019. Our pre-2019 data is too limited to include in the regression. We highlight some notable systems.

3.1.1. PERFORMANCE INCREASES RELIED ON AI SUPERCOMPUTERS USING MORE AND BETTER AI CHIPS

The annual performance increase of 2.5× resulted from two roughly equal factors: increased number of AI chips, and improved performance per chip.

First, the number of chips in the highest performing AI supercomputers increased by 1.6× annually (Figure 6). In January 2019, Oak Ridge's Summit had the highest chip count, with 27,648 NVIDIA V100s.[8] By March 2025, xAI's Colossus had the highest chip count of all known systems with 200,000 NVIDIA H100s and H200s.[9] Including pre-2019 systems in the regression would probably result in a lower growth rate. However, we cannot reliably do so because our data collection only goes back to 2019.

Second, the computational performance per chip in the most performant AI supercomputers increased by 1.6× annually. Three chip generations are notable in our study period. Between 2019 and 2021, NVIDIA's V100 was the most prominent chip, making up more than 90% of installed performance. In 2021, NVIDIA's A100 gained prominence and became the most prevalent chip by 2023, with AMD's MI250X and Google's TPU v4 making up minority shares.[10] In 2023, NVIDIA's H100 became more widespread, exceeding 50% of total performance in our dataset by July 2024.

The 1.6× (90% CI: 1.5–1.7) improvement in computational performance per chip of leading AI supercomputers is slightly faster than the general trend of AI chip performance improving 1.28× per year (90% CI: 1.24–1.32) for FP32 and 1.38× per year (90% CI: 1.28–1.48) for FP16 (Rahman, 2025; Hobbhahn et al., 2023). This difference likely stems from AI

---

[8]Summit was the system with the highest chip count within the top-10 leading AI supercomputers by performance. Tianhe-2, a Chinese system from 2013 had a higher AI chip count of 48,000, but it was not in the top-10 AI supercomputers by performance as of January 2019.

[9]Colossus phase 2 likely used 150k H100s and 50k H200s (Shilov, 2024a).

[10]Our coverage of TPUs is limited given Google exclusively uses them internally and hardly publicizes their deployment.

supercomputers primarily incorporating leading AI chips rather than average-performing ones.

AI chip quantity in AI supercomputers
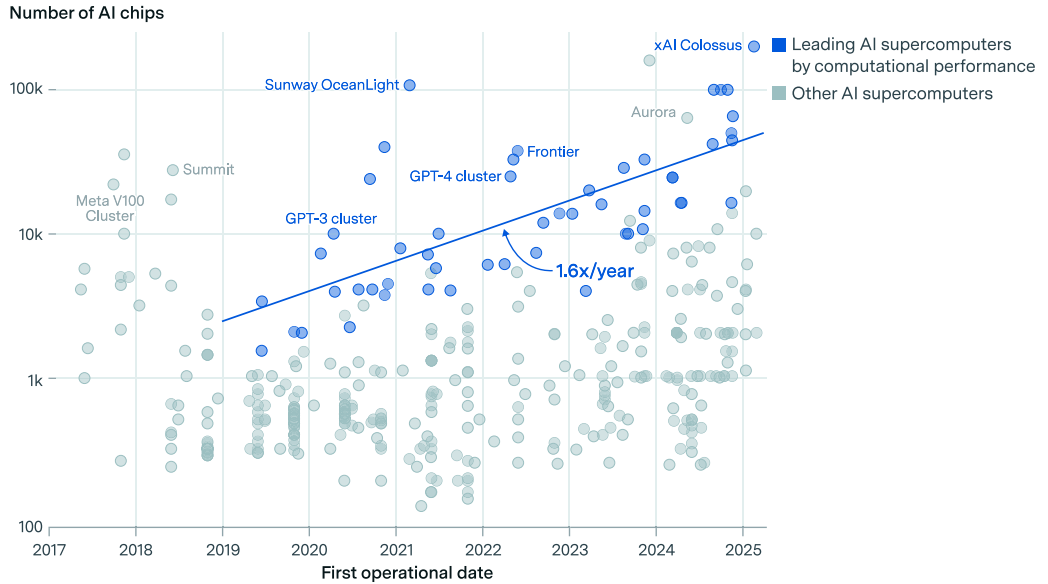
Figure 6: The number of AI chips in the leading AI supercomputers grew by $1.6\times$ per year (90% CI: 1.5–$1.8\times$). We start the regression in 2019, but gathered data further back to determine which 2019 AI supercomputers were in the top-10. Our pre-2019 data is too limited to include in the regression. See Section 2 for full methods.[11]

3.1.2. AI SUPERCOMPUTER PERFORMANCE INCREASED FASTER THAN TRADITIONAL SUPERCOMPUTERS

Benhari et al. (2024) found that the 64-bit performance of the largest Top500 supercomputer increased by $1.45\times$ per year between 1994 and 2023.[12] This growth rate makes top-10 AI supercomputers' performance increase significantly faster than the historic trend for the Top500's top machine. Two factors likely drive this divergence: AI-specific chips and faster investment growth.

First, AI chip performance has outpaced that of CPUs (Hobbhahn et al., 2023). This is because AI computing workloads have different properties than traditional computing, allowing AI chip designers to optimize performance for parallel matrix operations, which has led to AI chip performance advancing significantly faster than CPU performance (Hobbhahn et al., 2023).

Second, investment in AI supercomputers has increased more rapidly than investment in traditional supercomputers. The Top500 list was historically shaped by government-funded projects, which only slowly increased in budgets. However, our AI supercomputer dataset primarily captures systems owned by large companies, which have rapidly increased investment in AI supercomputers in the 2020s (Cottier et al., 2024).

3.1.3. AI SUPERCOMPUTERS IN PRIVATE INDUSTRY HAVE OUTPACED THOSE IN GOVERNMENT OR ACADEMIA

The performance of the leading AI supercomputers from companies grew by $2.7\times$ annually between 2019 and March 2025. Meanwhile, the performance of the leading AI supercomputers owned and funded by governments and academic institutions grew, significantly slower, by only $1.9\times$ annually (p = 0.022). The largest known public AI supercomputer, Lawrence Livermore's El Capitan, now only achieves 22% of the computational performance of the largest known industry AI supercomputer, xAI's Colossus. We discuss this shift from the public to the private sector in Section 4.4

---

[11]Note on Sunway OceanLight: This Chinese system is the first AI supercomputer to cross the 100,000 AI chip threshold. However, limited details are available, and we were unsure about whether to count the processors used as AI chips because they primarily serve as CPUs. Yet the chips support 16-bit operations and the system is part of the top-10 most performant AI supercomputers of its time (Shilov, 2023b). We thus decided to include it.

[12]They find that the doubling time is 1.87 years, which means performance increases by $2.35\times$ annually: $2^{(1/1.87)} = 1.45$.

Performance comparison of public vs private sector AI supercomputers
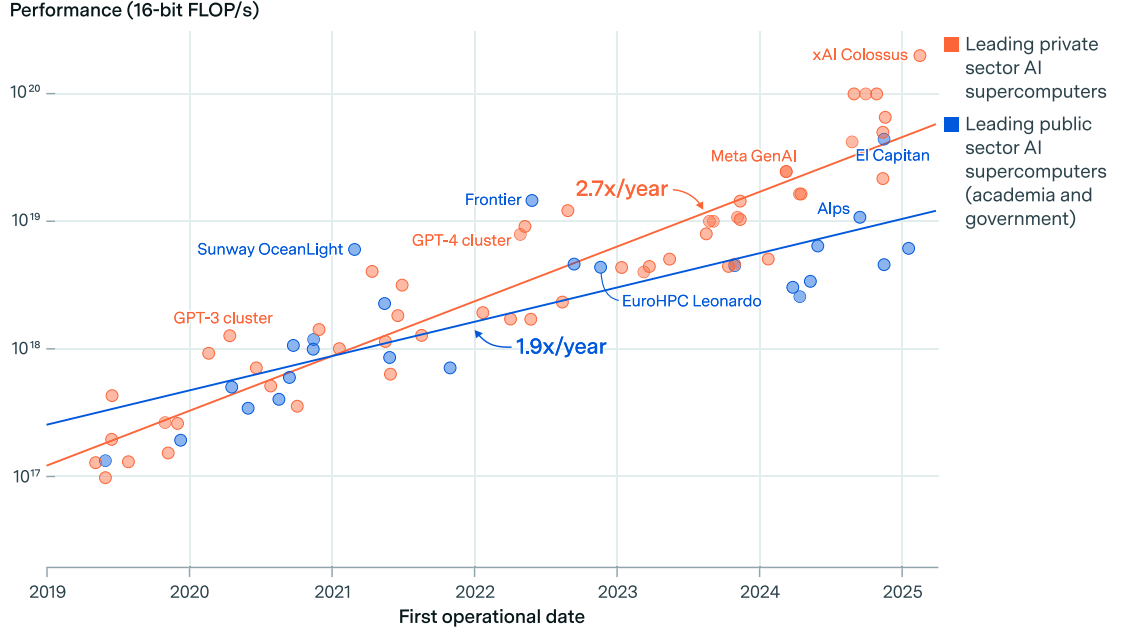
Figure 7: Performance of leading AI supercomputers owned by the private versus the public sector (government and academia). Leading public sector systems started out larger but have not kept pace with industry systems, which have grown at 2.7× annually (90% CI: 2.5–2.9×), while public sector systems have only grown at 1.9× annually (90% CI: 1.6–2.2×). Note that we exclude AI supercomputers funded and owned by both public and private institutions.

3.1.4. AI SUPERCOMPUTERS HAVE KEPT PACE WITH 4–5× ANNUAL GROWTH IN THE LARGEST TRAINING RUNS

Sevilla & Roldan (2024) found that the training compute for the largest AI models has grown by 4.2× per year (90% CI: 3.6–4.9×) between 2018 and 2024. This aligns with our observed AI supercomputer performance growth, after we account for increasing training durations.[13]

In Figure 8, we show the required computational performance for the largest AI training runs, and the performance of leading AI supercomputers in our dataset.[14] We consider only industry systems, which ran the vast majority of AI training runs (Besiroglu et al., 2024). To calculate the performance needed for training runs, we divide training compute in FLOP by the training duration in seconds, adjusted by an average performance utilization of 40% (Sevilla et al., 2022).

Between 2019 and 2025, the largest industry AI supercomputers consistently achieved 10× the computational performance required for the largest AI training runs (not including compute required for experiments before the final training run). While the systems required for the largest training runs have grown slightly faster than the leading AI supercomputers (3.4× vs 3.0×), we find no statistically significant difference in the two trends ($p = 0.18$). Hence, AI supercomputer growth has been consistent with the increase in training compute, as shown in Figure 9.

---

[13]Training durations of the top-10 largest AI training runs increased by 1.4× annually between 2019 and 2025 (Frymire, 2024).

[14]Epoch's notable AI model database reports training compute (in FLOP) independent of the precision. We thus assess the performance trend considering the highest performance across 32, 16, and 8-bit, which were the most commonly used precisions for AI training between 2019 and 2025.

Leading AI supercomputers vs performance needed for largest AI training runs
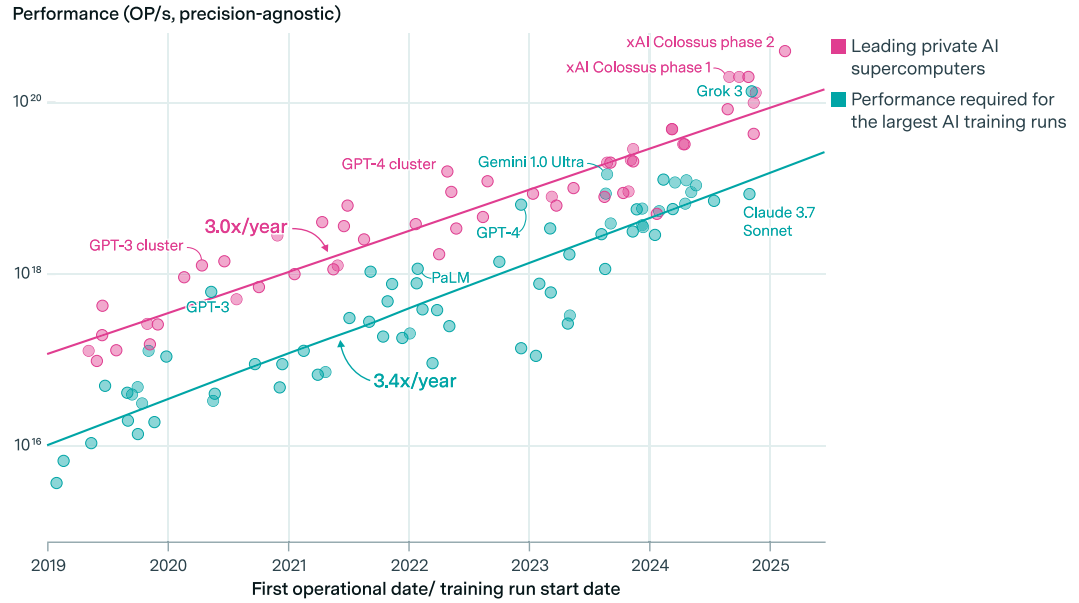
Figure 8: Computational performance of largest industry AI supercomputers and the performance required for the largest reported AI training runs (Epoch AI, 2025). To estimate the size of the AI supercomputer needed for these training runs, we assume a GPU utilization rate of 40% and use the stated training run durations when available, or an estimate from a regression on training durations of the largest AI models. We subtract the training duration from the publication date for notable models to better approximate when their training started. Given that the notable model dataset does not report the numerical precision used for training, we also report precision-agnostic OP/s for AI supercomputers, considering the highest performance available across 32, 16, and 8-bit number formats.

Breakdown of factors contributing to training compute growth
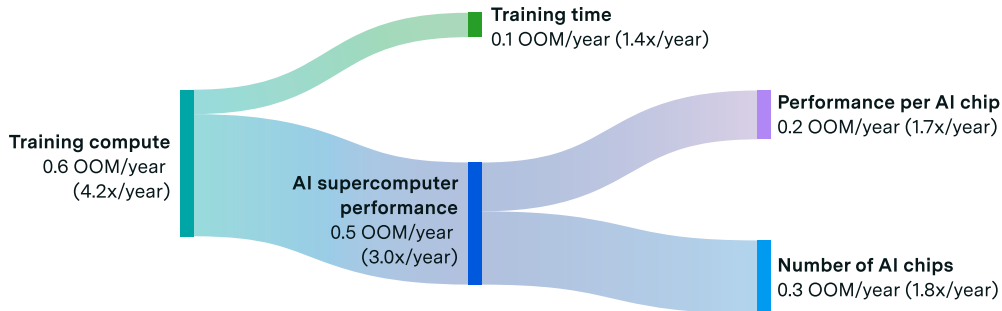
Figure 9: Overview of drivers of increasing training compute. OOM stands for orders of magnitude. AI supercomputer metrics are based on private sector systems and the highest computational performance across precisions.

3.2. Power requirements of the leading AI supercomputers doubled every 13 months

We assess the annual growth rate in power requirements of the leading AI supercomputers either based on reported power requirement or, if unavailable, by estimating the power requirement based on the number and type of AI chips, including additional IT infrastructure like CPUs, network switches, and data center supporting infrastructure like cooling and power conversion. For details on our power estimation, see Appendix B.4.

We find that the power need of the leading AI supercomputers increased by 2.0× each year between 2019 and 2025. In

January 2019, Summit at Oak Ridge National Lab had the highest power requirement with 13 MW.[15] In 2024, the first systems began to cross the 100 MW threshold and in March 2025, xAI's Colossus had the highest power requirement at an estimated 300 MW. For comparison, this is the equivalent of 250,000 U.S. households (U.S. Energy Information Administration, 2024).[16]

The rapid increase in power required for training frontier models is well documented (Fist & Datta, 2024; Sevilla et al., 2024; Pilz et al., 2025). We discuss whether this trend can continue in Section 4.2.3.

Power requirement of AI supercomputers
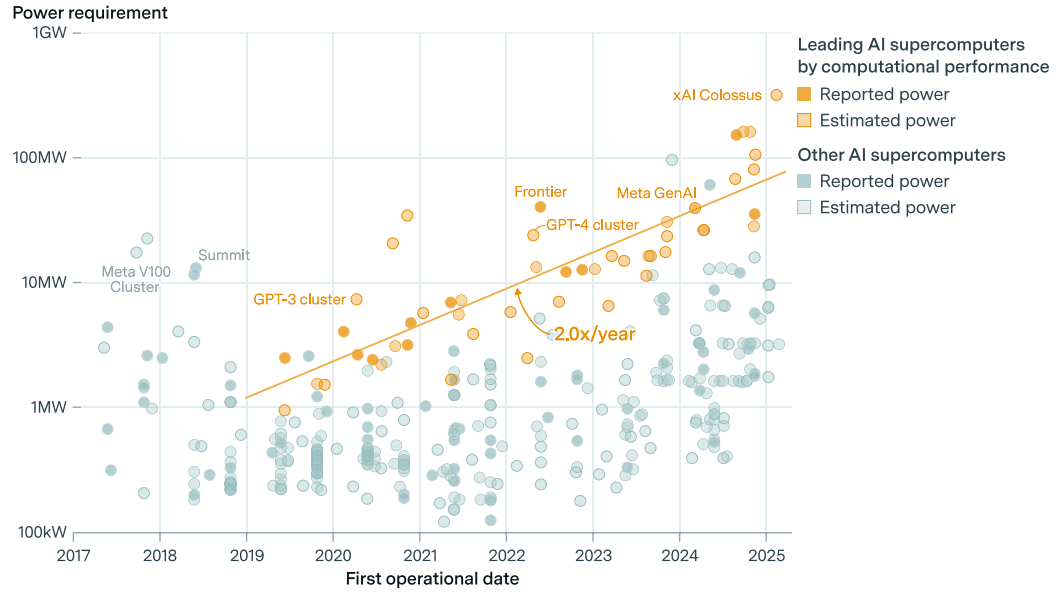
Figure 10: Peak data center power requirements of leading 10 AI supercomputers doubled every year (90% CI: 1.6–2.2× per year). We display reported power requirements whenever available. If not, we estimate capacity based on the number and type of chips used.

3.2.1. ENERGY EFFICIENCY OF THE LEADING AI SUPERCOMPUTERS IMPROVED BY $1.34\times$ PER YEAR

We calculate AI supercomputer energy efficiency in FLOP/s per watt (16-bit precision), including both hardware and data center power needs. To calculate efficiency, we divide the computational performance in FLOP/s by the reported or estimated data center power requirement in watts. Energy efficiency at the data center level includes servers, additional cluster components like networking switches, and supporting infrastructure like cooling and power conversion.

We find that between 2019 and 2025, AI supercomputer energy efficiency improved by $1.34\times$ every year (Figure 11). Holding computational performance constant, AI supercomputers required about 25% less energy per year.[17] This is roughly in line with the $1.31\times$ annual increase in energy efficiency of the most efficient supercomputers in the Top500 across the study period in Benhari et al. (2024).[18]

---

[15]Tianhe-2 had the highest power requirement in our dataset with 24 MW, but was not top-10 by performance.

[16]10,800 kWh /8760 h = 1.23 kW; 312 MW/ 1.23 kW = 250,000

[17]We are measuring energy efficiency as peak theoretical FLOP/s divided by required peak power for the AI supercomputer. This is different from energy efficiency in practice, which will be realized FLOP/s divided by average power consumption.

[18]Benhari et al. (2024) report a maximum value of 4.5×10^9 FLOP/s per watt for 2013 and 6.5×10^{10} FLOP/s per watt in 2023, implying a $1.31\times$ annual increase. Note that Benhari et al. (2024) report the energy efficiency of the most energy-efficient systems, whereas we report the energy efficiency of the most performant systems. However, the median in Figure 5 of their paper seems to track the maximum efficiency closely, implying that this trend is likely consistent throughout their data, including for the top-10 most performant systems.

Energy efficiency of AI supercomputers

Figure 11: The energy efficiency (16-bit FLOP/s per watt) of top-10 leading AI supercomputers increased by $1.34\times$ per year between 2019 and 2025 (90% CI: 1.25–$1.43\times$). Adoption of new chips was the primary driver of energy efficiency improvements, with data center infrastructure efficiency only playing a minor role. We use reported power requirement whenever available and estimated power otherwise. For detailed methods and limitations, see Appendix B.4.

Energy efficiency improvements for AI supercomputers can come from two sources: improvements in hardware efficiency and efficiency improvements in the data center infrastructure, such as cooling. Hardware efficiency improvements primarily stem from improvements in the AI chips, but also include improvements in other hardware such as CPUs, network switches, and storage.[19] We model improvements in the energy efficiency of the data center hosting the AI supercomputer by assuming they follow industry-wide trends in Power Usage Effectiveness (PUE) reported by Shehabi et al. (2024). PUE is the quotient of power supplied to hardware divided by power supplied to the data center. An ideal PUE of 1.0 would indicate that all power delivered to the data center goes directly to the hardware and no power is lost in voltage conversion or needed for cooling and other operations (Pilz & Heim, 2023).

Figure 11 shows significant improvements in energy efficiency each time a new AI chip becomes available. Meanwhile, PUE has improved more slowly and was already close to the ideal value of 1.0 in our estimate, causing efficiency improvements of less than 5% each year (Shehabi et al., 2024). Thus, energy efficiency improvements primarily resulted from AI supercomputers adopting more energy-efficient hardware.

3.3. The hardware cost of the leading AI supercomputers doubled every year

We analyze annual growth in the hardware cost for leading AI supercomputers based on either publicly reported cost figures or—if those are unavailable—by estimating the total hardware cost, based on the quantity of chips used and publicly available price data. We further include the estimated cost of additional hardware such as CPUs and network switches, but we do not model power generation or data center construction costs. We apply an inflation adjustment to all values to show costs in January 2025 dollars. Our cost estimates significantly diverge from the values reported by owners, but this could be because reported values primarily come from public projects that often get higher discounts on hardware purchases.[20]

We find that the hardware cost of the leading AI supercomputers increased by $1.9\times$ every year between 2019 and 2025. Our limited pre-2019 data indicates that hardware costs of more than $100 million were not uncommon before our study period,

---

[19]Note our estimation assumes a fixed ratio of AI chip power requirement to total IT power requirement and thus does not account for efficiency improvements in AI supercomputer components that are independent of efficiency improvements in AI chips.

[20]We estimate our hardware cost data is within $3\times$ the actual hardware cost in 90% of cases. See Appendix B.4.1 for a longer discussion of limitations and precision of our cost estimates.

with Oak Ridge National Lab's Summit costing about $200 million in 2025 USD. The most expensive AI supercomputer as of March 2025 was xAI's Colossus with an estimated hardware cost of $7 billion.

Cost of AI supercomputers
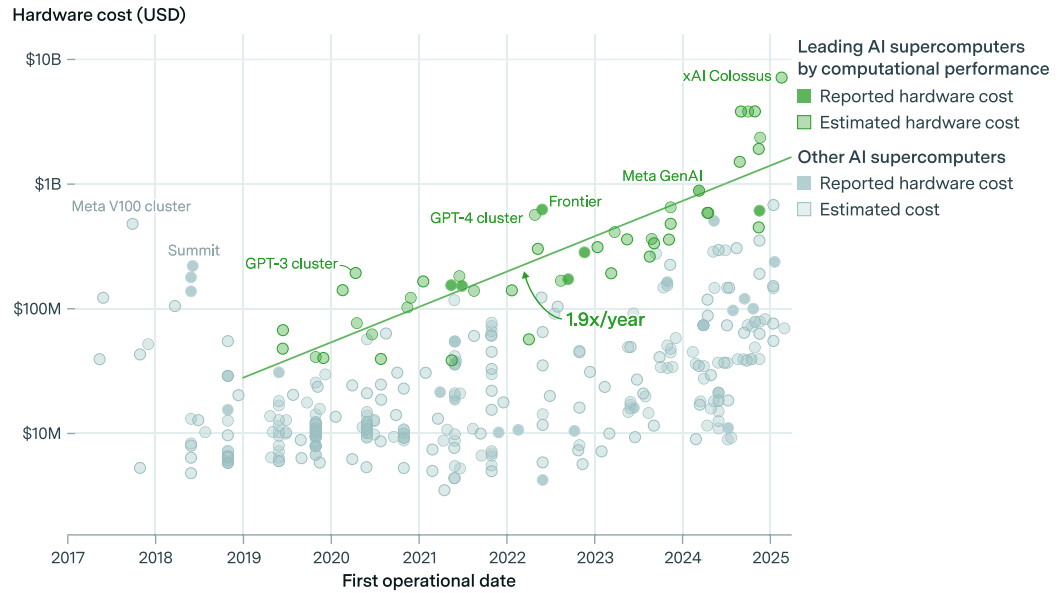
Hardware cost (USD)

Figure 12: The hardware cost of leading AI supercomputers (by 16-bit performance) when first operational has grown at a rate of 1.9× (90% CI: 1.8–2.1× per year) from 2019 to 2025. We use reported costs or, if unavailable, model costs using a hardware acquisition cost formula adapted from Cottier et al. (2024). We adjust all values for inflation to show 2025 USD. For a full explanation of the methods, see Appendix B.5

The 1.9× annual growth in hardware costs for leading AI supercomputers is slower than the 2.4× (90% CI: 2.0–2.9×) annual increase in total training costs reported by Cottier et al. (2024). This difference is due to two factors: First, training durations for frontier models have been extending by 1.4× annually (Frymire, 2024), meaning training runs use the same AI supercomputer for longer, which increases the amortized cost even if the AI supercomputer cost stays the same. Second, research personnel costs are a substantial and increasing fraction of AI development, but do not impact the hardware cost of AI supercomputers (Cottier et al., 2024).

3.4. Limitations of our data coverage

Before analyzing the distribution of AI supercomputers across sectors and countries, we emphasize two important limitations in our dataset:

a) **We only capture between 10 and 20% of all AI supercomputers that fall within our definition.** Specifically, we estimate our dataset covers about 10% of all relevant AI chips produced in 2023 and 2024 and about 15% of the chip stocks of the largest companies at the start of 2025. Our dataset covers about half of all systems used in the 25 largest training runs in Epoch AI (2025) as of March 2025. The low coverage means our data has limited precision, and a single system being added can significantly change the overall distribution.

b) **The level of coverage likely significantly varies across sectors, chip types, and companies.** For instance, we capture about half of Meta's total AI supercomputer performance while we capture none of Apple's AI supercomputers. We also likely cover government AI supercomputers much better than industry systems, since governments tend to be much more transparent about their projects.

Given these limitations, we focus on the distribution of AI supercomputers across sectors and countries because both provide reliable insights despite our low coverage: The shift in ownership from public to private sector is a large and robust effect across our entire dataset. Our country-level data is likely robust because we were able to cross-check it against other data (see Appendix C.3). Meanwhile, we do not analyze distributions across specific AI chip types or individual companies, as these would be more susceptible to the coverage biases in our dataset.

3.5. Companies now own the majority of AI supercomputers

For each AI supercomputer in our dataset, we classify the owner into one of three categories:

- Private: The owner is a company.
- Public: The owner is a government entity or a university.
- Public/Private: The AI supercomputer has several owners belonging to both sectors or if a private project received more than 25% of the total funding from a government.

We find that the share of private sector compute rapidly increased from less than 40% in 2019 to about 80% in 2025 (Figure 13), while the share of public AI supercomputers rapidly decreased from about 60% in 2019 to about 15% in 2025.[21] Our data may even underestimate this shift, given that companies are less likely to publish data on their systems than public owners. However, note that public sector entities may still be able to access private sector AI supercomputers, given many are available through cloud services. In Section 4.1 we discuss how increased economic importance of AI development and deployment likely led to the rapid increase in private sector share.

Share of aggregate AI supercomputer performance by public/private sector over time
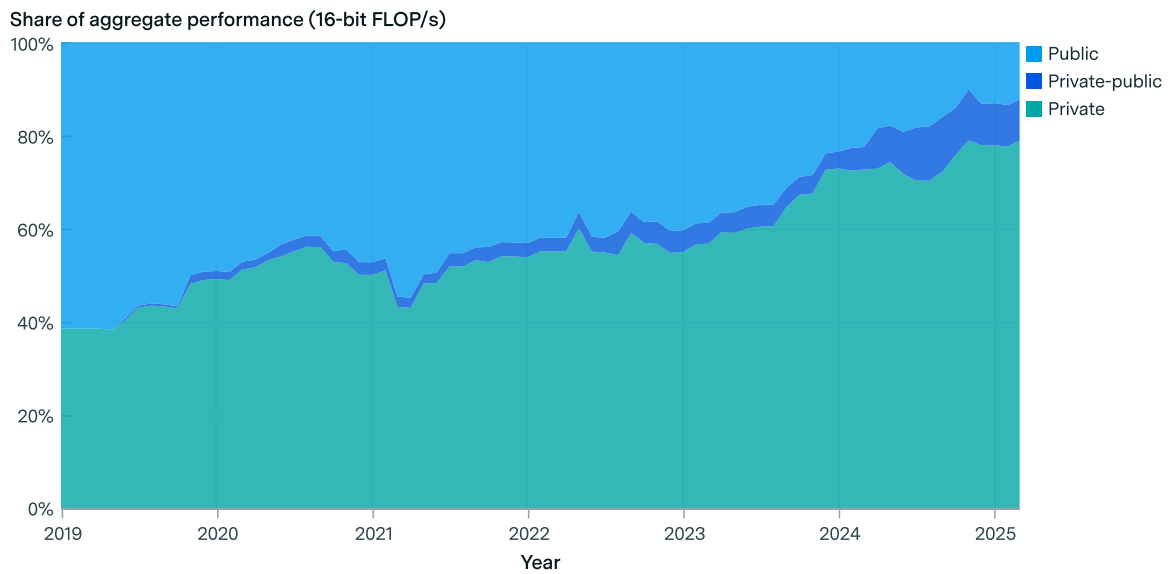
Figure 13: Relative performance shares of public and private sectors based on the owner of the AI supercomputer. An AI supercomputer may have several owners (e.g., if it was a collaborative project or if a government funded an industry project).

3.6. The United States accounts for the majority of global AI supercomputer performance, followed by China

When analyzing the distribution across countries, we find that the United States accounted for about 70% of the computational performance in our dataset at the start of 2019, while China accounted for about 20% (Figure 14).[22] Between 2019 and 2022, the Chinese share grew considerably, reaching about 40% at the start of 2022, although we are unsure if this reflects a real trend or is an artifact of our low data coverage. China's share has since diminished; in March 2025, the United States hosts around 75% of AI supercomputers by performance while China has around 15%.

---

[21]The figure below shows the trend in 16-bit precision. When considering performance across precisions, the trend is similar, with private owners making up about 85% of all AI supercomputers in 2025.

[22]Note that physical location of an AI supercomputer does not directly determine access, given many of our systems are available through cloud services. Furthermore, location also does not necessarily determine ownership since AI supercomputers sometimes belong to owners headquartered in other countries.

Share of aggregate AI supercomputer performance by country over time
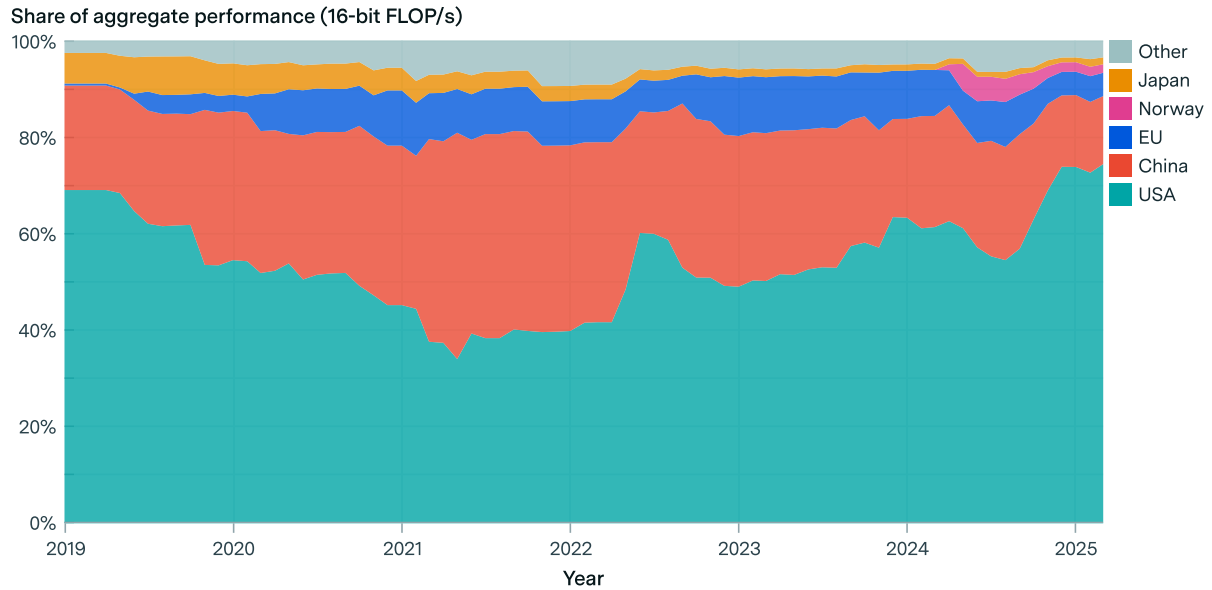
Our dataset covers an estimated 10–20% of global aggregate AI supercomputer performance as of March 2025. While coverage varies across companies, sectors, and hardware types due to uneven public reporting, we believe the overall distribution remains broadly representative. Future country shares may change dramatically as exponential growth continues in both AI chip performance and production volume.

Figure 14: Share of aggregate 16-bit computing power by country over time from AI supercomputers in our dataset. We are visualizing all countries that held a more than 3% share at some point in time. See Appendix C.1.1 for a discussion of our data coverage.

As of March 2025, all operational U.S.-based AI supercomputers in our dataset have a combined performance of 850,000 H100-equivalents (9.1×10^{20} FLOP/s), followed by China with 110,000 H100-equivalents (1.9×10^{20} FLOP/s) and the European Union with 50,000 H100-equivalents (5.6×10^{19} FLOP/s) (Figure 15). Total computational performance in the United States is thus almost 9 times larger than in China and 17 times larger than the total performance in the European Union.

Aggregate AI supercomputer performance in our dataset by country (as of March 2025)
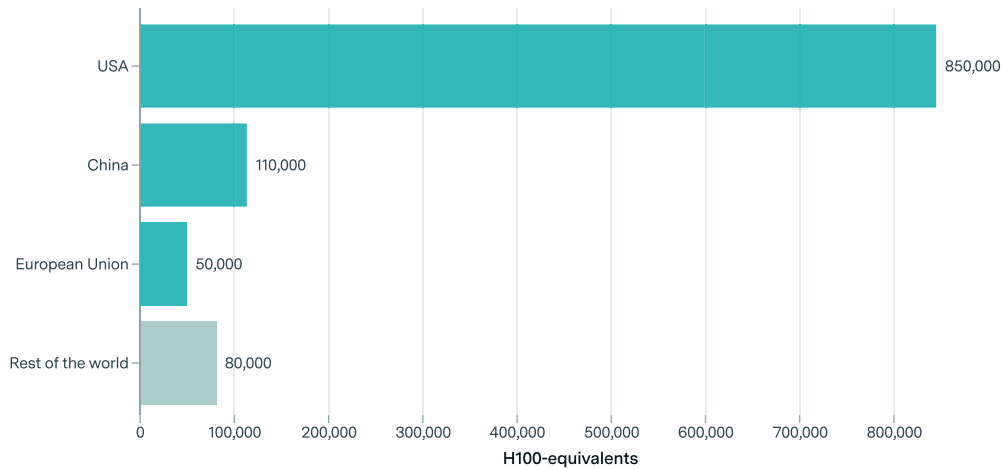
Our dataset covers an estimated 10–20% of global aggregate AI supercomputer performance as of March 2025, so the true compute capacity per county is likely 5–10x higher. While coverage varies across companies, sectors, and hardware types due to uneven public reporting, we believe the overall distribution remains broadly representative. Future country shares may change dramatically as exponential growth continues in both AI chip performance and production volume.

Figure 15: Total AI supercomputer performance by country in H100-equivalents. To convert a system's performance to H100-equivalents, we first take the performance in the lowest precision its AI chips support, considering 32-bit, 16-bit, and 8-bit. We then divide by the 8-bit performance of the H100.

4. Discussion

In this section, we first discuss what caused the rapid growth of AI supercomputer performance and resource needs. We then extrapolate these trends until 2030 and briefly discuss whether the growth in number of chips, power, and hardware cost can continue. We further discuss the geopolitical implications of AI supercomputer distribution across countries and how the increased industry share of AI supercomputers may impact AI research.

4.1. Rapid growth in AI compute both relied on and enabled the increasing economic importance of the AI industry

The rapid growth in AI supercomputer performance we observe has been primarily driven by a surge in AI investment. While traditional improvements in chip design and manufacturing have contributed to this growth (Roser et al., 2023; Hobbhahn et al., 2023), AI supercomputers have grown much faster than traditional AI supercomputers (Section 3.1.2). This acceleration reflects a fundamental shift in the primary use case of AI supercomputers from academic tools for scientific discovery to industrial machines running economically valuable workloads.

In 2019, the largest AI supercomputers were dominated by government supercomputers like the U.S. Department of Energy's Summit and Sierra. These systems were designed to handle a variety of workloads across different scientific domains and advance foundational research (Oak Ridge National Laboratory, undated). However, in the early 2020s, companies increasingly used AI supercomputers to train AI models with commercial applications, such as OpenAI's GPT-3 and GitHub's Copilot integration (Brown et al., 2020; Dohmke & GitHub, 2021). These demonstrations of AI capabilities led to a significant increase in investment in AI, creating a record demand for AI chips (Our World in Data, 2024; Samborska, 2024; Richter, 2025).

As investments in AI increased, companies were able to build more performant AI supercomputers with more and better AI chips. This created a reinforcing cycle: increased investment enabled better AI infrastructure, which produced more capable AI systems, which attracted more users and further investment. The growth of AI supercomputers was therefore both a result of increased funding and a cause of continued investment as AI supercomputers demonstrated their economic value.

Table 1: Historical data and extrapolation of trends based on the current largest AI supercomputer and historical growth rates described in Section 3. Using the growth rates of only industry-owned AI supercomputers would lead to higher extrapolated values. Extrapolated values are rounded to not imply precision.

DATE	LEADING AI SUPERCOMPUTER	PERFORMANCE (16-BIT FLOP/S)	H100-EQ [†]	NUMBER OF AI CHIPS	POWER	HARDWARE COST (2025 USD)
JUNE 2019	OAK RIDGE SUMMIT	3.46×10^{18}	3,492	28K	13MW	$200M
JUNE 2020	OAK RIDGE SUMMIT	3.46×10^{18}	3,492	28K	13MW	$200M
JUNE 2021	SUNWAY OCEANLIGHT	5.95×10^{18}	6,008	108K	N/A	N/A
JUNE 2022	OAK RIDGE FRONTIER	1.44×10^{19}	14,566	38K	40MW	$600M
JUNE 2023	OAK RIDGE FRONTIER	1.44×10^{19}	14,566	38K	40MW	$600M
JUNE 2024	META GENAI 2024A	2.43×10^{19}	24,576	25K	40MW	$900M
MARCH 2025	XAI COLOSSUS	1.98×10^{20}	200K	200K	300MW	$7B
JUNE 2026	*Extrapolated*	5×10^{20}	500K	300K	600MW	$14B
JUNE 2027	*Extrapolated*	1×10^{21}	1M	500K	1GW	$25B
JUNE 2028	*Extrapolated*	3×10^{21}	3M	800K	2GW	$50B
JUNE 2029	*Extrapolated*	8×10^{21}	8M	1.3M	5GW	$100B
JUNE 2030	*Extrapolated*	2×10^{22}	20M	2M	9GW	$200B

[†] Here, we define H100-equivalents (H100-eq) as the AI supercomputer's 16-bit performance divided by the H100's 16-bit performance. This is different than elsewhere in the paper, where we defined it in terms of maximum performance over 8, 16, or 32 bit. H100-equivalents is not a standardized measurement, and should be used only to get a general sense of the scale.

4.2. Can the observed trends continue?

In Section 3.1.4, we conclude that AI supercomputers have kept pace with the 4–5× annual growth in compute in the largest AI training runs.[23] This section will discuss what it would mean for each of the trends in chips, hardware cost, and power needs to continue until 2030.[24]

4.2.1. THE LARGEST AI SUPERCOMPUTER COULD NEED TWO MILLION CHIPS BY 2030

If the number of AI chips continues increasing by 1.6× every year, the largest AI supercomputer in 2030 will require about 2 million AI chips (Table 1). Sevilla et al. (2024) estimated that AI chip production could increase by 1.3× to 2× annually until 2030. Extrapolating from present-day chip production[25] this indicates a production of 7.4M to 144M AI chips annually in 2030.[26] If the largest AI supercomputer used 2 million AI chips in 2030, it would need between 1% and 27% of global annual AI chip production, indicating this scale is feasible if AI chip production continues growing at the estimated rate.

4.2.2. THE LARGEST AI SUPERCOMPUTER COULD HAVE A HARDWARE COST OF ABOUT $200B BY 2030

If the hardware cost of the leading AI supercomputers continues to increase at a rate of 1.9× annually, the leading system's hardware cost in 2030 will be about $200B (in 2025 USD). This is in addition to the cost of the data center facility, which is likely about $10B per GW, adding a further $90B to the acquisition cost (Pilz & Heim, 2023).

Current AI infrastructure is already close to this scale: In 2025, Microsoft announced plans to spend $80B on AI infrastructure globally, and AWS announced plans to spend more than $100B (Smith, 2025; Gonsalves, 2025). Meanwhile, OpenAI

---

[23]At least when considering only industry AI supercomputers and considering performance across precisions.

[24]Our extrapolations do not model deviations from the current training duration growth. If the duration of the largest AI training runs continues to increase by 1.4× annually, the largest training runs in 2030 may last 16 months, exceeding the optimal training duration according to (Sevilla, 2022). If training run duration stops increasing, AI supercomputers would have to grow at a faster rate to sustain a 4–5× annual increase in training compute for the largest AI models.

[25]Public sources estimate that NVIDIA shipped about 500k H100s in 2023 and 2 million in 2024, for a total of 2.5 million H100s (Nolan, 2023; Shilov, 2023c). However, some analysts (Garreffa, 2024) estimate NVIDIA produced up to 1.5 million H100s in Q4 of 2024. Assuming NVIDIA produced about 1M H100s on average per quarter in 2024 yields a total of 4.5 million H100s. NVIDIA produces the majority of all AI chips (Sastry et al., 2024).

[26]For the low-end range, we consider an annual production of 2 million AI chips, growing by 1.3× every year: 2M *1.3^5 = 7.4M. For the high-end range, we consider an annual production of 4.5 million AI chips, growing by 2× every year: 4.5M * 2^5 = 144M

announced plans to spend up to $500B on the Stargate project over four years (OpenAI, 2025). These announcements are compatible with $200 billion hardware costs for a single project by 2030, especially as AI investment is projected to continue growing (Zoting, Shivani, 2025; IDC, 2025; Grand View Research, 2024).

4.2.3. THE LARGEST AI SUPERCOMPUTER COULD NEED 9 GW OF POWER BY 2030

If AI supercomputer power requirements continue growing at a rate of 2.0× every year, the leading AI supercomputer will need about 9 GW of power in 2030 (Table 1). This is slightly higher than Sevilla et al. (2024)'s extrapolation of 6 GW and matches Pilz et al. (2025)'s estimate for the AI supercomputer running the largest training run in 2030.

The largest data center campuses today have a capacity of hundreds of MW, and as of early 2025, no existing campus exceeding 1 GW has been publicly reported (Pilz & Heim, 2023). While a 2 GW AI supercomputer in 2028 is likely feasible, a system with a capacity of 9 GW by 2030 would require as much power as 9 nuclear reactors can generate, and would likely face severe permitting and equipment supply chain challenges, as well as other potential challenges such as local community opposition (Pilz et al., 2025).[27] As they struggle to secure adequate power, companies may increasingly use decentralized training techniques that allow them to distribute a training run across AI supercomputers in several locations. Some notable training runs, including Google DeepMind's Gemini 1.0 and OpenAI's GPT-4.5, were reportedly already trained across several AI supercomputers (Moss, 2023; OpenAI, 2025; The White House, 2025).

4.2.4. CONCLUSION: POWER CONSTRAINTS WILL LIKELY BE THE MAIN CONSTRAINT TO CONTINUED GROWTH

Power constraints will likely become the primary bottleneck for AI supercomputer growth, driving a shift toward distributed training across multiple sites. This evolution could change how we measure AI training capabilities—from focusing on individual AI supercomputers to assessing companies' aggregate compute capacity. While chip production and hardware cost trends appear sustainable through 2030, the continuation of all these trends ultimately depends on AI applications delivering sufficient economic value to justify the massive investments required for infrastructure expansion.

4.3. U.S. dominance in global AI supercomputer distribution

This section discusses that U.S. dominance likely resulted from leading in related industries, and will likely continue, given stated U.S. policy and U.S. control of key AI chip production chokepoints.

4.3.1. U.S. DOMINANCE RESULTED FROM DOMINANCE IN CLOUD COMPUTING AND AI DEVELOPMENT

According to our data, around 75% of all AI supercomputer performance is currently based in the United States (Figure 14). How did the United States develop such a dominant position in AI supercomputers, while countries that used to play a prominent role in public supercomputing, like the UK, Germany, or Japan, declined in importance?

U.S. dominance was likely a direct result of AI supercomputers becoming increasingly commercialized and dominated by companies (instead of governments or academia), which were primarily based in the United States due to dominance in previous technologies. This advantage is evident in cloud computing infrastructure, where in 2019, the top three leading U.S. cloud companies, AWS, Microsoft, and Google alone made up 68% of global market share (Gartner, 2020). American companies also played leading roles in key AI advances, including in recommender systems, scientific applications like AlphaFold, and LLM chatbots like ChatGPT (Dong et al., 2022; Jumper et al., 2021; OpenAI, 2022). Overall, American companies were involved in developing 338 of the 476 notable AI models and trained 18 of the 25 largest AI models by training compute recorded by (Epoch AI, 2025). While limited reliable data on global market shares in AI applications exists, record user growth may indicate that U.S. companies also lead in total number of users (Hu, 2023).

4.3.2. THE UNITED STATES WILL LIKELY CONTINUE LEADING IN AI SUPERCOMPUTERS

The United States dominates not only AI development and cloud provision, but also the design of AI chips, and several inputs to semiconductor manufacturing (Sastry et al., 2024). The U.S. government has previously used its dominance in AI chips to impose export controls on AI chips and key equipment to China, and introduced an AI diffusion framework that puts conditions on the export of AI chips to countries that are not close U.S. allies (Allen, 2022; Heim, 2025).

---

[27] In 2025, the U.S. government began programs to support large-scale data center campuses and several companies have already announced plans to build multi-GW data centers (Moss, 2024; Skidmore & Swinhoe, 2024). Still, no known industrial facilities currently require several GW of power, indicating that this amount of power may be challenging to secure (Sevilla et al., 2024).

At the same time, some challenges could limit U.S. dominance in AI supercomputers:

- **Power requirements:** AI's power demand is massively increasing, both in terms of the power needed for AI supercomputers and in terms of the overall number of AI chips deployed, primarily for inference. (Pilz & Heim, 2023). The United States is facing significant challenges to add enough power generation capacity to sustain the current rate of AI data center growth (Pilz et al., 2025; Fist & Datta, 2024; Mahmood et al., 2025).
- **Investment in sovereign infrastructure by foreign governments:** Some governments have begun investing in local AI infrastructure, such as France (Reuters, 2025), the United Kingdom (UK Department for Science, Innovation and Technology, 2025), Saudi Arabia (Benito, 2024), and the UAE (Allen et al., 2025). However, most of these projects are small compared to leading U.S. AI supercomputers. Furthermore, given U.S. control of AI chip production, the United States could block chip access if these projects were threatening U.S. computing dominance.
- **Competition from China:** The Chinese government and Chinese companies are heavily investing in AI infrastructure, but being unable to import leading U.S. AI chips, the country relies on inferior U.S. or domestically produced AI chips. Limited AI chip access makes it more costly to establish significant AI supercomputers, and limits the total number of projects in China (Scanlon, 2025; Lin & Heim, 2025). So far, Chinese efforts at indigenous production of AI chips have been severely hampered by the inability to produce or import crucial equipment like DUV and EUV lithography machines that are extremely challenging to produce (Grunewald, 2023; Allen, 2022).

To summarize, the United States leads in AI model development and cloud computing and controls key chokepoints in the semiconductor supply chain. Combined with stated government policy to advance U.S. AI leadership, this leads us to conclude that the United States will likely continue leading in AI supercomputers for at least another six years (The White House, 2025).

4.4. Consequences of increased private sector dominance

Our finding that companies own a growing share of AI supercomputers matches a previously reported trend: AI research is increasingly dominated by large companies rather than academic or government groups. Besiroglu et al. (2024) found a stark decline in academic institutions' share of large-scale machine learning models, from approximately 65% in 2012 to just 10% in 2023.

The shift away from public ownership of AI supercomputers is likely due to their increased economic importance (Section 4.1), which has rapidly increased private AI investments. More investment allowed companies to build systems as expensive as xAI's Colossus, which had an estimated hardware cost of $7B. Meanwhile, the most expensive government projects, Frontier and El Capitan, cost only $600M each. Additionally, governments usually build only a small number of systems for research purposes. Meanwhile, major tech companies often build dozens of AI supercomputers, given that they are not just training larger models, but also serving millions of users around the world.

This shift from public to private ownership of AI supercomputers produces two significant consequences for AI research: restricted access for academic researchers and diminished visibility into AI development and deployment.

Limited access for academic researchers: The concentration of AI supercomputers in industry reduces access to frontier compute resources for academic researchers, who historically have contributed to AI progress and provided independent evaluation and scrutiny (Besiroglu et al., 2024). The ownership of systems does not inherently determine compute access because researchers can rent AI supercomputers through cloud companies (Heim & Egan, 2023). However, renting large quantities of AI chips—beyond a few thousand—for even short durations can still be prohibitively expensive for academic researchers, compelling them to rely on smaller, less powerful models (Lohn, 2023).

Lack of visibility: As companies now operate the leading AI supercomputers, they have become the main driver of frontier AI progress, relegating government and academic labs to a supporting role. Because companies are often less public about their research, governments may increasingly struggle to track capability gains in AI models (Besiroglu et al., 2024). Additionally, given the importance of compute for AI development and deployment, the scale and number of a nation's top AI supercomputers are increasingly tied to competitiveness in AI. With companies controlling most systems, governments increasingly lack data on the extent of their national AI infrastructure, hampering policymakers' ability to craft a coherent strategy for technological competition.

One option for governments to increase visibility into AI development and deployment and better understand national competitiveness could be to require companies to report key data about their infrastructure, such as the performance of their largest AI supercomputers and the total extent of their infrastructure (Sastry et al., 2024). Governments could also collect

intelligence on the AI compute capacity of other countries, allowing them to better understand their competitive position and potentially making it easier to verify potential future international agreements on AI (Baker, 2023; Sastry et al., 2024).

5. Conclusion

We compiled a dataset of 500 AI supercomputers between 2019 and 2025 and found that performance, number of chips, power requirements, and hardware cost have all grown exponentially. The rapid performance growth of AI supercomputers, combined with increasing training durations, has enabled a 4–5× annual increase in training compute for frontier AI models, which has fueled significant advances in AI capabilities and driven further investment in infrastructure. If trends continue, the leading AI supercomputer in 2030 could have a hardware cost of more than $200 billion and incorporate over 2 million AI chips. However, the projected power requirement of 9 GW would be challenging to secure in a single location, likely forcing companies to adopt decentralized training approaches across multiple sites.

Our data also reveals key trends in AI supercomputer ownership, with companies increasing their share of total AI supercomputer performance from 40% in 2019 to more than 80% in 2025. This finding emphasizes the previously observed increasing compute divide between industry and academia. The United States hosts approximately 75% of global AI supercomputer performance and will likely maintain this dominance through its control over the AI chip supply chain.

To conclude, AI supercomputers have been a key driver of AI progress and represent a central component of the AI supply chain (Sastry et al., 2024). Our analysis provides valuable information about AI supercomputers' growth patterns, distribution, and resource requirements. Such information will be increasingly important for policymakers, and more generally for understanding the trajectory of AI.

APPENDIX

A. Review of existing data sources

A.1. The Top500 list and its limitations for AI supercomputers

The Top500 list has been the primary leaderboard for tracking supercomputer performance since its inception in 1993. It ranks systems based on their performance in solving linear equations using the LINPACK benchmark (Dongarra, 1987). While this benchmark has provided a consistent, long-term method for comparing traditional high-performance computing (HPC) systems, it has several significant limitations when applied to AI supercomputers:

- Participation in the Top500 list is voluntary, leading to significant gaps in reporting. Companies, particularly cloud providers, which own many of the largest AI supercomputers, face limited incentives to report their AI supercomputers. Running the LINPACK benchmark diverts valuable supercomputer and engineer time from more economically valuable uses like AI training or deployment. Instead of reporting to the Top500, companies sometimes independently publish promotional blog posts about their systems (Langston, 2020; Meta, 2022; AWS, 2023), while often maintaining ambiguity about the number and size of their largest systems to avoid giving competitors unnecessary information about their strategies. Additionally, Chinese owner stopped reporting any systems to the Top500 list in 2022, presumably to reduce scrutiny and avoid U.S. sanctions (Shah, 2024).
- LINPACK is not an AI benchmark. It measures performance on linear equations requiring high-precision 64-bit number formats (Dongarra, 1987), while modern AI workloads run on much lower precision formats (16-bit, 8-bit, or even 4-bit for some inference workloads[28]). While performance on different precision formats was formerly highly correlated, the introduction of tensor cores for lower precision formats on AI accelerators led to drastically faster performance increases in these formats (Hobbhahn et al., 2023; Rahman & Owen, 2024). This divergence means LINPACK performance does not accurately capture a supercomputer's performance for AI workloads.[29] New benchmarks like HPL-MxP and ML-Perf better capture AI-relevant performance but have not been widely adopted (Luszczek, 2024; Mattson et al., 2020).

Besides the Top500, no major datasets of supercomputers exist, meaning that previous analyses of supercomputers, such as Hochman (2020), Tekin et al. (2021) and Chang et al. (2024) have exclusively relied on the Top500 list. While these analyses offer useful insights into changes in components, performance, and energy efficiency of traditional supercomputers, the limitations of the Top500 lists discussed above mean the observed trends do not adequately capture AI supercomputers.

A.2. Commercial databases of AI supercomputers

Some analysts like SemiAnalysis and The Information have private databases of AI supercomputers that are available for paid subscribers. Furthermore, some companies such as Omdia offer trackers of AI chip shipments (SemiAnalysis, 2024; The Information, 2025; Galabov et al., 2025). These databases are typically focused on providing business intelligence. Thus, they do not assess historical trends and may not capture data from non-industry sources. Furthermore, these databases usually do not disclose their methods and sources and do not make the analysis of their data publicly available.

B. Detailed Methods

B.1. Data collection process

We relied on systematic Google searches and publicly available datasets to find potential AI supercomputers. For each potential AI supercomputer, we conducted an additional search to find and verify all relevant publicly available data about it.

Search methodology:

a) We used the Google Search API to search for terms such as "AI supercomputer" and "GPU cluster" in consecutive 12-day windows (1-1-2019–1-3-2025). We additionally conducted year-by-year country searches (e.g., "Albania AI

---

[28]Or even 4-bit precision for some inference workloads (Ashkboos et al., 2023).

[29]For instance, Microsoft's Eagle and Japan's Fugaku have comparable performances on LINPACK (5.6×10^{17} FLOP/s vs 4.4×10^{17} FLOP/s), but given that Fugaku does not contain any GPUs or other chips optimized for low-precision performance, they diverge by almost an order of magnitude on FP8 performance (2.9×10^{19} FLOP/s vs 4.3×10^{18} FLOP/s) (Lee, 2023; Riken Center for Computational Science, undated).

supercomputer").

- Although our study period begins in 2019, we also conducted a similar, pared-down Google search for January 2016–January 2019 in order to be able to determine which AI supercomputers were in the top 10 by computational performance at the start of 2019. For this, we reduced our search terms by roughly 80% to lower the number of records to look through.

b) We parsed the top results with the Beautiful Soup Python package and used GPT-4o via the OpenAI API to extract system names and chip counts of any AI supercomputers mentioned.

c) We grouped entries by name in a spreadsheet, deduplicated, verified all potential AI supercomputers manually, and added those that fit our inclusion criteria to our dataset.

d) Find additional details about the Google Search methods in Section B.2.

Additional sources:

a) Top500 list, inferring AI chip counts from reported accelerator cores.

- Many systems in the Top500 did not contain AI chips; however, those that did usually listed the 'Accelerator/Co-Processor' type and the total number of 'Accelerator/Co-Processor Cores.' Since we knew the number of cores for each AI chip model, we calculated the implied AI chip count for the system by dividing the number of cores by the cores per AI chip. We verified this method by checking it for AI supercomputers in the Top500 with previously known AI chip counts.
- We considered all Top500 entries from June 2014 to November 2024 (but included only those that qualified for our inclusion criteria between 2017 and 2025).

b) Epoch AI's notable AI models dataset.

c) Published compilations of Chinese AI supercomputers (redacted, please reach out).

d) A small number of entries from a project on sovereign compute resources led by Aris Richardson (publication forthcoming).

e) MLCommons Results.

f) gpulist.ai (last accessed January 2025).

g) Articles and newsletters shared by colleagues, such as from SemiAnalysis, Transformer, and Import AI.

Remaining components:

- We built our initial dataset via Google Alerts for the keyword "AI supercomputer" (June 2023–Aug 2024)[30].
- Two Chinese-language analysts conducted targeted searches of systems in China and Hong Kong (see Appendix B.3).
- Our main data collection focused on AI supercomputers that first became operational between 2019 and 2025. However, we also included AI supercomputers that became operational between 2017 and 2019 if they met the standard inclusion criteria, or if they were operational before 2017 and were at least 1% as large as the largest known supercomputer in January 2017.
- We collected various additional sources for details on specific supercomputers using the Perplexity API.
- For over 500 key supercomputers, an Epoch staff member did an additional verification of the entry (marked as true in the 'Verified Additional Time' field). This focused on systems that were especially large for their time, most Chinese systems, and any outliers.

A full, up-to-date documentation of all fields in our dataset is available at: `https://epoch.ai/data/ai-supercomputers-documentation`

B.2. Google search methodology

We conducted automated Google searches spanning from January 2019 to March 2025 for consecutive 12-day windows, using various keywords related to AI supercomputers. For each search term, we collected different amounts of results based on their utility in finding relevant information:

- "AI Supercomputer": 30 Google results
- "AI Supercomputer cluster": 30 Google results
- "AI Supercomputer news": 20 Google results
- "AI Supercomputer cluster news": 20 Google results

---

[30]Rose Hadshar and Angelina Li contributed additional entries during the Epoch FRI Mentorship Program 2023.

- "GPU Cluster": 20 Google results
- "Compute Cluster": 10 Google results
- "V100 Cluster": 10 Google results
- "A100 Cluster": 10 Google results
- "H100 Cluster": 10 Google results

We parsed all websites using the BeautifulSoup (2025) Python library and used GPT-4o from the OpenAI API to search for information on all mentioned AI supercomputers (see prompt below).

Our searches yielded over 20,000 unique websites, resulting in approximately 2,500 potential AI supercomputer mentions after deduplication. For each unique AI supercomputer, we used the Perplexity API to collect additional data sources (see prompt below).

GPT-4O PROMPT FOR INITIAL EXTRACTION

Here is the text from a webpage that potentially contains some information about AI supercomputers. Please list the names of any AI supercomputer clusters that are listed in this article, separated by semicolons if there are multiple. If you know the company/organization name that owns/runs it, you should write the supercomputer name as the company/organization name, followed by the name of the cluster. If the cluster does not have a name, simply refer to it with 'UNNAMED' and include any identifiable information given. Please include any information about the number and type of AI chips (e.g. GPUs or TPUs) in square brackets after the cluster name. Say '[NOINFO]' if there is no information in the article about chip type or quantity. For example, a response might look like 'OpenAI Stargate [NOINFO]; Frontier [37,632 AMD MI250X]; Microsoft UNNAMED Arizona H100s [50,000 NVIDIA H100s]'. You should only list AI supercomputer clusters and associated chip information, nothing else. If there are no supercomputer clusters mentioned in the article, just reply with 'None'. If you can't access or read the article, just reply with 'Could not access article'. However, this should be rare, and mainly only happen if the article is paywalled. Do not mention any other details. Article text: {TEXT HERE}

PERPLEXITY PROMPT FOR DETAILED INFORMATION

Tell me all the details you can about the {SUPERCOMPUTER NAME} supercomputer, including but not limited to: What type of AI accelerator chips (eg GPUs, TPUs, etc) do they use (be as specific about the exact type of chip as possible)? How many do they have, if any? When was it completed, or when is it expected to be completed? When was it first announced? What is the timeline for any updates/iterations to this supercomputer? Where is it located? (be as specific as possible) How many AI FLOP/s could it do? Who operates it? Who uses it? Who owns the supercomputer? Please list several organizations if it is a joint partnership, and list if these organizations are or part of government, academia, industry, or something else? Are there multiple supercomputers that could go by roughly this name? Have there been different versions/iterations of this supercomputer?

B.3. Approach for finding Chinese AI supercomputers

We decided to redact our approach to finding Chinese AI supercomputers and avoid providing identifying information about them throughout the paper to preserve data sources. We take this step as a precautionary measure because Chinese websites cited in public reports have been redacted or replaced with malware in the past (Wei, 2023).

If you would like to request access to our methodology for Chinese AI supercomputers, please contact Konstantin at kfp15@georgetown.edu.

B.4. Power requirements

We calculated the peak power demand for each AI supercomputer with the following formula:

$$\text{Chip TDP} \times \text{number of chips} \times \text{system overhead} \times \text{PUE}$$

We collected Thermal Design Power (TDP) for most chips when publicly available, though we did not find the TDP for some Chinese chips and custom silicon such as Google's TPU v5p. We considered both primary and secondary chips when counting the number and types of chips. We used a $2.03\times$ multiplier for non-GPU hardware to account for system overhead (additional power needed for other server components like CPUs, network switches, and storage), based on NVIDIA DGX H100 server specifications (NVIDIA, 2025). We also factored in Power Usage Effectiveness (PUE), which is the ratio of

total data center power use to IT power use (with a minimum value of 1). According to the 2024 United States Data Center Energy Usage Report (Shehabi et al., 2024), specialized AI datacenter facilities had an average PUE of 1.14 in 2023, which is 0.29 lower than the overall national average of 1.43. We adjusted the trend for all datacenter facilities to estimate the average PUE of AI datacenters by subtracting 0.29 from the overall values reported by Shehabi et al. (2024) (Table 2).

Table 2: AI data center power usage effectiveness (PUE) over time, adapted from (Shehabi et al., 2024).

YEAR	2016	2017	2018	2019	2020	2021	2022	2023	2024	2025
PUE	1.31	1.29	1.26	1.22	1.20	1.18	1.17	1.14	1.12	1.10

The full formula we use is:

$$
\begin{aligned}
Power = [&(\textit{Primary AI chip TDP} \times \textit{Primary AI chip quantity}) \\
+&(\textit{Secondary AI chip TDP} \times \textit{Secondary AI chip quantity})] \\
\times&\ \textit{Server overhead factor} \times \textit{Datacenter PUE}
\end{aligned}
$$

We base some of the reported power values in our dataset on the top 500 list. However, the list reports average power utilization during the benchmark, rather than peak power requirement. To determine peak power, we compare peak and average power for supercomputers where we have both, find that they differ on average by a factor of 1.5, and scale all the Top500 reported power figures by this factor. We then multiply by the PUE in the given year to find peak power demand for the entire system.

B.4.1. LIMITATIONS WITH OUR POWER DATA

We rely on owner-reported power estimates for 15% of the AI supercomputers in our dataset. These reported figures lack standardization—some may represent only critical IT load at theoretical maximum utilization, while others include complete data center infrastructure overhead (accounting for power conversion losses and cooling requirements).

For the remaining 85% of systems, we estimate the power requirements as detailed in the previous section. A key limitation of our current approach is the application of a uniform $2.03\times$ multiplier for all chip types to account for additional system hardware. Future analyses would benefit from developing chip-specific overhead multipliers that better reflect the varying cluster-level power requirements across different AI chip and cluster architectures.

To check for consistency between reported and estimated power values, we plotted the correlation below (Figure 16). The correlation coefficient of 0.97 indicates our values are highly correlated.
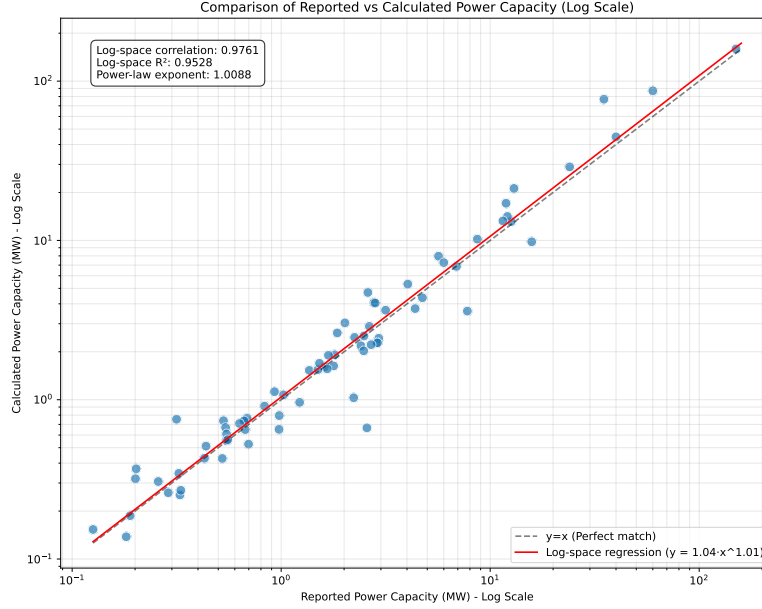
Figure 16: Comparison of power requirements for AI supercomputers that reported it, versus our calculations of power requirements based on chip type and count.

Note that our methods assess theoretical peak power usage when all the processors are fully utilized and not power consumption. The average power consumption of an AI supercomputer is usually only a fraction of its peak.

B.5. Hardware cost

We use the publicly reported total hardware cost of the AI supercomputer in our analysis whenever it is available. When it is unavailable, we estimate this cost based on the chip type, quantity, and public chip prices. The procedure used to estimate costs is adapted from Cottier et al. (2024). Using Epoch's dataset of hardware prices, we select the latest known price of the chips used in the AI supercomputer, from before the system's first operational date. For each type of chip, we multiply the cost per chip by the number of chips, multiply by factors for intra-server and inter-server overhead, and then sum these costs if there are multiple types of chips. Intra-server cost overhead was estimated in Cottier et al. (2024) for the NVIDIA P100 (1.54×), V100 (1.69×), and A100 (1.66×), based on known DGX and single-GPU prices near release. We use the mean of these factors (1.64×) for all chips, to estimate server prices, including interconnect switches and transceivers. Then, we adjust for the cost of server-to-server networking equipment, which was estimated to be 19% of final hardware acquisition costs.

Additionally, we apply a discount factor of 15% to the final hardware cost of the AI supercomputer to account for large purchasers of AI chips often negotiating a discount on their order. We discuss limitations with this estimate and our cost data in the next section.

Our final formula for estimating hardware cost is as follows:

$$
\begin{aligned}
\textit{Hardware acquisition cost} = &\left[(\textit{Primary AI chip cost} \times \textit{Primary AI chip quantity})\right. \\
&+\left.(\textit{Secondary AI chip cost} \times \textit{Secondary AI chip quantity})\right] \\
&\times \textit{Intra-server overhead} \times \textit{Inter-server overhead} \times \textit{Discount factor}
\end{aligned}
$$

In this formula, our intra-server overhead, or "chip-to-server" factor, is 1.64×, our inter-server overhead, or "server-to-cluster" factor, is 1.23×, and our discount factor is 0.85×.

Notably, our cost figures refer only to the hardware acquisition cost of the AI supercomputer, and not costs required for maintenance, electricity, or the cost of the datacenter hosting it.[31]

---

[31]A 2025 estimate of the cost of datacenters puts them at $11.7 million per MW. This could be combined with our power requirement estimates to get an estimate of hardware plus datacenter acquisition (Cushman & Wakefield, 2025).

All cost values are adjusted for inflation into 2025 USD, using the producer price index for the Data Processing, Hosting, and Related Services industry, reported by the Federal Reserve Bank of St. Louis (U.S. Bureau of Labor Statistics, 2025). We divided pre-2025 cost figures by the price index value at its closest reported date and multiplied by the price index value in January 2025. Our trends and forecasts refer to values in 2025 USD.

B.5.1. LIMITATIONS WITH OUR HARDWARE COST DATA

Our cost data for AI supercomputers has several important limitations:

1. We found reported cost figures for only a limited subset of AI supercomputers, with data predominantly from public sector systems rather than industry deployments.

2. The reported figures may diverge from true costs in multiple ways.

   - They sometimes represent planned contract costs rather than final realized expenditures.
   - Contract figures may bundle additional expenses, such as multi-year operational costs, that should be excluded from our analysis.
   - When uncertainty about the precise meaning of reported costs is too high, we excluded the data, though some ambiguity likely remains.

3. We also encountered challenges with estimating hardware costs based on chip quantities and prices.

   - Our price dataset lacks information for some GPU types, particularly custom silicon, though it does cover most common GPUs.
   - Google does not sell TPUs, so our price data for them is based on comparison of their performance and manufacturing costs with those of NVIDIA chips that have similar technical specifications.
   - Most GPU suppliers do not publish wholesale prices, forcing us to rely on third-party retailer prices and reports from experts that can vary significantly by vendor and time.
   - We use the most recent listed price for each GPU, but prices fluctuate substantially with market conditions, so our limited time-series data means some AI supercomputer costs may be mismatched with the prices actually paid for the chips.

4. Given limited data, we assume that all AI supercomputers have the same overhead costs, but this is unlikely, particularly for systems built five years ago.

5. The discount factor is another significant source of uncertainty. Price negotiations generally occur privately, making reliable estimates difficult, and discounts vary substantially by supplier, purchaser, chip type, and time. For simplicity and due to data limitations, we apply a constant 15% discount rate across all AI supercomputers, but we expect the true rate to vary significantly by AI supercomputer. We selected this rate because it best aligns with the difference between our cost estimates and reported costs, and stated estimates of discount rates.[32] However, as stated above, our reported cost data is itself biased. Our universal discount rate likely overestimates costs for major purchasers like U.S. national labs[33] and the largest GPU buyers while underestimating costs in other scenarios.

As a consequence of these limitations, we estimate that a 90% confidence interval for the true hardware cost value is +/- 0.5 orders of magnitude (within a factor of $\sim 3\times$) of our estimate.

B.6. Forecasts

We extrapolate our observed trends by using the leading AI supercomputer as of March 2025 (xAI's Colossus) and assuming trends continue until 2030. E.g., for the number of chips, we assume 200,000 chips in March 2025 and multiply this number by $1.6\times$ each year. Note our approach to extrapolations is simplistic and can only provide rough estimates for future values.

B.7. Figures and regressions

For all figures and regressions, we filtered the dataset as follows:

---

[32]Citi Analysts imply that Microsoft received a 33% discount compared to other purchasers, who paid what we would count as the full price (Shilov, 2024b). If these groups buy equal amounts of chips, this implies an average discount of 16% (Morgan, 2021).

[33]NextPlatform implies that the Oak Ridge National Lab Summit supercomputer got close to a 50% discount on the cost of their GPUs, and that industry partners have historically paid (Morgan, 2024) $1.5\times$ to $2\times$ more for chips than National Labs.

1. We excluded 99 AI supercomputers where the "Exclude" field is marked. 85 of these systems are outside of our definition because they do not meet our performance threshold. We also excluded 14 systems for other reasons, such as because we decided the chips they used did not qualify as AI chips.

2. We further excluded 92 AI supercomputers marked as "Possible duplicates". (We try to only mark systems as potential duplicates if we think there is a >25% chance they are a duplicate.)

3. We further excluded 36 AI supercomputers where "Single cluster" is marked as "No" or "Unclear".

4. We excluded 15 AI supercomputers where "Certainty" is lower than "Likely".

5. We excluded 113 AI supercomputers where "Status" is "Planned", i.e., systems that were not yet operational as of March 2025.

In total, we include 470 out of the 825 systems in our dataset in the analysis. Of these, 389 became operational in 2019 and after.

For all regressions, we consider the 57 AI supercomputers that were in the top-10 by 16-bit FLOP/s and became operational between 1-1-2019 and 1-3-2025.[34]

For our distribution figures we consider all 470 systems remaining after filtering, including those that became operational before 2019. We exclude AI supercomputers that were superseded by newer entries after the newer entry's first operational date.

B.8. Adequately representing performance gains from using lower precision units

Values in calculations for AI training (such as model weights, gradients, and updates) can be represented in different precisions. This is analogous to how you may represent the same number as "$15,228,349,053.84" or "$15 billion", depending on the context. In this example, the first representation has a much higher precision than the second, but it also takes more memory to store.

Until the 2010s, AI training primarily used relatively high-precision 32-bit number formats but moved to 16-bit representation in the late 2010s[35] and currently seems to be moving to 8-bit, thanks to new hardware supporting these precisions and algorithmic innovations to use the new number formats efficiently (Huang et al., 2020; NVIDIA, 2023). Given that working with values in lower precisions requires less memory and computations, AI chips offer much faster performance for calculations in lower precisions.

The shift in precision used for training in our study period makes it challenging to adequately display performance trends in our data.

- If we showed the highest available performance across these three precisions (Max OP/s)[36] it may seem like AI supercomputers that supported 8-bit precision in the early 2020s were more powerful than they actually were in practice, since 8-bit precision was not widely used to train AI models then.[37] If we used this precision-agnostic trend for our forecasts, we would further imply that shifts to lower precisions will continue, but we cannot make any claims about whether or not that will be the case.[38]
- Instead, we limit our analysis to performance in 16-bit precision (16-bit OP/s), which 92% of the AI supercomputers included in our analysis support.[39] However, we acknowledge that only considering 16-bit performance does not adequately show the performance gains AI companies achieved by moving to lower precision.

In practice, we find that trends in a) Max OP/s and b) 16-bit OP/s are mostly consistent. We thus use 16-bit OP/s as the

---

[34]In some figures we specify that we are showing trends for the 59 AI supercomputers that were in the top-10 considering highest performance across 32, 16, and 8-bit precisions.

[35]Micikevicius et al. (2017) is an early example of mixed-precision training which moved the most computationally expensive operations to 16-bit.

[36]OP/s stands for operations per second.

[37]Specifically, we are unsure when 8-bit training first became widespread. Developers usually do not report what precisions they use to train their models, making it difficult to assess when newly available formats were widely adopted.

[38]Specifically, we are unsure when 8-bit training first became widespread. Developers usually do not report what precisions they use to train their models, making it difficult to assess when newly available formats were widely adopted.

[39]For comparison, 96% of AI supercomputers have a performance for Max OP/s (performance across 32, 16, and 8-bit precisions) The remaining AI supercomputers either lack performance data or we only found a performance for 64-bit precision.

default for our trend analysis and forecasts, but discuss the Max OP/s trend whenever it converges.[40]

Meanwhile, we decided to use Max OP/s for our inclusion criteria, i.e., to select whether or not a given system has at least 1% of the performance of the leading operational AI supercomputer.

We include an overview table showing all metrics in each of 16-bit FLOP/s, 8-bit OP/s, and Max OP/s in Appendix D.1.

C. Limitations

This section summarizes some overall limitations of our data. We discuss limitations with specific parts of our data in the methods section (Appendix B).

C.1. Summary of limitations

C.1.1. WE LIKELY ONLY COVER ABOUT 10-20% OF ALL AI SUPERCOMPUTERS WITHIN OUR DEFINITION

We use four references to assess our coverage:

- **Coverage by chip production:** Our dataset likely covers 20–37% of all NVIDIA H100s produced until 2025, about 12% of all NVIDIA A100s produced, and about 18% of all AMD MI300X produced. Meanwhile, we estimate we cover less than 4% of Google's TPUs and very few custom AI chips designed by AWS, Microsoft, or Meta. We also only cover about 2% of NVIDIA chips designed to be sold in China (including the A800, H800, and H20). Our average coverage of the six chip types we assessed is 11%.
- **Coverage by company:** The coverage of different companies varies considerably, from 43% for Meta and 20% for Microsoft to 10% for AWS and 0% for Apple. The coverage of Chinese companies is particularly poor. Our average coverage of 8 major companies is 15%.
- **Coverage of total 16-bit FLOP/s in China:** Between end of 2020 and end of 2024 we cover between 10-20% of total Chinese 16-bit FLOP/s based on an estimate by IDC (2025).
- **Coverage of largest training runs:** Our dataset contains a matching AI supercomputer for about half of the largest training runs as of March 2025 reported by Epoch AI (2025). However, we only find official confirmation that the system was used for the specific training run for one-third of all models. Coverage of Chinese training runs is slightly better compared to all training runs.

Overall, we estimate we cover between 10 and 20% of all AI supercomputers as of early 2025. For more details on our coverage, see Appendix C.3.

C.1.2. WE LACK DATA FOR KEY PROPERTIES

- We cannot reliably determine when an AI supercomputer was first operational. In most cases, we use the date an AI supercomputer was first reported as existing as the "first operational" date. However, owners may sometimes wait several months before publicly announcing their AI supercomputer, or they may announce a system even if it is not yet available. We expect that most of our "first operational" dates will be a few weeks to a few months later than the real date the AI supercomputer came online.
- We sometimes need to make assumptions about basic system facts. For instance, owners sometimes report vague chip quantities such as "EC2 UltraClusters are comprised of more than 4,000 latest NVIDIA A100 Tensor Core GPUs" (AWS, 2020), or "With thousands of MI300X GPUs available, clusters of any size can be deployed for reliable, high-performance computing." (Vultr, 2024). To include such AI supercomputers, we try to make reasonable estimates of the system's chips and performance and explain our reasoning in the notes field.
- Our data is incomplete. Some fields in our dataset are only filled for a fraction of systems, such as reported power requirement, reported hardware cost, and location. However, our data captures key statistics like performance and first operational date for more than 95% of all AI supercomputers that are included in our dataset.

C.1.3. KEY REASONS FOR LOW COVERAGE

Why do we only cover 10–20% of all AI supercomputers? The following factors contribute to our low data coverage:

---

[40]Notationally, we generally refer to 16-bit performance as FLOP/s (instead of "OP/s"), since this is is more common terminology.

a) Companies often choose not to report their AI supercomputers publicly. While companies may benefit from increased public and investor attention when they publish information on large AI supercomputers, they may also prefer to keep this information private to maintain ambiguity about their competitive position.

b) Companies may only report their largest AI supercomputers. A large fraction of all chips are sold to hyperscalers that have more limited incentives to publish information about their AI supercomputers. While they may benefit from publishing information about their largest systems, they have no incentives to publish about the number and size of smaller AI supercomputers.

c) Even if an owner publishes information about an AI supercomputer, our search methods may not find it, especially if the information is published in a language other than English or Chinese.

d) Chinese companies may try to avoid scrutiny from U.S. regulators, both for chips that they legally imported, such as NVIDIA's A800 and H800, as well as illegally imported chips like NVIDIA's A100 and H100. Chinese companies may have smuggled more than 100,000 AI chips last year (Grunewald, 2025). See Appendix C.2.4 for a longer discussion.

C.2. Detailed limitations

This section discusses some of the limitations of our data and analysis in more detail.

C.2.1. DEFINING AI SUPERCOMPUTERS IS CHALLENGING

Ideally, our dataset would only capture systems that can efficiently run large-scale AI training workloads. However, it is difficult to develop a practical definition that captures only such systems based on limited publicly available data. Additionally, some companies, including Google DeepMind and OpenAI, have used AI chips distributed across multiple data center campuses to train large models (Moss, 2023; Dickson, 2025). To adequately include relevant AI supercomputers, we considered the following four definitions:

a) AI chips within a single building
b) AI chips on a single data center campus
c) AI chips within a fixed proximity (e.g., 2 or 5 miles)
d) No distance limit; an AI supercomputer is any system capable of training large models.

We decided to use definition (b), given the following considerations: The single building (a) may miss cases where well-connected accelerators span multiple buildings on the same campus. A fixed proximity definition (c) is not feasible in practice since we do not know the precise physical location of most of the AI supercomputers in the dataset. Finally, a functional definition (d) is difficult to scope because assessing if a given AI supercomputer meets certain thresholds for performance, connectivity, and integrated operation requires data on network architecture and connections between AI supercomputers that public reports almost never provide. At the same time, we think it is useful to include AI supercomputers that meet the theoretical performance threshold but lack adequate network infrastructure, given it is comparatively easy to retrofit the networking equipment (see Appendix C.2.2).

We thus adopt the contiguous campus definition (b), where accelerators on a contiguous campus linked by high-bandwidth networks operate as a single AI supercomputer. However, there are two remaining limitations to this definition:

- **Limited data:** Public reports seldom include details on facility boundaries or network topology, making it hard to verify the contiguous nature of a campus.[41] When we are unsure if a reported system may span several campuses, we mark the field "Single Cluster" as "Unclear" (20 entries). We mark the "Single Cluster" field as "No" if we think the report most likely refers to a decentralized system (8 entries).
- **Decentralized training:** Our dataset currently does not capture the fact that AI developers may use multiple AI supercomputers for a training run. To assess which AI supercomputers may be most suitable for decentralized training, we would need additional information on the network bandwidth between them.

C.2.2. THEORETICAL PERFORMANCE DOES NOT NECESSARILY CORRESPOND TO USEFULNESS FOR LARGE-SCALE TRAINING

Systems may lack sufficient networking for efficiently running AI training. Public performance figures do not guarantee efficient large-scale training. Some AI supercomputers may suffer from inadequate networking, which can reduce utilization

---

[41]We found it particularly challenging to verify this for reports from companies and for AI supercomputers in China.

and prolong training runs (Narayanan et al., 2021). However, systems with inadequate networking infrastructure can easily be upgraded by changing the network fabric, usually at a fraction ($\sim$10–20%) of the total AI supercomputer cost (Lepton AI, 2024).

Performance on AI training depends on the software stack. Our analysis compares theoretical performance across hardware types. In practice, actual performance depends on the software stack and how well the hardware supports it. For instance, despite having a higher theoretical performance, SemiAnalysis assessed that AMD's MI300X is less useful for large-scale AI training than NVIDIA's H100 (Patel et al., 2024). This software ecosystem gap becomes especially significant when evaluating AI supercomputers across different hardware platforms, as systems based on Chinese AI chips may not achieve their theoretical potential without the mature software infrastructure that NVIDIA's CUDA provides.

Theoretical performance does not fully capture AI inference performance. Our database focuses on systems suitable for AI training. A system's computation performance is not a good proxy for how well it can run AI inference workloads. NVIDIA's H20, for instance, delivers comparable inference performance to the H100 on certain workloads despite having only 1/7th the raw computational power, due to its high memory bandwidth. We recommend differentiating between FLOP/s (or OP/s for 8-bit and lower) when assessing training capabilities and memory bandwidth in Byte/s when assessing inference and long-context capabilities.

C.2.3. LIMITATIONS WITH OUR CHINESE DATA

Despite involving Chinese speakers in our data collection, we encountered several significant challenges in gathering comprehensive data on Chinese AI supercomputers.

1. **Official announcements often lack key data**, such as information on chip type and quantity. Furthermore, reported performance values often do not include precision.
2. **Sources sometimes report aggregate data for several AI supercomputers.** Computing zones that consist of several separate data center campuses sometimes report total computing capacity at an aggregate level rather than breaking down by individual AI supercomputers.
3. **Different conventions.** Chinese sources sometimes use different metrics and reporting standards than Western conventions, sometimes reporting the number of server racks that we cannot easily convert to chip numbers.

While we encounter similar issues for AI supercomputers in other countries, they are particularly common in China. However, we estimate that our database covers 10–20% of Chinese AI supercomputer performance, which is similar to our coverage estimate for U.S. data (see Appendix C.3).

C.2.4. CHINESE OWNERS MAY HAVE BECOME MORE SECRETIVE ABOUT THEIR AI SUPERCOMPUTERS, BUT THIS HAS NOT IMPACTED OUR DATA COVERAGE

In the late 2010s and early 2020s, Chinese supercomputer announcements frequently led to U.S. sanctions, with companies like Sugon, Phytium, and several national supercomputing centers being added to the Entity List due to concerns about military use of these systems (U.S. Bureau of Industry and Security, 2019; U.S. Department of Commerce, 2021). This is likely what caused China to release less information about its AI supercomputers. In 2022, China stopped submitting any systems to the Top500 list (Chik, 2022).

In October 2022, the U.S. first introduced export controls on AI chips and semiconductor manufacturing equipment with the goal of slowing down Chinese advances in AI (Allen, 2022). These export controls were strengthened in October 2023 and December 2024 by fixing loopholes and further restricting Chinese import of chip manufacturing tools (Dohmen & Feldgoise, 2023; Allen, 2024). Furthermore, to reduce chip smuggling, the United States introduced the AI Diffusion Framework in early 2025, requiring additional countries to file for a license to import U.S. AI chips (Heim, 2025). These actions may have incentivized Chinese owners to further increase secrecy about their AI supercomputers to reduce scrutiny from the United States, particularly if they deployed smuggled AI chips.

However, the effects of increased Chinese secrecy on our data coverage are limited. While we see a decrease in the number of Chinese systems added to our database in 2021 and 2022, the number of Chinese systems increased again in 2024 (Figure 17). Comparing the aggregate performance in our database with IDC (2025)'s estimate of total 16-bit FLOP/s in China indicates that our coverage was consistently between 10 and 20% of Chinese performance (see Table 5).

Number of AI supercomputers added each year in the United States and China
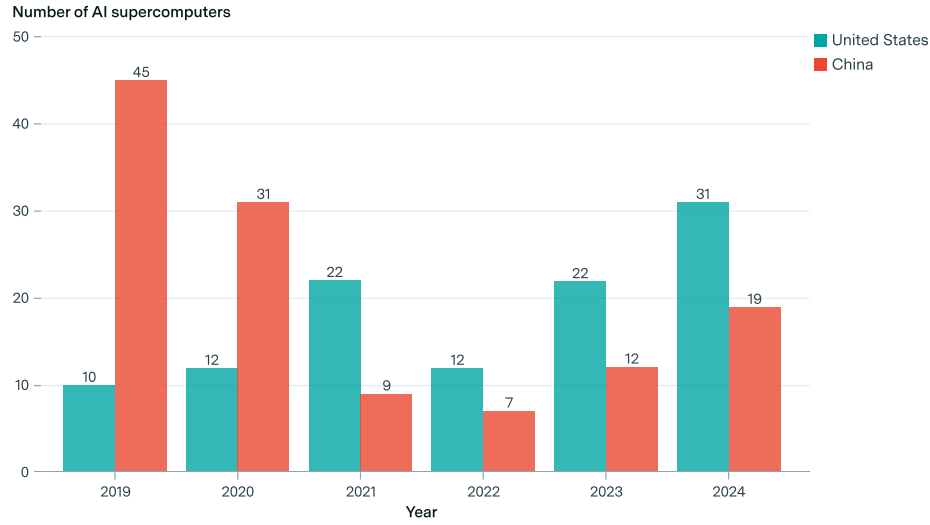
Figure 17: Number of Chinese and U.S. systems added each year.

C.3. Comparing our data with public reports

To assess what fraction of AI supercomputer capacity we capture in the dataset and how our coverage differs between chip types and companies, we compare our data to four sources of public information:

- Estimates of the total production of AI chips.
- Estimates of the total AI chip stock of companies.
- An estimate of the total 16-bit FLOP/s in China by IDC (2025).
- The fraction of the largest publicly known AI models that were likely trained on an AI supercomputer in our dataset.

C.3.1. ESTIMATING THE COVERAGE OF ALL AI SUPERCOMPUTERS BASED ON TOTAL CHIP PRODUCTION

One relevant reference point for our coverage is what fraction of total production we cover for different chip types (Table 3). While some AI chips may be sold to individuals and small research groups, we expect that the vast majority of all AI chips will be used in AI supercomputers that would fall within our definition.

Table 3: Variation of coverage by chip type based on public reports of AI chip production until 2025. Note that the public estimates may include chips in AI supercomputers that are not yet operational or that are otherwise outside of our inclusion criteria. Our full dataset includes potentially existing and planned systems and has a higher coverage. Note that we explicitly search for the H100, A100, and V100 in our automated methodology. This may marginally increase our coverage of these three chip types compared to others.

CHIP TYPE	PUBLIC ESTIMATE	DATASET	IMPLIED COVERAGE
H100/H200	2.5M − 4.5M[42]	830K	36.5% − 20.3%
A100	1.5M − 3M[43]	234K	16.1% − 8.1%
H20	1M[44]	−[45]	0%
H800/A800	>200K[46]	2K	<1.5%
AMD MI300	400K[47]	72K	18%
GOOGLE TPUs	>4M[48]	95K	<4%
OTHER CUSTOM SILICON	?[49]	4K	?%
TOTAL	9.6 − 13.1M	1.2M	9.2% − 12.5%

Based on the public sources used in the table, our dataset covers between 20% and 37% of all NVIDIA H100s produced until late 2024.[50] However, coverage is much worse for NVIDIA's H20, A800 and H800, Google's TPUs, and other custom silicon chips. The average coverage is about 10%. (Note that the table above only includes confirmed operational AI supercomputers. Our dataset also contains planned AI supercomputers that make up another 920k H100s and 33k MI300X. Some of those may include chips already included in the production volume estimates.)

Table 3 reveals that our dataset likely covers H100, A100, and MI300 equally well, whereas coverage of Google's TPUs and other custom silicon chips is significantly worse. This is expected, given that NVIDIA and AMD sell their chips to a wide range of customers, incentivizing them to report about successful projects to attract more customers. Meanwhile, Google and other hyperscalers only deploy their chips within the company, offering limited incentives to publish more than a few large AI supercomputers.

C.3.2. COVERAGE BY COMPANY

Another reference point for our coverage is comparing our chip numbers to the publicly reported numbers of chips acquired by different companies (Table 4). We expect that hyperscalers deploy most of their AI chips in AI supercomputers covered by our definition, since even when primarily running inference workloads, they usually deploy thousands of AI chips in the same data center. (Note that the March 2025 inclusion threshold was at 2,000 H100-equivalents but was below 1,000 H100-equivalents until August 2024.)

Table 4: Public reports of number of chips owned for various companies at the end of 2024 and comparison with our dataset[51]

COMPANY	PUBLIC CLAIM	OUR DATASET	IMPLIED COVERAGE
META	350K H100S	149K	42.8%
MICROSOFT	475K – 855K H100[52]	118K	14% – 25%
AWS	200K H100S (IN 2024)	–	0%
GOOGLE	170K H100S (IN 2024)	8K	4.7%
APPLE	180K H100S[53]	–	0%
COREWEAVE	175K GPUS[54]	57K	22.8%
BYTEDANCE	310K HOPPERS[55]	8K	3%
TENCENT	230K HOPPERS (IN 2024)	–[56]	0%
TOTAL	2.09M – 2.47M	0.34 M	13.8 – 16.3%

Note: Public estimates cannot be verified and only serve as an approximate assessment of coverage. Some sources are inconsistent with others.

[42]Public sources estimate that NVIDIA shipped about 500k H100s in 2023 and 2 million in 2024, for a total of 2.5 million H100s (Nolan, 2023; Shilov, 2023c). However, Garreffa (2024) estimates NVIDIA produced up to 1.5 million H100s in Q4 of 2024. Assuming NVIDIA produced about 1M H100s on average per quarter in 2024 yields a total of 4.5 million H100s.

[43]Reports on how many A100s NVIDIA produced are limited, but the company reportedly shipped 500k in Q3 2023 (Shilov, 2023a). The A100 was first produced in 2020 and likely reached peak production in 2023 before demand reduced in 2024. It thus seems plausible that NVIDIA produced between 1.5 – 3 million A100s until 2025.

[44]Financial Times (2023)

[45]We capture 30k H20s that DeepSeek likely owns, but exclude these from the analysis because we are uncertain if they are in the same location.

[46]Public reports indicate Chinese companies spent $5 billion on NVIDIA H800 and A800 in 2023 (Pires, 2023a), indicating at least 200k of these chips imported (conservative estimate assuming $25k average price per chip (Champelli et al., 2024) .)

[47]AMD to ship up to 400,000 new AI GPUs in 2024 (Chen & Chan, 2023).

[48]Google's internal TPU production likely reached 2 million TPUs in 2023 (Martin, 2024), although public data is severely limited, given Google does not sell TPUs to outside companies. Assuming a similar production in 2024, there would be at least 4 million TPUs.

[49]Microsoft, AWS, and Meta all developed their own custom silicon AI chips deployed in-house (Borkar et al., 2024; AWS, undated; Tal et al., 2024), but we were unable to find trustworthy public estimates of the total numbers.

[50]We do not account for H100s produced in 2025, since these would unlikely be installed in any systems before our March 1st cutoff.

Table 4 shows that our coverage differs considerably between companies. While we cover almost half of Meta's H100s, we cover only 5% of Google's and none of Apple's H100s. Our data is particularly limited for Chinese hyperscalers. However, Table 4 does not consider AI supercomputers we cover based on reported performance, but for which we lack the specific chip type. This is especially common for Chinese systems.

C.3.3. COVERAGE OF CHINESE DATA

To assess data coverage of AI supercomputers in China, we compare the aggregate 16-bit performance of all Chinese systems in our database to the total Chinese 16-bit performance published in a 2025 report by market intelligence firm International Data Corporation (IDC, 2025). We find that we cover between 10 and 20% of Chinese 16-bit performance between the end of 2020 and the end of 2024 (Table 5). Not all 16-bit performance would likely fall under the definition of our database, so actual coverage of AI supercomputers is likely somewhat higher.

Table 5: FP16 Performance

	OUR DATA	**IDC**	**IMPLIED COVERAGE**
2020	1.05×10^{19}	7.50×10^{19}	14%
2021	1.88×10^{19}	1.55×10^{20}	12%
2022	3.46×10^{19}	2.60×10^{20}	13%
2023	4.18×10^{19}	4.17×10^{20}	10%
2024	1.46×10^{20}	7.25×10^{20}	20%

We were unable to find reliable total performance estimates for other countries, so we had to limit our coverage analysis by FLOP/s to Chinese data.

C.3.4. COVERAGE OF AI SUPERCOMPUTERS USED IN THE LARGEST TRAINING RUNS

To check how well our dataset covers the AI supercomputers used for known large training runs, we check which of the 25 largest training runs in Epoch AI's notable AI models dataset (as of 1 March 2025) correspond to AI supercomputers in our dataset. (Note that our dataset uses the models dataset as a data source. To avoid circularity we distinguish between systems reported independently from the training run and systems included in our dataset based exclusively on the reports of the training run.)

We find that for about half of the largest AI training runs, we capture an AI supercomputer that could have plausibly been used or was confirmed to be used in the training run (Figure 18; Table 6).

Our data coverage is slightly better for Chinese AI supercomputers, where we find plausible AI supercomputers for about two thirds of all reported models (Figure 18; Table 7).

---

[51]Note we only include systems in our analysis if we are confident they exist in a single site rather than a distributed system. E.g., AWS announced 20k H100 clusters in 2023, but did not explicitly say whether or not those were on the same data center campus.

[52]Microsoft likely made up 19% of total 2023 NVIDIA revenue (Fox, 2024). We assume they maintained a 19% share of revenue throughout 2024, and bought a mix of NVIDIA data center products that is approximately equal to NVIDIA's sales mix. Based on estimates for H100 shipments in our previous section, this indicates Microsoft owns between 475k and 855k H100s.

[53]~2,500 servers in 2023 and 20,000 servers in 2024 * 8 GPUs per server = 180k.

[54]Estimate, given the claim that most of the 250,000 total GPUs said to be H100s and some H200s (Morgan, 2025).

[55]About 50k H100 in 2023 and 240k in 2024 (Pires, 2023b; Alexsandar K, 2024).

[56]We identified two Tencent AI supercomputers but were unable to identify the performance or hardware used.

Coverage of AI supercomputers used for the largest AI training runs

Top-25 AI training runs (all countries)   Chinese top-25 AI training runs
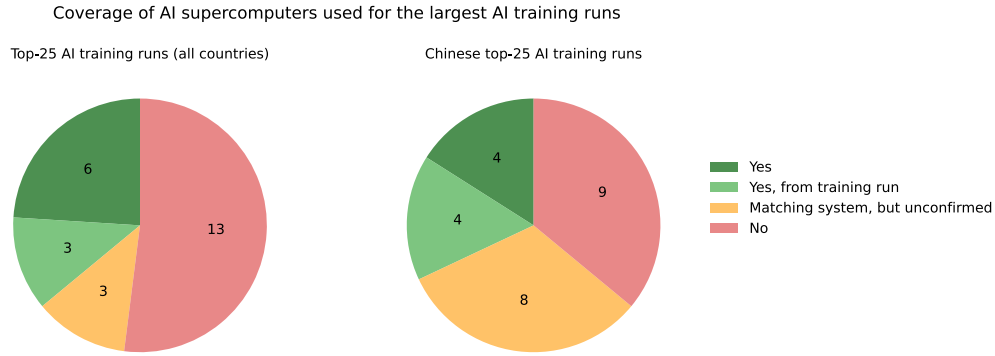

Figure 18: Coverage of AI supercomputers used for the largest AI training runs according to Epoch AI's notable models dataset. "Yes, from training run" indicates we cover the AI supercomputer, but only based on reports about the training runs itself. "Matching system, but unconfirmed" means an AI supercomputer in our dataset was likely used by the model developer but we find no public reports on whether or not the system was actually used for the training run.

Table 6: Coverage of largest AI training runs (all countries) according to Epoch AI's notable model dataset

TRAINING RUN	COVERED	NOTE
GROK-3	Yes	Trained on Colossus in Memphis, Tennessee
GEMINI 1.0 ULTRA	Yes, from training run	
GPT-4O	No	
LLAMA 3.1-405B	Yes	Presumably trained on Meta GenAI 2024a or 2024b (Oldham et al., 2024)
CLAUDE 3.5 SONNET	No	
GLM-4-PLUS	No	
CLAUDE 3.7 SONNET	No	
GROK-2	Matching AI supercomputer, but unconfirmed	Trained on the Oracle Cloud. "Oracle OCI Supercluster H100s" matches the description of the training details (Trueman, 2024)
DOUBAO-PRO	No	
GPT-4 TURBO	No	Possibly trained on same AI supercomputer as GPT-4, but no confirmation
MISTRAL LARGE 2	No	
GPT-4	Yes	Likely trained on Iowa AI supercomputer (O'Brien & Fingerhut, 2023). Entered in the dataset as "Microsoft GPT-4 cluster"
NEMOTRON-4 340B	Matching AI supercomputer, but unconfirmed	"NVIDIA CoreWeave Eos-DFW" appears to match the training description
CLAUDE 3 OPUS	No	
GEMINI 1.5 PRO	No	We capture several systems from Google, but none were likely used for this model
GLM-4 (0116)	No	
MISTRAL LARGE	Yes	Likely used Leonardo
ARAMCO METABRAIN AI	No	
INFLECTION-2	Yes, from training run	
INFLECTION-2.5	No	We capture several of Inflection's systems, but none were confirmed
REKA CORE	Yes, from training run	
LLAMA 3.1-70B	Yes	Presumably trained on Meta GenAI 2024a or 2024b (Oldham et al., 2024)
LLAMA 3-70B	Yes	Trained on Meta GenAI 2024a or 2024b (Oldham et al., 2024)
QWEN2.5-72B	Matching AI supercomputer, but unconfirmed	
GPT-4O MINI	No	

36

Table 7: Coverage of AI supercomputers used for the largest AI training runs in China according to Epoch AI (2025) as of March 2025.

MODEL	COVERED
GLM-4-PLUS	No
DOUBAO-PRO	No
GLM-4 (0116)	No
QWEN2.5-72B	Matching AI supercomputer, but unconfirmed
TELECHAT2-115B	Matching AI supercomputer, but unconfirmed
DEEPSEEK-V3	Yes
DEEPSEEK-R1	Yes
MEGASCALE (PRODUCTION)	Yes, from training run
SENSECHAT	Yes
QWEN2.5-32B	Matching AI supercomputer, but unconfirmed
HUNYUAN-LARGE	No
QWEN2-72B	Matching AI supercomputer, but unconfirmed
YI-LARGE	No
DEEPSEEK-V2.5	Matching AI supercomputer, but unconfirmed
YI-LIGHTNING	Yes, from training run
QWEN1.5-72B	Matching AI supercomputer, but unconfirmed
QWEN-72B	Matching AI supercomputer, but unconfirmed
XVERSE-65B-2	No
HUNYUAN	No
LUCA 2.0	No
QWEN2.5-CODER (32B)	Matching AI supercomputer, but unconfirmed
BLUELM 175B	No
ERNIE 3.0 TITAN	Yes
MEGASCALE (530B)	Yes, from training run
XTRIMOPGLM -100B	Yes, from training run

D. Additional data

D.1. Overview of trends in different precisions and by sector.

Table 8: Overview of key trends between 2019 and March 2025. Square brackets indicate the 90% confidence interval. Note 8-bit trend is only starting in July 2021.[57]

LEADING AI SUPERCOMPUTERS (INCLUDING BOTH PUBLIC AND PRIVATE)		
	16-BIT OP/S	**8-BIT OP/S**	**MAX OP/S**
PERFORMANCE GROWTH	**2.54** [2.35–2.74]	**2.60** [2.31–2.93]	**2.55** [2.34–2.78]
NUMBER OF CHIPS	**1.60** [1.45–1.78]	**1.69** [1.47–1.94]	**1.46** [1.29–1.64]
PERFORMANCE PER CHIP	**1.60** [1.49–1.71]	**1.54** [1.42–1.67]	**1.77** [1.62–1.94]
HARDWARE COST	**1.92** [1.76–2.11]	**1.99** [1.72–2.30]	**1.76** [1.58–1.97]
COST-PERFORMANCE	**1.36** [1.29–1.42]	**1.37** [1.29–1.45]	**1.51** [1.43–1.60]
POWER	**1.95** [1.77–2.15]	**2.12** [1.85–2.42]	**1.78** [1.60–1.99]
ENERGY EFFICIENCY	**1.34** [1.25–1.43]	**1.26** [1.20–1.32]	**1.51** [1.39–1.63]

LEADING PRIVATE AI SUPERCOMPUTERS		
	16-BIT OP/S	**8-BIT OP/S**	**MAX OP/S**
PERFORMANCE GROWTH	**2.69** [2.47–2.92]	**3.17** [2.78–3.61]	**3.00** [2.76–3.27]
NUMBER OF CHIPS	**1.82** [1.66–2.00]	**2.14** [1.85–2.47]	**1.83** [1.65–2.03]
PERFORMANCE PER CHIP	**1.50** [1.44–1.57]	**1.48** [1.36–1.61]	**1.65** [1.55–1.76]
HARDWARE COST	**2.06** [1.88–2.26]	**2.39** [2.09–2.73]	**2.05** [1.86–2.26]
COST-PERFORMANCE	**1.33** [1.28–1.39]	**1.32** [1.26–1.39]	**1.47** [1.41–1.54]
POWER	**2.16** [1.98–2.35]	**2.57** [2.26–2.93]	**2.16** [1.96–2.37]
ENERGY EFFICIENCY	**1.27** [1.23–1.31]	**1.23** [1.19–1.28]	**1.40** [1.34–1.46]

LEADING PUBLIC AI SUPERCOMPUTERS		
	16-BIT OP/S	**8-BIT OP/S**	**MAX OP/S**
PERFORMANCE GROWTH	**1.86** [1.60–2.15]	**1.79** [1.46–2.19]	**1.90** [1.63–2.22]
NUMBER OF CHIPS	**1.21** [0.98–1.50]	**1.20** [0.96–1.49]	**1.11** [0.89–1.38]
PERFORMANCE PER CHIP	**1.56** [1.34–1.82]	**1.48** [1.31–1.67]	**1.75** [1.45–2.11]
HARDWARE COST	**1.40** [1.25–1.57]	**1.34** [1.09–1.65]	**1.38** [1.20–1.58]
COST-PERFORMANCE	**1.41** [1.28–1.56]	**1.48** [1.32–1.66]	**1.51** [1.32–1.73]
POWER	**1.41** [1.17–1.70]	**1.38** [1.10–1.74]	**1.31** [1.07–1.61]
ENERGY EFFICIENCY	**1.38** [1.19–1.61]	**1.33** [1.19–1.47]	**1.56** [1.28–1.90]

D.2. Chip types in our dataset

This section covers some additional statistics about the chip types in our dataset. Given the high variance in coverage of different companies and chip types, this data is likely not representative of the broader field.

The majority of chips captured in our dataset are NVIDIA's Hopper, Ampere and Volta chips (Figure 19; Table 9). When grouping similar chips together (such as the H100 and H200), our dataset contains quantities of 27 unique chips. The table below shows all chip types that contribute more than 10,000 chips to our dataset in aggregate. (Note some entries capture chip type but without a known quantity; those are excluded from the table.)

Note that we explicitly search for the H100, A100, and V100 in our automated methodology. This may somewhat increase our coverage of these three chip types compared to others.

---

[57]We assess this trend only after 50 AI supercomputers in our dataset support 8-bit precision.
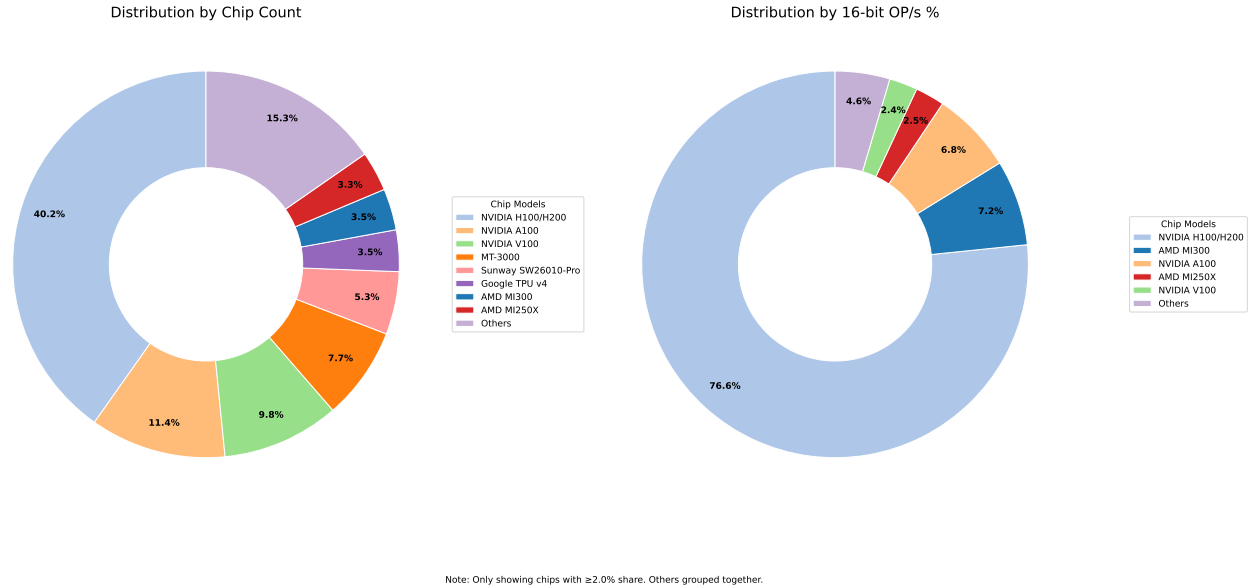
Figure 19: Chip types in operational AI supercomputers in our dataset; by chip count (left) and by FLOP/s (right). Showing only shares above 2%.

Table 9: Table including all chips totaling more than 10,000 in our dataset. Note that this only includes AI supercomputers where chip identity is known. Our dataset also includes AI supercomputers with chip quantities where the chip type is unknown; these account for 4% of total performance. Performance share is in 16-bit FLOP/s

CHIP MODEL	TOTAL NUMBER	PERFORMANCE SHARE (%)
NVIDIA H100/H200	830,000	76.6
NVIDIA A100	235,000	6.8
NVIDIA V100	203,000	2.4
MT-3000	160,000	0.0
SUNWAY SW26010-PRO	109,000	0.6
GOOGLE TPU V4	72,000	1.8
AMD MI300	72,000	7.2
AMD MI250X	69,000	2.5
SHENSUAN-1	64,000	0.2
INTEL MAX 1550	64,000	0.3
NVIDIA TESLA K20X	26,000	0.0
GROQCHIP LPU V1	20,000	0.2
NVIDIA TESLA K40C	10,000	0.0

When sorting chips by manufacturer, we find that NVIDIA chips make up about 75% of total performance in our dataset (Table 10; Figure 20). This is consistent with NVIDIA's 2024 AI chip market share of about 70 - 95%. Chinese-designed chips make up less than 2% of the performance in our dataset. However, for Chinese AI supercomputers, we disproportionally lack information about hardware. Thus, a significant share of the unknown chips are likely of Chinese origin.

Table 10: Table including all chips totaling more than 10,000 in our dataset. Note this only includes AI supercomputers where chip identity is known. The dataset also includes AI supercomputers with chip quantities where the chip type is unknown; these account for 4% of the dataset by performance. Note that the performance share is in 16-bit performance.

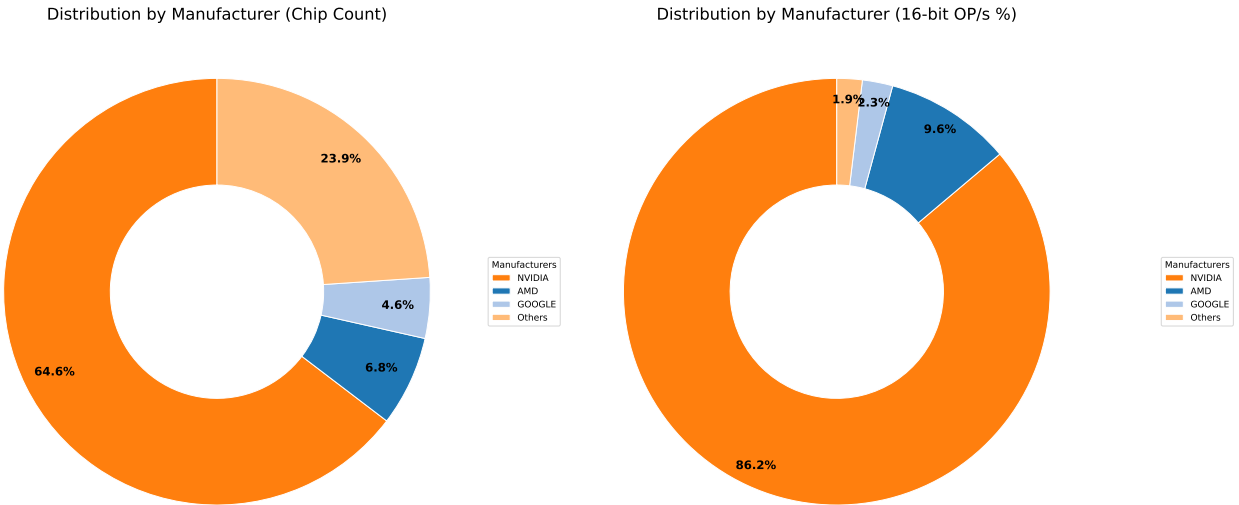| MANUFACTURER	TOTAL	PERFORMANCE SHARE (%)
NVIDIA	1,334,962	86.1
AMD	141,120	9.6
SUNWAY	108,544	0.6
GOOGLE	94,856	2.3
INTEL	67,744	0.5
SHENSUAN	64,000	0.2
GROQ	19,725	0.2
OTHER	215,584	0.0

Figure 20: Manufacturer distribution by chip count (left) and by performance (right). We only show manufacturers that contributed at least 2% of total performance in our dataset.

This is consistent with the hardware recorded in Epoch AI (2025), where the vast majority of models with known hardware were trained on NVIDIA chips (Table 11). However, note that the dataset does not focus on hardware, and therefore, the coverage is incomplete.

Table 11: Share of models trained on hardware produced by different manufacturers. Note that hardware information is not recorded for most models.

MANUFACTURER	NUMBER OF MODELS	SHARE
UNKNOWN	226	50.8%
NVIDIA	144	25.6%
GOOGLE	72	16.2%
HUAWEI	2	0.4%
AMD	1	0.2%

References

Alexsandar K. Microsoft Acquired Nearly 500,000 NVIDIA "Hopper" GPUs This Year, December 2024. URL https://www.techpowerup.com/330027/microsoft-acquired-nearly-500-000-nvidia-hopper-gpus-this-year. Accessed 15-04-2025.

Allen, G. C. Choking off China's Access to the Future of AI, October 2022. URL https://www.csis.org/analysis/choking-chinas-access-future-ai. Accessed 20-04-2025.

Allen, G. C. Understanding the Biden Administration's Updated Export Controls, December 2024. URL https://www.csis.org/analysis/understanding-biden-administrations-updated-export-controls. Accessed 20-04-2025.

Allen, G. C., Adamson, G., Heim, L., and Winter-Levy, S. The United Arab Emirates' AI Ambitions, January 2025. URL https://www.csis.org/analysis/united-arab-emirates-ai-ambitions. Accessed 20-04-2025.

Ashkboos, S., Markov, I., Frantar, E., Zhong, T., Wang, X., Ren, J., Hoefler, T., and Alistarh, D. Quik: Towards end-to-end 4-bit inference on generative large language models. *arXiv:2310.09259*, 2023. URL https://arxiv.org/abs/2310.09259.

AWS. Amazon Web Services (AWS) - Cloud Computing Services, 2020. URL https://pages.awscloud.com/amazon-ec2-p4d.html. Accessed 15-04-2025.

AWS. Project Ceiba – Largest AI Super Computer Co-Built with NVIDIA, 2023. URL https://aws.amazon.com/nvidia/project-ceiba/. Accessed 15-04-2025.

AWS. AI Accelerator - AWS Trainium - AWS, undated. URL https://aws.amazon.com/ai/machine-learning/trainium/. Accessed 15-04-2025.

Baker, M. Nuclear arms control verification and lessons for AI treaties. *arXiv:2304.04123*, 2023. URL https://arxiv.org/abs/2304.04123.

BeautifulSoup. beautifulsoup4, 2025. URL https://pypi.org/project/beautifulsoup4/. Accessed 15-04-2025.

Benhari, A., Trystram, D., Dufossé, F., Denneulin, Y., and Desprez, F. Green HPC: An analysis of the domain based on Top500. *arXiv:2403.17466*, 2024. URL https://arxiv.org/pdf/2403.17466.

Benito, A. Saudi Arabia launches $100 Billion AI initiative to lead in global tech, October 2024. URL https://www.cio.com/article/3602900/saudi-arabia-launches-100-billion-ai-initiative-to-lead-in-global-tech.html. Accessed 20-04-2025.

Besiroglu, T., Bergerson, S. A., Michael, A., Heim, L., Luo, X., and Thompson, N. The compute divide in machine learning: A threat to academic contribution and scrutiny? *arXiv:2401.02452*, 2024. URL https://arxiv.org/abs/2401.02452.

Borkar, R., Wall, A., Pulavarthi, P., and Yu, Y. Azure Maia for the era of AI: From silicon to software to systems — Microsoft Azure Blog, 2024. URL https://azure.microsoft.com/en-us/blog/azure-maia-for-the-era-of-ai-from-silicon-to-software-to-systems/. Accessed 15-04-2025.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. URL https://arxiv.org/abs/2005.14165.

Champelli, P., Niiler, E., and Fitch, A. The Nvidia Chips Inside Powerful AI Supercomputers, 2024. URL https://www.wsj.com/tech/ai/nvidia-chip-technology-artificial-intelligence-006e29d4. Accessed 15-04-2025.

Chang, J., Lu, K., Guo, Y., Wang, Y., Zhao, Z., Huang, L., Zhou, H., Wang, Y., Lei, F., and Zhang, B. A survey of compute nodes with 100 TFLOPS and beyond for supercomputers. *CCF Transactions on High Performance Computing*, 6(3): 243–262, 2024. URL https://link.springer.com/article/10.1007/s42514-024-00188-w.

Chen, M. and Chan, R. AMD to ship up to 400,000 new AI GPUs in 2024, say sources, 2023. URL https://www.digitimes.com/news/a20231205PD217/amd-ai-gpu-2024-us-china-chip-ban.html. Accessed 15-04-2025.

Chik, H. No sign of China's new supercomputers among world's Top500. South China Morning Post, November 2022. URL https://www.scmp.com/news/china/science/article/3180337/no-sign-chinas-new-supercomputers-among-worlds-top500. Accessed 20-04-2025.

Cottier, B., Rahman, R., Fattorini, L., Maslej, N., Besiroglu, T., and Owen, D. The rising costs of training frontier AImodels. *arXiv:2405.21015*, 2024. URL https://arxiv.org/pdf/2405.21015v2.

Cushman & Wakefield. Data Center Development Cost Guide 2025, 2025. URL https://cushwake.cld.bz/Data-Center-Development-Cost-Guide-2025/8-9/. Accessed 15-04-2025.

Dickson, B. GPT-4.5 for enterprise: Are accuracy and knowledge worth the high cost?, January 2025. URL https://venturebeat.com/ai/gpt-4-5-for-enterprise-do-its-accuracy-and-knowledge-justify-the-cost/. Accessed 20-04-2025.

Dohmen, H. and Feldgoise, J. A Bigger Yard, A Higher Fence: Understanding BIS's Expanded Controls on Advanced Computing Exports, October 2023. URL https://cset.georgetown.edu/article/bis-2023-update-explainer/. Accessed 20-04-2025.

Dohmke, T. and GitHub. Introducing GitHub Copilot: Your AI Pair Programmer, June 2021. URL https://github.blog/news-insights/product-news/introducing-github-copilot-ai-pair-programmer/. Accessed 19-04-2025.

Dong, Z., Wang, Z., Xu, J., Tang, R., and Wen, J. A brief history of recommender systems. *arXiv:2209.01860*, 2022. URL https://arxiv.org/abs/2209.01860.

Dongarra, J. J. The linpack benchmark: An explanation. In *International Conference on Supercomputing*, pp. 456–474. Springer, 1987. URL https://link.springer.com/chapter/10.1007/3-540-18991-2_27.

Epoch AI. Data on Machine Learning Hardware, 10 2024. URL https://epoch.ai/data/machine-learning-hardware. Accessed: 18-04-2025.

Epoch AI. Data on Notable AI Models. Epoch AI Data Hub, 2025. URL https://epochai.org/data/notable-ai-models. Accessed 20-04-2025.

Erdil, E. and Besiroglu, T. Algorithmic progress in computer vision. *arXiv:2212.05153*, 2022. URL https://arxiv.org/abs/2212.05153.

Financial Times. Nvidia to make $12bn from AI chips in China this year despite US controls, 2023. URL https://www.ft.com/content/b76ef55b-21cd-498b-ac16-5660908bb8d2. Accessed 15-04-2025.

Fist, T. and Datta, A. How to Build the Future of AI in the United States, October 2024. URL https://ifp.org/future-of-ai-compute/. Accessed 20-04-2025.

Fox, M. A single customer made up 19% of Nvidia's revenue last year. UBS thinks it's Microsoft, 2024. URL https://www.businessinsider.com/nvidia-stock-mystery-customer-microsoft-ubs-revenue-h100-gpu-chips-2024-5. Accessed 15-04-2025.

Frymire, L. The length of time spent training notable models is growing. Epoch AI, 2024. URL https://epoch.ai/data-insights/training-length-trend. Accessed 20-04-2025.

Galabov, V., Sukumaran, M., and Lewis, A. Data Center Server Market Insights and Forecast. Technical report, Omdia, 2025. URL https://omdia.tech.informa.com/collections/afcei005/data-center-server-market-insights-and-forecast. Accessed 20-04-2025.

Garreffa, A. Analyst says NVIDIA Blackwell GPU production volume will hit 750K to 800K units by Q1 2025, 2024. URL https://www.tweaktown.com/news/100980/analyst-says-nvidia-blackwell-gpu-production-volume-will-hit-750k-to-800k-units-by-q1-2025/index.html. Accessed 15-04-2025.

Gartner. Gartner Says Worldwide IaaS Public Cloud Services Market Grew 37.3% in 2019, August 2020. URL https://www.gartner.com/en/newsroom/press-releases/2020-08-10-gartner-says-worldwide-iaas-public-cloud-services-market-grew-37-point-3-percent-in-2019. Accessed 20-04-2025.

Gonsalves, A. Amazon to spend $100B on AWS AI infrastructure. TechTarget, February 2025. URL https://www.techtarget.com/searchenterpriseai/news/366619057/Amazon-to-spend-100B-in-AWS-AI-infrastructure. Accessed 20-04-2025.

Grand View Research. AI Infrastructure Market Size, Share & Growth Report, 2030, 2024. URL https://www.grandviewresearch.com/industry-analysis/ai-infrastructure-market-report. Accessed 15-04-2025.

Grunewald, E. Introduction to AI Chip Making in China. Institute for AI Policy and Strategy, July 2023. URL https://www.iaps.ai/research/ai-chip-making-china. Accessed 20-04-2025.

Grunewald, E. AI Chip Smuggling is the Default, not the Exception. AI Policy Bulletin, January 2025. URL https://www.aipolicybulletin.org/articles/ai-chip-smuggling-is-the-default-not-the-exception. Accessed 20-04-2025.

Heim, L. Understanding the Artificial Intelligence Diffusion Framework, January 2025. URL https://www.rand.org/pubs/perspectives/PEA3776-1.html. Accessed 20-04-2025.

Heim, L. and Egan, M. Accessing Controlled AI Chips via Infrastructure-as-a-Service (IaaS): Implications for Export Controls. Center for the Governance of AI, November 2023. URL https://cdn.governance.ai/Accessing_Controlled_AI_Chips_via_Infrastructure-as-a-Service.pdf.

Ho, A., Tamay, B., Erdil, E., Owen, D., Rahman, R., Guo, Z. C., Atkinson, D., Thompson, N., and Sevilla, J. Algorithmic Progress in Language Models, 2024. URL https://epoch.ai/blog/algorithmic-progress-in-language-models. Accessed 20-04-2025.

Hobbhahn, M., Heim, L., and Aydos, B. Trends in Machine Learning Hardware, 2023. URL https://epoch.ai/blog/trends-in-machine-learning-hardware. Accessed 20-04-2025.

Hochman, T. Building Baseload: Reforming Permitting for AI Energy Infrastructure, 2020. URL https://www.thefai.org/posts/building-baseload-reforming-permitting-for-ai-energy-infrastructure. Accessed 20-04-2025.

Hu, K. ChatGPT sets record for fastest-growing user base - analyst note. Reuters, February 2023. URL https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/. Accessed 20-04-2025.

Huang, Q., Bao, B., Liang, Y., Guo, X., Huang, M., Tekur, C., and Carilli, M. Introducing native PyTorch automatic mixed precision for faster training on NVIDIA GPUs, September 2020. URL https://pytorch.org/blog/accelerating-training-on-nvidia-gpus-with-pytorch-automatic-mixed-precision/. Accessed 20-04-2025.

IDC. Artificial Intelligence Infrastructure Spending to Surpass the $200bn USD Mark in the Next 5 years, According to IDC. Technical report, International Data Corporation, February 2025. URL https://www.idc.com/getdoc.jsp?containerId=prUS52758624. Accessed 20-04-2025.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. Highly accurate protein structure prediction with Alphafold. *Nature*, 596(7873):583–589, 2021. URL https://www.nature.com/articles/s41586-021-03819-2.

Khan, S. and Mann, R. AI Chips: What They Are and Why They Matter. Georgetown's Center for Security and Emerging Technology, April 2020. URL https://cset.georgetown.edu/publication/ai-chips-what-they-are-and-why-they-matter/.

Langston, J. Microsoft announces new supercomputer, lays out vision for future AI work - Source, 2020. URL https://news.microsoft.com/source/features/ai/openai-azure-supercomputer/. Accessed 15-04-2025.

Lee, J. Microsoft Azure Eagle is a Paradigm Shifting Cloud Supercomputer, 2023. URL https://www.servethehome.com/microsoft-azure-eagle-is-a-paradigm-shifting-cloud-supercomputer-nvidia-intel/. Accessed 15-04-2025.

Lepton AI. The Missing Guide to the H100 GPU Market, January 2024. URL https://blog.lepton.ai/the-missing-guide-to-the-h100-gpu-market-91ebfed34516. Accessed 20-04-2025.

Lin, A. and Heim, L. Deepseek's Lesson: America Needs Smarter Export Controls, February 2025. URL https://www.rand.org/pubs/commentary/2025/02/deepseeks-lesson-america-needs-smarter-export-controls.html. Accessed 20-04-2025.

Lohn, A. Scaling AI — Cost and Performance of AI at the Leading Edge. Georgetown's Center for Security and Emerging Technology, December 2023. URL https://cset.georgetown.edu/publication/scaling-ai/.

Luszczek, P. Results, November 2024. URL https://hpl-mxp.org/results.md. Accessed 15-04-2025.

Mahmood, Y., Byrd, C., Somani, E., Pilz, K. F., and Heim, L. Possible Options for Unlocking and Securing U.S. Energy for AI Production, March 2025. URL https://www.rand.org/pubs/working_papers/WRA3883-1.html. Accessed 20-04-2025.

Martin, D. Google Was Third Biggest Data Center Processor Supplier Last Year: Research, 2024. URL https://www.crn.com/news/components-peripherals/2024/google-was-third-biggest-data-center-processor-supplier-last-year-research. Accessed 15-04-2025.

Mattson, P., Cheng, C., Diamos, G., Coleman, C., Micikevicius, P., Patterson, D., Tang, H., Wei, G.-Y., Bailis, P., Bittorf, V., et al. Mlperf training benchmark. *Proceedings of Machine Learning and Systems*, 2:336–349, 2020. URL https://arxiv.org/abs/1910.01500.

Meta. Introducing the AI Research SuperCluster — Meta's cutting-edge AI supercomputer for AI research, 2022. URL https://ai.meta.com/blog/ai-rsc/. Accessed 15-04-2025.

Micikevicius, P., Narang, S., Alben, J., Diamos, G., Elsen, E., Garcia, D., Ginsburg, B., Houston, M., Kuchaiev, O., Venkatesh, G., et al. Mixed precision training. *arXiv:1710.03740*, 2017. URL https://arxiv.org/abs/1710.03740.

Morgan, T. P. Stacking Up AMD MI200 Versus Nvidia A100 Compute Engines, December 2021. URL https://www.nextplatform.com/2021/12/06/stacking-up-amd-mi200-versus-nvidia-a100-compute-engines/. Accessed 15-04-2025.

Morgan, T. P. Energy Giant Eni Boosts Its HPC Oomph By An Order Of Magnitude, 2024. URL https://www.nextplatform.com/2024/01/24/energy-giant-eni-boosts-its-hpc-oomph-by-an-order-of-magnitude/. Accessed 15-04-2025.

Morgan, T. P. CoreWeave's 250,000-Strong GPU Fleet Undercuts The Big Clouds, 2025. URL https://www.nextplatform.com/2025/03/05/coreweaves-250000-strong-gpu-fleet-undercuts-the-big-clouds/. Accessed 15-04-2025.

Moss, S. Training Google's Gemini: TPUs, multiple data centers, and risks of cosmic rays. DCD, December 2023. URL https://www.datacenterdynamics.com/en/news/training-gemini-tpus-multiple-data-centers-and-risks-of-cosmic-rays/. Accessed 20-04-2025.

Moss, S. Microsoft & OpenAI consider $100bn, 5GW 'Stargate' AI data center - report, 2024. URL https://www.datacenterdynamics.com/en/news/microsoft-openai-consider-100bn-5gw-stargate-ai-data-center-report/. Accessed 15-04-2025.

Narayanan, D., Shoeybi, M., Casper, J., LeGresley, P., Patwary, M., Korthikanti, V. A., Vainbrand, D., Kashinkunti, P., Bernauer, J., Catanzaro, B., Phanishayee, A., and Zaharia, M. Efficient large-scale language model training on gpu clusters using megatron-lm, 2021. URL https://arxiv.org/pdf/2104.04473.

Nolan, T. Nvidia h100: Are 550,000 gpus enough for this year?, August 2023. URL https://www.hpcwire.com/2023/08/17/nvidia-h100-are-550000-gpus-enough-for-this-year/. Accessed 19-04-2025.

NVIDIA. NVIDIA H100 Tensor Core GPU. NVIDIA Developer Blog, March 2023. URL https://www.nvidia.com/en-us/data-center/h100/. Accessed 20-04-2025.

NVIDIA. Introduction to NVIDIA DGX H100/H200 Systems, April 2025. URL https://docs.nvidia.com/dgx/dgxh100-user-guide/introduction-to-dgxh100.html#power-specifications. DGX H100/H200 User Guide, Power Specifications section. Last updated April 10, 2025.

Oak Ridge National Laboratory. Summit, undated. URL https://www.olcf.ornl.gov/olcf-resources/compute-systems/summit/. Accessed 20-04-2025.

O'Brien, M. and Fingerhut, H. Artificial intelligence technology behind ChatGPT was built in Iowa — with a lot of water, 2023. URL https://web.archive.org/web/20250220064324/https://apnews.com/article/chatgpt-gpt4-iowa-ai-water-consumption-microsoft-f551fde98083d17a7e8d904f8be822c4. Accessed 15-04-2025.

Oldham, M., Lee, K., and Gangidi, A. Building Meta's GenAI Infrastructure, 2024. URL http://web.archive.org/web/20240828225612/https://engineering.fb.com/2024/03/12/data-center-engineering/building-metas-genai-infrastructure/. Accessed 15-04-2025.

OpenAI. Introducing ChatGPT. OpenAI Blog, November 2022. URL https://openai.com/index/chatgpt/. Accessed 20-04-2025.

OpenAI. Announcing the Stargate Project. OpenAI Blog, January 2025. URL https://openai.com/index/announcing-the-stargate-project/. Accessed 20-04-2025.

Our World in Data. Annual global corporate investment in artificial intelligence, by type, 2024. URL https://ourworldindata.org/grapher/corporate-investment-in-artificial-intelligence-by-type. Accessed 20-04-2025.

Patel, D., Nishball, D., and Knuhtsen, R. Mi300x vs H100 vs H200 Benchmark Part 1: Training – CUDA Moat Still Alive, January 2024. URL https://semianalysis.com/2024/12/22/mi300x-vs-h100-vs-h200-benchmark-part-1-training/. Accessed 20-04-2025.

Pilz, K. and Heim, L. Compute at Scale: A Broad Investigation Into the Data Center Industry. *arXiv:2311.02651*, 2023. URL https://arxiv.org/abs/2311.02651.

Pilz, K. F., Mahmood, Y., and Heim, L. AI's Power Requirements Under Exponential Growth: Extrapolating AI Data Center Power Demand and Assessing Its Potential Impact on U.S. Competitiveness. *RAND Corporation*, (RR-A3572-1), 2025. URL https://www.rand.org/pubs/research_reports/RRA3572-1.html.

Pires, F. Chinese Companies Spend $5 Billion on Nvidia GPUs for AI Projects, 2023a. URL https://www.tomshardware.com/news/chinese-companies-spend-big-on-nvidia-gpus-for-ai-projects. Accessed 15-04-2025.

Pires, F. China's ByteDance Has Gobbled Up $1 Billion of Nvidia GPUs for AI This Year, 2023b. URL https://www.tomshardware.com/news/chinas-bytedance-has-gobbled-up-dollar1-billion-of-nvidia-gpus-for-ai-this-year. Accessed 15-04-2025.

Rahman, R. The computational performance of machine learning hardware has doubled every 2.2 years. Epoch AI, April 2025. URL https://epoch.ai/data-insights/peak-performance-hardware-on-different-precisions. Accessed 20-04-2025.

Rahman, R. and Owen, D. Performance improves 12x when switching from FP32 to tensor-INT8. Epoch AI, January 2024. URL https://epoch.ai/data-insights/hardware-performance-trend. Accessed 20-04-2025.

Reuters. Details of 110 Billion Euros in Investment Pledges at France's AI Summit. Reuters, January 2025. URL https://www.reuters.com/technology/artificial-intelligence/details-110-billion-euros-investment-pledges-frances-ai-summit-2025-02-10/. Accessed 20-04-2025.

Richter, F. Infographic: Nvidia's AI-Fueled Rally Hasn't Been Without Hiccups. Statista, January 2025. URL https://www.statista.com/chart/32358/nvidia-share-price/. Accessed 20-04-2025.

Riken Center for Computational Science. About Fugaku — RIKEN Center for Computational Science, undated. URL https://www.r-ccs.riken.jp/en/fugaku/about/. Accessed 15-04-2025.

Roser, M., Appel, C., and Ritchie, H. What is Moore's Law? Our World in Data, 2023. URL https://ourworldindata.org/moores-law. Accessed 20-04-2025.

Samborska, A. Investment in generative AI has surged recently. Bloomberg, June 2024. URL https://ourworldindata.org/data-insights/investment-in-generative-ai-has-surged-recently. Accessed 20-04-2025.

Sastry, G., Heim, L., Belfield, H., Anderljung, M., Brundage, M., Hazell, J., O'Keefe, C., Hadfield, G. K., Ngo, R., Pilz, K., Gor, G., Bluemke, E., Shoker, S., Egan, J., Trager, R. F., Avin, S., Weller, A., Bengio, Y., and Coyle, D. Computing power and the governance of artificial intelligence. *arXiv:2402.08797*, 2024. URL https://arxiv.org/abs/2402.08797.

Scanlon, N. Beyond DeepSeek: How China's AI Ecosystem Fuels Breakthroughs, February 2025. URL https://www.lawfaremedia.org/article/beyond-deepseek--how-china-s-ai-ecosystem-fuels-breakthroughs. Accessed 20-04-2025.

SemiAnalysis. Datacenter industry model. Commercial database, 2024. URL https://semianalysis.com/datacenter-industry-model/. Accessed 20-04-2025.

Sevilla, J. The Longest Training Run, 2022. URL https://epoch.ai/blog/the-longest-training-run. Accessed 15-04-2025.

Sevilla, J. and Roldan, E. Training Compute of Frontier AI Models Grows by 4-5x per Year, 2024. URL https://epoch.ai/blog/training-compute-of-frontier-ai-models-grows-by-4-5x-per-year. Accessed 20-04-2025.

Sevilla, J., Heim, L., Hobbhahn, M., Besiroglu, T., Ho, A., and Villalobos, P. Estimating Training Compute of Deep Learning Models, 2022. URL https://epoch.ai/blog/estimating-training-compute. Accessed 15-04-2025.

Sevilla, J., Besiroglu, T., Cottier, B., You, J., Roldán, E., Villalobos, P., and Erdil, E. Can AI Scaling Continue Through 2030?, August 2024. URL https://epoch.ai/blog/can-ai-scaling-continue-through-2030/. Accessed 20-04-2025.

Shah, A. Top500: China Opts Out of Global Supercomputer Race, 2024. URL https://thenewstack.io/top500-chinas-supercomputing-silence-aggravates-tech-cold-war-with-u-s/. Accessed 15-04-2025.

Shehabi, A., Smith, S. J., Hubbard, A., Newkirk, A., Lei, N., Siddik, M. A., Holecek, B., Koomey, J. G., Masanet, E. R., and Sartor, D. A. 2024 united States Data Center Energy Usage Report. Technical report, Lawrence Berkeley National Laboratory, 2024. URL https://eta-publications.lbl.gov/sites/default/files/2024-12/lbnl-2024-united-states-data-center-energy-usage-report.pdf. Accessed 20-04-2025.

Shilov, A. Nvidia sold half a million H100 AI GPUs in Q3 thanks to Meta, Facebook — lead times stretch up to 52 weeks: Report, 2023a. URL https://www.tomshardware.com/tech-industry/nvidia-ai-and-hpc-gpu-sales-reportedly-approached-half-a-million-units-in-q3-thanks-to-meta-facebook. Accessed 15-04-2025.

Shilov, A. China's secretive Sunway Pro CPU quadruples performance over its predecessor, allowing the supercomputer to hit exaflop speeds, 2023b. URL https://www.tomshardware.com/tech-industry/supercomputers/chinas-secretive-sunway-pro-cpu-quadruples-performance-over-its-predecessor-allowing-the-supercomputer-supercomputer-to-hit-exaflop-speeds. Accessed 15-04-2025.

Shilov, A. Nvidia to reportedly triple output of compute gpus in 2024: Up to 2 million h100s, October 2023c. URL https://www.tomshardware.com/news/nvidia-to-reportedly-triple-output-of-compute-gpus-in-2024-up-to-2-million-h100s. Accessed 19-04-2025.

Shilov, A. xAI Colossus supercomputer with 100K H100 GPUs comes online — Musk lays out plans to double GPU count to 200K with 50K H100 and 50K H200. Tom's Hardware, December 2024a. URL https://www.tomshardware.com/tech-industry/artificial-intelligence/xai-colossus-supercomputer-with-100k-h100-gpus-comes-online-musk-lays-out-plans-to-double-gpu-count-to-200k-with-50k-h100-and-50k-h200. Accessed 20-04-2025.

Shilov, A. Nvidia's H100 AI GPUs cost up to four times more than AMD's competing MI300X, February 2024b. URL https://www.tomshardware.com/tech-industry/artificial-intelligence/nvidias-h100-ai-gpus-cost-up-to-four-times-more-than-amds-competing-mi300x-amds-chips-cost-dollar10-to-dollar15k-apiece-nvidias-h100-has-peaked-beyond-dollar40000. Accessed 15-04-2025.

Skidmore, Z. and Swinhoe, D. Meta announces 4 million sq ft, 2GW Louisiana data center campus, 2024. URL https://www.datacenterdynamics.com/en/news/meta-announces-4-million-sq-ft-louisiana-data-center-campus/. Accessed 15-04-2025.

Smith, B. Microsoft to invest $80 billion in AI infrastructure. Microsoft Blog, January 2025. URL https://blogs.microsoft.com/on-the-issues/2025/01/03/the-golden-opportunity-for-american-ai/. Accessed 22-04-2025.

Tal, E., Viljoen, N., Coburn, J., and Levenstein, R. Our next-generation Meta Training and Inference Accelerator, 2024. URL https://ai.meta.com/blog/next-generation-meta-training-inference-accelerator-AI-MTIA/. Accessed 22-04-2025.

Tekin, A., Tuncer Durak, A., Piechurski, C., Kaliszan, D., Aylin Sungur, F., Robertsén, F., and Gschwandtner, P. State-of-the-art and trends for computing and interconnect network solutions for HPC and AI. Technical report, 2021. URL https://prace-ri.eu/wp-content/uploads/State-of-the-Art-and-Trends-for-Computing-and-Interconnect-Network-Solutions-for-HPC-and-AI.pdf. Accessed 20-04-2025.

The Information. AI Data Center Database, 2025. URL https://www.theinformation.com/projects/ai-data-center-database. Accessed 20-04-2025.

The White House. Executive Order on AI Infrastructure Development. The White House Briefing Room, February 2025. URL https://bidenwhitehouse.archives.gov/briefing-room/presidential-actions/2025/01/14/executive-order-on-advancing-united-states-leadership-in-artificial-intelligence-infrastructure/. Accessed 20-04-2025.

Top500. Top500 List. Online database. URL https://www.top500.org. Undated. Accessed 20-04-2025.

Trueman, C. xAI's Memphis Supercluster has gone live, with up to 100,000 Nvidia H100 GPUs, 2024. URL https://web.archive.org/web/20241009045341/https://www.datacenterdynamics.com/en/news/xais-memphis-supercluster-has-gone-live-with-up-to-100000-nvidia-h100-gpus/. Accessed 15-04-2025.

UK Department for Science, Innovation and Technology. AI Opportunities Action Plan. Technical report, UK Government, March 2025. URL https://www.gov.uk/government/publications/ai-opportunities-action-plan/ai-opportunities-action-plan. Accessed 20-04-2025.

U.S. Bureau of Industry and Security. Addition of Entities to the Entity List and Revision of an Entry on the Entity List. Technical report, U.S. Department of Commerce, June 2019. URL https://www.federalregister.gov/documents/2019/06/24/2019-13245/addition-of-entities-to-the-entity-list-and-revision-of-an-entry-on-the-entity-list. Accessed 20-04-2025.

U.S. Bureau of Labor Statistics. Producer Price Index by Industry: Data Processing, Hosting and Related Services, 2025. URL https://fred.stlouisfed.org/series/PCU518210518210. Accessed 15-04-2025.

U.S. Department of Commerce. Commerce Adds Seven Chinese Supercomputing Entities to Entity List for their Support to China's Military Modernization, and Other Destabilizing Efforts, April 2021. URL https://www.commerce.gov/news/press-releases/2021/04/commerce-adds-seven-chinese-supercomputing-entities-entity-list-their. Accessed 20-04-2025.

U.S. Energy Information Administration. Frequently Asked Questions (FAQs), 2024. URL https://www.eia.gov/tools/faqs/faq.php?id=97&t=3. Accessed 20-04-2025.

Vultr. Pioneering the Future of AI with AMD Instinct™ MI300X GPUs, Broadcom, and Juniper Networks — Vultr Blogs, 2024. URL https://blogs.vultr.com/Lisle-data-center. Accessed 15-04-2025.

Wei, L. U.S. Think Tank Reports Prompted Beijing to Put a Lid on Chinese Data - WSJ, 5 2023. URL https://www.wsj.com/world/china/u-s-think-tank-reports-prompted-beijing-to-put-a-lid-on-chinese-data-5f249d5e. Accessed 18-04-2025.

Zoting, Shivani. Artificial Intelligence (AI) Infrastructure Market Size, Share, and Trends 2025 to 2034. Technical report, Precedence Research, January 2025. URL https://www.precedenceresearch.com/artificial-intelligence-infrastructure-market. Accessed 20-04-2025.