INDUSTRY REPORT

Capacity Gap Analysis for AI Processing

Version 1.2 – 28 October 2024

Abstract

AI is becoming the most significant paradigm shift in history, with a direct impact on the global economy. Generative AI, in particular, is expected to greatly enhance productivity within work processes. Some studies estimate that Generative AI could contribute between $2.6 trillion and $4.4 trillion annually—for comparison, the EU's entire GDP in 2023 was $17.1 trillion. This Industry Report estimates the current and projected computing needs for generative AI training and inference in 2024 and 2030, while identifying key challenges and potential solutions to meet these demands. The main challenges in AI development include data availability, scalability limitations of centralized systems, power constraints, and challenges in accelerator manufacturing. By 2030, creating new distributed and decentralized systems for AI training—leveraging a continuum of HPC, cloud, and edge resources—will be a critical aspect for meeting the processing demands of AI training and the low-latency requirements of AI inference. This underscores the urgent need for strong collaboration between supercomputing, cloud computing, and edge computing, as well as the development of a management and orchestration platform to create cloud-edge environments that address the demands of AI processing workflows and the high-performance, low-latency requirements of their components.

(The numbers in this Report illustrate trends or patterns, rather than precise or exact figures)

Contents

Citation: OpenNebula Systems (2024) 'Industry Report: Capacity Gap Analysis for AI Processing', Version 1.2 (28 October 2024). Published online at OpenNebula.pro

1. Computational Complexity of Generative AI

AI infrastructure is costly because the underlying algorithmic problems are extremely computationally intensive. While the exact number of operations required for training and inference in transformer-based models varies depending on the specific model, a fairly accurate rule of thumb states that it depends primarily on two factors: the number of parameters (i.e. the weights of the neural networks) in the model and the number of input and output tokens (i.e. the data used in the training dataset).

Inference: A forward pass of a transformer model requires approximately `2*m*p` floating-point operations, where `p` is the number of parameters in the model and `m` is the number of input/output tokens in the sequence.

Training: Requires about `6*n*p` floating-point operations, as the backward pass involves four additional operations. Here, `p` is the number of model parameters, and `n` represents the number of tokens in the training dataset.

The following Table shows an estimation for some models:

Model	**Year**	**Parameters** (billion: 10^9)	**Training Data - Tokens -** (billion: 10^9)	**Training Compute Needs** (flops)	**Inference Compute Needs** (flops) (*)
BERT	October 2018	0.34	0.0033	6.7 x 10^15	680 x 10^9
GPT-J	June 2021	6	402	14.5 x 10^21	12 x 10^12
GPT-2	February 2019	1.5		1 x 10^21	3 x 10^12
GPT-3	June 2020	175	300	309.6 x 10^21	358 x 10^12
GPT-4	March 2023	1,760	1,000	10,560 x 10^21	3,500 x 10^12
Llama 2 (**)	July 2023	69	2,000	828 x 10^21	114 x 10^12
Llama 3.1(**)	July 2024	405	15,000	36,450 x 10^21	810 x 10^12

(*) Inference with 1024 tokens
(**) This is a very conservative computational complexity model; there are models such as Meta's LLaMA whose compute requirements are even higher.

AI training and inference have distinct execution profiles, each requiring different infrastructure architectures to meet their specific needs. Training relies on offline, specialized, centralized, and tightly-coupled HPC-like architectures, consisting of high-performance nodes interconnected by low-latency networks. In contrast, inference demands interactive, general-purpose distributed edge environments with low-latency connections to end users. While training will be typically carried out in the **cloud** or on **supercomputers**, inference will primarily occur on-demand on **edge** computing platforms, often managed by telecom operators.

The **model used in this report offers a simplified evaluation** of the two ends of the spectrum within the broad landscape of training related to Generative AI and AI/ML in general. On one end, the largest foundation models are incredibly resource-intensive and costly, meaning that only a handful of companies can afford to train them at scale, as is the case today. On the other end of the spectrum, inference is a high-throughput application that is widely utilized across various industries. In the middle, there is an opportunity for a broader range of companies to engage in fine-tuning foundation models—whether

starting with smaller or progressively larger ones—or developing smaller models tailored to specific domains. These organizations can leverage their own and/or domain-specific datasets to fine-tune these models, making specialized AI solutions more accessible and affordable.

Moreover, we must consider that, in some cases, it may be necessary to perform not only inference but also training on foundation models at the edge to meet data privacy and security requirements. In these cases, resources are constrained, so a balance between the model size (to fit edge resources) and its accuracy must be achieved. As AI continues to evolve, the **cloud-edge continuum will play a critical role**, not only in combining processing capacity across distributed HPC and cloud systems but also in addressing the diverse needs of different types of AI systems.

2. Current Computing Needs for Generative AI in 2024

> A state-of-the-art LLM model in 2024 requires approximately **10^25 flops for training** and **10^15 flops for inference**.

This estimation outlines the capacity that HPC and cloud infrastructures must provide to support the training of top LLM models, and the capacity required from edge nodes to handle Inference runs.

2.1. Processing Gap for AI Training in 2024

To estimate the performance of HPC/Cloud systems for AI training, we assume that training runs typically last a minimum of four months (approximately 10^7 seconds).

> The performance of a computing infrastructure for AI training in 2024 should reach 10^18 flop/s, **equivalent to 1 exaflop/s.**

Regarding **High Performance Computing**, the TOP500 ranking highlights the 500 most powerful non-distributed computer systems in the world, updated twice a year using the Linpack benchmark. These systems are designed for tightly-coupled, large-scale scientific and engineering applications with computational demands similar to AI training.

As of June 2024, the Frontier Supercomputer in the United States remains the leading HPC system globally, achieving a peak performance of **1.7 exaflop/s**. Frontier is hosted at Oak Ridge National Laboratory and was the first supercomputer to break the exaflop barrier. In the European Union, the LUMI Supercomputer in Finland is the top HPC system, with a peak performance of **0.5 exaflop/s**. LUMI is part of the EuroHPC initiative and plays a crucial role in Europe's supercomputing capabilities and the new AI Factories strategy.

In **Cloud Computing**, CoreWeave is estimated to be operating 50,000 GPUs with a peak performance of **16 exaflop/s**, making it a major player in the high-performance computing (HPC) and AI infrastructure space. In September 2024, Oracle announced the world's largest, first AI zettascale supercomputer in the Cloud. This system will feature 131,072 Blackwell GPUs, which is more than three times the number of GPUs in the Frontier Supercomputer and over six times more GPUs than other hyperscalers, and will be able to scale up to **2.4 zettaflop/s**. Oracle did not confirm when its Blackwell-powered service will come online considering the chips are currently being produced.

In the European Union, the cloud provider Scaleway may now be operating around 5,000 GPUs with a peak performance of **1.6 exaflop/s**, further strengthening Europe's presence in AI and cloud computing.

Moreover, CoreWeave has committed to investing an additional $2.2 billion in Europe to address the growing demand for AI infrastructure, bringing its total investment in the region to **$3.5 billion**. This investment mirrors the scope of funding under projects like IPCEI-CIS. These investments will certainly strengthen the growing dominance of US cloud providers in the European market.

These numbers serve as **estimates to give a broad sense of potential capability**. We are comparing HPC systems, general-purpose clouds, and specialized GPU-accelerated clouds based on their peak performance. However, **the reality is that these systems are built with different architectural models to meet specific workload profiles**. HPC supercomputers are designed for the efficient execution of large-scale simulation codes and feature low-latency interconnects to minimize communication overhead. In contrast, general-purpose clouds are optimized for high-throughput virtualized servers, while GPU-accelerated clouds are built to accelerate massively parallel applications. Moreover, the performance achieved when running AI training tasks tends to be higher than that for scientific simulations, as deep learning model training and inference primarily use FP16 precision rather than the FP64 precision typically required in scientific computing.

In many cases, the **actual performance in practical AI training scenarios remains speculative**. The lack of detailed public information on the GPU models and whether the GPUs can be combined into a single cluster for model training introduces significant uncertainty. Real-world performance often depends on architectural and technical limitations, making the actual outcomes harder to predict.

This report does not examine the **supply chain** for the core components of processing infrastructure. NVIDIA, which controls 90% of the AI GPU market, has also acquired Infiniband, the leading provider of low-latency networks for HPC and cloud infrastructure. In addition, NVIDIA owns CUDA, the dominant development platform, and is increasingly expanding its influence up the technology stack. Major cloud providers, with privileged access to GPUs, are heavily investing in AI and are developing their own proprietary accelerators.

2.2. Processing Gap for AI Inference in 2024

To estimate the performance of edge systems for AI inference, we assume a minimum latency of 1 second.

> The performance of the edge nodes in 2024 for AI Inference should be 10^15 flop/s. This is **1 petaflop/s.**

This shows the need for edge nodes with GPU devices; for example, at least 4 NVIDIA A100 devices.

3. Projected Computing Needs for Generative AI in 2030

Several studies have analyzed the top 10 LLM models (i.e. those released by OpenAI, Google DeepMind and Meta) by compute, and have estimated that their **training processing needs have increased by a factor of 4-5x/year since 2010**. This is consistent with the tables presented in Section 1. An x4 annual growth in AI training compute outpaces some of the fastest technological expansions in recent history. For example it surpasses the peak performance growth of TOP500 supercomputers (x2 energy every 2-3 years), and even the data growth (x2 every 2-3 years).

If we assume that training processing needs will continue to grow at a rate of x4 per year through 2030, the computing **demand for AI training will be 10,000 times higher** by then. Considering that the model size and number of training tokens should scale at an equal rate, compute **demand for AI Inference will be 100 times higher**. In other words, by 2030, as stated by Epoch AI, *it will be very likely possible to train models that*

exceed GPT-4 in scale to the same degree that GPT-4 exceeds GPT-2 in scale. This means that by the end of the decade we might see advances in AI as drastic as the difference between the rudimentary text generation of GPT-2 in 2019 and the sophisticated problem-solving abilities of GPT-4 in 2023.

A state-of-the-art LLM model in 2030 is projected to require approximately **10^29 flops for training** and **10^17 flops for inference.**

The analysis does not take into account the impact of emerging AI approaches, such as **Explainable AI** (XAI), which enables human users to understand and trust the results and outputs produced by machine learning models. It also overlooks advancements like **Multimodal Learning**, where models are trained using diverse inputs—such as images, audio, and structured data—simultaneously. Furthermore, the integration of transformer architectures into other domains, such as time series analysis with models like **Temporal Fusion Transformers**, is another key development that remains unaddressed.

3.1. Peak Scenario - Processing Gap for AI Training in 2030

To estimate the performance of HPC/Cloud systems for AI training, we assume that training runs typically last a minimum of four months (approximately 10^7 seconds).

In the peak scenario, the performance of a computing infrastructure for AI training in 2030 is projected to reach 10^22 flop/s, **equivalent to 10 zettaflop/s.**

3.2. Conservative Scenario - Processing Gap for AI Training in 2030

While projections suggest a x10,000 increase in compute demand, this figure could be significantly lower if we take into account several key factors:

- There is a **trend toward longer durations of AI Training runs**. Since 2010, the length of training runs has increased by 20% per year, which would be on trend to x3 larger training runs by 2030.
- Processing costs can be reduced by **optimizing the use of existing computers**. This can be achieved by making smarter algorithms that use less compute for the same output. In 2020, OpenAI estimated that training a 2012 model required 44 times less compute in 2019 than it did originally in 2012. This corresponds to algorithmic efficiency doubling every 16 months. This gives approximately a x33 increase between 2024 and 2030.

Overall, this x100 reduction in compute demand results in a corresponding x100 overall increase.

In the conservative scenario, the performance of a computing infrastructure for AI training in 2030 is projected to reach 10^20 flop/s, **equivalent to 0.1 zettaflop/s.**

One trend that could reduce computing demand is the use of **smaller models**. Smaller models can be more efficient to train and deploy, making them an attractive solution for industries aiming to balance performance with sustainability and cost-efficiency. According to Gartner, *by 2027, more than 50% of the GenAI models that enterprises use will be specific to either an industry or business function — up from approximately 1% in 2023. These domain models can be smaller, less computationally intensive and lower the hallucination risks associated with general-purpose models.*

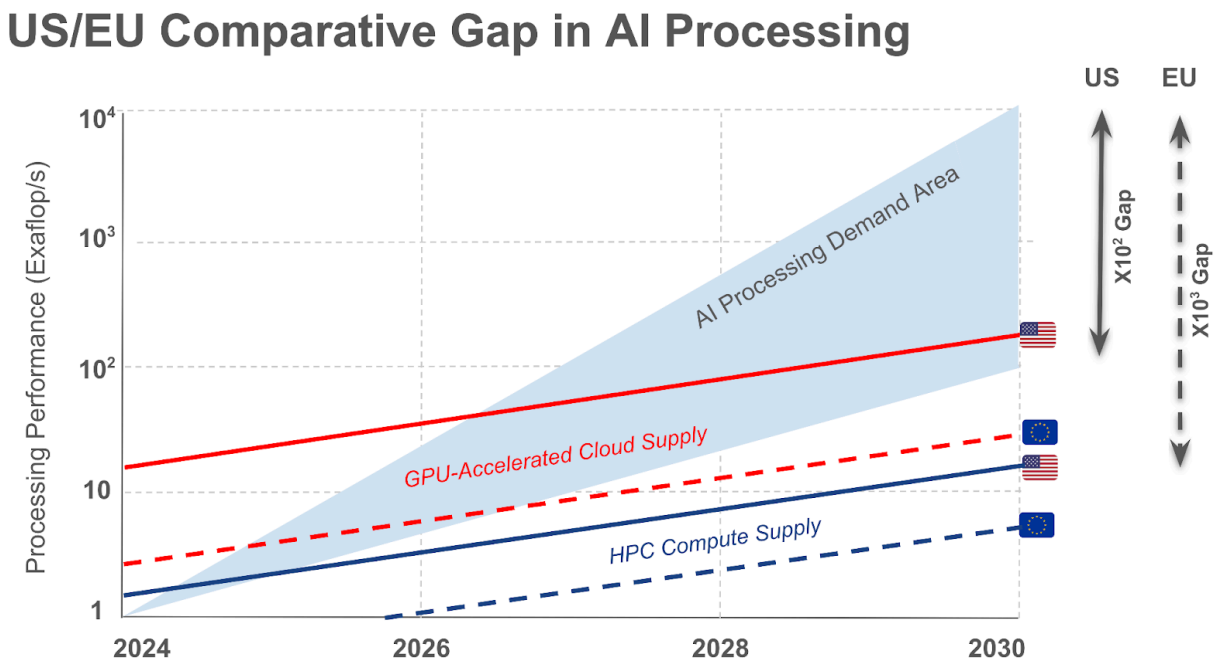
Another important aspect to consider is the **potential AI bubble** that many analysts are forecasting. Studies indicate a significant gap between the revenue expectations implied by the AI infrastructure

build-out and the actual revenue growth in the AI market. While projected revenue expectations range from $600 billion to $1 trillion, current revenues are under $100 billion. For example, OpenAI's revenue stands at approximately $3.4 billion, and the combined AI-related revenue for Google, Microsoft, Apple, and Meta is estimated to be around $10 billion. Current generative AI demand may not justify infrastructure investments at a scale 100 times higher than OpenAI's earnings.

3.3. Challenges in AI Training

While global data is estimated to double every 2-3 years, **language modeling datasets are expanding at an even faster rate of 2.9x per year**. The largest models today already rely on datasets containing tens of trillions of words (refer to the tables in Section 1). It's estimated that the total stock of publicly available human-generated text amounts to around 300 trillion tokens.If current growth trends continue, language models are projected to fully use this stock of available text between 2026 and 2032. This timeline could shorten further with the possibility of overtraining, with models that use fewer parameters and more data. The challenges in bridging the gap between the increasing demand for larger language models and the finite stock of publicly available text include Data Availability and Exhaustion, Data Quality and Curation, Synthetic Data Generation, and Privacy and Ethical Considerations. Open data will play a pivotal role in overcoming these obstacles and driving the development of future AI systems.

AI Training in 2030 will require systems able to deliver between 100 and 10,000 exaflop/s. The performance of the top TOP500 supercomputer has doubled every two years over the past decade. By 2030, **the top HPC and GPU-accelerated cloud systems in the US are expected to deliver up to 20 exaflop/s and 200 exaflop/s, respectively, while in the EU, they are projected to reach 5 exaflop/s and 20 exaflop/s**. This represents a roughly 10x increase in performance. In the peak scenario, the US and EU are expected to present a x100 and x1000 performance gap, respectively. If we plan to rely on centralized supercomputers for AI processing, **we will need significant technological and architectural advancements to boost their performance** between x10 and x1,000 compared to the average evolution seen in recent decades.



Source: "*Industry Report: Capacity Gap Analysis for AI Processing*", OpenNebula Systems (October 2024).

Regarding power consumption, various studies indicate that data centers' electricity consumption is projected to grow by 5% annually until 2030, resulting in an increase of 1.5 to 2 times current levels by that time. We will need **between x50 and x5.000 times more power** than is currently used today. However, this requirement could be reduced by **conducting research on more efficient hardware, optimized software programming, and energy-efficient data centers**, among other innovations.

Regarding **supply chain**, NVIDIA, with a 90% market share, saw explosive growth in 2023, shipping approximately 3.76 million data-center GPUs. This marks an increase of over 1 million units compared to 2022, when shipments totaled 2.64 million units, representing a 40% annual increase. If this trend continues, shipments could increase x10 by 2030. Given that flop/s per dollar doubles approximately every 2.5 years, the performance of a GPU is expected to be 10 times higher by 2030. However, to meet the growing demand for AI accelerators in the peak scenario, we would need an additional x100 increase beyond current production levels. This highlights the **urgent need to design and manufacture new accelerators**, as the required GPU production far exceeds current capabilities.

In this analysis, we have not considered any **paradigm shifts** in compute development, such as quantum computing, which could create an entirely new market structure and much higher compute capacity.

3.4. Processing Gap for AI Inference

In order to estimate the performance of Edge systems for AI Inference, we assume that minimum latency is 1 second.

> The performance of the edge nodes in 2024 for AI Inference is projected to reach 10^17 flop/s. This is **100 petaflop/s.**

By 2030, edge nodes operated by telecom providers should deliver a x100 increase in performance compared to today's capabilities. This may be lower, only x3 and 3 petaflops, if we also consider the potential improvements in algorithm efficiency.

4. The Strategic Relevance of the Distributed Cloud-Edge Continuum

From a strategic perspective, the most effective way to meet future AI processing needs is through the development of new distributed and decentralized systems. Leveraging a continuum of HPC, cloud, and edge resources will be crucial for addressing the intensive processing demands of AI training and the low-latency requirements of AI inference. This includes:

- Developing **new open and decentralized AI models and applications** that spread workloads across multiple data centers, regardless of their physical proximity, to meet the future challenges of AI processing, while enabling a more secure, resilient, and potentially fairer AI ecosystem.
- Accelerating the creation of **a high-performance, distributed cloud-edge continuum** across, that will be able to meet the growing demands of both AI and ultra-low-latency applications in the future.
- Developing dedicated **low-latency networks to interconnect and federate HPC systems** is essential for executing tightly-coupled AI training models.

As part of the IPCEI-CIS initiative, OpenNebula Systems, in collaboration with other leading cloud computing players, is developing a management and orchestration platform to create cloud-edge

environments that meet the needs of AI processing workflows and the high-performance and low-latency requirements of their components.

ONEnextgen (UNICO IPCEI-2023-003) is supported by by the Ministerio para la Transformación Digital y de la Función Pública through the UNICO IPCEI Program and co-funded by the European Union's NextGenerationEU instrument through the Recovery and Resilience Facility (RRF)