FIDE PARTNERS

From Megawatt to Gigawatt: AI's demand on tomorrow's Data Centres

How AI will reshape the hyperscale Data Centre landscape

August 2023

Author:
Javier Semprún

FIDE PARTNERS

From Megawatt to Gigawatt: AI's demand on tomorrow's Data Centres

Tel: +34 910 244 113
info@fidepartners.com
https://fidepartners.com/es/

August 2023

How AI will reshape the hyperscalers Data Centre landscape

Inside

WHITEPAPER

As AI advances, this whitepaper highlights the surge in demand for data centres, driven by the computational needs of AI, particularly with models like GPT-4. Projections estimate the AI market's capacity requirement at 1GW, potentially reaching 8GW by 2026. The rise of gigawatt-scale facilities may pose challenges to national power grids, requiring green power solutions.

How Data Centres Evolved: From In-house to AI-Powered Giants

In-house

Owned and maintained internally by the company it supports

Colocation

Owned by an operator, selling space, power and cooling to multiple customers

Hyperscale-Cloud

Large facilities, owned by hyperscalers and supporting their activities

HPC

Facilities, supporting mining and other HPC processes, which require extensive amounts of power

AI

Large facilities, on-site power generation, very high-density setups, liquid cooling

The evolution of data centres can be represented by distinct eras, marked by disruptive shifts in the computational power demanded by businesses and enabled by technological advancements.

Each era represents a pivotal moment where data centres have adapted and expanded to meet the escalating demands of firms, pushing the boundaries of what is possible in terms of processing capabilities.

In-house data centres gave way to colocation sites as digitalisation and IT infrastructure requirements grew, which then exploded in size and number as hyperscalers began expanding their cloud services.

As businesses seek to harness the power of AI for transformative insights and innovations, the new era for data centres is poised to be defined by artificial intelligence (AI). This era will witness the emergence of massive data centres that can handle the computational demands of artificial intelligence.

AI's Power Surge: How models like GPT-4 redefine Data Centre demands

AI programs are trained by exposing them to vast amounts of data and desired outputs, allowing them to learn patterns and make predictions. This training process involves complex mathematical algorithms and requires powerful hardware, including graphics processing units (GPU), to train and run AI models. Leading players have trained state-of-the-art AI models on supercomputers: for example, OpenAI's GPT-4 was trained on 16,000 to 25,000 GPUs.

Once trained, AI programs can answer user queries by applying the learned knowledge to new data, a process named 'inference'.

To support AI training and inference, data centres must provide ample computing power, storage capacity, and efficient cooling systems to handle the demanding AI workloads. With widespread AI adoption, the requirements for large-scale data centres will increase drastically.

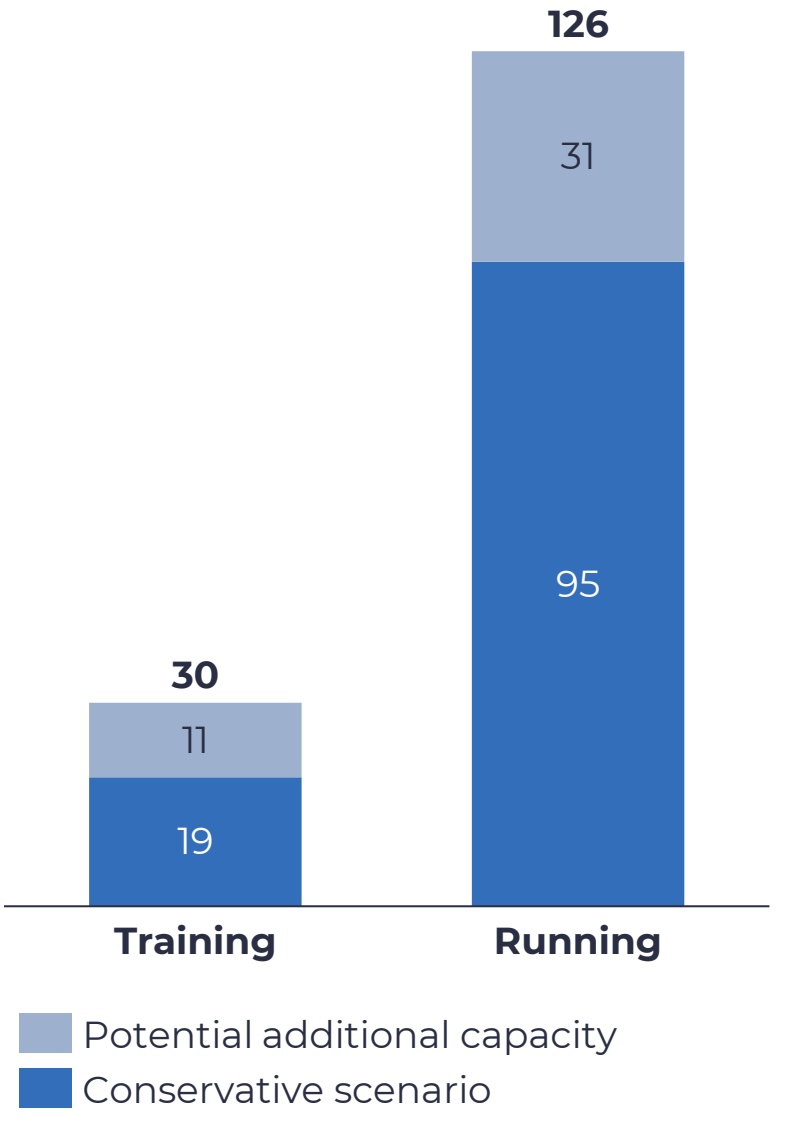
To estimate the requirements needed for GPT-4's inference workloads, we can use BLOOM as a comparable model.

BLOOM is a large language model (LLM) developed by the Hugging Face community and BigScience that is similar in size to GPT-4 with 176bn parameters compared to 175bn.

Based on published data on the energy consumption of BLOOM, as well as the user base evolution and characteristics of ChatGPT, we estimated the current data centre capacity required to train and run GPT-4.

Our estimate for training capacity for GPT-4 ranges from 19MW to 30MW and from 95MW to 126MW for its inference requirements based on the estimated number of users and the number of queries sent per user daily.

Estimated capacity, required for GPT-4
(MW)



AI's Power Surge: Estimating the impact on global Data Centre demand

ChatGPT is the most popular AI application today, and there are many others created by the likes of Google, DeepMind, Meta and Nvidia, among others, generating text-text, text-image, text-voice, text-video, image-image, and multimodal (combination) applications.

As of today, the total capacity requirement for AI programs has likely passed 1GW. High growth will persist as AI integration outpaces new system efficiencies.

AI integration will skyrocket as Big Tech and hyperscalers embed AI tools into their existing applications. Notably **Microsoft**, with 345 million paid seats, is in the process of fully integrating OpenAI's models into its entire Office suite. **Google** is already incorporating AI tools across Workspace, of which there are over three billion users globally. **Zoom** recently announced a partnership with OpenAI that enables AI-generated summaries and message drafts for its 300 million daily users. **Adobe Acrobat** unveiled its AI co-pilot for Photoshop with image-image capabilities, available for over 100 million daily users. **AWS**, the market leader in the Cloud market, offers various AI tools for text, images, voice and videos.

For instance, call centres and customer support services are primed to use text-voice and voice-text tools.

We are on the precipice of mass AI integration and, consequentially, mass AI adoption. Furthermore, improved functionality of future models will enable more complex user queries, increasing the load per query.

On the other hand, **efficiency gains in AI models** – there are many projects underway in both the private sector and academia – will somewhat moderate the growing data centre capacity requirements for AI. Of those, some focus on developing specialised applications for AI programs. Although each requires less capacity, more of these models will be needed to satisfy demand. Additionally, a set of specialised models may run less efficiently than a single generalised model.

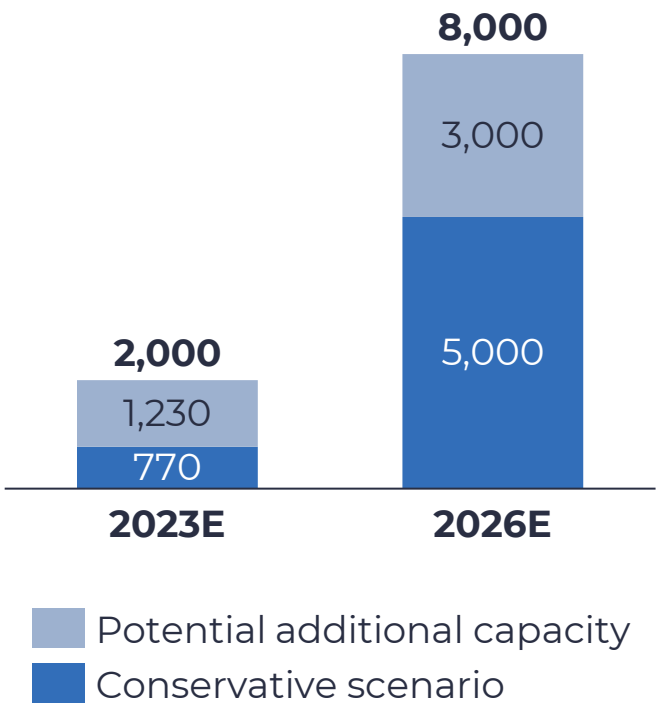
Other researchers are investigating shortcuts (e.g., reducing decimal places computed) with limited impact on processing capabilities. There are several other projects with significant potential; however, thus far, the only surefire ways to significantly enhance processing capabilities are to add more parameters and data. Over the last couple of years, Big Tech has added hundreds of billions of parameters to its models, necessitating ever more capacity for training and running programs.

AI's Power Surge: Estimating the impact on global Data Centre demand

Using our estimates for ChatGPT as a base, we model the capacity requirement for today's broader AI market, roughly at 1GW.

Considering future AI integration and efficiency gains, we estimate the capacity requirement to reach about 6.5 GW by 2026. The upper and lower bounds of our confidence interval are shown in the graph.

Estimated capacity, required for the AI market
(MW)



Gigawatt scale: The next wave of Hyperscale-Data Centres

Data centre capacity will need to expand remarkably and will not correspond to traditional builds. To take advantage of efficiencies of scale, new hyperscale data centre deployments may land in the order of magnitude higher, near 1GW of capacity, rather than the current hundreds of MW.

It is likely that a GW-sized facility will put too much of a strain on a country's power grid, and therefore, will need to build its own generation for primary (and possibly even secondary) supply, with the country's power grid serving as a backup. The power generation may need to be green to be politically feasible. This may involve small modular nuclear reactors, hydrogen fuel cells, and solar or wind power combined with traditional hydrocarbon turbines.

As of today, latency is not a critical issue, but rather the value derived from the output. Efforts in LLM development are concentrated on tackling larger and more complex queries and providing more sophisticated answers. The implication is that the facilities do not need to be located near large cities. Selecting remote areas to build these independent data centre ecosystems will likely prove easier in securing zoning permits and meeting air quality requirements.

Additional considerations include cooling and rack capacity. Most data centres today are built with air cooling; however, the power density will require a water or liquid system. In terms of rack capacity, the racks may reach about 35kW-50kW each, with the data halls built to suit.

Gigawatt scale: The next wave of Hyperscale-Data Centres

Beyond the foreseeable horizon, leading players may move to Tier II locations, as many have already done for the cloud. As edge nodes to the cloud, these secondary locations will cache small parts of the AI model, rather than fully replicating it, to improve the quality of service and optimise network traffic.

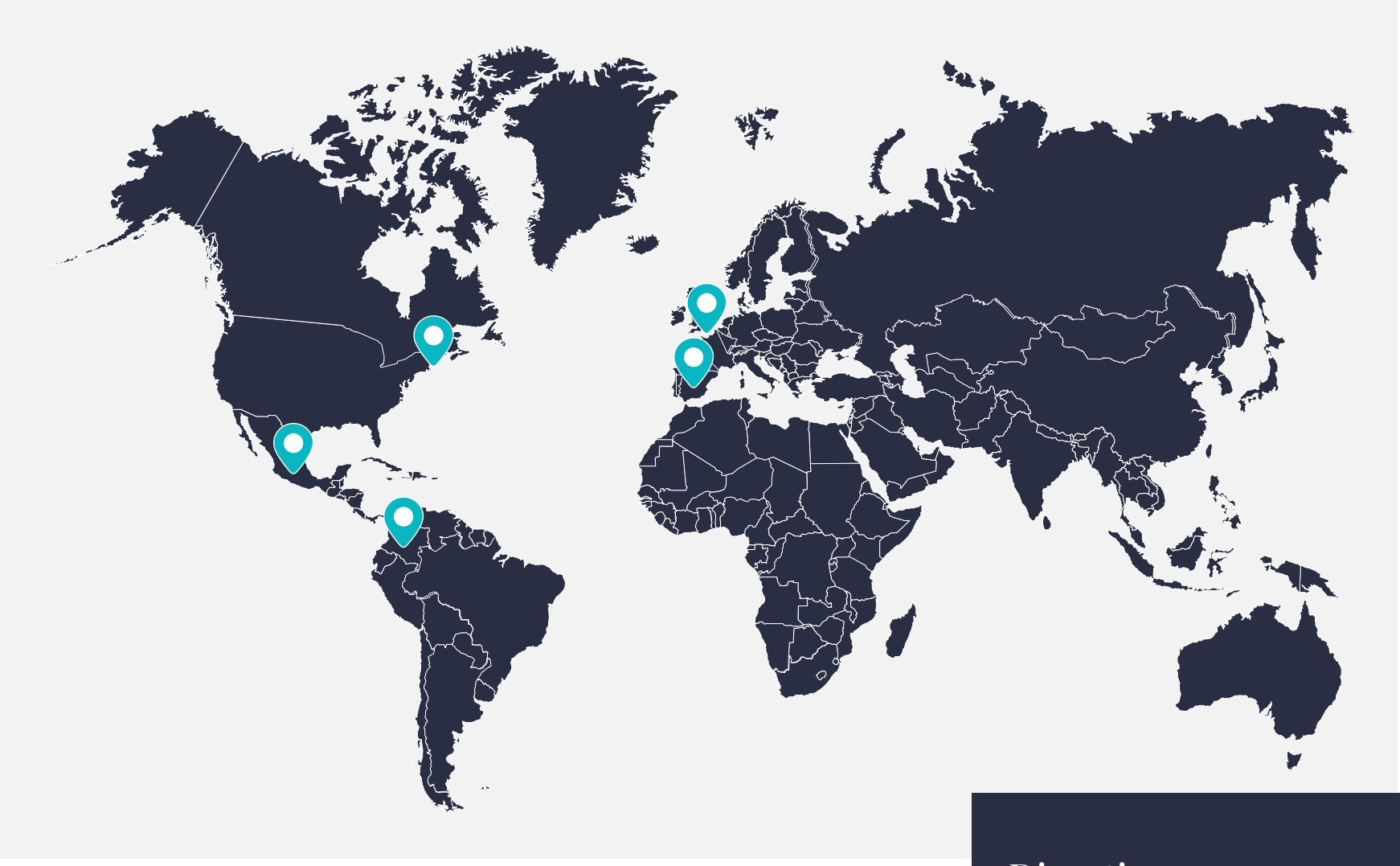
In the near term, data centres will likely become true industrial hubs with their own digital and power infrastructures.

These developments will likely entail joint ventures between data centres and power generation and distribution entities.

Some countries and regions are likely to benefit more from the deployment of these sites. Initially, most will occur in the US, where all the leading tech players are based. Players like Microsoft have been rumoured to already be in the process of building the first AI-focused sites.

In Europe, the Nordics seem like the best fit as the availability of renewable power, the colder weather and the experience of data centre players in the cryptocurrency industry (which has similar power density and site design requirements) place them ahead of other European hubs. In the traditional FLAP-D (Frankfurt, London, Amsterdam, Paris and Dublin) markets, France, with abundant nuclear power seems to be better positioned than its competing markets.

About Fide Partners



Our experience

Fide Partners is a team of experts with broad and extensive experience in digital infrastructure projects worldwide.

Our consultants have advised DC operators and investors in several assignments in primary and regional hubs. **We have completed assignments in over 35 countries, covering more than 480 facilities totalling 1GW in IT capacity**

Join us in shaping the future of data centres.

If you are interested in the topic, please leave a comment or reach us via email.

Javier Semprún Henkart

Telephone:
+34 910 244 113

Mail:
javier.semprun@fidepartners.com

Web:
https://fidepartners.com

Direction:

London:
Aldwych House
London, WC2B 4HN
United Kingdom

Madrid:
C/Don Ramón de la Cruz, 6, 1º
28001 - Madrid
Spain

Bogotá:
Carrera 11A #98-50
Ofc. 704, Edificio Punto99
110221, Bogota
Colombia

Boston:
50 Milk Street,
Planta 15, C.P. 02109.
Boston, MA

Mexico