FIDE PARTNERS

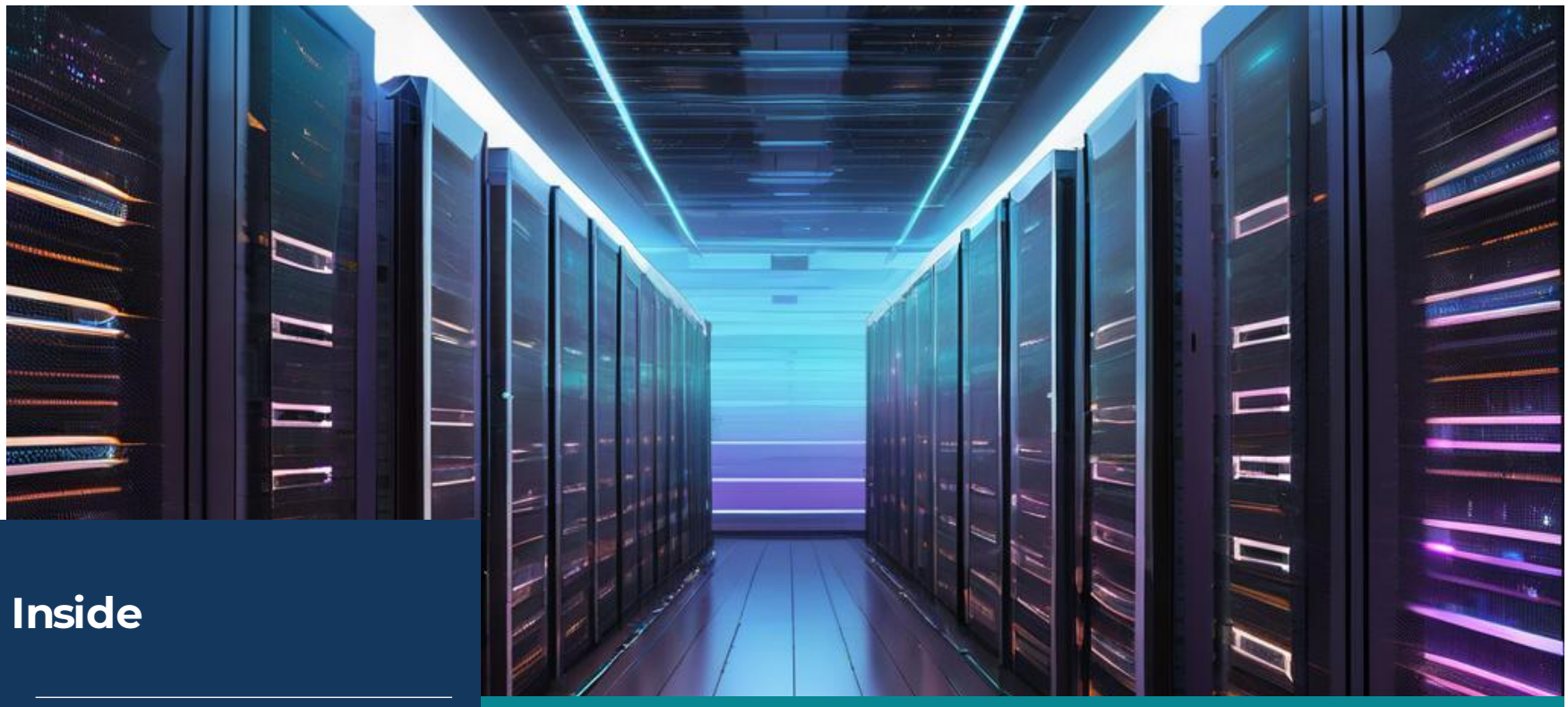
AI Training and Inference Impact on Data Centers

https://fidepartners.com/
Tel: +1 617 359 8114;
+34 910 244 113
info@fidepartners.com

July 2025

AI's Impact on Data Centers: Training vs. Inference

Authors: Jim Andrew, Freddy Farah Abuchaibe, Miguel Caldas



Inside

As inference grows, data centers that are near customers are becoming crucial for delivering low-latency, high-efficiency processing to end-users

The rapid advancement of Artificial Intelligence (AI) is significantly reshaping the global data center landscape. As AI technologies become increasingly integral across various industries, the demand for data centers equipped to handle AI workloads is experiencing unprecedented growth.

This whitepaper explores the data center requirements in distinct industry markets by comparing AI training and inference technical requirements. Furthermore, through industry-specific examples, we highlight how AI-driven applications are influencing data center roll-outs, strategies and investments.

AI Training vs. Inference and Its Implications for Global Data Center Infrastructure

Introduction

The data center market has experienced accelerating growth in recent years, driven by the rapid evolution of AI technology increasing demand for higher data consumption, on top of growing demand for cloud services, internet penetration, 5G consolidation, and digitalization trends across industries.

This surge has led to substantial investments in new data center construction and facility upgrades in major markets worldwide. The rapid penetration of AI in many applications is expected to accelerate this demand further, as AI-driven applications require greater computational power and infrastructure scalability.

The emergence of new AI models, such as DeepSeek in late 2024, has marked a potential inflexion point on how industries could leverage its possible use cases at a lower cost and processing needs.

AI training and inference are the primary drivers of this increasing AI-driven demand, sparking debates about the power consumption associated with each and how the data center market will adapt to meet these needs.

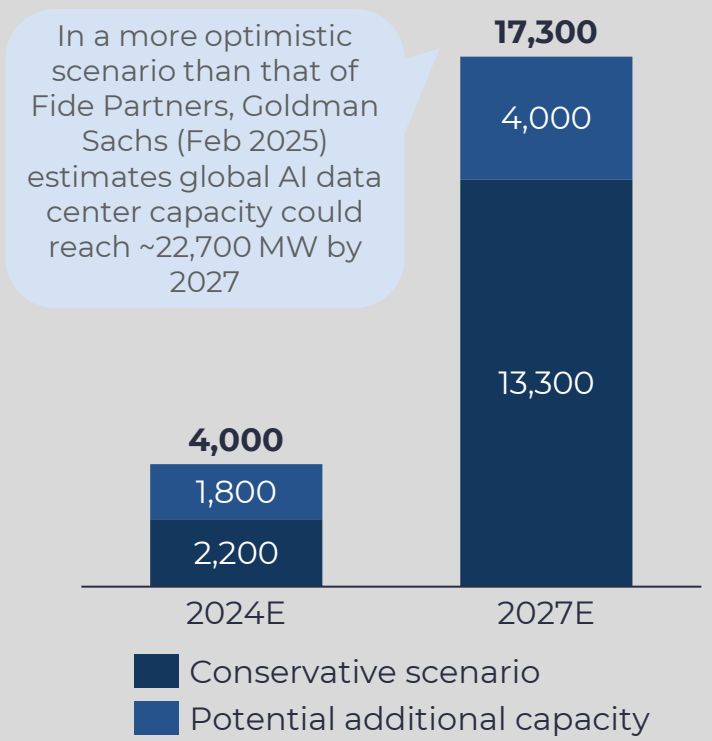
As demand for AI workloads continues to grow, the distinction between locations and facilities based on whether they support AI training or inference will be a key factor in market development. Due to latency requirements and to control costs, AI inference for some use cases is expected to be delivered near customers, while AI training may have greater flexibility in site selection.

AI training involves feeding a machine learning model large datasets and adjusting its parameters to minimize error. While traditional approaches rely on labeled data, advancements in self-supervised learning, particularly with transformer-based Neural Networks, enable models to learn from unstructured data by creating auxiliary tasks that uncover inherent patterns. AI training is characterized by more complex and computationally intensive processes, which require massive processing power, typically using GPUs, TPUs or AI accelerators, as large-scale data processing is required.

AI inference is the process of using a trained AI model to make predictions on new data. It is less computationally demanding than training, requiring fewer resources, but it still benefits from optimized hardware. AI inference often requires low latency, especially in applications such as real-time recommendations across many applications.

Estimated global capacity required for the AI market

(MW)



Sources: Fide Partners; Goldman Sachs, Feb 2025 — "AI to Drive 165% Increase in Data Center Power Demand by 2030

AI demand is expected to continue growing but may depend on several key factors

Future demand for data center capacity remains challenging to predict, as it depends on several unpredictable factors.

The pace of AI adoption across industries, the diversity of chip architectures and their power consumption, energy availability and sustainability initiatives, technological advancements, regulatory considerations, and AI's specific compute requirements are expected to play crucial roles in shaping demand.

Likewise, the capacity of data centers to absorb this demand will depend on how these factors evolve.

"*And these AI data centers, if you will, are improperly described. They are, in fact, AI factories. You apply energy to it, and it produces something incredibly valuable.*"

- NVIDIA Co-Founder & CEO – Jensen Huang at COMPUTEX 2025-05/25

AI Training vs. Inference: A Technical Breakdown

AI models process data through two main approaches: full-scale training and inference-based adaptation, each with distinct computational and cost requirements.

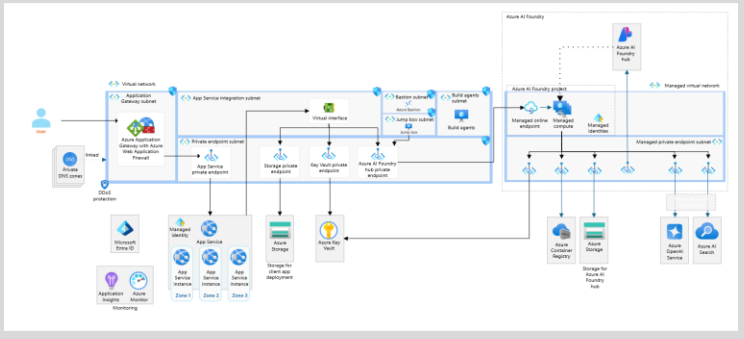
Training a large AI model like GPT-4 has historically required around 25,000 NVIDIA A100 GPUs, consuming 50 MW and over 100 days for a single cluster. However, more recent deployments, such as xAI's Grok, are reportedly using GPU clusters exceeding 100,000 units. These setups demand high-density racks (80–120 kW per rack), liquid cooling, and high-speed interconnects, like InfiniBand or NVLink. Total training costs can exceed several hundred million dollars, driven by infrastructure, energy, and GPU hardware (typically priced at USD20,000–USD40,000 per unit). This type of AI training is commonly conducted in hyperscale data centers operated by Google, AWS, Microsoft, and Meta. For example, Meta is projected to deploy 1.3 million GPUs by 2025, with spending plans reaching up to USD65bn. Increasingly, specialized GPU-as-a-Service providers such as CoreWeave and Lambda Labs are emerging as scalable alternatives for training and fine-tuning AI models, making AI technology available for smaller focused models created by new entrants.

AI inference typically operates with lower hardware and power needs. Deployments like DeepSeek, Mistral, and LLaMA rely on pre-trained models that currently operate efficiently with ~2,000 NVIDIA H800 GPUs, typically in a 15–40 kW per rack setup. As the market shifts toward next-generation Blackwell GPUs (B100/B200), inference performance is expected to improve further while maintaining manageable power envelopes. This enables flexibility in colocation and edge data centers where efficiency and latency matter. Inference instance costs range from millions to tens of millions of dollars. Techniques like retrieval-augmented generation (RAG) and parameter-efficient fine-tuning (PEFT) enable adaptation without costly retraining.

Key use cases for AI training include large language model (LLM) development, where massive datasets are used to understand, generate, and process human language. Other examples include medical imaging and drug discovery, which depend on AI models to analyze datasets for disease detection and pharmaceutical research. There are also national security agencies that use AI training for large-scale simulations, handling massive datasets to predict complex scenarios.

AI inference, on the other hand, is widely used in real-time applications like customer service chatbots and virtual assistants, where models handle conversations efficiently without massive GPU clusters. For example, inference could be widely used in healthcare diagnosis for AI-assisted radiology to detect diseases in X-rays, CT scans, and MRIs. In finance, fraud detection systems could analyze transaction patterns in real time. E-commerce and streaming platforms use AI inference for personalized recommendations, optimizing user experiences dynamically.

Azure ML Architecture



Sources: Microsoft

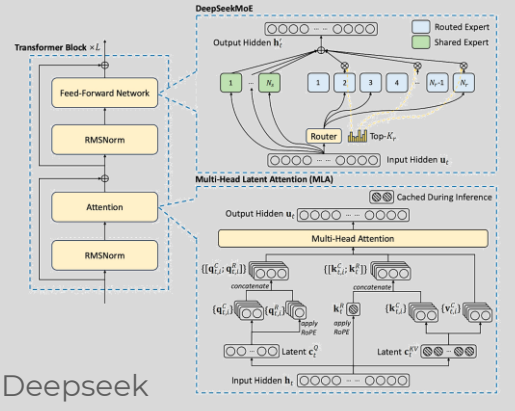
AI training architectures rely on large-scale distributed computing, using thousands of GPUs interconnected via high-bandwidth InfiniBand and NVLink to minimize latency. Techniques like tensor and pipeline parallelism optimize memory and compute efficiency.

Platforms like Azure enhance scalability with custom accelerators and liquid-cooled GPU clusters, reducing data bottlenecks and accelerating deep learning convergence, as also seen in models like GPT-4.

Deepseek employs a hierarchical mixture of experts (MoE) framework, dynamically routing tokens through specialized subnetworks. This allows it to activate only 1:16 experts per layer, using 37bn of 671bn parameters at a time.

By leveraging context-aware gating and cross-layer attention fusion, Deepseek optimizes efficiency, reducing redundant activations while maintaining high inference throughput and training scalability.

DeepSeek V3 Architecture



Sources: Deepseek

AI Training and Inference is expected to Drive Demand Across Data Centers of All Scales

Inference is expected to drive the demand for edge data centers in urban areas

The integration of AI into data center operations is expected to require significant infrastructure adjustments to meet the distinct demands of AI training and inference. As AI positions itself globally and penetrates different industries, power density requirements are expected to increase substantially.

As of 2020, the average power density demanded was 8kW per rack, while in 2024, power density per rack had more than doubled to 17kW and AI training racks are now demanding up to +100kW per rack. These adjustments encompass enhanced computational power, optimized cooling systems, strategic data center placement, interconnection and networking, and sustainable energy practices.

As inference scales, infrastructure funds are targeting Tier 2 and colocation data centers (2–5 MW) to host GPU workloads.

With newer, more efficient GPUs like NVIDIA's L4 and H100, and the upcoming Blackwell and Vera Rubin architectures, edge and regional facilities can now support 15–40 kW racks suitable for real-time AI processing. These setups enable low-latency inference without requiring the scale of hyperscale data centers.

Facilities for AI training will require high-density computing infrastructure to handle large-scale data processing, leading to advanced cooling systems and sustainable energy solutions.

On the other hand, AI inference data centers are expected to prioritize low-latency, energy-efficient hardware optimized for real-time processing, often leveraging edge computing to minimize transport delays.

However, as AI models become more complex and resource-intensive, inference power density is rising. This may drive the deployment of GPUs and accelerators in edge facilities to meet low-latency demands.

AI data center training vs. inference main priorities

Category	Training	Inference
Power density	Requires high-density racks at 80-120kW supporting GPU clusters	Typically operates at 15-40 kW per rack with variable utilization, though power density is rising as AI models grow more complex
Latency	Situated at long distances from data sources and requires only GPU clusters within 50m of each other for optimal networking	Requires strategic positioning near population centers as it must maintain <100ms latency to users depending on the modality
Cooling	Requires liquid cooling for >60kW racks and significant PUE improvements	May use hybrid cooling systems with traditional architectures
Scale	Requires dedicated power infrastructure, often 200MW+ per facility, land availability and, ideally, renewable power	Distributed power needs across multiple locations, typically +1MW per facility
Location	Strategically located in secondary markets with optimal power availability and lower costs	Close to urban areas with high populations to ensure low latency and operational reliability

High priority ● ◔ ◑ ◕ ○ Low priority

The Rise of AI Inference in Urban Areas and Tier 2 and 3 Edge Markets

AI inference markets are expected to be driven by proximity to end users

The growing demand for AI inference near urban areas is driven by the need for low-latency processing, cost efficiency, and localized AI applications. Unlike AI training, which is best suited for hyperscale data centers due to its high computational requirements, inference benefits from being closer to end-users. This proximity enables real-time AI-powered services across industries where ultra-low latency is critical, such as autonomous vehicles, financial trading, and smart city infrastructure. In contrast, industries such as industrial automation and large-scale manufacturing, often located outside urban centers, benefit from edge inference processing in Tier 2 and Tier 3 markets, where AI-powered predictive maintenance and supply chain optimization enhance operational efficiency.

Deploying inference workloads in urban data centers may reduce data transport costs and network congestion by minimizing reliance on centralized hyperscale hubs. Industries such as retail, logistics, and smart infrastructure rely on real-time AI-driven decision-making, making urban inference hubs essential for optimizing customer experiences, financial transactions, and public services. However, industries that operate outside major urban areas, such as industrial manufacturing, agriculture, and remote energy facilities, benefit from inference processing at the edge in Tier 2 and Tier 3 markets, where AI-powered automation and real-time monitoring can operate independently of urban infrastructure.

This shift also aligns with sustainability goals, as inference is expected to be less power-intensive than training and can be strategically deployed in areas with renewable energy sources and lower infrastructure costs. Texas (USA), for example, is experiencing a significant surge in the data center market, driven by relatively low energy costs and a deregulated energy market, making it a strategic hub for AI infrastructure. The expansion of AI inference facilities in urban centers not only reduces the carbon footprint of AI workloads but also optimizes bandwidth usage by decreasing network congestion on primary routes.

From an infrastructure perspective, urban areas may be more expensive to scale as land is expected to have higher prices and to be scarcer. However, the availability of local computing resources ensures that inference processing can be performed closer to data generation points, minimizing delays and improving efficiency.

Additionally, network optimization and latency reduction remain key drivers of AI inference expansion in urban environments. Low-latency inference is critical for applications such as autonomous transportation, financial trading, and smart city services, where even milliseconds of delay can impact decision-making.

Ultimately, the shift toward decentralized AI inference represents more than just an operational adjustment as it signals a fundamental transformation in how AI services are deployed. By expanding AI inference capabilities close to end users, data centers of all sizes will ensure that AI-powered applications scale efficiently, reducing network transport costs, improving sustainability, and delivering real-time intelligence where it is needed most.

AI inference impact overview by sector

Smart cities and 5G	Governments and enterprises	Healthcare facilities	Industrial automation and manufacturing	Retail and e-commerce
As smart city initiatives and 5G networks continue to expand, AI inference will be a fundamental enabler of connected mobility, energy grids, and security systems.	Entities in urban areas are investing in AI-driven urban planning, traffic management, and emergency response systems, reinforcing the importance of localized inference hubs as a cornerstone of future data infrastructure.	Facilities in regional markets can leverage AI-powered medical imaging diagnostics, patient monitoring, and drug discovery, allowing hospitals and research centers to process data locally rather than relying on distant hyperscale data centers.	Often located in Tier 2 and 3 cities to reduce costs, these sectors benefit from AI-driven predictive maintenance, robotics, and quality control. Edge AI inference enables real-time decision-making on production lines while minimizing downtime	AI inference plays a critical role in fraud detection, dynamic pricing, and customer personalization, ensuring that businesses deliver optimized experiences without unnecessary data transit delays.

AI Inference: Powering Real-Time Decision-Making in Healthcare, Finance and Banking

How data centers will enable AI-driven inference in industry hubs

Two of the most rapidly growing use cases for AI inference are in healthcare and finance, where real-time decision-making is critical. As AI adoption expands, demand for low-latency, high-efficiency inference solutions is surging in major global hubs, creating new opportunities for edge and medium-sized data centers strategically located near hospitals and financial districts.

Medical imaging is one of the most promising areas for AI inference, enabling faster and more accurate analysis of X-rays, CT scans, and MRIs. AI models trained on vast datasets can assist radiologists by detecting abnormalities earlier and with greater precision, improving patient outcomes. Unlike AI training, which requires immense computational resources and hyperscale data centers, AI inference operates efficiently on smaller-scale infrastructure, making it ideal for hospitals, clinics, and medical research institutions that require real-time results.

A prime example of where this technology is gaining traction is the Texas Medical Center in Houston, the world's largest medical complex, housing more than 60 hospitals and medical institutions. Deploying AI inference solutions in edge and mid-size data centers close to this hub allows for faster processing, reduced latency, and improved integration with existing healthcare workflows. Other major medical hubs, such as Boston's Longwood Medical Area, California's Stanford Medicine cluster, and London's Harley Street medical district, also present significant opportunities for AI-driven diagnostics.

As demand grows, edge and medium-sized data centers (ranging from 1 to 20 MW) are becoming critical in supporting AI inference. These facilities provide high-speed connectivity and specialized compute resources while maintaining a compact footprint compared to hyperscale data centers.

In the financial and banking sector, AI inference is revolutionizing fraud detection, cybersecurity, and risk management, where real-time decision-making can prevent billions in losses. AI models analyze vast volumes of transactional data, user behavior, and market activity in real time, enabling financial institutions to detect anomalies and prevent threats before they escalate.

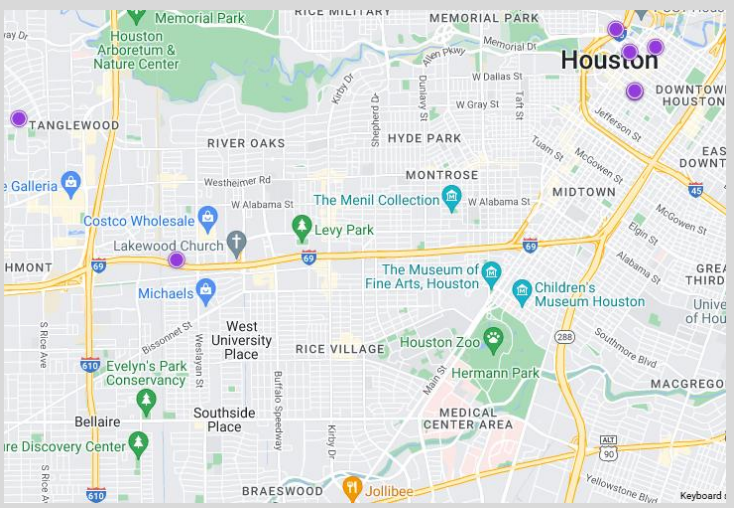
While high-frequency trading (HFT) remains a notable use case, where milliseconds can define success or failure, the broader application of AI to security and compliance is becoming even more critical.

To support these workloads, edge and regional data centers near financial hubs provide the low-latency environments needed for AI-driven analysis. Rather than relying on hyperscale data centers optimized for training, financial institutions increasingly use edge and medium-sized data centers (1 to 20 MW) located near financial districts.

A key example is New York City's Financial District, home to America's major financial institutions. The region hosts over 190 data centers within 100 miles, 48 of them within just 2 miles. Facilities like Equinix offer 1.7MW of IT capacity, while larger centers like Sabey support up to 18 MW. These facilities are optimized for inference tasks, reducing latency and strengthening cybersecurity and fraud detection capabilities to support faster and more secure decision-making.

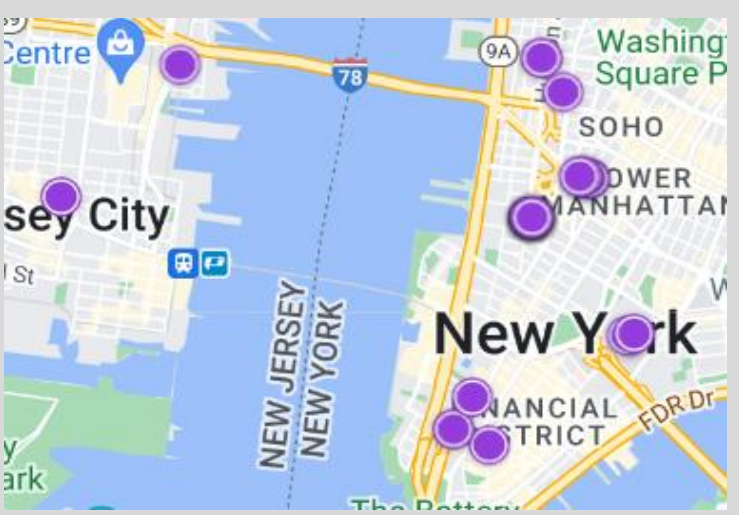
As AI adoption accelerates, healthcare, finance and banking, among other industries, will continue to drive demand for distributed, high-performance AI inference solutions. Edge and medium-sized data centers near these industry hubs will play a critical role in enabling real-time AI applications, ensuring that healthcare providers can deliver faster diagnostics and financial institutions can combat fraud and optimize cybersecurity with greater efficiency.

Data Centers near Texas Medical Center



Sources: Datacenters.com ● Data center locations

Data Centers near NY Financial District



Sources: Datacenters.com ● Data center locations

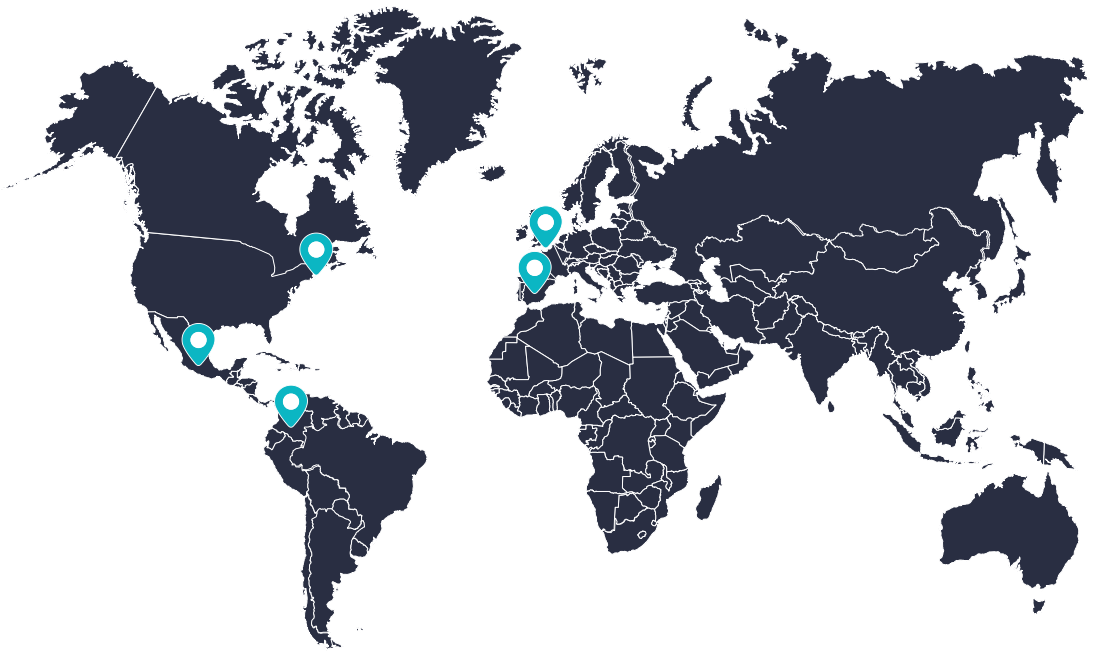
FIDE PARTNERS

Conclusion: The Future of AI-Driven Data Center Evolution

The rapid integration of AI into industries worldwide is reshaping the data center landscape, driving unprecedented demand for both high-performance AI training clusters and low-latency AI inference infrastructure. As AI models grow more sophisticated, the distinction between training and inference has become crucial in determining how and where computational resources are deployed.

AI training remains the domain of hyperscale data centers, requiring massive GPU clusters, high-density power consumption, and advanced cooling solutions. However, AI inference is fueling demand for edge and mid-sized data centers closer to end-users, enabling real-time processing for applications in finance, healthcare, and beyond. The expansion of AI inference infrastructure in urban hubs, financial districts, medical clusters, and Tier 2 and 3 edge markets underscores the need for highly connected, power-efficient, and strategically located facilities.

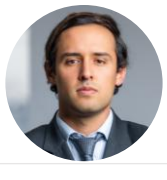
As AI adoption accelerates, data centers must evolve to meet these growing demands. The future of AI-driven infrastructure will depend on a balance between centralized hyperscale hubs for training and distributed edge locations for inference. This shift will redefine industry strategies, investment priorities, and the global footprint of data centers, shaping the next era of digital transformation.



Contact:

Jim Andrew
Partner
Mail: jim.andrew@fidepartners.com
LinkedIn: Jim Andrew

Freddy Farah Abuchaibe
Senior Consultant
Mail: freddy.farah@fidepartners.com
LinkedIn: Freddy Farah Abuchaibe



Miguel Caldas
Consultant
Mail: miguel.caldas@fidepartners.com
LinkedIn: Miguel Caldas

Direction:

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

London:
Aldwych House
London, WC2B 4HN
United Kingdom

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

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

Mexico