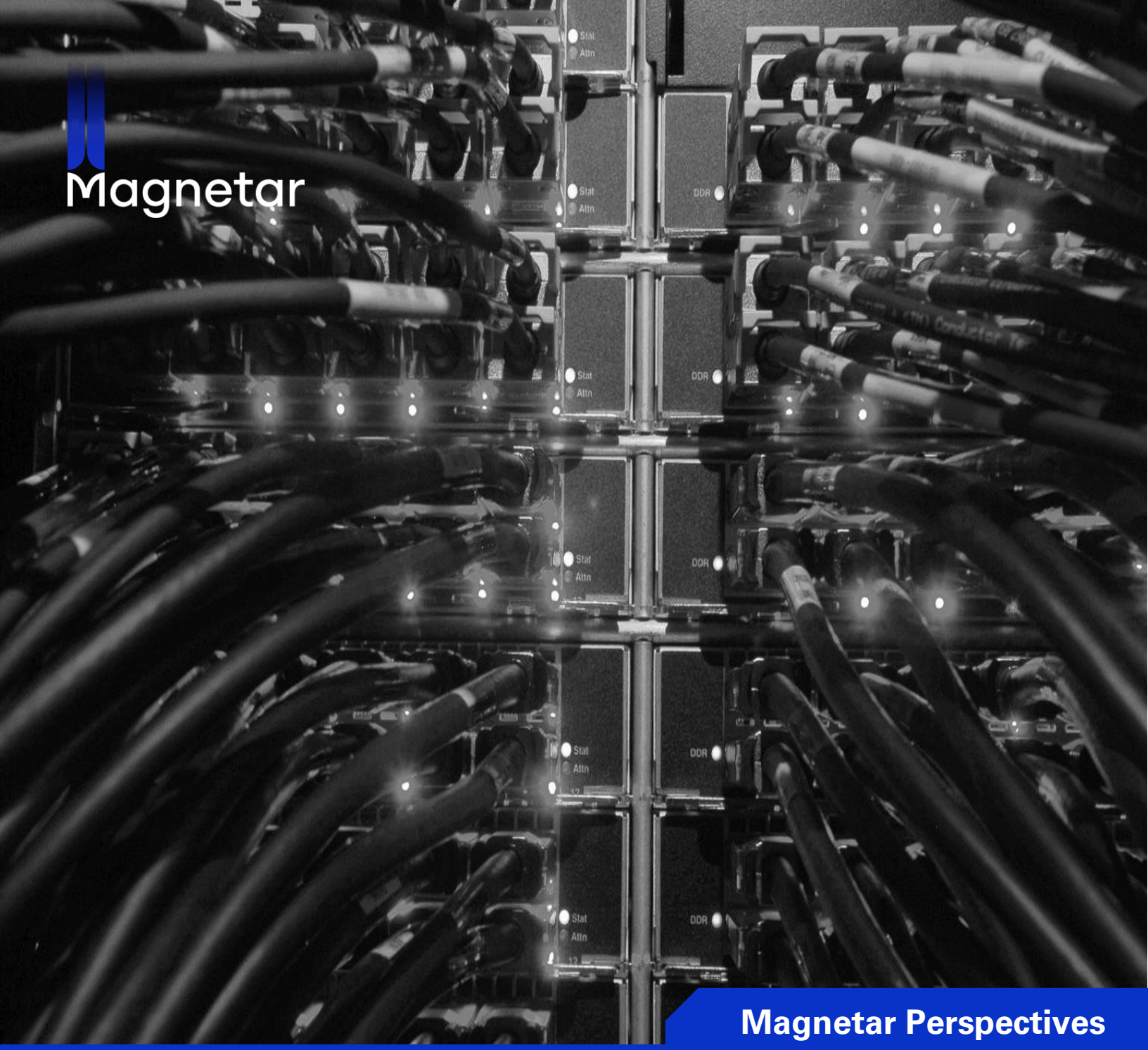


Magnetar Perspectives

Investing in AI Infrastructure - Part I: The Data Center of the Future

By Magnetar Ventures

December 2024

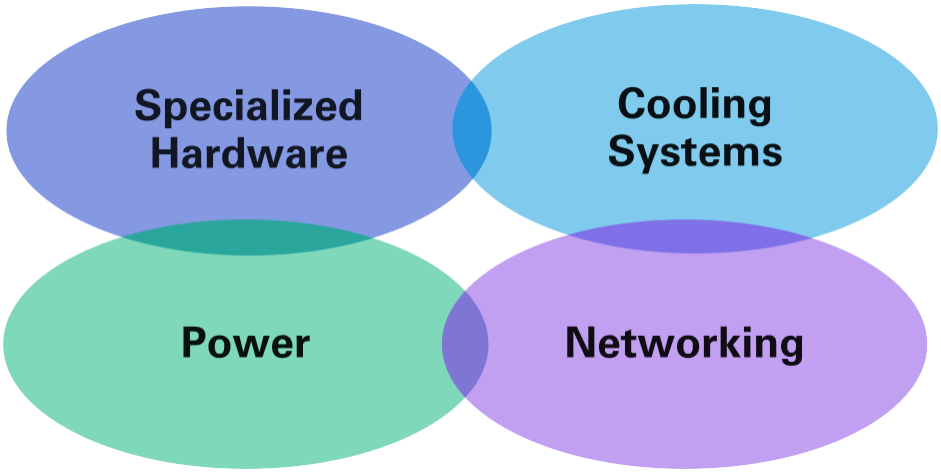
Introduction: AI Infrastructure Investing

The Future of AI Infrastructure

The rapid evolution of Generative AI platforms, fueled by breakthroughs in machine learning and artificial intelligence (AI), is driving unprecedented demand for the infrastructure that supports these technologies. Central to this infrastructure are data centers—the digital hubs where models are trained and deployed. However, the data centers of yesterday, designed for general computing, are ill-equipped to handle the demands of modern AI workloads.

Generative AI, particularly language models and image generation systems, require **High-Performance Computing (HPC)** environments capable of handling massive data throughput and computational power. To meet these requirements, traditional data centers must adopt significant advancements in hardware, power, cooling, and networking solutions. However, implementing these changes is not only capital-intensive but also demands deep expertise in both AI and infrastructure. Moreover, the vast infrastructure expansion required to keep pace with the growing demands of the AI revolution is immense. In this paper, we will further explain the computational demands of AI and their broader implications on AI infrastructure, underscoring why investing in next-generation datacenters is necessary. We will also explore the complexity of these capital-intensive investments and the specialized skillset required for success.

High Performance Computing Environment



The Unique Computational Demands of AI

The Expanding Scale of AI Models

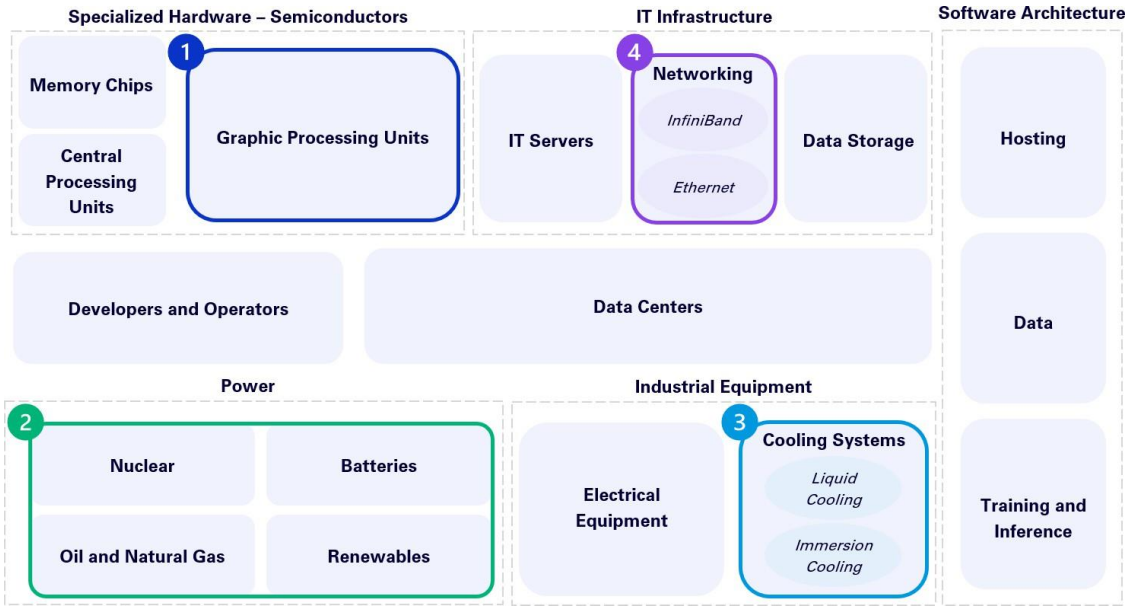
To understand the infrastructure required for AI, it's crucial to recognize the computational demands that modern AI models impose. A language model is a machine learning system designed to predict and generate coherent language, relying on internal variables called **Parameters** to fine-tune its performance. As the number of parameters increases, the model's

ability to refine its predictions and generate nuanced language improves. A model with more than 110 million parameters is typically classified as a **Large Language Model (LLM)**.[1] [2]

In recent years, LLMs have expanded rapidly, with the number of parameters increasing exponentially. OpenAI's GPT-2, released in 2019, contains 1.5 billion parameters, followed by GPT-3 in 2020 with 175 billion. By 2023, GPT-4 was estimated to surpass 1 trillion parameters. Each additional parameter results in more calculations during both **Training** and **Inference**, demanding greater computational capacity. As these models continue to grow in scale and require even more GPUs for training and inference, the supporting infrastructure must advance in tandem, necessitating substantial improvements in data center capabilities to accommodate the escalating computational demands and required HPC environment.[3]

The Requirements and Challenges of Implementing HPC Environments

HPC environments required for AI differ significantly from the environments of traditional data centers, as they operate on a larger scale and demand more advanced infrastructure. This includes specialized hardware like GPUs, power supplies capable of supporting intense computational demands, and enhanced cooling systems. Additionally, high uptime and low latency are critical for ensuring continuous, real-time processing and smooth operations, making advanced networking solutions equally essential for AI data centers. Implementing and maintaining these complex systems requires a highly skilled workforce with specialized engineers playing a crucial role in the successful deployment and operation of such intricate infrastructure. The diagram below shows how these requirements fit into the broader AI data center value chain[4] :



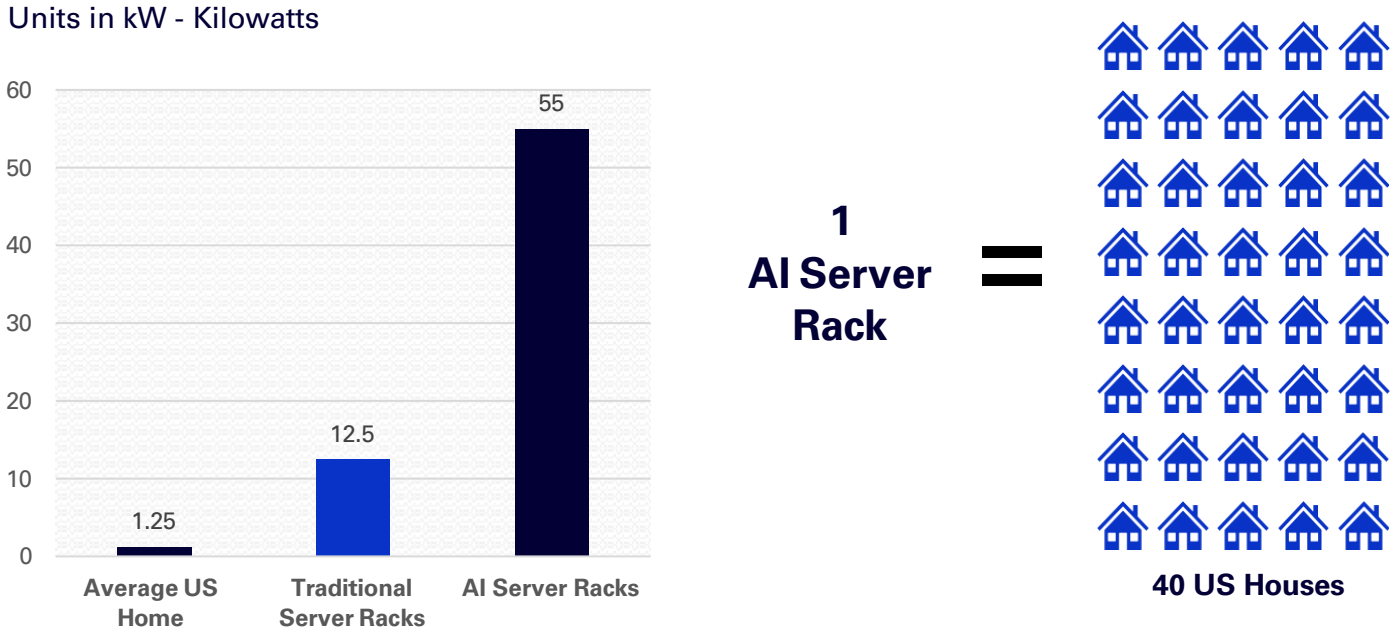
1. **Hardware (GPUs)**: The specialized hardware needed for AI, specifically high-end GPUs such as the NVIDIA's Hopper and soon to be released Blackwell series, is a significant

[1] Britannica
[2] Google Developers
[3] Expert Beacon
[4] Inspired by Generative Value

departure from the general-purpose CPUs used in traditional data centers. GPUs for AI are engineered for parallel processing, a critical requirement for efficiently training and running massive models. NVIDIA currently holds the dominant position with an 88% market share, but other companies such as AMD, SambaNova, Cerebras, and a long tail of new chip companies are also vying for market share.[5]

2. **Power**: AI data centers require significantly more power compared to traditional data centers due to the HPC environments needed to support AI workloads. Traditional server racks typically consume around 12-13 kW of power per rack, whereas AI server racks that house specialized hardware like GPUs, can consume 50-60 kW per rack. This considerable increase in power demand highlights the greater infrastructure needs of AI data centers.[6] For example, the average American household uses approximately 1.25 kW of energy. This means that a single AI server rack could demand the same amount of energy as about 40 U.S. homes – *Figure 1*. As AI workloads expand, new hyperscale data centers are being designed to consume over 1 GW (1,000,000 kW) of power with some server racks expected to require up to 100 kW. At these utilization levels, AI data centers could account for 20-25% of global energy consumption, highlighting the shift in power demand and the evolving infrastructure necessary to support these advancements.[7]

Figure 1: Power Consumption of AI Server Racks



3. **Cooling Systems**: The computational intensity of AI workloads generates significant heat making efficient cooling solutions essential: cooling accounts for about 30% of a data center's power consumption. Traditional air-cooling systems are becoming insufficient as

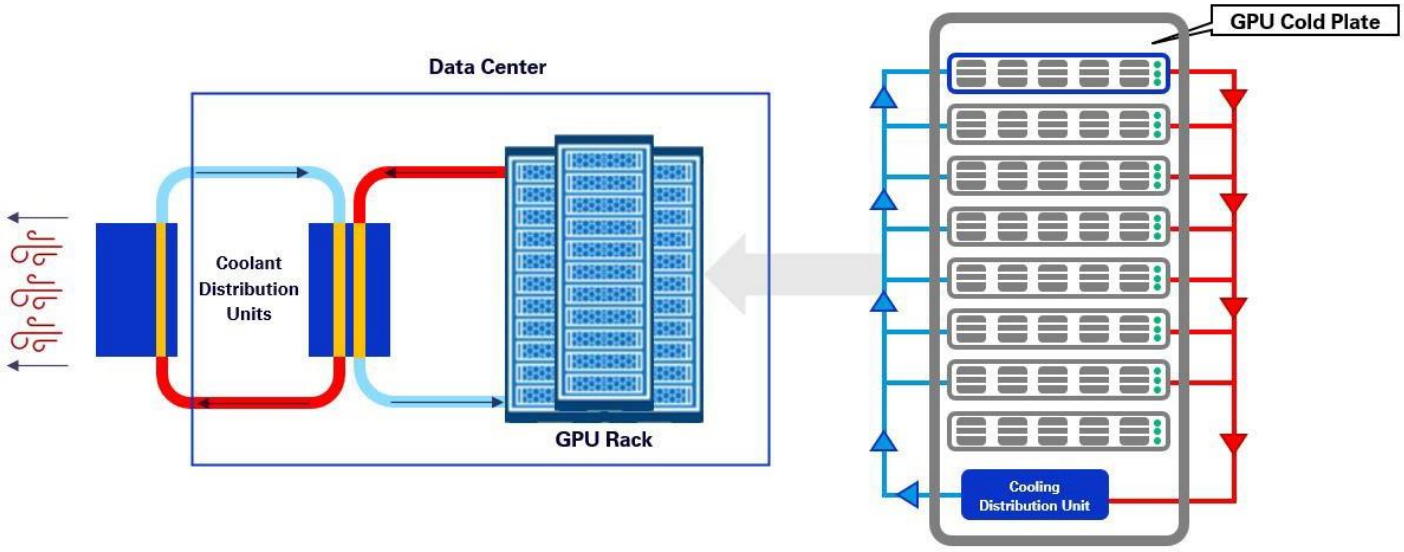
[5] John Peddie Research
[6] Creative Strategies
[7] Data Center Dynamics

power densities rise and heat generated in rack cabinets increase driving a shift towards **Liquid Cooling** and **Immersion Cooling**.

- **Liquid Cooling**: This technique circulates coolant through cold plates attached directly to GPU chips absorbing heat more efficiently than air cooling. Liquid cooling systems are more expensive and complex but offer better performance in dense AI environments.
- **Immersion Cooling**: A cutting-edge method where GPUs and other components are submerged in non-conductive liquid. This approach offers the highest cooling efficiency but comes with steep costs and complexity in installation and maintenance. As density increases, immersion cooling may become more prevalent in high-end HPC data centers.

While more complex and costly, both liquid cooling and immersion cooling can reduce cooling power consumption by up to 60% and are better suited for handling the high-power densities required by AI. Hyperscale providers are at the forefront of adopting these cooling methods.[8]

Figure 2: Liquid Cooling System



The Figure on the left represents the datacenter at large while the figure on the right provides a close-up view of individual server racks, highlighting the internal cooling mechanisms that operate within each rack.

4. **Networking**: AI workloads require high-speed, low-latency networking solutions to handle the immense volumes of data transferred between GPUs and other processing units. Technologies like InfiniBand and high-speed Ethernet lead in this space; InfiniBand, with

[8] Enconnex

its ultra-low latency, high throughput, and in-network computing capabilities, optimizes data flow and accelerates AI model training and inference.[9] Ethernet is also widely used for AI workloads due to its cost-effectiveness and scalability, especially with new AI-optimized Ethernet solutions. To further improve networking efficiency, innovations like **Photonics** and **Optical Networking** are being explored. These technologies aim to enhance bandwidth and reduce latency by using light to transfer data, significantly boosting throughput and lowering power consumption in AI-driven data centers.[10]

Ensuring high-speed, efficient networking alone is not enough; AI data centers must also achieve high **Uptime**, the percentage of time a system remains fully operational without interruptions. In mission-critical AI environments, continuous operation is essential. Tier 3 data centers offer 99.982% uptime, while Tier 4 facilities provide an even higher reliability with 99.995% uptime. These tiers are particularly suited for AI workloads due to the **Redundancy**, **Fault Tolerance**, and reliability needed to manage the computational demands of advanced AI infrastructure. [11]

5. **Personnel Expertise:** One of the key bottlenecks in deploying large-scale AI infrastructure is the shortage of qualified data center engineers. Each component of a data center, from the computational hardware to the cooling systems, is complex and requires deep technical knowledge. As the demand for AI-driven data centers continues to grow, so too does the need for highly trained professionals who can design, install, and maintain these systems efficiently. Without this specialized talent, the pace of scaling AI infrastructure will inevitably slow, creating delays in meeting the operational and computational demands of AI workloads. Addressing this talent gap is crucial to ensuring the continued expansion and efficiency of AI infrastructure at scale.

Navigating the Complexity of AI Infrastructure

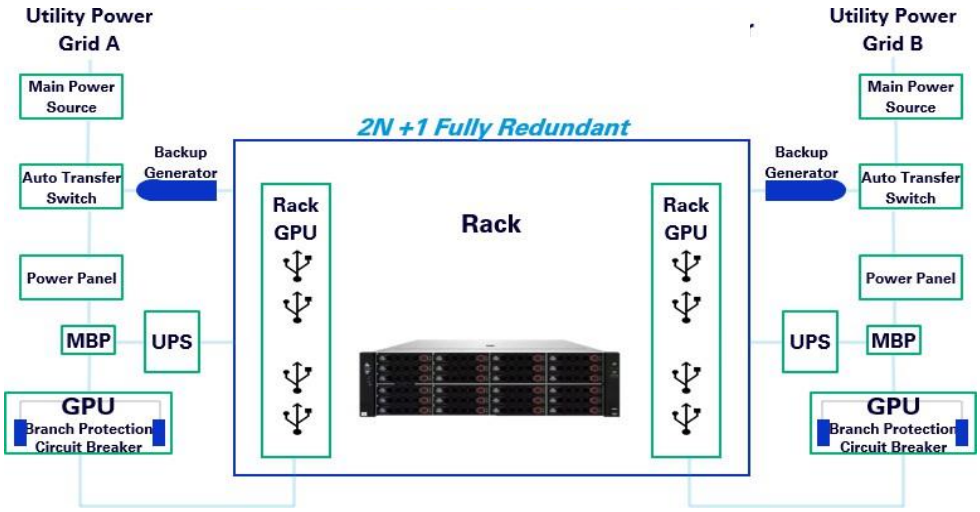
AI infrastructure is a highly complex ecosystem that demands specialized knowledge and expertise to manage effectively. At the core are GPUs, which perform the massive parallel computations essential for AI model training. Unlike consumer GPUs, those used in HPC for AI are larger, far more advanced, and rely on **High-Bandwidth Memory (HBM)** to enable high data transfer rates and efficient computation—adding both sophistication and intricacy. Meeting power requirements for AI infrastructure involves more than just supplying high capacity; it requires designing systems with multiple redundant pathways, continuous monitoring, and advanced distribution methods to ensure uninterrupted service. Hyperscale data centers, for instance, rely on multi-layered power architectures, such as **2N + 1 Power Redundancy** in Tier 4 facilities, to prevent outages and maintain high uptime despite hardware failures.

[9] Nvidia

[10] Broadcom

[11] Colocation America

Figure 3: Typical Tier 4 Data Center



Growing power density also necessitates advanced cooling solutions, such as liquid and immersion cooling systems, which come with complex installation and operational requirements. For instance, liquid cooling systems alone require detailed engineering to manage compressors, pumps, and submerged networks, all of which underscore the need for specialized maintenance and expertise to ensure these systems operate efficiently. Networking infrastructure is equally crucial as AI workloads rely on low-latency, high-throughput connections to handle vast volumes of data, pushing Hyperscalers to invest in advanced networking solutions.

Together, these elements illustrate a multifaceted AI infrastructure that requires deep knowledge of hardware, power, cooling, and networking to manage effectively and to support the continued evolution of AI technologies.

Capital Intensity of AI Infrastructure

The capital required to build and sustain AI infrastructure is immense, particularly as AI workloads demand specialized, high-performance data centers. A single hyperscale AI data center with a capacity of 1,000 MW (1,000,000 kW) exemplifies the scale of these costs. AI server racks alone, which consume around 60 kW each, represent a significant expense. Assuming each rack holds approximately 8 GPUs, and with GPUs priced between $30,000 and $40,000, the cost for GPUs alone ranges from $4 billion to $6 billion[12].

Beyond hardware, substantial investment is also required to support the operational needs of these data centers. Cooling systems, essential to manage the heat generated by AI workloads, can add another $1 billion to $2 billion in installation costs alone, with liquid cooling systems

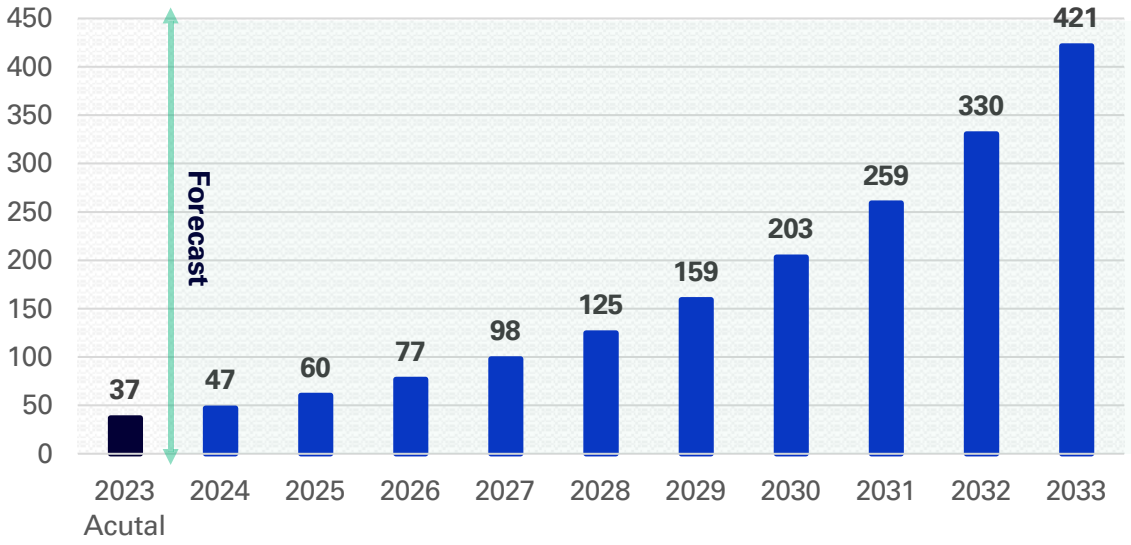
[12] Nvidia

priced between $1,000 and $2,000 per kW of cooling capacity.[13] Power distribution and networking infrastructure contribute additional expenses, as these data centers must be equipped with advanced power systems and high-throughput networking to handle the massive data flow required by AI.

Given these substantial costs and the rapidly growing demand, the AI infrastructure market is positioned for substantial and exponential growth. As shown in Figure 5, the market is projected to grow at a CAGR of 27.53% over the next decade, expanding to more than 11 times its current size by 2033. This impressive trajectory underscores the vast investment opportunities within AI infrastructure.[14]

Figure 4: AI Infrastructure Market Size

2023 Actual + Forecast (Billions)



Conclusion: AI Infrastructure Investment

The rapid expansion of AI, especially with advancements in Generative AI and Large Language Models, is driving unprecedented demand for advanced infrastructure. Traditional data centers are no longer sufficient to meet the computational, power, and cooling requirements of these AI workloads, opening significant investment opportunities in HPC environments, specialized GPUs, and advanced networking solutions like InfiniBand. As AI workloads continue to scale, hyperscale data centers require substantial power capacity, with utilities projected to invest billions to meet these demands. Ensuring uptime and sustainability also hinges on power redundancy, robust backup systems, and renewable energy integration.

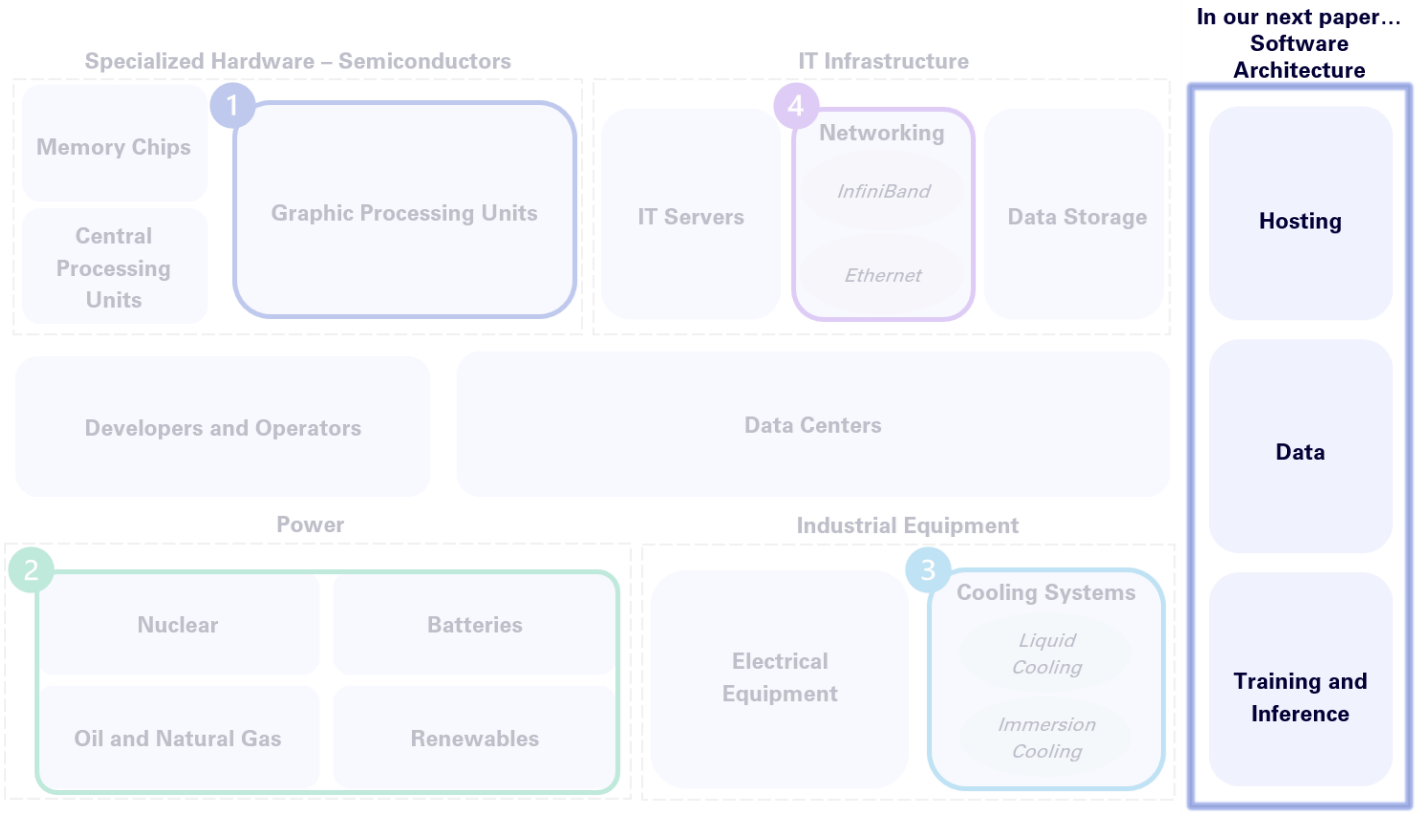
Furthermore, cooling technologies such as liquid and immersion cooling are now essential as AI environments generate immense heat, requiring substantial capital for efficient thermal management. Implementing and managing this advanced infrastructure demands highly

[13] Econnex
[14] Precedence

specialized talent, particularly engineers skilled in AI-specific hardware, power, and networking solutions. Without this expertise, the complexities of deploying and maintaining AI data centers can become significant obstacles.

Ultimately, only informed investors who grasp the nuances of AI infrastructure will be able to capitalize fully on these emerging opportunities. The specialized nature of AI hardware, networking, and power systems requires a sophisticated understanding of the market to navigate successfully. Those equipped with the right knowledge and talent will be well-positioned for long-term growth as AI continues to evolve.

Stay tuned and join us in exploring these themes in part II of this series, Investing in AI Infrastructure - Part II: The SW Architecture layer.



GLOSSARY

Large Language Models: an AI system trained on vast text data to generate and understand human language.

High Performance Compute: Use of powerful computing systems, often clusters of servers or supercomputers, to perform complex tasks that require substantial processing power, such as simulations, data analysis, and AI model training.

Parameters: Internal values the model adjusts during training to learn patterns in data and make predictions, such as generating the next word in a sentence.

Training: Process where a model learns patterns from data by adjusting its internal parameters to improve accuracy.

Inference: Process where a trained model makes predictions or generates outputs based on new, unseen data.

GPUs: Specialized processors designed for parallel processing, commonly used to accelerate tasks such as rendering graphics and training AI models by handling large amounts of data simultaneously.

Liquid Cooling: System that uses liquid coolant to absorb and remove heat from components, such as GPUs, offering more efficient cooling than traditional air-based methods

Uptime: Amount of time a system, such as a server or data center, is operational and available without interruptions.

Redundancy: The duplication of components or processes to ensure reliability and continuous operation in case of failures.

Fault Tolerance: The ability of a system to continue functioning correctly even when part of its components fails, ensuring uninterrupted performance and minimal downtime.

Low-Latency: Fast response time between input and output, critical for real-time applications

Ethernet: Networking technology that connects devices in a local area network for data transmission through wired connections

InfiniBand: Networking technology used primarily in data centers and supercomputing, offering low-latency and high-bandwidth communication for interconnecting servers, storage, and other computing resources

Photonics: A technology that uses light to transmit and process information.

Optical Networking: Communication technology that uses light transmitted through fiber optic cables to enable high-speed data transfer over long distances with minimal signal loss, commonly used in internet and telecommunications infrastructure.

High Bandwidth Memory: Type of fast, high-performance memory used in GPUs and other processors. It stacks memory chips vertically and uses a wide bus to provide much higher data transfer speeds compared to traditional memory, making it ideal for tasks requiring large amounts of data processing

2N + 1 Power Redundancy: Power backup system design that provides double the required capacity (2N) plus an additional backup (+1) to ensure continuous operation. This setup ensures that even if both primary and secondary power sources fail, there is an extra layer of backup to prevent downtime.

Sources:

(1) Britannica - https://www.britannica.com/topic/large-language-model
(2) Google Developers - https://developers.google.com/machine-learning/resources/intro-llms
(3) Expert Beacon - https://expertbeacon.com/gpt-4-parameters/
(4) Generative Value - A Primer on AI Data Centers - by Eric Flaningam
(5) John Peddie Research - https://www.jonpeddie.com/news/shipments-of-graphics-add-in-boards-decline-in-q1-of-24-as-the-market-experiences-a-return-to-seasonality/
(6) Creative Strategies - https://creativestrategies.com/research/data-center-evolution-ai-changing-datacenter-design-strategies/
(7) Data Center Dynamics - https://www.datacenterdynamics.com/en/marketwatch/navigating-ais-growing-impact-on-data-center-power/
(8) Enconnex - https://blog.enconnex.com/data-center-cooling-costs-and-how-to-reduce-them
(9) Nvidia - https://info.nvidia.com/accelerate-ai-workloads-with-infiniband-webinar.html
(10) Broadcom - https://www.broadcom.com/blog/optimizing-the-network-for-ai-workloads
(11) Colocation America - https://www.colocationamerica.com/data-center/tier-standards-overview
(12) Nvidia - https://www.barrons.com/livecoverage/nvidia-gtc-ai-conference/card/nvidia-ceo-says-blackwell-gpu-will-cost-30-000-to-40-000-l0fnByruULe4RAdr4kPE
(13) Enconnex - https://blog.enconnex.com/data-center-cooling-costs-and-how-to-reduce-them
(14) Precedence - https://www.linkedin.com/pulse/artificial-intelligence-ai-infrastructure-market-prathamesh-sakpal-tgqnf/

IMPORTANT DISCLOSURES

The views expressed in this commentary are the views of Magnetar Capital's Ventures investment team ("Ventures Investment Team"). The views expressed reflect the current views of the Ventures Investment Team as of the date hereof, and neither the Ventures Investment Team nor Magnetar Capital (together with its affiliates, "Magnetar") undertake any responsibility to advise you of any changes in the views expressed herein.

This piece is for informational purposes only and is not, and may not, be relied on in any manner as legal, tax, investment, accounting or other advice, or as an offer to sell, or a solicitation of an offer to buy, any security or instrument in or to participate in any trading strategy with any Magnetar fund, account or other investment vehicle (each a "Fund"), nor shall it or the fact of its distribution form the basis of, or be relied on in connection with, any contract or investment decision. If such offer is made, it will only be made by means of an offering memorandum (collectively with additional offering documents, the "Offering Documents"), which would contain material information (including certain risks of investing in such Fund) not contained in this document and which would supersede and qualify in its entirety the information set forth in the document. Any decision to invest in a Fund should be made after reviewing the Offering Documents of such Fund, conducting such investigations as the investor deems necessary and consulting the investor's own legal, accounting and tax advisers to make an independent determination of the suitability and consequences of an investment in such Fund. In the event that the descriptions or terms described herein are inconsistent with or contrary to the descriptions in or terms of the Offering Documents, the Offering Documents shall control. None of Magnetar, its funds, nor any of their affiliates makes any representation or warranty, express or implied, as to the accuracy or completeness of the information contained herein and nothing contained herein should be relied upon as a promise or representation as to past or future performance of a Fund or any other entity, transaction, or investment.

Recipients should bear in mind that past performance does not predict future returns and there can be no assurance that a Fund will achieve comparable results, implement its investment strategy, achieve its objectives or avoid substantial losses or that any expected returns will be met.

Investment concepts mentioned in this commentary may be unsuitable for investors depending on their specific investment objectives and financial position. Tax and regulatory considerations, margin requirements, commissions and other transaction costs may significantly affect the economic consequences of any transaction concepts referenced in this commentary and should be reviewed carefully with one's investment and tax advisors.

All information in this commentary is believed to be reliable as of the date on which this commentary was issued and has been obtained from public sources believed to be reliable. No representation or warranty, either express or implied, is provided in relation to the accuracy or completeness of the information contained herein.

This commentary discusses broad market, industry, or sector trends, or other general economic, market, regulatory or political conditions and should not be construed as research, investment advice, or any investment recommendation.

Unless otherwise stated, references to general venture initiatives, priorities or practices are not intended to indicate that Magnetar has materially contributed to such actions and such initiatives, priorities, or practices are subject to change, even materially, over time. This content is provided for informational purposes only and there is no guarantee that Magnetar will invest in similar opportunities in the future.

Third-Party Information. Certain of the information contained in this piece has been obtained from sources outside Magnetar and could prove to be incomplete or inaccurate and is current only as of any specific date(s) noted therein. Magnetar makes no representations as to the accuracy or completeness of such information contained in this piece and neither Magnetar nor any of its affiliates takes any responsibility for, and has not independently verified, any such information. In particular, you should note that, since many investments may be unquoted, net asset value figures in relation to funds may be based wholly or partly on estimates of the values of such funds' investments provided by the originating banks of those underlying investments or other market counterparties, which estimates may themselves have been subject to no verification or auditing process or may relate to a valuation at a date before the relevant net asset valuation for such fund, or which have otherwise been estimated by Magnetar.

No Assurance of Investment Return. Prospective investors should be aware that an investment in a Fund is speculative and involves a high degree of risk. There can be no assurance that a Fund will achieve comparable results, implement its investment strategy, achieve its objectives or avoid substantial losses or that any expected returns will be met. A Fund's performance may be volatile. An investment should only be considered by sophisticated investors who can afford to lose all or a substantial amount of their investment. A Fund's fees and expenses may offset or exceed its profits. There can be no assurance that SRT initiatives will be successful. A decision to invest should take into account the objectives and characteristics of the relevant Fund as set out in more detail in the applicable Offering Documents.

Trends. There can be no assurances that any of the venture trends described herein will continue or will not reverse. **Past performance does not predict future returns.**

Forward-Looking Statements. This commentary may contain "forward-looking statements within the meaning of the safe harbor provisions of the Private Securities Litigation Reform Act of 1995". You can identify these forward-looking statements by the use of words such as "outlook," "indicator," "believes," "expects," "potential," "continues," "may," "will," "should," "seeks," "approximately," "predicts," "intends," "plans," "scheduled," "estimates," "anticipates," "opportunity," "leads," "forecast" or the negative version of these words or other comparable words. Forward-looking statements are not historical facts. All forward-looking statements are subject to various factors, including, without limitation, economic and market conditions, changing levels of competition within certain industries and markets, changes in interest rates, changes in legislation or regulation, and other competitive, governmental, and regulatory factors affecting Magnetar's operations, the applicable Fund's or potential vehicle's operations, and the operations and performance of any companies/securities/instruments identified herein, any or all of which could cause actual results to differ materially from estimated or projected results. The forward-looking statements speak only as of the date hereof, and we undertake no obligation to publicly update or review any forward-looking statement, whether as a result of new information, future developments or otherwise.