Amphenol
NETWORK SOLUTIONS™

Design Data Center Infrastructure for Generative AI

White Paper

Written by: Charles Su (PhD), *Senior Optical Engineer*

August 30, 2023

Contents

Figures

1. Overview

Generative AI can significantly impact and benefit various aspects of an enterprise. Recently, generative AI technologies powered by models like GPT-3, have shown remarkable potential in transforming the way businesses operate and engage with their customers. Here we just name some applications of generative AI for enterprise:

- *Software Development and Networking*: Generative AI can assist IT teams by automatically generating code snippets, scripts, and configurations for various tasks in software development and networking. This can speed up the development process and ensure consistency in code quality.
- *Troubleshooting and Issue Resolution:* AI-powered solutions can analyze problem descriptions and provide potential solutions, helping IT teams troubleshoot and resolve issues more efficiently.
- *Process Automation:* Businesses can leverage generative AI to automate routine tasks and workflows, freeing up human resources for more complex and creative tasks.
- *Training and Onboarding:* Generative AI can create training materials and interactive guides, facilitating the onboarding process for new employees and providing continuous learning resources.
- *Documentation and Knowledge Management:* AI can generate documentation, FAQs, and other knowledge resources, ensuring that important information is easily accessible to both employees and customers.
- *Project Management and Planning:* AI can assist in project management tasks by analyzing data, generating reports, and providing insights that aid decision-making.
- *Call Centers and Customer Service:* AI-powered virtual assistants can handle customer inquiries, provide information, and even resolve common issues, enhancing the customer service experience.
- *Data Analytics:* Generative AI can help in generating reports, visualizations, and even generating insights from complex datasets, aiding data-driven decision-making.
- *Content Creation:* From writing articles to social media posts, AI can generate content quickly, adhering to specified styles and tones.

- *Design and Development:* AI can assist in designing graphics, user interfaces, and even prototypes based on input criteria.
- *Predictive Maintenance:* In industrial settings, generative AI can analyze sensor data to predict and prevent equipment failures, optimizing maintenance schedules.

The adoption of generative AI in these areas can lead to increased efficiency, improved customer experiences, reduced human error, and the ability to tackle complex tasks more effectively.

Fig.1 Data Center [1]

2. Will existing data center infrastructures be able to handle the growing workload from generative AI?

The scale and complexity of modern generative models, like GPT-4 with potentially over a trillion parameters, pose significant challenges in terms of computing power, infrastructure, and optimization. Usually, the generative AI process has two phases: training the large language models (LLMs) that form the core of generate AI systems, and operating the application with these trained LLMs.

Fig. 2 An AI/ML Data Center [2]

Phase 1: Training the Large Language Models (LLMs)

Training massive language models involves processing enormous datasets through complex neural networks. This process requires substantial computational resources, typically in the form of high-performance computing (HPC) clusters equipped with GPUs (Graphics Processing Units) or TPUs (Tensor Processing Units). These clusters run for extended periods, often for weeks or months, to fine-tune the model's parameters through iterative processes.

Phase 2: Inference and Operational Phase

Once the models are trained, they are deployed for real-world use in various applications. In this operational phase, the infrastructure requirements change. Low-latency access becomes crucial, as users expect quick responses. This often calls for a distributed infrastructure spread across different geographical locations to reduce latency. This requirement can be challenging to address with a traditional centralized cloud model, necessitating the establishment of data centers in multiple locations.

Scaling Challenges and Cost Implications

As models become more complex, the computational demands increase exponentially. Distributing the computation across multiple accelerators and servers is a common approach to handling these demands. However, this can lead to higher costs due to the need for more hardware and infrastructure to support such distributed computing setups.

Additionally, the memory capacity and performance requirements of these models can quickly outpace the capabilities of even the most powerful cloud-based GPUs and TPUs. This presents challenges in optimizing hardware and software to efficiently run the algorithms, as well as potentially requiring specialized hardware architectures to accommodate these demands.

Given the ongoing trend toward even larger models with more parameters, it's likely that addressing these computational challenges will continue to be a priority. Innovations in hardware architecture, such as more memory-efficient designs and accelerators specifically tailored for AI tasks, might play a crucial role in mitigating some of these challenges. Furthermore, optimizing model architectures and training techniques to achieve better results with fewer parameters could also help alleviate the strain on infrastructure.

Challenges When Designing Infrastructure Data Centers for AI

Planning and designing machine learning infrastructure for data centers, especially for HPC and generative AI workloads, requires a deep understanding of transmission protocols, network topologies, and physical connectivity. The decisions made in these areas directly impact the system's performance, scalability, and resiliency, and play a vital role in achieving efficient and effective machine learning operations.

- **Transmission Protocols:**
  Ethernet is a widely used and flexible network protocol known for its scalability and cost-effectiveness. However, its latency and throughput might not be optimal for training AI applications. To address this, Remote Direct Access Memory (RDMA) protocols supported by technologies like InfiniBand offer high-speed data transfers with low latency, making them suitable for AI/ML workloads [3].

  RDMA over Converged Ethernet (RoCE) is an emerging protocol [4] that aims to combine the scalability and cost-effectiveness of Ethernet with the low-latency advantages of RDMA. It's being considered as a potential protocol of choice for hyperscale data centers supporting AI applications.

- **Topologies:**
  The topology of a data center network greatly affects performance, reliability, and cost. InfiniBand is managed differently than Ethernet and can offer advantages in terms of performance and low latency. When designing a network for AI/ML clusters, particularly for large-scale models like LLMs, choosing the right topology is crucial. The design trade-offs between performance, reliability, and cost must be carefully considered.

  Scalability is a key consideration, especially in the context of training large language models. InfiniBand clusters can be scaled by adding additional switches, necessitating a flexible and easily expandable fiber infrastructure.

- **Physical Connectivity:**
  The physical connectivity within a data center involves the actual cables, connectors, and pathways that enable communication between different hardware components. It's important to have reliable and high-quality physical connectivity to ensure data can flow without interruptions or delays. As data centers grow in scale, managing physical connectivity becomes more challenging, and maintaining efficient airflow and cable

management is essential to avoid thermal issues and performance bottlenecks. Active optical cables (AOCs) and pluggable optics are commonly used for short-reach connections inside data centers. AOCs offer simplicity in installation, as they don't require intricate connector cleaning and inspection skills. However, they lack the flexibility of pluggable transceivers.

Pluggable transceivers provide more flexibility, making it easier to upgrade network links over time without the need to replace the entire cable. This is particularly advantageous for accommodating future data rate increases.

Balancing these considerations while accounting for the limited power and cooling capacity of data centers is a complex task. Efficiently distributing power to components while managing heat generation is critical to prevent thermal throttling and ensure consistent performance. Designers often need to make trade-offs between hardware performance, power consumption, and cooling solutions to optimize the system's overall efficiency.

For generative AI, the computational demands are substantial, and designing hardware that can handle these demands without overheating is a significant challenge. Advanced thermal management techniques, such as liquid cooling and innovative airflow designs, are being explored to address this issue.

The Implementation of Data Center with AI Clusters

When it comes to implementing AI clusters within data centers, there are unique challenges and architecture considerations. The distinction between front-end and back-end networks and the demands of AI workloads are crucial to address for efficient and effective operations. Below is a breakdown of the key points:

- **Front-End and Back-End Network Architecture:**
  Traditionally, data centers have been organized into front-end and back-end networks. The front-end network handles conventional computing workloads, while the back-end network caters to the demands of AI clusters. AI clusters require a new architecture due to the specific requirements of GPU servers and the connectivity demands.
- **GPU Server Connectivity:**
  GPU servers used in AI clusters require significantly more interconnectivity between servers. While the number of servers per rack might be limited due to power and heat constraints, the connectivity demands increase. This can lead to a higher amount of inter-rack cabling compared to traditional data centers.

- **NVIDIA DGX Platform:**
  NVIDIA's DGX H100 [5] is an example of an enterprise AI server architecture. It showcases the connectivity demands of AI clusters. The DGX H100 has multiple high-speed ports for switches, storage, and management. In a DGX SuperPOD, which contains multiple GPU servers, the number of fiber and copper links grows significantly due to the scale of the deployment. To meet these challenges, the HyperX™ Optical Distribution Frame solution[6] from Amphenol Network Solutions is engineered to safely and efficiently manage and scale large amounts of fiber patching, up to 3,456 fibers per frame for example.

- **Latency Sensitivity:**
  AI and machine learning algorithms, like high-performance computing, are extremely sensitive to network latency. Even small reductions in latency can have a substantial impact on the overall performance and cost of running large training models. Latency reduction becomes a priority, as AI training time is valuable, and any delay can lead to significant costs.

- **Limited Reach Considerations:**
  AI clusters are designed with proximity in mind. Servers within an AI cluster are expected to be close together to minimize latency. To maintain low-latency links, most connections should be limited to relatively short distances, often within 100 meters. This design constraint ensures that data can be exchanged quickly and efficiently between servers

  It is very important to balance connectivity, latency, and power considerations when designing AI cluster architecture. The demands of AI workloads necessitate a different approach compared to traditional computing clusters, and as AI continues to drive innovation, optimizing network design and latency management will remain essential for realizing the full potential of these systems.

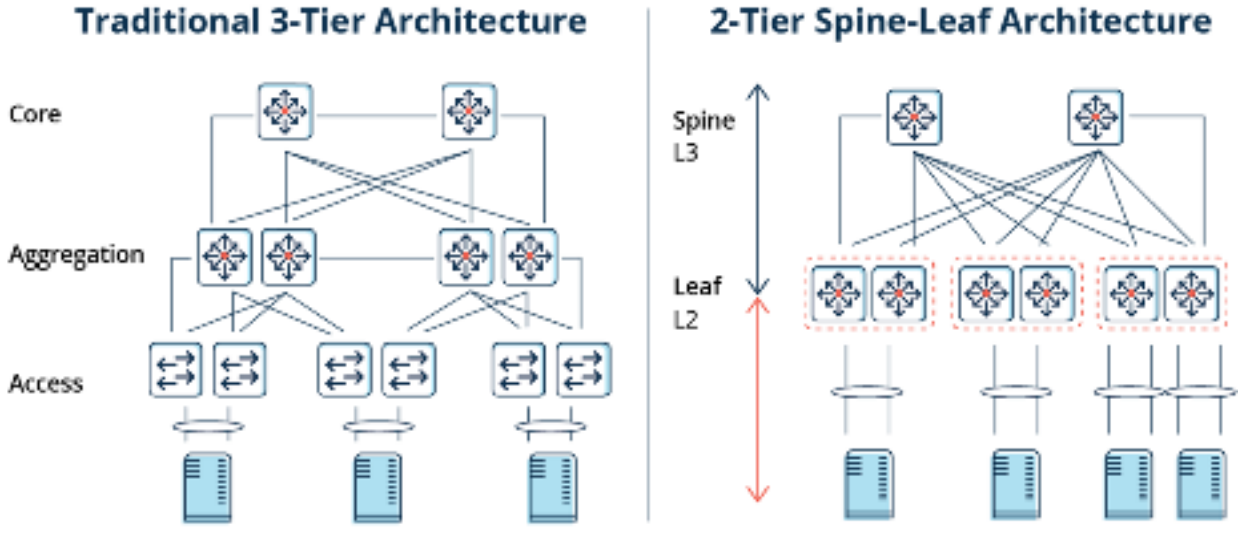
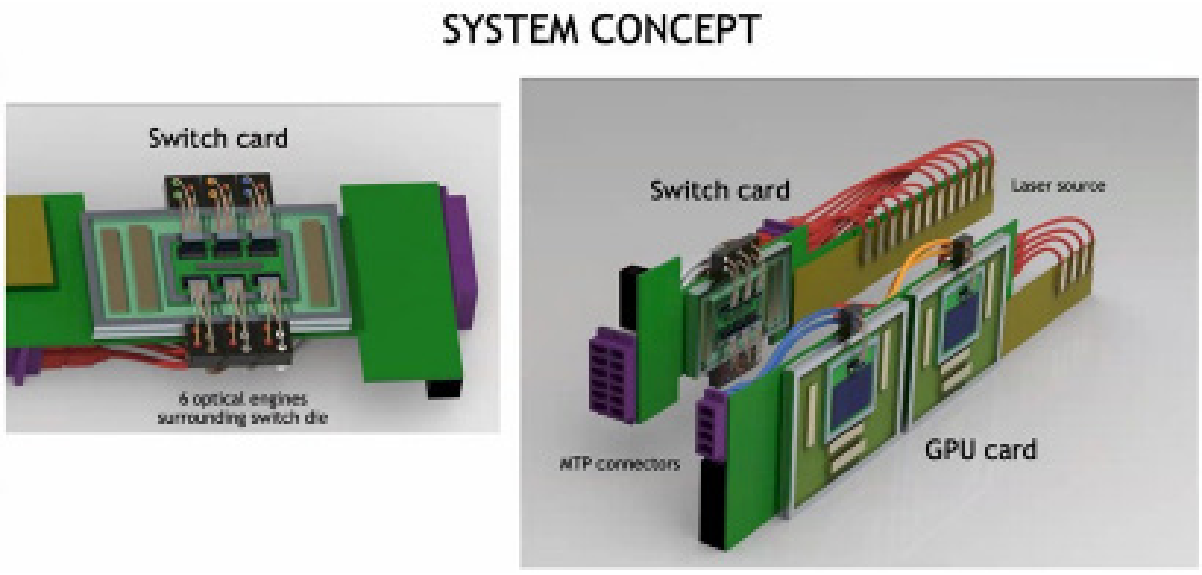


Fig. 3 Transforming traditional 3-tier DC to the 2-tier DC with AI cluster [7]



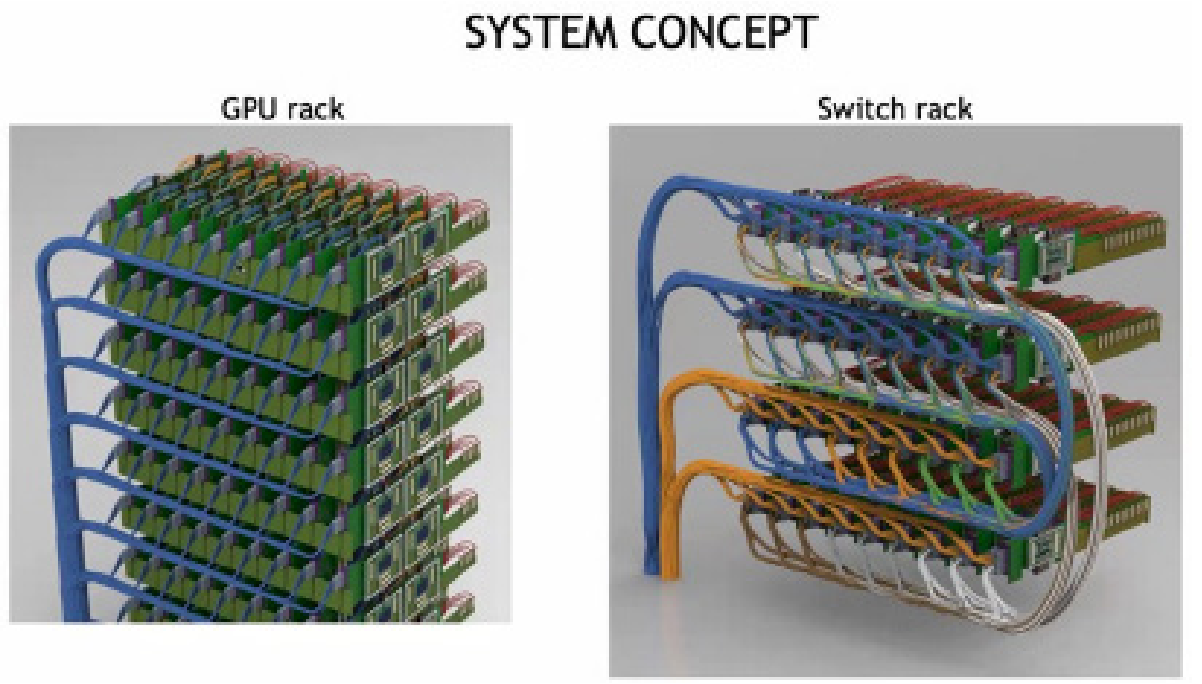


Fig.4 Nvidia optical interconnect concept from GPU to Switch rack [8]

3. Summary

Incorporating the following considerations into data center planning and implementation is crucial to effectively harness the power of AI, especially in the context of generative AI where the computational demands are particularly high. The evolution of data center architecture to accommodate these advancements will play a significant role in shaping the future of AI technologies and applications.

- *Denser Optical Connections:* The data center landscape is evolving to support the next generation of AI and ML workloads. This evolution includes denser optical connections to handle the increased demands for high-speed data transfers and low-latency communication. These denser connections are essential for interconnecting the numerous processors required for training and executing advanced AI models.
- *Network Designs and Transmission Protocols:* Traditional network designs might not be optimal for AI and ML clusters. New network designs, transmission protocols like RDMA, and technologies like RoCE are being adopted to ensure efficient data movement and minimize latency, both critical for AI training and inference.
- *Cabling Considerations:* Careful planning of cabling infrastructure is crucial. Proper cabling within AI clusters can result in significant cost savings, reduced power consumption, and faster installation times. Well-designed cabling can also contribute to improved efficiency and overall performance, particularly when handling large-scale AI training workloads.
- *Cost and Power Efficiency*: Optimizing cabling, network connectivity, and other data center components isn't just about improving performance; it's also about increasing cost and power efficiency. Efficient data movement and reduced latency can result in faster training times, which can translate to cost savings considering the substantial expenses associated with AI model training.
- *Benefits of AI in Networking:* AI itself is playing a role in enhancing data center networking. AI-driven analytics can help identify network bottlenecks, predict failures, and optimize network configurations, further improving the overall efficiency and reliability of AI/ML workloads.
- *Future-Proofing for AI:* Adapting data center infrastructure to support AI/ML workloads is a strategic move. As AI technology continues to advance and models become even larger and more complex, having a flexible and robust infrastructure in place will position organizations to fully leverage the benefits of AI-driven insights and innovation.

At Amphenol Network Solutions [9], we partner with our customers to help design and manufacture the components that can address AI/ML requirements in data centers. We can offer products and solutions engineered for enabling high speed and power efficiency in your data center architectures.

Please visit our website https://amphenol-ns.com/Solutions for more details about our unique solutions for your data center applications.

4. References

[1] https://news.microsoft.com/de-ch/2021/06/10/new-offers-from-the-swiss-data-centers-of-the-microsoft-cloud/

[2] https://www.datacenterdynamics.com/en/analysis/making-an-ai-write-about-data-centers/

[3] https://www.nvidia.com/en-us/data-center/nvlink/

[4] https://en.wikipedia.org/wiki/RDMA_over_Converged_Ethernet

[5] www.nvidia.com/en-us/data-center/dgx-platform

[6] https://amphenol-ns.com/Our-Products/Fiber-Distribution/HyperX

[7] https://medium.com/@shihlin-chang/transforming-data-centers-with-ai-f86a4d03c566

[8] https://www.nextplatform.com/2022/08/17/nvidia-shows-what-optically-linked-gpu-systems-might-look-like/

[9] https://amphenol-ns.com/

About the Author

Charles Su (PhD)
Senior Optical Engineer

Charles Su, PhD, is a seasoned Senior Optical Engineer with over 20 years of experience in the telecom industry. Specializing in optical fiber components and systems, he has demonstrated strong leadership capabilities, successfully leading teams and delivering impactful solutions to address complex challenges. With a deep understanding of fiber optics technologies, Charles Su is renowned for his forward-thinking approach to next-generation product development. He has a passion for understanding and meeting customer needs, consistently developing innovative solutions that exceed expectations. Committed to driving customer success, Charles Su leverages his extensive industry knowledge and dedication to continuous improvement to deliver exceptional results.

About Amphenol

At Amphenol Network Solutions, we are driven by a passion for innovation and a relentless commitment to creating customized solutions that seamlessly integrate with your unique requirements. With our deep understanding of fiber optic technology, we specialize in creating tailored solutions that anticipate and adapt to the rapidly evolving demands of your network. Through our responsive support, unwavering commitment, and ongoing collaboration, we ensure that our solutions are ready to deliver superior performance and reliability.

509.926.6000

getinfo@amphenol-ns.com

www. amphenol-ns.com