Enterprise Strategy Group™ by TechTarget

WHITE PAPER

Enabling Ethernet-powered Solutions for GenAI

The Importance of Open Networking

By Bob Laliberte, Principal Analyst
Enterprise Strategy Group

January 2024

Contents

AI Infrastructure Is Growing Rapidly

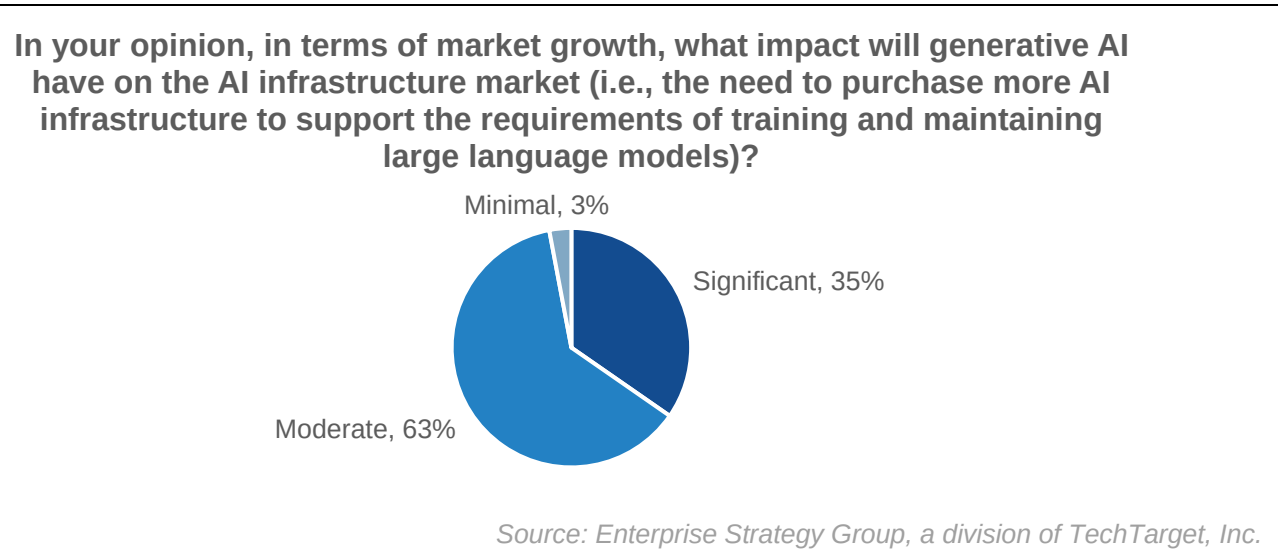
Globally, generative AI (GenAI) has triggered a tsunami of interest and activity. In fact, TechTarget websites have witnessed over 900% growth in search activities related to GenAI in 2023. It is important to note that this interest extends beyond just interest. Service providers have been early adopters of this technology, with many expanding their portfolio of services to include GPU-as-a-service offerings, and large enterprises are building out private GenAI infrastructure for internal use cases such as consumer analytics and supply chain and inventory management. Indeed, many corporate boards and C-level executives have already created initiatives for applying GenAI to their businesses processes. Furthermore, at the most recent Microsoft Ignite conference, GenAI leader Nvidia's CEO Jensen Huang predicted that GenAI will have a significant impact, stating, "It's bigger than PC. It's bigger than mobile. It's going to be bigger than the internet."[1]

According to TechTarget's Enterprise Strategy Group (ESG), it is easy to understand why organizations are so eager to deploy GenAI solutions. ESG research indicates the expected benefits from AI to include enhanced insights, improved revenue and profitability, faster decision-making speeds, enhanced customer experiences, and improved operational efficiency.[2]

It is also clear that these GenAI initiatives will require organizations to adopt new infrastructure, software, and services to support these initiatives. But those environments can vary greatly, as noted by Jeff Clarke, Vice Chairman and Chief Operating Officer at Dell Technologies. "GenAI is far from a one-size-fits-all model. It requires an end-to-end solution, the right infrastructure, a data plan, software and services that work seamlessly to support workloads across clouds, on-prem and at the edge."

ESG research illustrated that more than 9 out of 10 (97%) organizations believe there will be significant or moderate growth of AI infrastructure due to GenAI (see Figure 1).[3] This will be required to support both the front-end (user) and back-end (GPU) environments to ensure robust GenAI environments.

Figure 1. Expected Growth on AI Infrastructure Market From GenAI



Source: Enterprise Strategy Group, a division of TechTarget, Inc.

[1] Source: CRN, "Microsoft Ignite 2023: Nvidia CEO Huang Says Microsoft Is Now 'More Collaborative And Partner-Oriented'," November 2023.
[2] Source: Enterprise Strategy Group Complete Survey Results, *Navigating the Evolving AI Infrastructure Landscape*, December 2023.
[3] Ibid.

Further reinforcing the desire to embrace GenAI, organizations are going beyond researching the topic and are making plans to deploy GenAI environments, with research highlighting that the vast majority of respondents (92%) plan to do so in the next 12 months.[4]

To do this, organizations need specialized infrastructure designed to handle the specific requirements of GenAI, especially for the back-end GPU environment. However, deploying completely new technology can present challenges on many different levels.

Challenges Moving to New Technology

Deploying any new technology can be challenging for IT, even when it is a simple replacement of an existing technology. Brand-new technologies and/or architectures can be much more difficult to deploy. Unfortunately, GenAI requires new architectures, which require new compute, storage, and network infrastructures, especially for back-end GPU environments. This will not only require more infrastructure, but also, more importantly, carefully architected systems to accommodate the massive connectivity requirements across GPU clusters. Typical 50 Gigabit Ethernet (GbE) or 100 GbE top-of-rack (ToR) connections with 400 GbE uplinks would cause significant congestion and delays for large language models and put the entire initiative at risk.

When asked about the biggest challenges that organizations face when implementing generative AI solutions, survey respondents highlighted several issues, including employee expertise and skills, technical complexity, the inability to integrate with existing or legacy systems, and cost—among many other challenges related to data quality, ethical considerations, and transparency (see Figure 2).[5]

Figure 2. Top Challenges of GenAI

Source: Enterprise Strategy Group, a division of TechTarget, Inc.

[4] Ibid.
[5] Source: Enterprise Strategy Group Complete Survey Results, *Beyond the GenAI Hype: Real-world Investments, Use Cases, and Concerns*, August 2023.

It shouldn't be a surprise that the top challenge is a lack of skills and expertise, particularly for an emerging technology like generative AI. Most organizations will not have the resources with the requisite skills to assess, design, and implement large-scale GenAI infrastructure, especially the performance-intensive back-end environments.

Technical complexity can also impact GenAI deployments, as some solutions leverage proprietary technology, such as InfiniBand networks, which are typically reserved for high-performance computing (HPC) environments. As a result, there are a limited number of resources with the appropriate skill sets. This is especially true for enterprises and hyperscalers that standardized on Ethernet networks. Proprietary solutions can also be more difficult to integrate into any existing monitoring or orchestration platforms, requiring additional skills, hardware, and software. Another consideration when leveraging a proprietary solution is lead times. Given the complications over the past few years with the supply chain, organizations might be reluctant to choose solutions available from a single provider.

Due to these challenges, organizations also struggle with the associated high costs of implementing new GenAI solutions, especially proprietary ones that lock them into a specific vendor as they scale. The time it takes to evaluate and design a solution can be quite lengthy if there is a lack of reference designs and architectures.

Organizations Need an Open and Robust GenAI Infrastructure

Given these considerations, organizations need to look toward open solutions to help accelerate the deployment of GenAI infrastructure. Organizations will need to create new front-end environments that enable user interactions via a web-based interface and that are focused on ease of use and access. The back-end infrastructure is very different from traditional or even HPC environments and would need to support large language models (LMMs) that are powered by GPU clusters capable of consuming vast amounts of data. These back-end infrastructure environments are critical for a successful GenAI project.

Ideally, these solutions should be:

- **Comprehensive.** Organizations looking to deploy GenAI solutions need complete solutions for both front-end and back-end environments to accelerate adoption. These solutions would include the appropriate compute (including GPU clusters), storage, and networking for both environments. In addition to the infrastructure, these solutions require comprehensive automation and monitoring tools for not only the initial configuration and ongoing management but also to assist with fabric optimization and performance fine-tuning.
- **Highly performant.** For the network, this means deploying nonblocking fabrics with reliable delivery, high bandwidth, and low latency. This is the reason why the Ultra Ethernet Consortium (UEC) was created as part of the Linux Foundation's Joint Development Foundation, bringing together companies for industrywide cooperation in the development of Ethernet specifications and software APIs that empower AI environments with next-level performance, scalability, reliability (via the RoCE v2 protocol, for example), and interoperability.[6]
- **Pre-tested and proven.** To accelerate the adoption of these new GenAI environments, the ability to deploy a comprehensive solution that has been tested and proven to work effectively can help to avoid common deployment pitfalls. These solutions eliminate much of the research, analysis, and design time, enabling organizations to achieve their objectives and real value from their GenAI environments faster.

[6] Ultra Ethernet Consortium.

- **Open and extensible.** This would include leveraging merchant silicon and Ethernet fabrics as opposed to proprietary network technologies. GenAI environments require as much network performance as possible, but from open—not proprietary—standards. To accomplish this, the UEC will ensure that Ethernet can play a significant role in GenAI environments. In addition, organizations can take advantage of commercially available open source network operating systems such as SONiC (Software for Open Networking in the Cloud). It should be noted that both SONiC and UEC projects are hosted by the Linux Foundation, which eases industry collaboration and innovation.

Enterprise Strategy Group research highlights that organizations looking to modernize on-premises data centers cited leveraging hyperscale solutions on premises as their top action.[7]

- **Augmented with professional services.** The ability to accelerate the time to value for GenAI solutions will be aided by partners that can provide relevant expertise and experience. This would include the ability to conduct the appropriate assessments, architect the designs, and implement solutions in a timely manner. This could also include fully managed services and technical blueprints or validated designs.
- **Scalable.** With most organizations just getting started with GenAI, initial deployments might be limited in size but will need to scale up to accommodate increased requirements. Therefore, it will be imperative that GenAI infrastructure and, more specifically, the network environment can expand to support these needs.
- **Energy-efficient.** GPU-based solutions require massive amounts of power. Because of this, organizations need to take any steps possible to reduce the amount of power consumed. The latest-generation silicon technology that optimizes throughput-to-power ratios should be used. Higher-speed switches can consume less rack space, power, and cabling for a more cost-effective, ecologically friendly solution. In addition to reducing power, the ability to provide sustainability reports will also aid operations and management teams.
- **Software driven.** Focusing on software accelerates the pace of innovation, especially for software developed in open environments, as it is not predicated on a single vendor but rather potentially dozens of organizations contributing to its innovation.

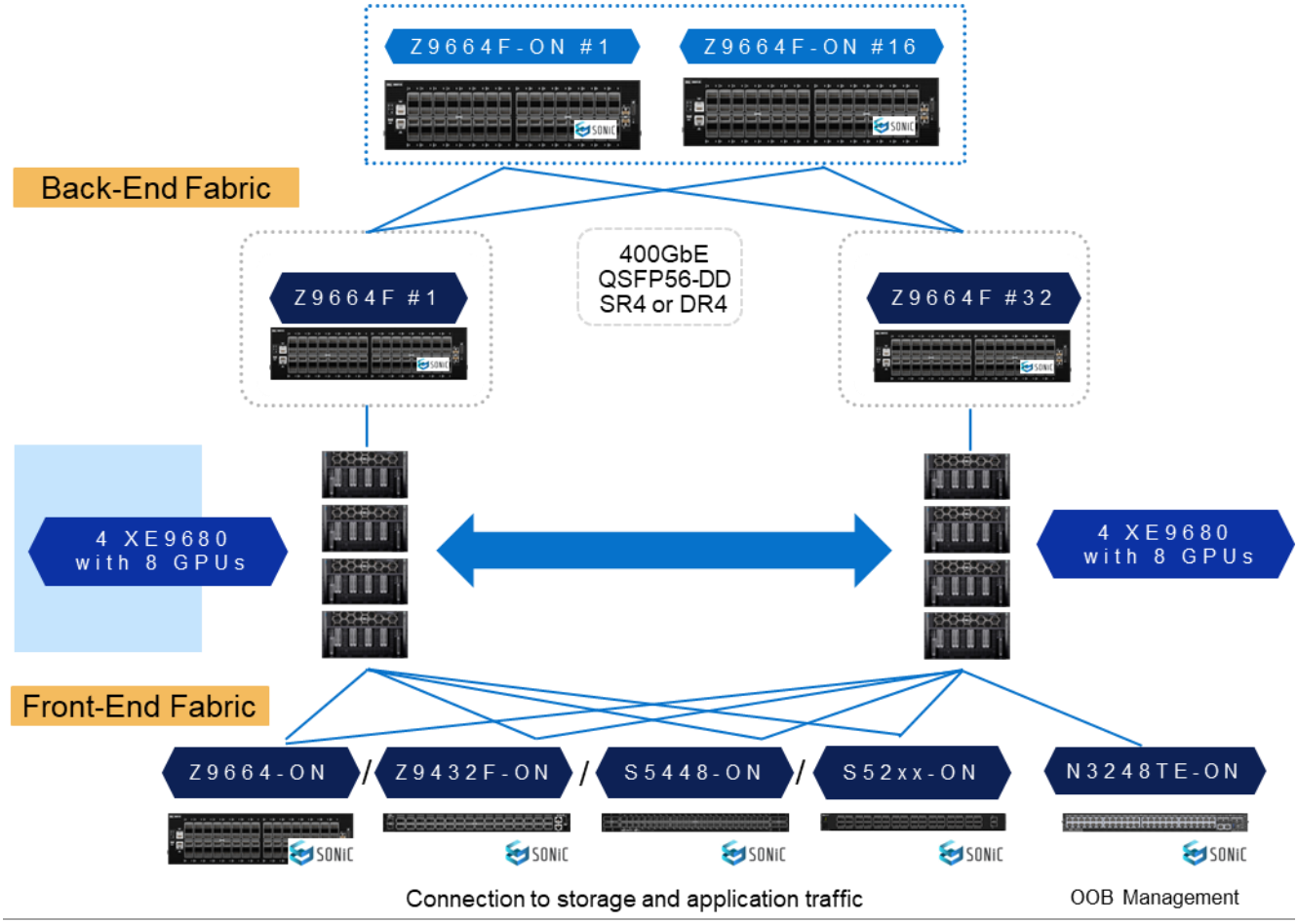
Dell Technologies Delivers Open Ethernet-powered Solutions for GenAI

Dell Technologies has been providing comprehensive and open infrastructure solutions for AI, modeling, and HPC environments for a number of years. It is leveraging its prior experience to enable GenAI infrastructure solutions for both front-end (application traffic, storage access, general network) and back-end (GPU fabric) environments that include compute, storage, and networking.

One of the keys to enabling a high-performance GenAI solution is a proven and open AI network fabric, as illustrated in Figure 3.

[7] Source: Enterprise Strategy Group Research Report, *2023 Technology Spending Intentions Survey*, November 2022.

Figure 3. Comprehensive AI Network Fabric Solutions



Source: Dell Technologies.

The Dell Technologies GenAI solutions include:

- **Modular compute systems.** Based on Dell PowerEdge XE servers and the company's experience serving the AI, modeling, and HPC market, these servers are acceleration-optimized for such environments. With options for air or liquid cooling as well as the number of GPUs, coupled with a focus on inference or training of LLMs, Dell has the right form factor and high-performance solution to match your GenAI compute needs. Compute environments are part of a validated design and architecture solution for GenAI.
- **AI-focused storage.** Dell has a range of storage options available based on the workload requirement, including PowerScale, Elastic Cloud Storage, and ObjectScale solutions. Ethernet-based PowerScale OneFS storage enables streaming reads and writes to quickly access data for AI workloads and improves AI modeling capability. Dell cites that PowerScale is field-tested with more than 1,000 customers running GPU workloadsM on them. As a result, there are numerous Dell Validated Design solutions based on these experiences. The wide range of options are all Energy Star-certified as well.

- **Next-gen Ethernet fabrics.** Centered on the Dell PowerSwitch and using next-generation silicon, such as the Broadcom Tomahawk 4, this open-network hardware can provide up to 51.2 Tbps with shared packet buffering. Commercially available as the PowerSwitch Z-series, the Z9664F-ON 64-port switch and Z9432F-ON 32-port switch can scale to support thousands of nodes. Additionally, Dell Technologies is a member of the UEC and will contribute to extending the applicability of Ethernet to power GenAI environments.
- **Software-driven architectures.** Dell Technologies remains committed to providing open networking solutions for network operating systems, orchestration, and monitoring in GenAI environments. For the network operating system, Dell Technologies has embraced and hardened SONiC, providing the global support, scale, and features required by large enterprises. The latest Enterprise SONiC Distribution by Dell Technologies (version 4.2) provides advanced support for AI environments that includes RDMA over Converged Ethernet version 2 (RoCE v2), enhanced hashing, and cut-through switching. The upcoming version 4.3 release provides enhancements for load balancing and mapping. All SONiC releases are tested and validated across the Z-series portfolio. The releases are also tested against Dell's third-party application partner ecosystem.
- **Provide services to accelerate adoption and optimization.** In addition to the 24/7 global support, Dell Technologies has professional services experts with proven experience to enable organizations to properly assess, design, and implement comprehensive GenAI solutions. Their ability to understand not only the network but also the compute and storage domains expedites the design process and reduces the risk of compatibility issues arising. These validated designs cover both inferencing and model customization, and there are services to cover data preparation and ingestion for GenAI pipelines. Dell also offers managed services to operate these AI environments.
- **Focus on sustainability.** Deploying GenAI environments at scale requires significant power resources. Dell's higher-speed switches in breakout mode require less rack space, power, and cabling. Leveraging the latest silicon technology enables servers, networking, and storage solutions to be as energy-efficient as possible. Being focused on energy efficiency enables organizations to reduce costs and energy consumption.

With these integrations, Dell Technologies is well positioned to deliver complete GenAI infrastructure solutions for both back-end and front-end environments.

Conclusion

The surge in GenAI interest and activity is driving organizations to evaluate solutions for their own environments. However, due to its recent popularity, most IT teams lack the expertise or experience to implement a solution in a timely manner. Also, to be fair, these GenAI infrastructures that require new architectures and technology are very complex. They must be carefully architected and provide a balanced system, so trying to source individual components and pull them together can be very risky. Because of this, organizations need to strategically partner to acquire the skills and tightly integrated solutions in order to ensure a successful GenAI environment.

However, organizations need to be mindful of comprehensive solutions that lock them into proprietary technology, especially as these environments scale. Open solutions can provide innovation, flexibility, and cost-effectiveness for large-scale GenAI environments. However, to ensure robust environments, it is also critical to ensure that these open solutions are fully tested, validated, and supported.

Dell Technologies provides complete GenAI solutions that incorporate all the infrastructure and software, including orchestration and management for both front-end and back-end environments. They also incorporate open compute, storage, and networking. Plus, organizations can leverage managed services, professional services, and fully validated designs and architectures that include Dell's partner ecosystem. These comprehensive yet modular solutions enable organizations to accelerate the deployment and value of GenAI solutions while reducing risk and ensuring greater operational efficiency.

About Enterprise Strategy Group
TechTarget's Enterprise Strategy Group provides focused and actionable market intelligence, demand-side research, analyst advisory services, GTM strategy guidance, solution validations, and custom content supporting enterprise technology buying and selling.

contact@esg-global.com
www.esg-global.com