ANALYST BRIEF

Sponsored by: Dell

Investment in AI processing infrastructure is growing at an accelerated pace. The good news is that for AI processing infrastructure, over 50% of systems will not be accelerated in 2024 and can run on standard servers and Ethernet networking.

GenAI Is Evolving at Record Speeds as Businesses Start Their GenAI Journey

February 2024

Written by: Brandon Hoff, Research Director, Enabling Technologies: Networking and Comm, and Vijay Bhagavath, Research Vice President, Cloud and Datacenter Networks

Introduction

IDC research provides predictions and underlying drivers that we expect to impact IT investments in 2024 and beyond. Technology leaders and their counterparts in the lines of business (LOBs) can use this document to guide their strategic planning efforts.

Operations teams have been capturing data, creating data lakes, and leveraging the cloud to store their data. Now, with the popularity of ChatGPT, the iPhone moment for generative AI (GenAI), operation teams know what they can do with their data sets. Since everybody knows the benefits of generative AI, operations teams are also facing additional pressure from investors, executives, and the market to implement an effective GenAI strategy. There are several technologies, a wide range of options, that can be leveraged to improve business operations and employee productivity, from GenAI to ML to digital twins and more. Successful implementation of the right technology will become an essential KPI for operations teams and the business overall.

AT A GLANCE

KEY TAKEAWAYS

Start planning GenAI infrastructure:

- » Integrate AI faster than others, first, by moving at least a copy of data on premises through an initiative to pull data back from the cloud.
- » Invest in understanding the value of the algorithms underpinning generative AI, AI, ML, and digital twins in the business, and prioritize based on business value.
- » Three steps: Deploy standard servers and Ethernet networking for GenAI evaluation. Scale out GenAI for enterprise-sized workloads as needed leveraging standard Ethernet. Rebalance workloads between on-premises and off-premises infrastructure to optimize capex and opex over the next three to five years.

Understanding GenAI Today

With the explosion in demand for AI, cloud service providers (SPs) and enterprises are building out their infrastructure at an accelerated pace. Cloud SPs are consuming a vast majority of AI accelerators and building out their own AI infrastructure, but these accelerators are expensive, driving up the cost of AI services at the cloud SPs. AI accelerators consist of GPUs, TPUs, FPGAs, ASSPs, and ASICs. These cloud SPs are building AI factories for massive workloads that span the needs of a wide range of companies and are mainly deployed at the largest IT environments in the world, which works out to be about nine companies.

On the other hand, GenAI infrastructure for enterprise-sized workloads can be built with standard systems with no acceleration required. IDC predicts that over 50% of GenAI systems will not be accelerated in 2024; therefore, anybody can start deploying their GenAI Infrastructure with standard servers and networking. GPUs are also available, for those who need them. There are multiple options for deploying AI infrastructure as well as multiple types of GenAI, AI, ML, and digital twins that will benefit different companies in different ways. There are benefits with running GenAI on standard servers since GenAI software stacks are generally supported. Companies that invest in their on-premises standard infrastructure will be able to advance their GenAI initiatives faster than others. It is essential that IT teams start to evaluate the various GenAI, AI, ML, and digital twin algorithms to identify which ones make the biggest impact to their business.

Benefits

GenAI and other foundational models are changing the game, taking assistive technologies to a new level, and bringing powerful capabilities to nontechnical users. GenAI has the potential to increase efficiency and productivity, open new opportunities for growth, lower costs, and provide a competitive advantage for companies that leverage it.

Building your own GenAI infrastructure kick- starts integrating this game-changing technology into business operations and builds onsite expertise into the GenAI technology stack. By prioritizing technology investments that build initial GenAI infrastructure on premises based on standard enterprise servers and Ethernet networking will create a time-to-market advantage for businesses leveraging this transformative technology.

Considerations

Act Now to Leverage GenAI

There is excitement surrounding GenAI, given the impressive results that ChatGPT and other models deliver, GenAI provides value, but the value will vary based on the source of proprietary data and the algorithms deployed. Board rooms, investors, and executives will be asking questions and looking to see how GenAI can help their business.

For businesses that have captured large amounts of unstructured proprietary data, GenAI promises to create original content from the proprietary existing data that is expected to help rewire the organization for continuous innovation. A crawl, walk, and run approach makes sense to understand what GenAI can bring to the business and how to move forward.

Building Your Initial GenAI Infrastructure on Ethernet

For enterprise-sized workloads, standard systems provide the performance that is needed to start the GenAI journey. Plus, basing GenAI infrastructure on standard servers and Ethernet networking allows the use of enterprise operating systems, enterprise management tools, and enterprise network management tools. Once the compute requirements for the LLMs that benefit the business are understood, compute performance can be improved with the right selection of GenAI and AI accelerators. The key is that the AI infrastructure needs to be well architected. A well-architected Fabric can support tens of AI compute nodes to thousands of AI compute nodes.

While there may be different options for networking for GenAI workloads, the ubiquitous, open, multivendor preferred option is Ethernet networking. Initial GenAI deployments can be supported with standard Ethernet networking available today for GenAI clusters.

Building GenAI Infrastructure with Ultra Ethernet

As each business starts in their walk and run phases of GenAI development, it may make sense to build their own scale-out GenAI infrastructure. Building a scale-out GenAI infrastructure requires two key additions: datacenter AI accelerators and AI networking. In typical scale-out GenAI infrastructure, eight datacenter GPUs are deployed on each server, and for each GPU, there is a high-speed NIC or DPU deployed to deliver high-performance networking.

A key requirement for a scale-out AI infrastructure is high-performance networking. For GenAI LLMs, the bottleneck in processing is the time the data spends in the network. For some workloads, time in network can be up to 60% of the processing time for a LLM, leaving the compute infrastructure idle as data moves between compute clusters. For the AI network, there is improved networking that is available today and delivered through the Ultra Ethernet Consortium that promises interconnect that is as performant as supercomputing networks, scalable to the cloud datacenter, and as cost effective and ubiquitous as Ethernet. AI networking is essential to address the growth in network demands of GenAI and HPC at scale. The good news is that the Ultra Ethernet Consortium is supported by most Ethernet switch vendors.

> "The GenAI datacenter Ethernet switching market in the enterprise segment is forecast to grow at a CAGR of 158.2%, from $41.9 million in 2023 to $1.0 billion in 2027," Vijay Bhagavath, IDC.

For performance, there are three key pieces of technology that are required: high-speed SerDes, PHYs, and Optics. These three technologies are used in Ethernet and other networking technologies, so essentially, there is no performance advantage for any specific networking technology. To achieve the highest performance from Ethernet, the InfiniBand Trade Association launched the RDMA over Converged Ethernet (RoCE) initiative and defined the RoCE protocol. RoCE is supported on standard datacenter switches, and there are additional improvements to boost performance such as high-radix Ethernet switching, cut through switching, load balancing, and higher bandwidth up to 800GbE (4 x 200GbE) links coming to market.

Initial testing of GenAI LLMs can provide early insight into the benefits that GenAI can bring to the business and assist in building a strategy for GenAI LLM, as well as what types of infrastructure would be required. Essentially, the software stack drives semiconductor requirements for the next step in the enterprise GenAI evolution. Understanding the software stack will assist in deploying an optimized hardware infrastructure.

Rebalancing On-Premises Versus Off-Premises Infrastructure as Semiconductor Costs Stabilize

As the supply for datacenter GPUs increases, more vendors will offer datacenter GPUs, more AI accelerators will become available, and more GenAI processing power will be available for on-premises deployments. In parallel, bottlenecks at the cloud service providers will disappear and costs would be expected to stabilize. Once this happens, three to five years out, rebalancing GenAI workloads between on-premises infrastructure and cloud infrastructure will optimize capex and opex.

Conclusion

GenAI is the breakthrough technology for AI. Businesses will need to have a GenAI strategy/plan that should kickoff now for enterprise-sized workloads to continue the journey to integrate this game-changing technology into enterprise operations.

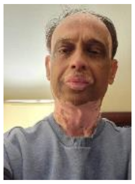
Demand is high, driving up component prices and cloud SP pricing. At the same time, IDC predicts that over 50% of GenAI systems will not be accelerated in 2024; therefore, anybody can start deploying their GenAI Infrastructure with standard servers and networking. GPUs are also available, for those who need them. There are multiple options for deploying AI infrastructure as well as multiple types of GenAI, AI, ML, and digital twins that will benefit different companies in different ways.

IDC's prediction is that businesses will bring back data from the cloud for GenAI processing to lower opex costs. Businesses will start their development and testing with GenAI on standard compute and Ethernet networking hardware and invest as they learn what LLMs work for their business and the value they can extract from their proprietary data.

> "The market has continued to underestimate the growth in GenAI, and IDC expects to see robust growth in GenAI infrastructure and semiconductors," Brandon Hoff, IDC.

Building out the infrastructure for GenAI LLM testing on off-the-shelf servers and enterprise Ethernet networks will unlock the value of GenAI for the enterprise.

About the Analysts

	Brandon Hoff, *Research Director, Enabling Technologies: Networking and Comm* Brandon Hoff leads IDC's networking and communications infrastructure within IDC's Enabling Technologies team. Mr. Hoff covers technology trends, workloads, products, vendors, supply chain, and end-user adoption strategies in enterprise IT and datacenters of web, cloud, and telecommunications service providers.
	***Vijay Bhagavath,*** *Research Vice President, Cloud and Datacenter Networks* Vijay Bhagavath provides actionable thought leadership and pragmatic insights on Cloud and Datacenter Networking markets and technologies. Vijay has a deep understanding of the overall networking market, technologies, product road maps, competitive differentiation, and deployment strategies, enabling him to provide insightful commentary and guidance for vendors, cloud providers, enterprise IT buyers and practitioners.

MESSAGE FROM THE SPONSOR

Bring AI to your data

Dell Technologies accelerates your journey from possible to proven by leveraging innovative technologies, a comprehensive suite of professional services, and an extensive network of partners.

- Simplified. Speed up time-to-results by combining strategic guidance and roadmaps with proven and validated solutions.
- Tailored. Get the most value of your data with infrastructure designed for your business needs.
- Trusted. Build your AI future on a secure foundation, protecting your data and intellectual property.

Deliver the best AI performance and simplify sourcing, deployment, and management of AI infrastructure designed for the Generative AI era — with technology, innovation, and Dell Technologies advantages to deliver smarter, faster outcomes.

For more information, visit www.dell.com/AI.

IDC Custom Solutions

The content in this paper was adapted from existing IDC research published on www.idc.com.

IDC Research, Inc.
140 Kendrick Street
Building B
Needham, MA 02494, USA
T 508.872.8200
F 508.935.4015
Twitter @IDC
idc-insights-community.com
www.idc.com

This publication was produced by IDC Custom Solutions. The opinion, analysis, and research results presented herein are drawn from more detailed research and analysis independently conducted and published by IDC, unless specific vendor sponsorship is noted. IDC Custom Solutions makes IDC content available in a wide range of formats for distribution by various companies. A license to distribute IDC content does not imply endorsement of or opinion about the licensee.

External Publication of IDC Information and Data — Any IDC information that is to be used in advertising, press releases, or promotional materials requires prior written approval from the appropriate IDC Vice President or Country Manager. A draft of the proposed document should accompany any such request. IDC reserves the right to deny approval of external usage for any reason.