HPE aruba networking

Brochure

An introductory guide to data center switching for generative AI

Simplify, speed, and prepare for the GenAI revolution

HPE GreenLake

What is generative AI?

Artificial intelligence is powering businesses across all industries and verticals and becoming an inevitable general-purpose need akin to a utility. AI-powered solutions are an essential technology for analyzing continuously arriving data using task-specific AI models, such as machine learning and natural language processing (NLP) to gain insights and power real-time decision making. These data and processing pipelines are operationalized along with traditional software and solutions.

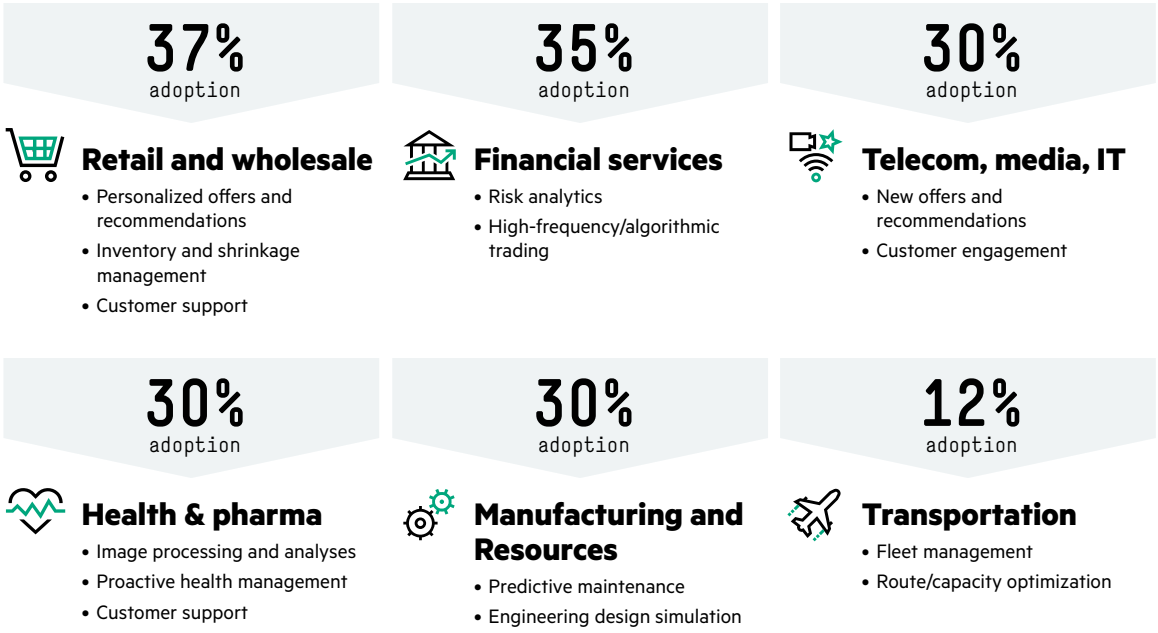

Figure 1. GenAI use cases by industry[1]

Generative pre-trained transformer (GPT) models, such as GPT-3, have gained significant attention for their ability to generate human-like text. They are being used in applications like content generation, question-answering, language translation, chatbots, and creative writing.

Large language models (LLMs), are a broader category of AI, used in a wide variety of applications. For example, they are employed in sentiment analysis, text summarization, language translation, text classification, and more. LLMs can be fine-tuned for specific industries, such as healthcare, finance, and customer support, to address domain-specific tasks.

The primary difference in use cases and applications is that GPT models, while versatile, are often celebrated for their text generation capabilities, while LLMs are utilized in a more diverse range of NLP tasks.

HPE GreenLake for LLMs is an example of an LLM that runs on an AI-native architecture that is uniquely designed to run a single, large-scale AI training and simulation workload at full computing capacity. This offering supports AI and high-performance computing (HPC) jobs on up to tens of thousands of CPUs or GPUs at once. This capability is effective, reliable, and efficient for training AI and creates more accurate models, allowing enterprises to speed deployment of novel use cases.

[1]Source: O'Reilly Survey (AI Adoption in the Enterprise — Production only). Mar 2022.

20%

of Ethernet data center switch ports will be connected to accelerated servers to support AI workloads by 2027[1]

100.8%

Compound annual growth rate (CAGR) for the worldwide GenAI data center switching market through 2027[2]

With bandwidth in AI growing, the portion of Ethernet switching attached to AI/ML and accelerated computing will migrate from a niche today to a significant portion of the market by 2027. We are about to see record shipments in 800Gbps based switches and optics as soon as products can reach scale in production to address AI/ML.[3]

— **Alan Weckel**, founder and technology analyst at 650 Group.

AI training vs. inferencing

Training an AI algorithm is the process of teaching a base algorithm how to make a correct decision by feeding it massive amounts of data. Once an algorithm is trained, you can send it out into the world to do its job. When a trained AI algorithm gives you an answer, that's called inferencing. It uses what it learned during training to draw conclusions or make predictions based on real world input. Machine learning inference is the process of using a pre-trained ML algorithm to make predictions. In machine learning, after a model has been trained on a dataset, it is deployed to make predictions or classifications on new, unseen data.

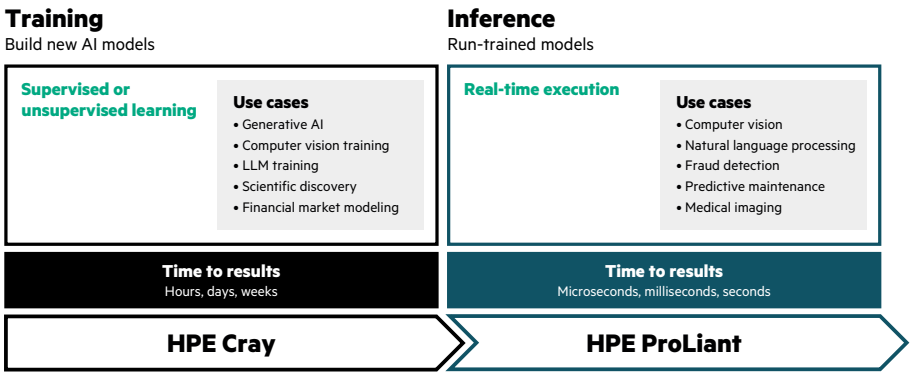

Figure 2. AI training vs. inferencing

Inference workloads tend to need GPU acceleration for large models but may use CPUs alone, depending on the application. Some inference applications may run on edge devices outside the data center.

Infrastructure requirements for GenAI

Large scale GenAI models require vast amounts of GPU-enabled compute to support terabit sized training data sets, supporting billions of parameters, accessed from ultra-low latency memory and storage systems. The network that supports these massive GenAI models is customized for optimized, power efficient, predictable performance across LLM multi-tenant workload environments.

Network technology required to support GenAI

Every aspect of the network, including compute, data processing units (DPUs), I/O, cabling and optics, acceleration software and the network itself — for example InfiniBand 400/800G Ethernet switches—fabric and topologies are highly tuned to support the overall system.

[1]Dell'Oro Group, 20 Percent of Ethernet Data Center Switch Ports Will Connect to AI Servers By 2027, According to Dell'Oro Group, July 2023.

[2]IDC, Worldwide Generative AI Datacenter Switching Forecast, 2023–2027: Insights on Cloud and Enterprise Buyer Journeys, November 2023.

[3]650 Group, Data Center AI Networking Surges over 100% Y/Y as InfiniBand and Ethernet Achieve Record Revenues in 1Q23; According to 650 Group, July 2023.

AI fabrics: InfiniBand vs. Ethernet

While InfiniBand has traditionally been used in supercomputers, HPC, and large-scale, public cloud GPU-accelerated GenAI deployments, 100/200/400G+ leaf/spine Ethernet fabrics have emerged as a viable alternative to these networks based on Ethernet's compelling price/performance advantages, latency improvements, scalability, and standards-based interoperability.

HPE Slingshot is one of the highest performing network fabrics in the market today. Slingshot is currently deployed at large scale in multiple supercomputers across the globe, including the first Exascale supercomputer, Frontier at Oak Ridge National Laboratory, the Perlmutter system for the National Energy Research Scientific Computing Center, and the upcoming Aurora 2 Exaflop system for Argonne National Lab.

Ethernet is also becoming a popular choice for AI clusters at large enterprises and tier 2/3 cloud providers based on familiarity with the tooling and management of Ethernet fabric connectivity with existing compute, storage, and security infrastructure.

AI back end vs. front end networks

Back-end networks are used to connect the AI servers/nodes together in a cluster. The back-end connectivity can also include dedicated storage if it is not converged with compute. Front-end networks are often built with traditional networking and switching used for general purpose servers/workloads and shared storage.

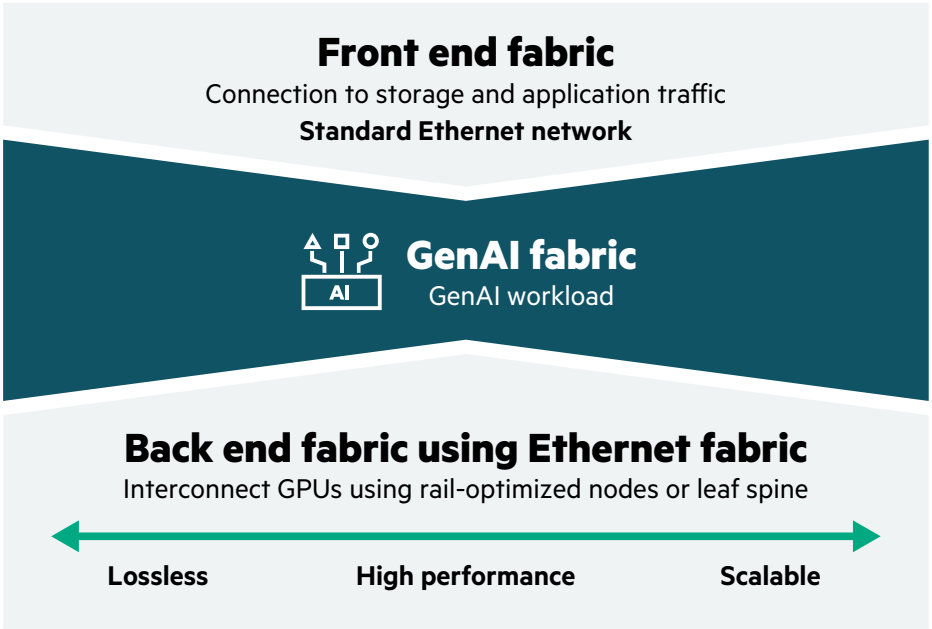

Figure 3. GenAI high level architecture

Lossless Ethernet fabrics provide a good alternative for an increasing number of smaller scale on-premises private enterprise AI deployments that require high-performance and low latency connectivity for AI training, inferencing and tuning.

GenAI back-end fabric network design types

Multi-layer Clos – The ToR switch interconnects the servers and provides the connectivity to the other racks via the aggregation switch. The spine provides connectivity to other pods. Best used for CPU heavy workloads.

Rail optimized – Based on GPU-centric clusters where each GPU has two different communications paths. One path is through the NVIDIA® NVSwitch (supports hi-bandwidth but short-range interconnect) and the other through the rail switches. The NVIDIA NVSwitch on the individual server creates a fast interconnect that forms a high-bandwidth domain. These rail switches are connected to spine switches to form a full-bisection, any-to-any Clos network topology.

Rail only – The rail-only network architecture is like the rail optimized but removes network connectivity across GPUs with different ranks in different rails. However, communication is still possible by forwarding the data through high-bandwidth domain.

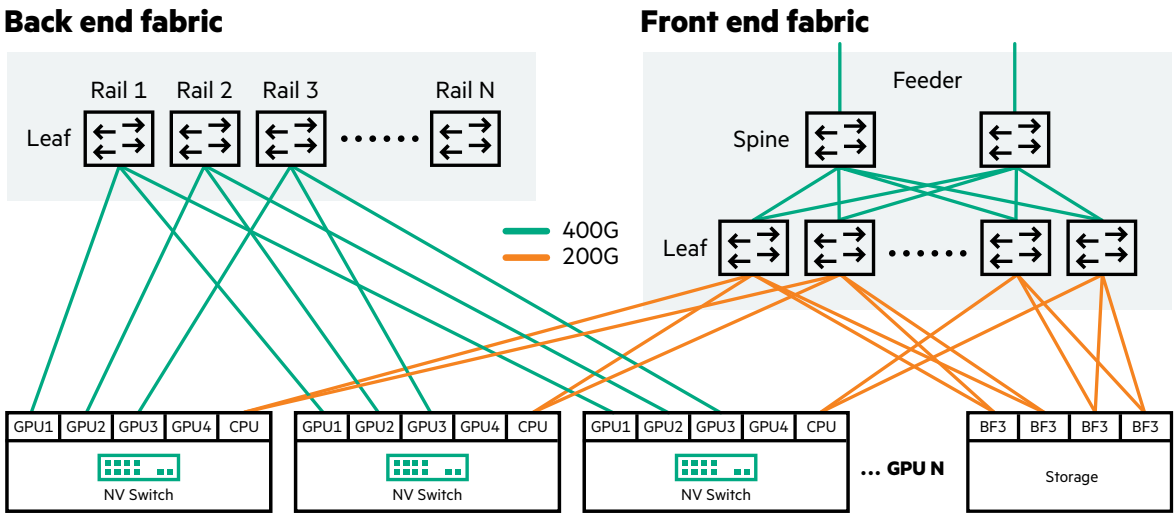

Figure 4. Rail-only architecture

Lossless fabrics help complete jobs faster

To satisfy the characteristics of AI workloads, AI fabrics need provide a “lossless fabric” — meaning data is sent with the acknowledgment that loss or deterioration of information could take place. It’s important to note that we are discussing tail latency, also known as network latency, not packet latency. While in HPC networks, packet latency can have a huge impact on application performance, for AI/ML workloads, what matters most is network latency or tail latency.

HPE Aruba Networking allows enterprises to capitalize on the ability of GenAI to deliver great digital experiences to customers and employees with unprecedented scalability, performance, and operational efficiency.

This is critical as any delayed flow can impede the progress of all nodes, resulting in having those very expensive accelerator nodes idle as they wait for the network to respond. It's not about the latency of the packet, but rather the time needed to complete the job.

Key technologies that help

- **Remote direct memory access over converged Ethernet version 2 (ROCEv2)** is a high-performance network computing technology that lets data transfer directly between the memory of two devices without having to involve a server CPU. It allows multiple packets to be transferred or routed simultaneously over a single connection, reducing latency and complexity as well as boosting throughput.
- **Explicit congestion notification (ECN)** enables a lossless Ethernet network by monitoring for network congestion or other situations where packets could get dropped and throttling back the network to keep that from happening.
- **Priority flow control** helps control congestion in Layer 3-based networks and plays an important role in overall congestion management.
- **Spine-leaf clos fabric design** refers to Ethernet switching fabrics that are built with 100/200/400 (moving to 800) Gbps connectivity from the server NIC to leaf and through the spine. These are the preferred choices for handling high data throughput and low latency requirements necessary for mission-critical AI applications.

Taken together, these technologies can give an Ethernet network the ability to prioritize certain sets of workloads, such as AI workloads that cannot tolerate any dropped packets and will always get network priority even if there's congestion.

Prepping your data center network for AI

Next-generation AI architectures require a dedicated network fabric that delivers a combination of high performance and low latency connectivity to ensure the fastest training, inferencing, and tuning model job completion times.

With early HPC and AI training networks, high speed, low latency, proprietary InfiniBand networks initially gained popularity for their fast and efficient communication between

servers and storage systems. Today, the open alternative is 100/200/400G+ leaf/spine Ethernet switching, which is gaining significant momentum for supporting the AI data center networking market and is expected to become the dominant technology.

Modern AI applications need high-bandwidth, lossless, low-latency, scalable, multi-tenant networks that interconnect hundreds or thousands of GPUs at high speed from 100G to 400G beyond. Ethernet-based networking fabrics provide the reliability and performance that AI workload clusters with hundreds to thousands of GPUs require. HPE Aruba Networking can help you design and build a dedicated AI network fabric to get you started.

HPE Aruba Networking AI-Ready data center switching

The HPE Aruba Networking CX 9300 Switch Series is a next-generation 25.6Tbps, 1U fixed configuration switch that supports 32 ports of 100/200/400GbE. The CX 9300 provides AI/HPC optimization features including low latency, lossless network quality of service (QoS) and connectivity characteristics that AI/HPC requires including ROCEv2, ECN, and PFC.

HPE Aruba Networking CX 6300M and 8325 switches also provide management and data center network connectivity to HPE/AI servers and storage. The CX 6300M can be used as a low cost GbE out-of-band management (OOBM) switch. The CX 8325 can be used as a server top of rack (ToR) rack switch with 10/25GbE and 40/100GbE connectivity from the server's virtual machines.

In summary, next-generation data center switching from HPE Aruba Networking allows enterprises to capitalize on the ability of GenAI to deliver great digital experiences to customers and employees with unprecedented scalability, performance, and operational efficiency.

Learn more at

arubanetworks.com/solutions/data-center-modernization/

Visit **ArubaNetworks.com**

Make the right purchase decision.
Contact our presales specialists.

Contact us

Hewlett Packard Enterprise

BRO_Data-center-switching-GenAI_A4_RB_050724 a00138537ENW