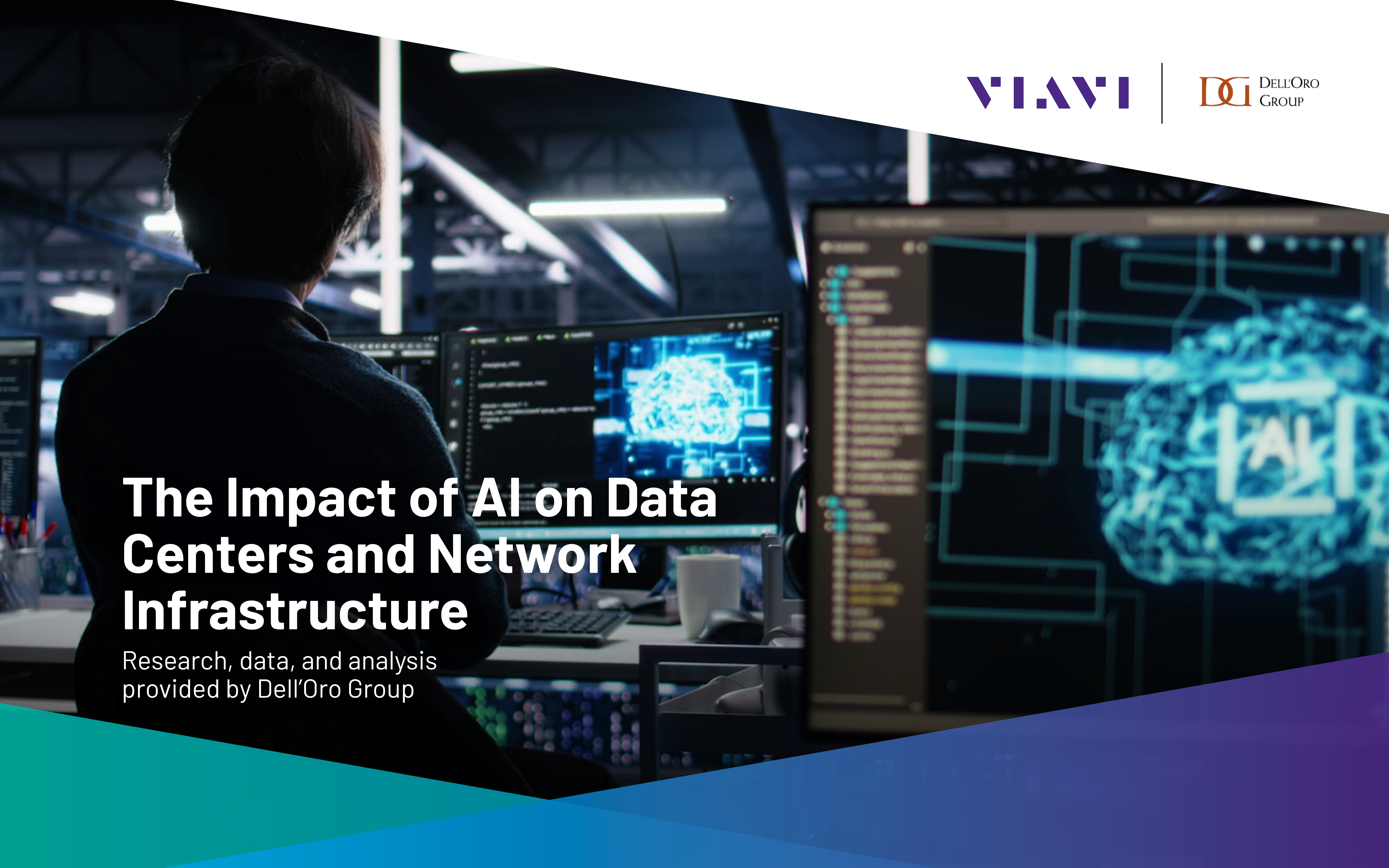
VIAVI | DELL'ORO GROUP



The Impact of AI on Data Centers and Network Infrastructure

Research, data, and analysis
provided by Dell'Oro Group

Artificial Intelligence (AI) is fundamentally transforming the design, operation, and scale of data centers and network infrastructure. As AI workloads—particularly large language models and machine learning systems—demand unprecedented compute power and low-latency data flow, organizations are evolving their digital infrastructure accordingly.

Key Impacts of AI on network and data center infrastructure include:

- **Rising Demand for Compute and Power:** AI workloads require high-performance GPUs, custom accelerators, and dense compute environments, significantly increasing power consumption and necessitating advanced cooling solutions such as liquid cooling.
- **AI-Optimized Operations:** AI is being used within data centers to automate operations, enhance energy efficiency, predict hardware failures, and optimize workload scheduling, improving uptime and reducing costs.
- **Next-Generation Network Infrastructure:** to support AI's scale, high-bandwidth, low-latency networking fabrics are becoming standard. AI-specific network architectures and faster interconnects are essential for distributed model training and real-time inference.
- **Edge Computing and 5G Integration:** the need for real-time AI inference is driving investment in edge infrastructure and 5G networks, enabling localized processing and minimizing latency for applications like autonomous systems and augmented reality.
- **Security and Resilience:** AI enhances security posture through real-time anomaly detection, adaptive policy enforcement, and intelligent threat response, improving network resilience against sophisticated attacks.
- **Architectural Shifts:** AI is accelerating the move toward hyperscale and composable data center architectures, enabling flexible allocation of compute, storage, and network resources tailored to dynamic AI workloads.

Preparing for the AI Infrastructure Era

Enterprises and cloud providers face mounting pressure to invest in scalable, energy-efficient infrastructure and adopt AI-driven operational models to remain competitive.

The convergence of AI, edge computing, and next-generation networking will shape the future of digital infrastructure. Through the following data and insights, Dell'Oro Group and VIAVI Solutions illustrate the scale and complexity of the challenges ahead — and how the industry can prepare to deliver on AI's growing promise.

All research, data, and analysis provided by Dell'Oro Group.
Primary analysts: Sameh Boujelbene and Baron Fung.

A New Era of AI Has Begun

> *The age of AI has begun.*
>
> *Artificial Intelligence is as revolutionary as mobile phones and the Internet.*
>
> *ChatGPT is the only demonstration of technology that struck me as revolutionary since the introduction of graphical user interface in 1980.*

BILL GATES

> *ChatGPT is just the start.*
>
> *We are at the iPhone moment of AI.*

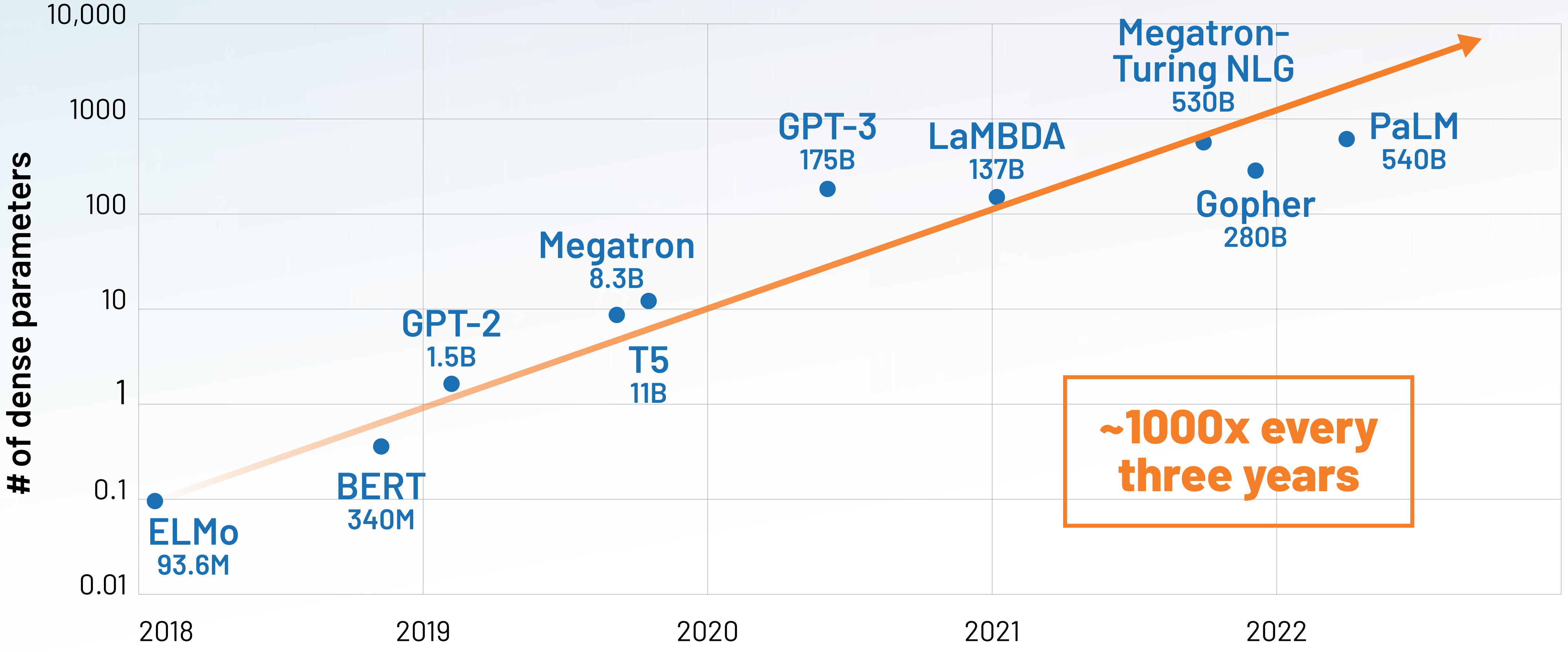
JENSEN HUANG

CEO, NVIDIA

The Scale of AI Applications

The size and complexity of AI models are growing at an unprecedented pace.

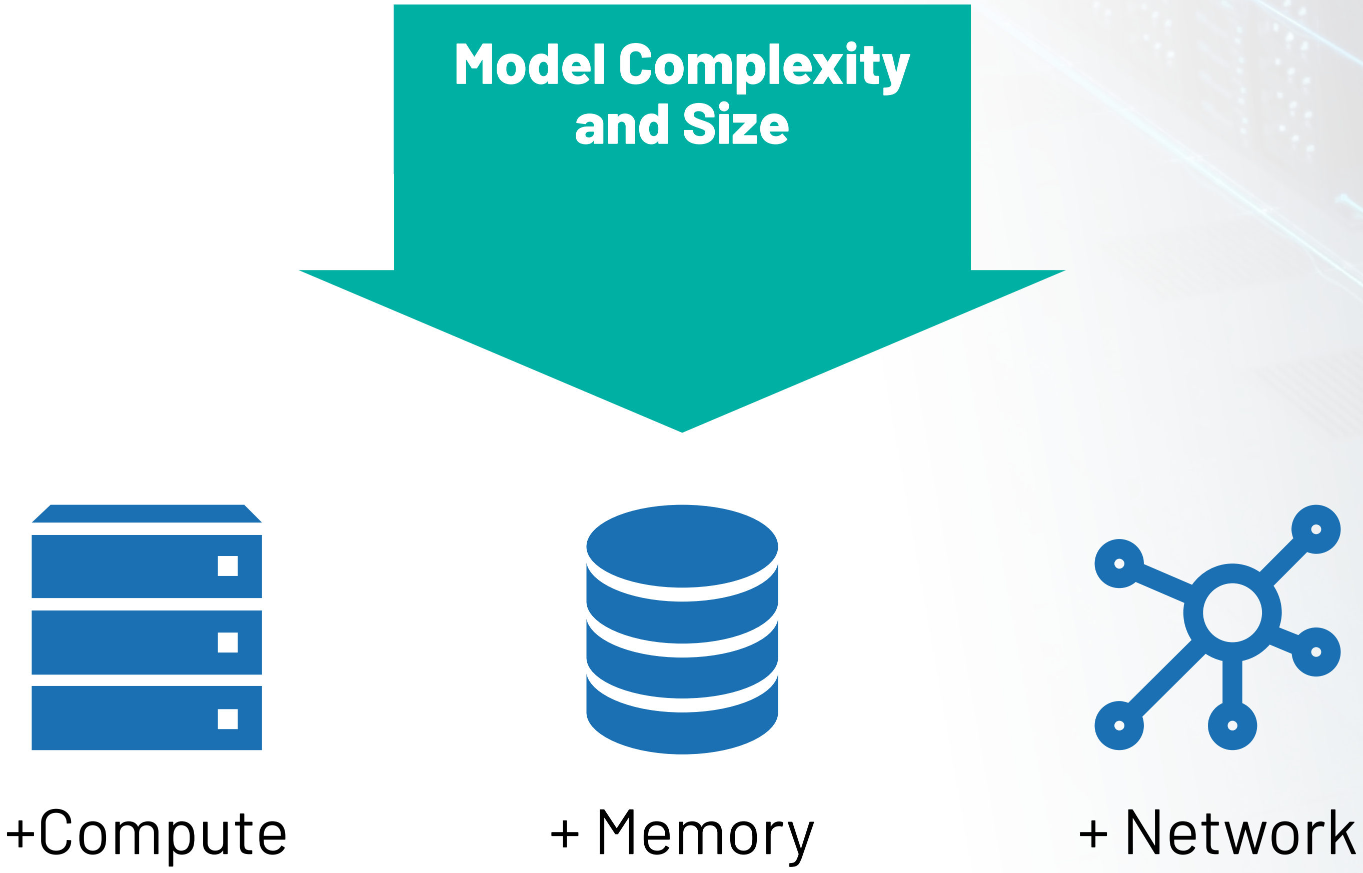
Generative AI and large language models are defined by the enormous number of parameters they manage — often reaching into the billions or even trillions. As shown in the accompanying chart, the number of parameters has been increasing nearly 1,000X every three years, driving a fundamental shift in how infrastructure is designed and deployed.



AI Applications and Infrastructure

The complexity of an AI application — measured by the number of parameters — directly affects the compute resources required to run it.

Larger models require significantly more GPUs and demand high-performance, low-latency network fabrics to connect them. As AI models grow, infrastructure must scale accordingly, influencing everything from server density to interconnect design.


Model Complexity and Size

+Compute

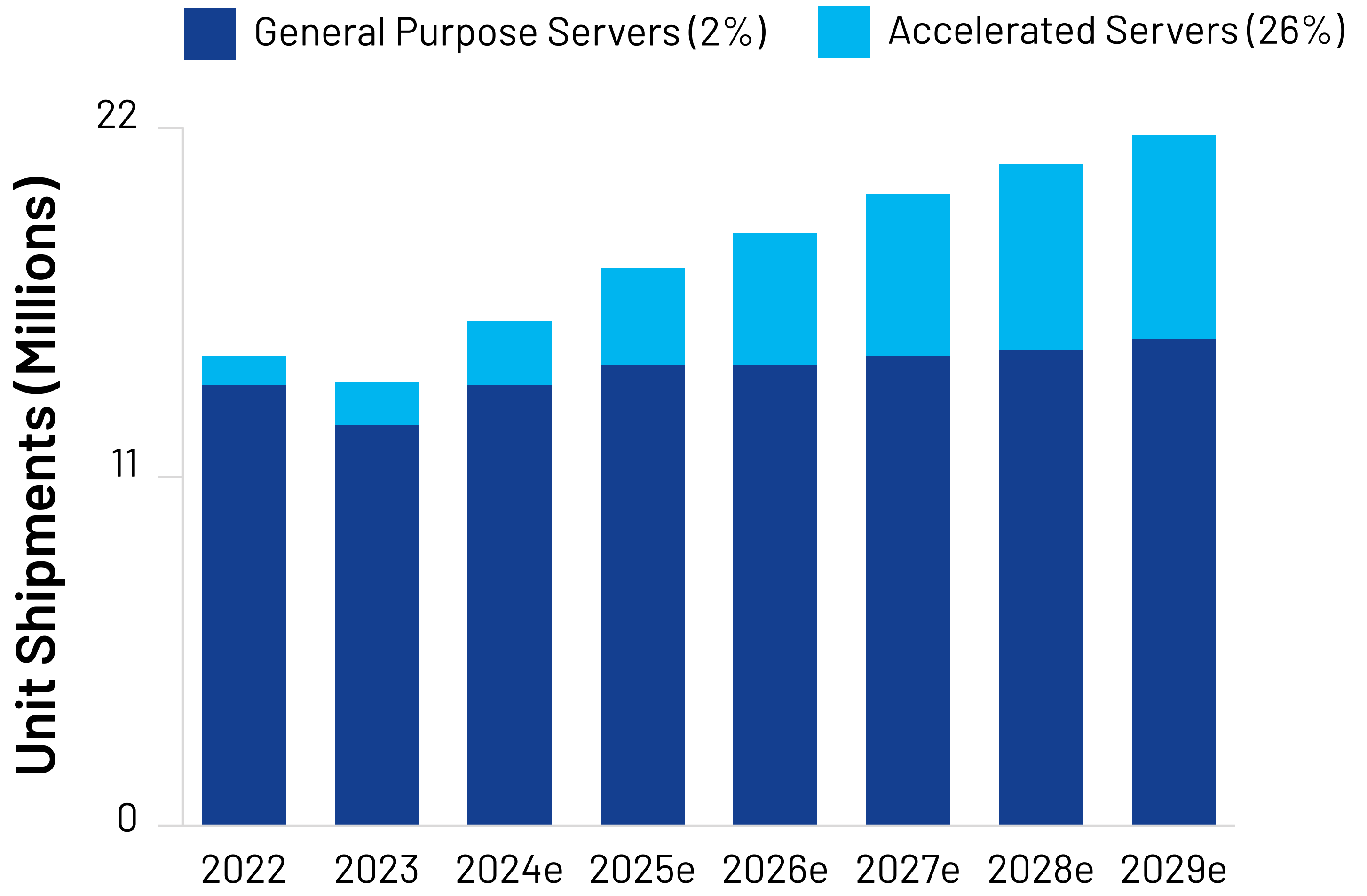
+ Memory

+ Network

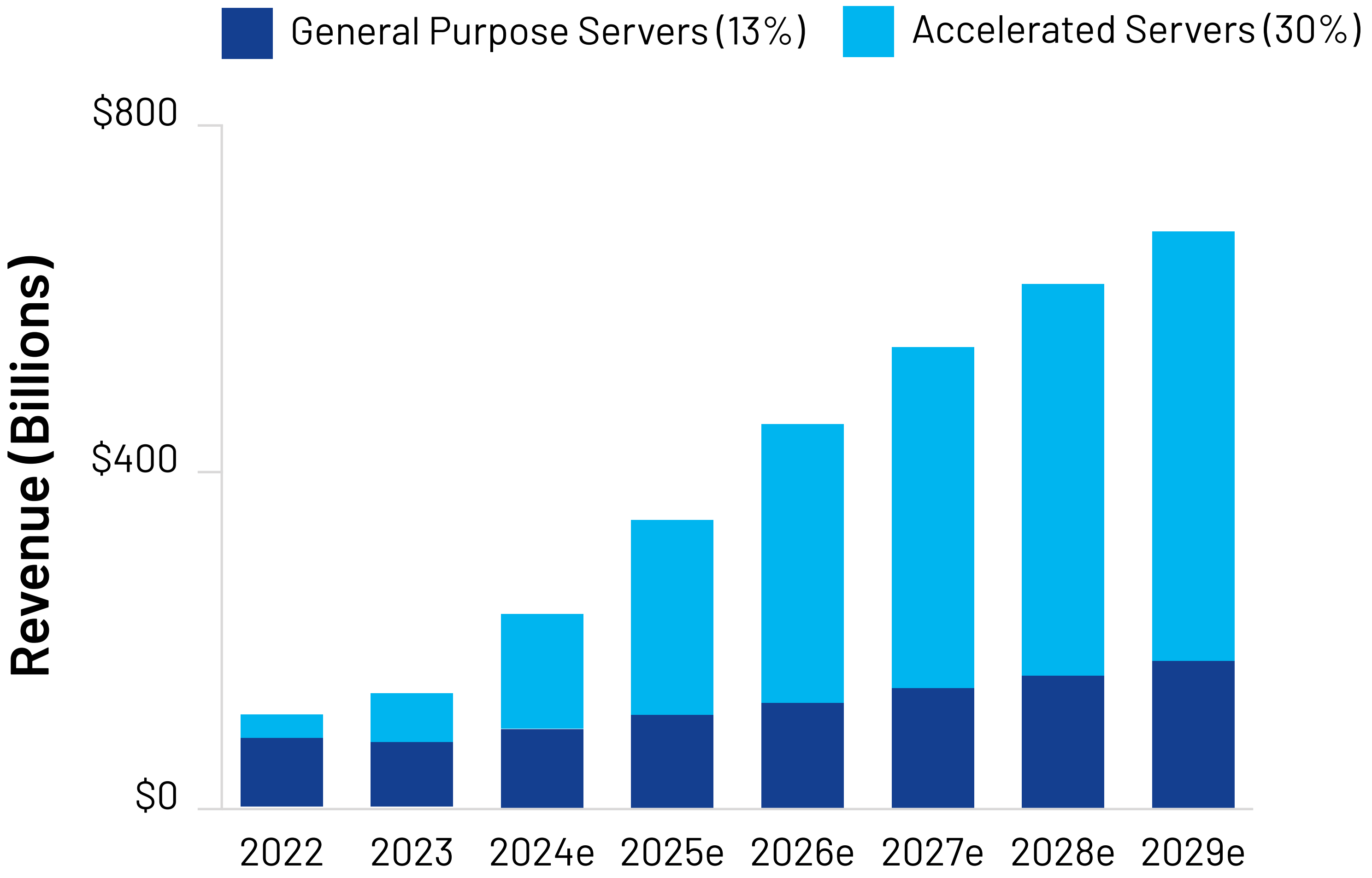
Accelerated Server Five-Year Forecast

AI infrastructure, also referred to as accelerated computing, will be a major driver of capital expenditure (CapEx) over the next five years.

Currently, AI servers make up only about 10% of total server units deployed in data centers (see left-hand chart), but this share is expected to double within five years. From a CapEx perspective (right-hand chart), AI servers represent a much larger share due to their higher cost — typically five to ten times more expensive than general-purpose servers.

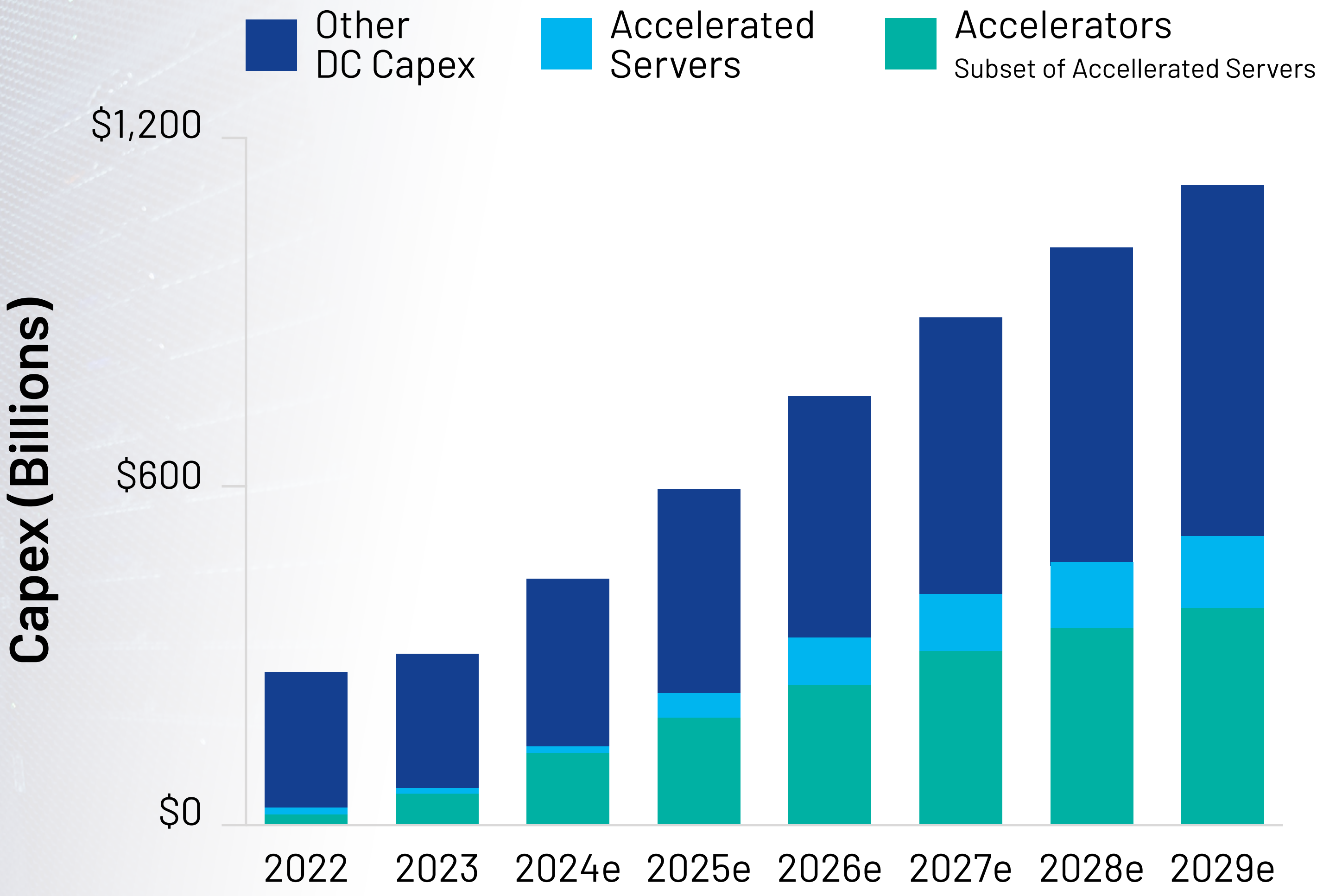


Total (6%) (5-year CAGR)



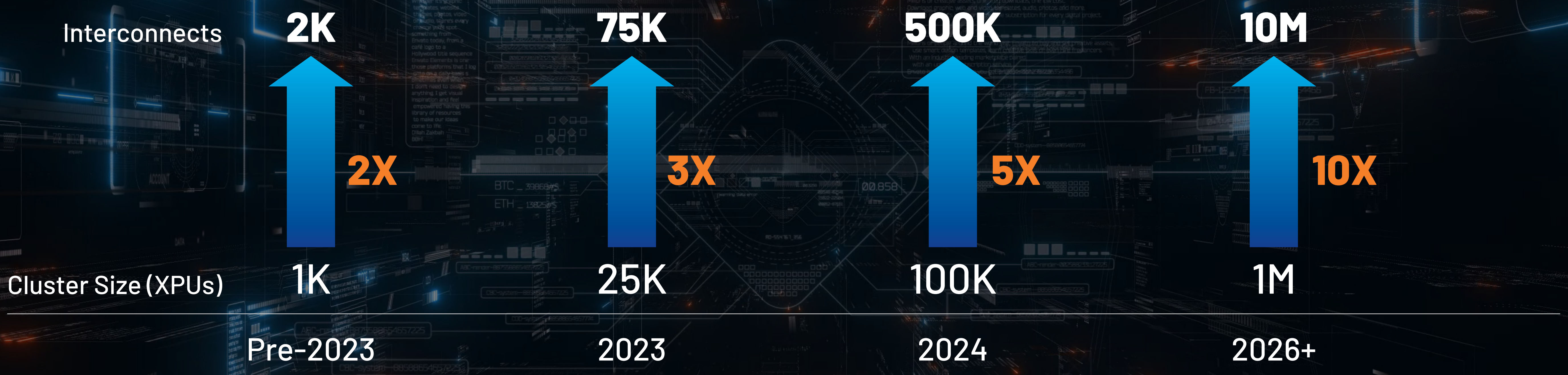
Total (24%) (5-year CAGR)

Data Center CapEx Five-Year Forecast



By 2029, **nearly half** of all data center CapEx is projected to be driven by accelerated servers. This underscores the increasing importance of AI-ready infrastructure and the need for strategic planning to manage investment and growth effectively.

AI Clusters Exploding in Size



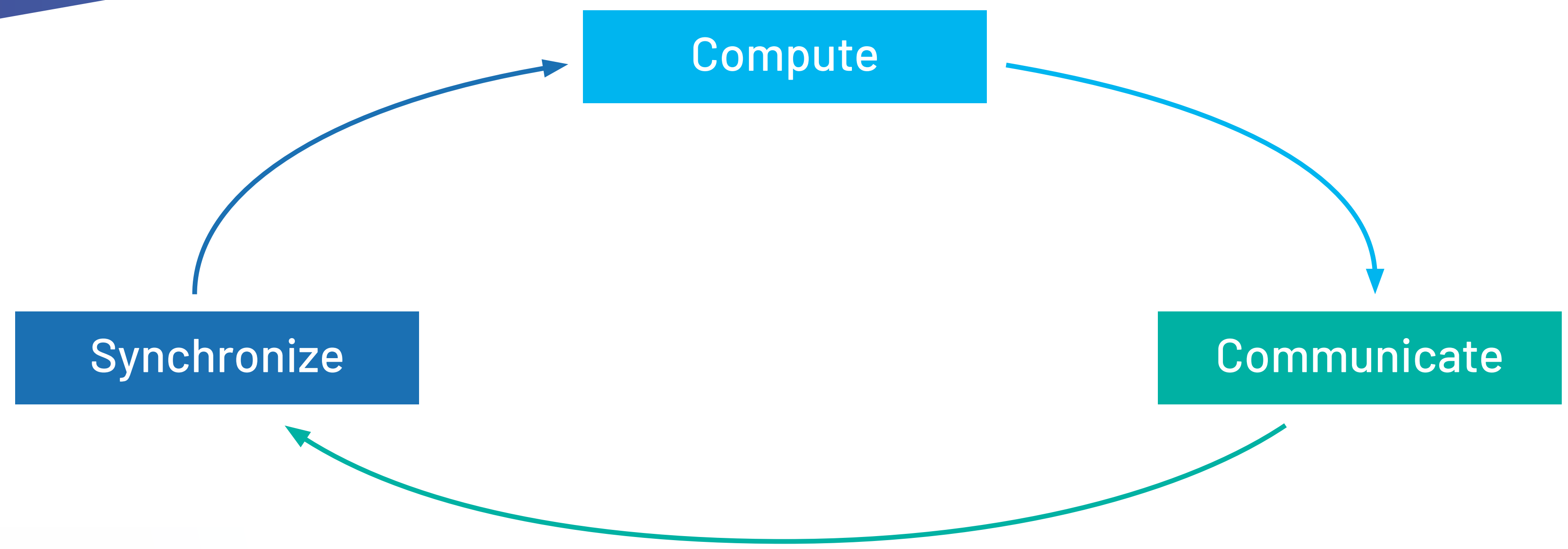
	Pre-2023	2023	2024	2026+
Interconnects	2K	75K	500K	10M
	2X	3X	5X	10X
Cluster Size (XPUs)	1K	25K	100K	1M

New AI Models:
From 100K+ to
1 million accelerators

Distributed Systems

THE NETWORK IS THE COMPUTER

AI Traffic Characteristics

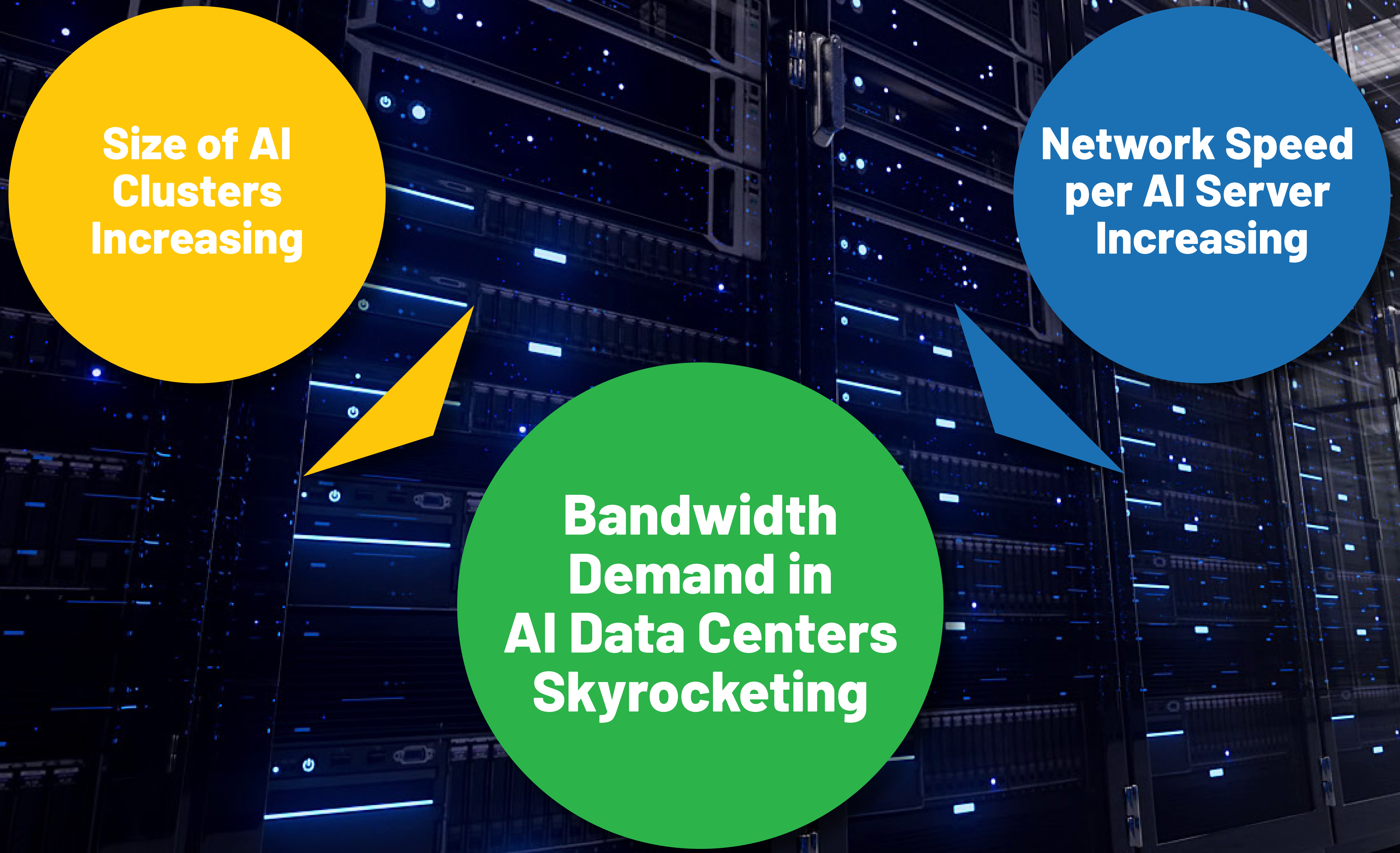


AI workloads create unique traffic patterns and networking requirements:

- Large volumes of "elephant flows" (high-bandwidth, long-lived data transfers)
- Highly compute- and data-intensive workloads
- Frequent short remote memory accesses
- Many nodes initiating transmission simultaneously
- Overall progress often constrained by any single delayed flow

- **Average cluster sizes are increasing rapidly:**
  - AI model sizes are growing 1,000X every three years
  - Cluster size (measured in accelerators) is quadrupling every two years
- **Bandwidth requirements per accelerator are increasing:**
  - From 200/400/800 Gbps today to 1 Tbps+ in the near future
- **AI traffic growth rates are accelerating:**
  - Up to 10X every two years in some large cloud provider networks

Drivers for Bandwidth Demand in AI Data Centers



Size of AI Clusters Increasing

Network Speed per AI Server Increasing

Bandwidth Demand in AI Data Centers Skyrocketing

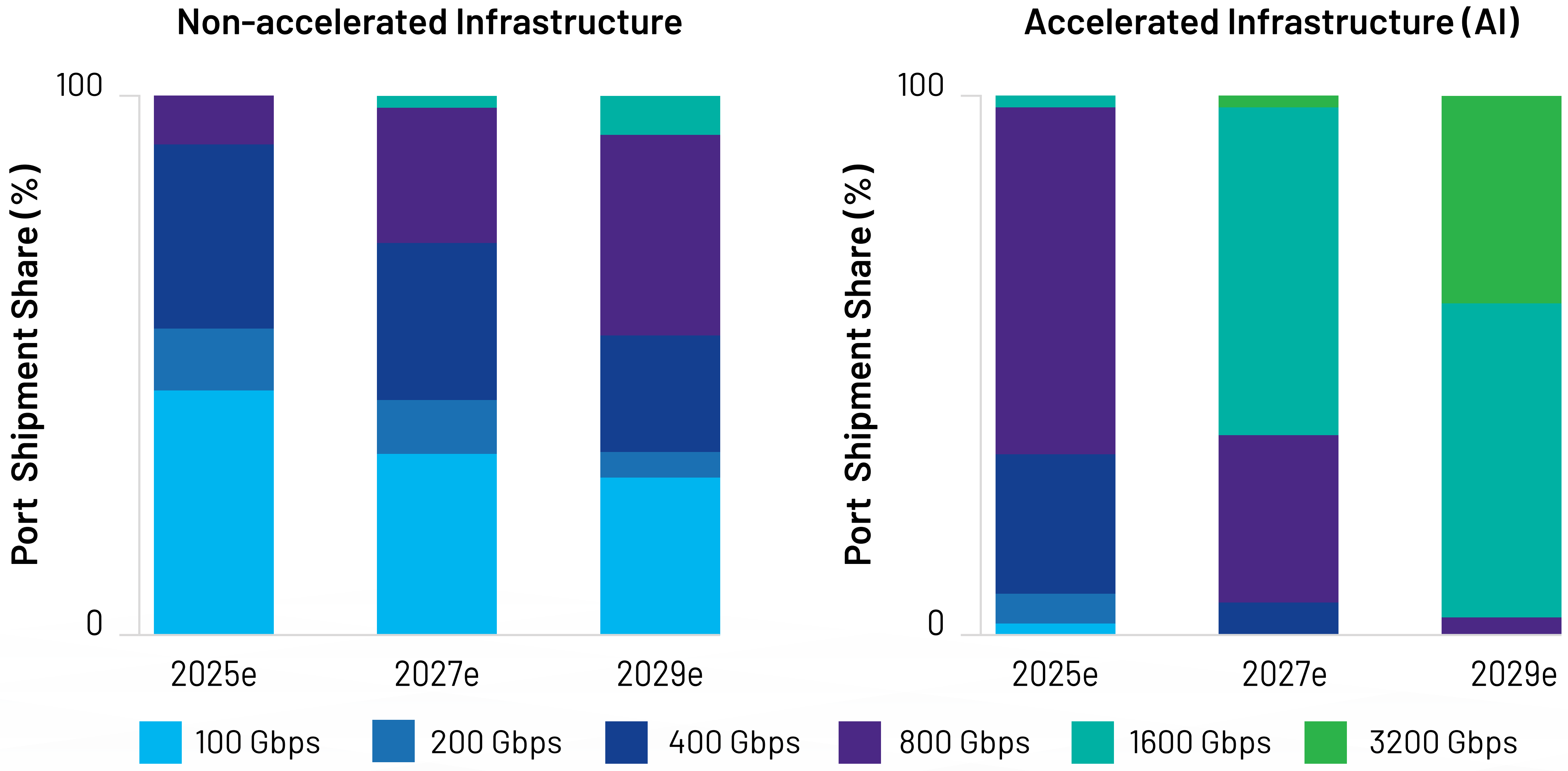
Switch Port Speed Transition in AI Data Centers (100 Gbps+)

Unlike traditional data center networks, which are compute-bound, AI data centers are increasingly network-bound — meaning network capacity is the limiting factor.

AI networks must operate near full utilization to maximize the performance of high-cost GPU resources. Back-end AI networks also have much shorter refresh cycles — about two years or less, compared to five years in traditional front-end networks.

As a result, switch port speeds are advancing rapidly:

- 2024: Majority at 800 Gbps
- 2027: Majority expected at 1,600 Gbps
- 2030: Majority projected at 3,200 Gbps



- Computer-bound
- Low network utilization rate (average of 50%)
- Refresh cycle is 5 years or more

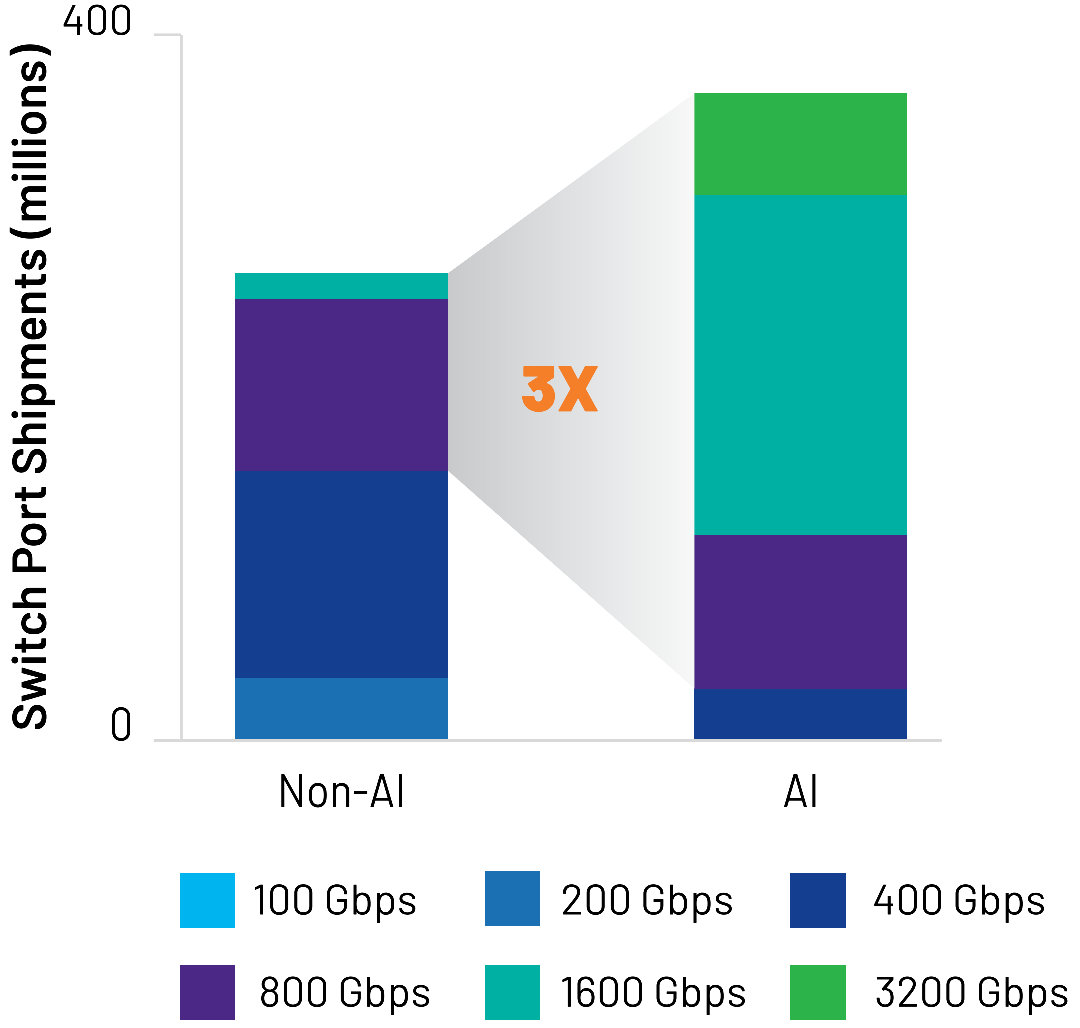
- Network bound
- Network needs to operate at nearly 100%
- Refresh cycle is 18 to 24 months

AI Clusters Drive Higher Optics Demand

The demand for high-speed switch ports in AI clusters is accelerating rapidly. Over the next five years, **AI data centers are expected to ship well over 300 million high-speed ports, with the majority at 1,600 Gbps.**

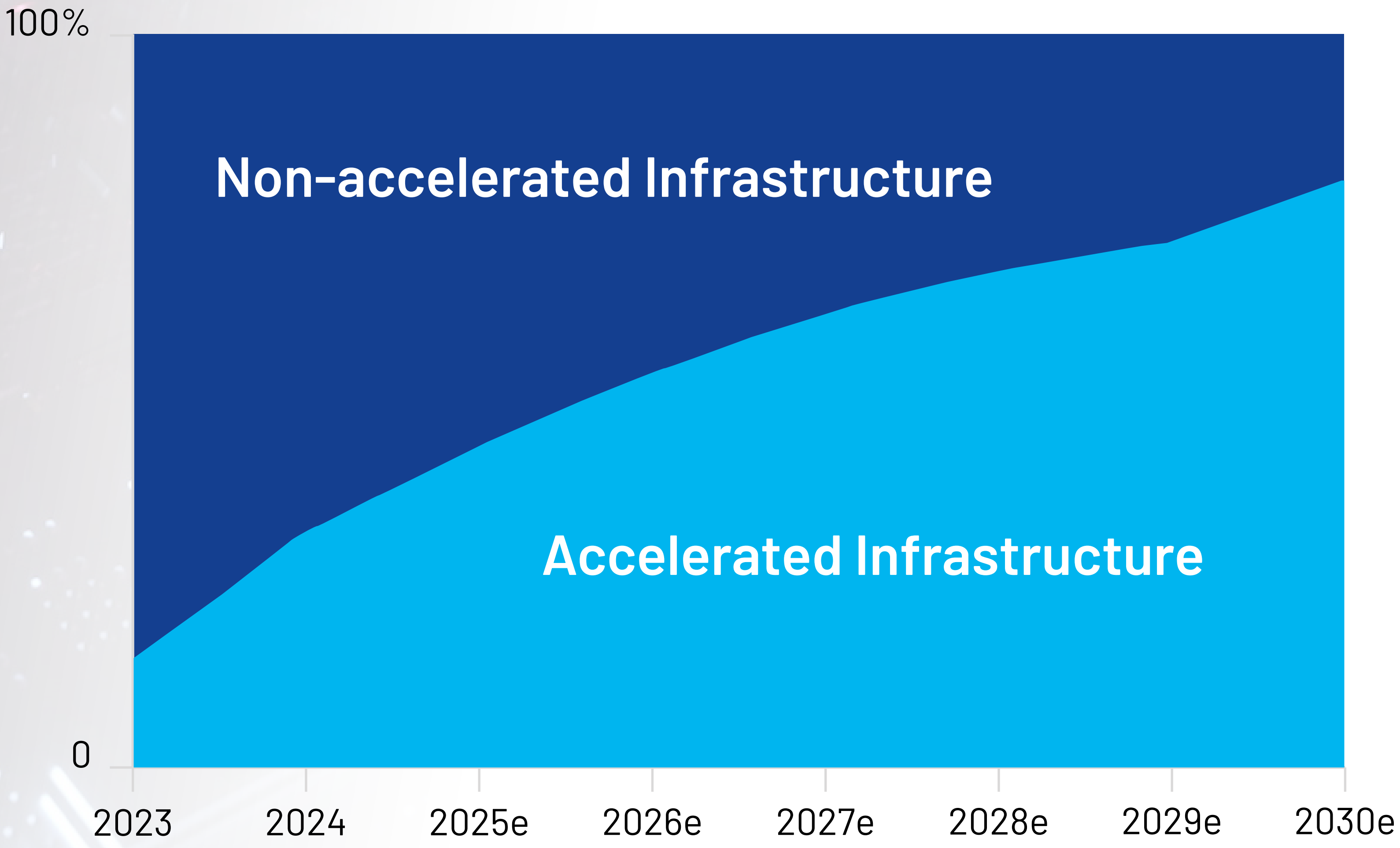
In comparison, non-AI data centers are projected to ship just over 100 million ports during the same period, with the bulk concentrated at 800 Gbps.

Cumulative Switch Port Shipments (2024-2029)



Bandwidth Requirements in AI Data Centers

Accelerated computing has become essential for AI's evolution, significantly improving the speed of complex computations and data processing tasks. These capabilities are increasingly critical not just for AI itself, but also for applications across industries that depend on AI and machine learning to deliver new capabilities and insights.



Key Takeaways

- **AI infrastructure buildouts continue to accelerate.**
- **By 2028, AI-driven networks are projected to nearly double the addressable market for data center switches.**

> *At this point, I'd rather risk building capacity before it is needed, rather than too late.*
>
> **MARK ZUCKERBERG**
> AUGUST 2024

> *In tech, when you are going through transitions like this, the risk of underinvesting (in AI) is dramatically higher than overinvesting.*
>
> **SUNDAR PICHAI**
> JULY 2024

VIAVI Solutions

VIAVI (NASDAQ: VIAV) is a global provider of network test, monitoring and assurance solutions for telecommunications, cloud, enterprises, first responders, military, aerospace and railway. VIAVI is also a leader in light management technologies for 3D sensing, anti-counterfeiting, consumer electronics, industrial, automotive, government and aerospace applications. Learn more at www.viavisolutions.com/ai and follow us on VIAVI Perspectives, LinkedIn and YouTube.

DELL'ORO GROUP

Dell'Oro Group, founded in 1995 and headquartered in Silicon Valley, is a boutique research firm specializing in telecommunications, security, enterprise networks, and data center infrastructure. They specialize in quarterly reports on market share, market size, and pricing, by macro region of the world, as well as five-year forecast reports, analyst judgment on technology trends, and market sizing with history and forecast. Dell'Oro Group also provide advanced research reports and special topic reports addressing developing trends. Industry executives, investors, and government agencies use Dell'Oro Group data for critical decision-making, and they have earned a "Gold Standard" reputation. Learn more at www.delloro.com

VIAVI Solutions

viavisolutions.com/ai