Juniper Networks

5 Key Data Center Trends for 2025

New currents in networking for AI, sustainability, AIOps, and more

Contents

In 2025, AI infrastructure shifts from aspiration to reality

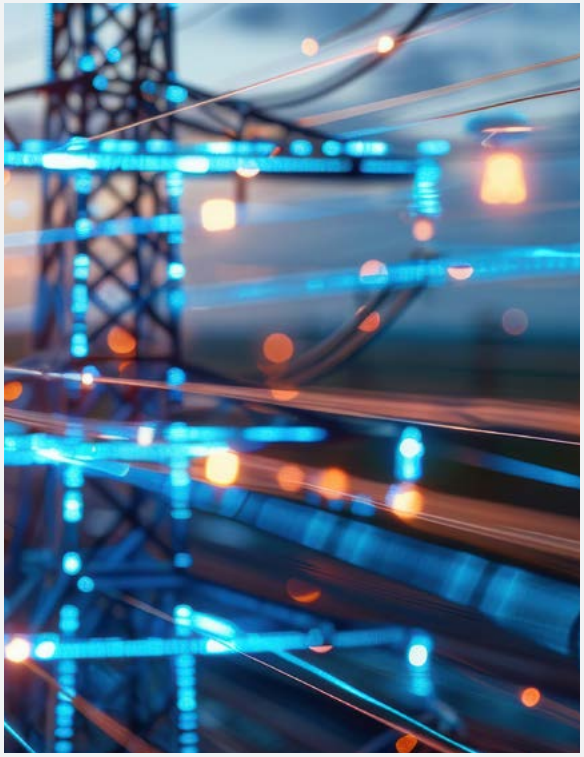
As the majority of organizations and employees move to implement AI, data center experts are rapidly evolving their approach to optimize the way they handle AI infrastructure.

Full production AI using enterprise-owned data center infrastructure is still an evolving trend, but some early adopters are now realizing they need to rethink how they handle elements like workload distribution, optimizing for inference, vendor mix, cooling, and more.

In 2025, organizations are getting more sophisticated and want more control over how they meet their AI goals by deploying new tools, new infrastructure, and new ideas that make end user experiences even richer. Here are five key trends driving innovation in the AI data center this year.

Key takeaways

Organizations face new challenges, like fitting the compute they need into the power they can access, and new opportunities, like agentic capabilities shifting AI into high gear.



$4.4T

Potential market value added

McKinsey predicts LLMs trained on corporate data could add $2.6-4.4 trillion in value across 63 business use cases.

15%

Of day-to-day work decisions automated

Gartner predicts that agentic AI will autonomously handle 15% of day-to-day work decisions by 2028.

3 NYCs

Equivalency of new power required by 2030

American Electric Power reports that they will need to connect three New York Cities' worth of electricity to the grid to satisfy customer demand by 2030, largely driven by new AI data centers.

01 Moving from RAG to riches

Moving from RAG to riches

Retrieval-Augmented Generation (RAG) will help large enterprises more easily customize off-the-shelf LLMs with their own data.

Inferencing systems deliver trained AI applications to end users and devices. Depending on the size of the model, inference can be deployed on a single GPU or server or as a multi-node deployment where the application is distributed across multiple servers for increased scale and performance.

RAG augments a pre-trained LLM by providing supplemental data obtained from a new data source. RAG helps enterprises deliver specific and accurate responses to customer or device queries, informed by their proprietary datasets, without the cost and headache of training their own LLM. This is important because most enterprises are currently sitting on unmined gold in the form or reams of corporate data. To extract maximum value from AI, enterprises need to effectively tap into their data.

But this data is often sensitive—especially in highly regulated industries—and it's typically hosted at enterprise, colocation, or edge data centers. Furthermore, end users don't want the extra second or two of lag that retrieving supplemental data from the cloud would potentially entail. For these reasons, RAG inference works best with an on-prem or hybrid cloud model.

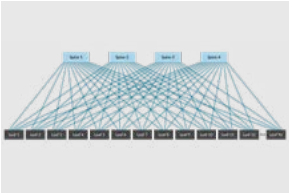
02 Open standards-based tech

An industry-wide shift will lead to open standards-based tech

A shift away from single vendor solutions is opening the supply chain, expanding enterprises' options and improving flexibility.

In the last quarter of 2024, we saw many customers who had originally opted for a single vendor solution reconsider their infrastructure needs. As their requirements came into focus and their comfort with the technology grew, the flexibility of open standards-based technology became more of a logical requirement.

Ethernet provides unparalleled flexibility, choice, and ease of operations. Enterprises still on the AI sidelines would be wise to build with open standards-based technology like Ethernet (versus InfiniBand) and leverage proven, multivendor management tools.



Flexibility

Platforms with intent-based automation and auto-generated configurations for multivendor switches are flexible enough to deploy new infrastructure quickly or upgrade legacy platforms with the latest fabric technology without service disruptions or training delays.

Choice

Any engineer comes Ethernet-ready, so operations teams can deploy consistent fabric technology across the entire data center and choose from any number of vendors and platforms.

Innovation

Ethernet is the most widely adopted and ubiquitous networking technology ever, and forthcoming developments like the Ultra Ethernet Consortium's Ultra Ethernet Transport (UET) protocol will optimize it even further for high-performance computing and AI networks.

03 AIOps takes center stage

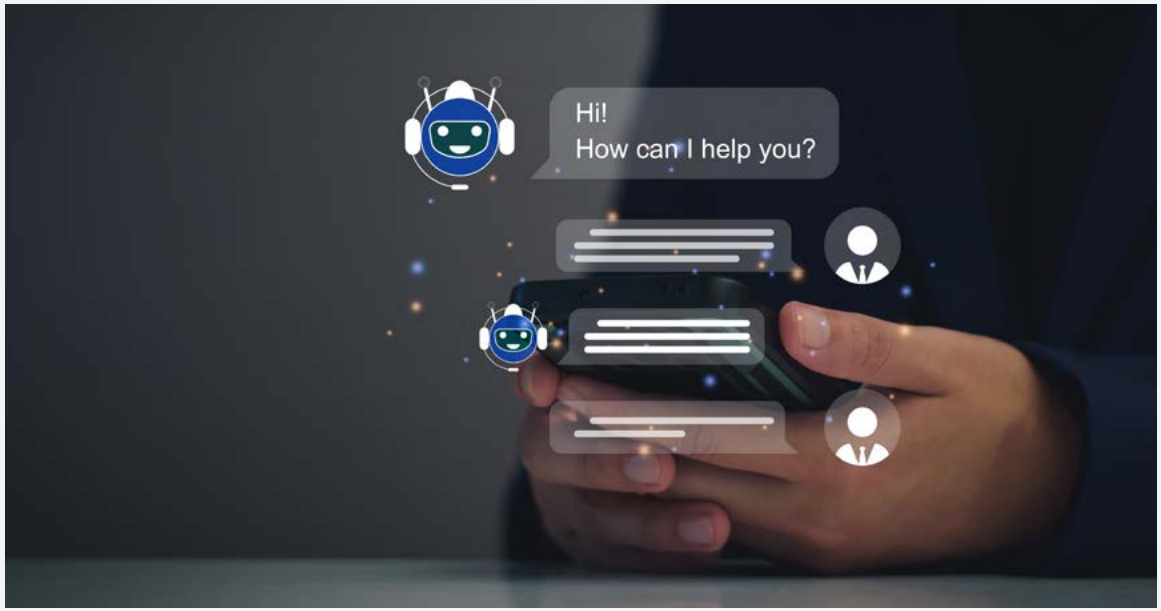
AIOps takes center stage

As network demands accelerate, AIOps will help optimize data center network performance for both traditional and new AI workloads.

AIOps will continue to play a growing role in predictive and proactive maintenance of data center networks, minimizing downtime, optimizing system health, and ultimately improving end user experiences.

As conversational chatbots emerge as a standard feature in automation and management tools, data center teams will interact with these systems easily using natural language. But many of today's natural language-based data center chatbots are currently more novelty than useful tool. In 2025, the leading DC chatbots will evolve into intelligent assistants—think junior engineer rather than inexperienced intern. Network agents will help human DC operators with coordination and administrative tasks, critical insights, and proactive recommendations.

These advancements will enable data centers to meet the growing demands of AI workloads with sophisticated operational intelligence.



04 AI goes agentic

In 2025, AI will go agentic

Agentic AI will do for executing workflows what GenAI does for creating content.

Imagine an AI that autonomously watches for extreme weather and reconfigures supply chains on the fly. Or an AI travel agent that books a hotel based on your preferences and income, and plans outings based on your preferred forms of recreation.

Late last year, Google rolled out Gemini 2.0, the closest step yet toward agentic AI—AI that can plan, reason, and take autonomous action based on user input—and more agentic-capable models will soon follow.

The use cases are countless, and the technology is still in its infancy, but this is where enterprises are likely to get the best ROI from AI. To ensure great end user experiences and responsive decision-making, agentic AI will need to pull together data from multiple sources with low latency. Enterprises looking to maximize their agentic AI's potential will need robust data center infrastructure (and data strategy) in place.

Agentic AI could dramatically improve AI's potential

Gartner

05 Power and cooling for the AI era

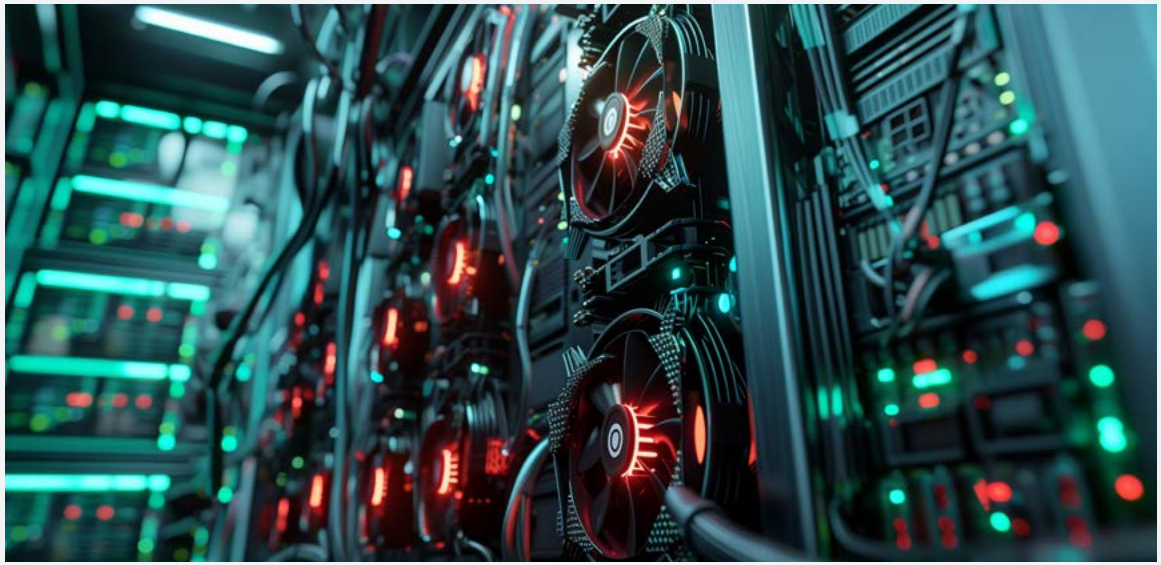
Organizations get creative about power and cooling

As the AI market explodes, operators look to new cooling methods and alternate energy sources to meet their sustainability goals.

AI infrastructure demands a tremendous amount of electricity—GPUs are power-hungry and require more cooling than traditional CPUs—and available power threatens to be a primary limiting factor for organizations hoping to build out their own AI clusters.

This year will bring more innovative ways to optimize power and water usage. Liquid cooling and closed-loop zero-evaporation cooling methods will continue to gain traction. In networking, customers will turn toward more power efficient optical modules, like linear pluggable optics (LPO) and linear receive optics (LRO), to achieve high performance with reduced energy consumption. AIOps will allow data centers to continuously monitor, analyze, and adjust cooling systems in real time.

Nuclear energy will also play a major role as organizations look to meet their growing compute demands without compromising their sustainability initiatives. The nuclear plant at Three Mile Island is planned to reopen to power Microsoft's AI and cloud computing data centers, and the tech giant has invested a billion dollars in nuclear energy in Wyoming. A handful of companies are building minireactors to power data centers, as well.



1.5 million

liters of water per day for a single hyperscale facility's cooling and humidification

Next steps

Get the AI infrastructure solution that lets your team shine.

Connect with an expert

Ready to learn how Juniper Networks can guide you on your networking for AI journey? Schedule a consultation with an expert.

Contact us →

Discover The NOW Way to Network for AI

Learn about Juniper's AI data center solutions, our capabilities practice and ecosystem of products, solutions, and expert consultative networking services.

Explore NOW →

More related resources

Find out how Juniper has turned clients' ideas, technologies, and opportunities into outcomes across a variety of industries and scenarios.

White paper: Networking the AI Data Center →

Video: AI Networking is Crazy (But Is It Fast Enough?) →

Juniper's Ops4AI Lab →

Get more insights

Get more insights into emerging innovations and trends in business, tech, artificial intelligence, and more delivered directly to your inbox.

Visit the feed →

juniper.net

Juniper Networks Inc.
1133 Innovation Way
Sunnyvale, CA 94089

7400203-001-EN January 2025