Data Center Vision: How Datacenter Infrastructure Will Evolve to Support AI and Accelerated Compute

Sean Graham
Research Director,
Cloud to Edge Datacenter Trends, IDC

Table of Contents

CLICK BELOW TO NAVIGATE TO EACH SECTION IN THIS DOCUMENT.

Executive Summary

AI is undergoing unprecedented innovation and investment. It will redefine business operations and drive transformative disruption. AI leverages accelerated computing, leading enterprises to develop new datacenter strategies, implement new technologies, and acquire new skills to support AI and the datacenters of the future.

Key Findings

AI as the new strategic workload

GenAI/AI will become a strategic workload that requires dedicated investments and architectural choices.

Role of the CIO

CIOs are at the intersection of business and technology and will need a strategic road map, a governance structure, and AI economic models that align IT investments with business goals.

Power density

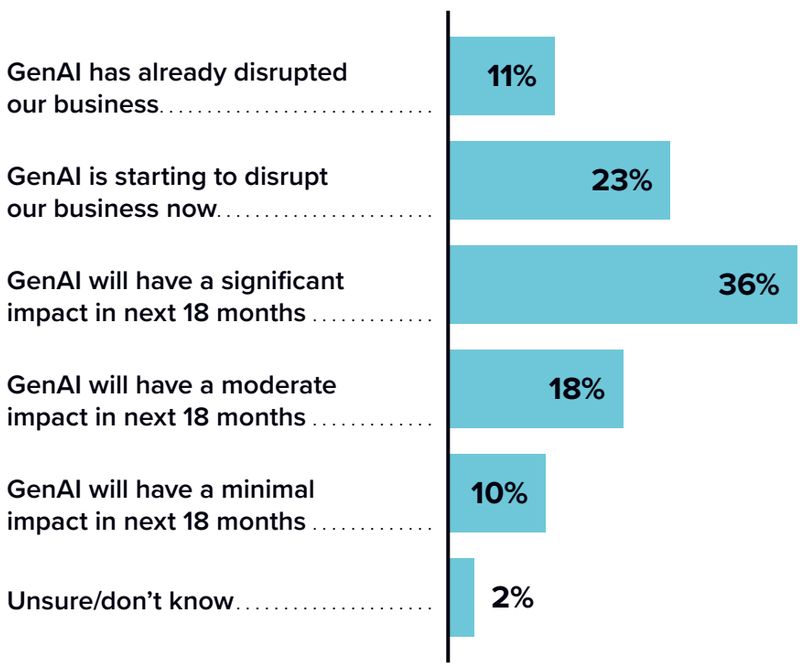
Power density in datacenters is increasing by orders of magnitude. The datacenter of the future will need to be fundamentally different from current designs to support AI.

GenAI's Impact on Business

Over one-third of tech leaders believe that GenAI is already disrupting their business, and 88% expect it to at least moderately impact their business by October 2025.

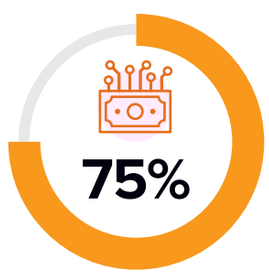
To what extent do you believe broad access to GenAI tools will disrupt your organization's competitive position?

(Worldwide)



By 2026, CIOs will:

lead the creation of a strategic road map for AI.



implement AI economic models, including tech acquisition costs

CIOs will play a strategic role by creating a comprehensive, strategic road map, a governance structure, and AI economic models that align IT investments with business goals to drive sustainable growth.

n = 889 (Worldwide), n = 220 (North America), n = 220 (Western Europe), n = 300 (Asia/Pacific); Source: IDC's *Future Enterprise Resiliency & Spending Survey Wave 4,* April 2024; Source: IDC's *2025 FutureScapes*

Enterprises Are Prioritizing Investment in HR, Legal, Supply Chain, Research and Development, Procurement, and Marketing Use Cases

Current Investment:

As competitors automate parts of their business with AI, they will start to see productivity gains that may improve their financial performance — organizations will need to ensure they innovate at the same rate.

HR

Talent Acquisition
Resume Selection

Marketing

Content Marketing
Local Content Translation

Marketing

Web Marketing
Personalized Digital Assistants

Marketing

Advertising
Personalized Promotions

Procurement

Contract Management
Contract Summarization

Marketing

Marketing Ops
Instantaneous Insights

Procurement

Spend Analytics
Spend Report Creation

Sales

Guided Selling
Generative Recommendations

Procurement

Risk Management
Persona-Based Dashboards

Marketing

Advertising
Video Creation from Scripts

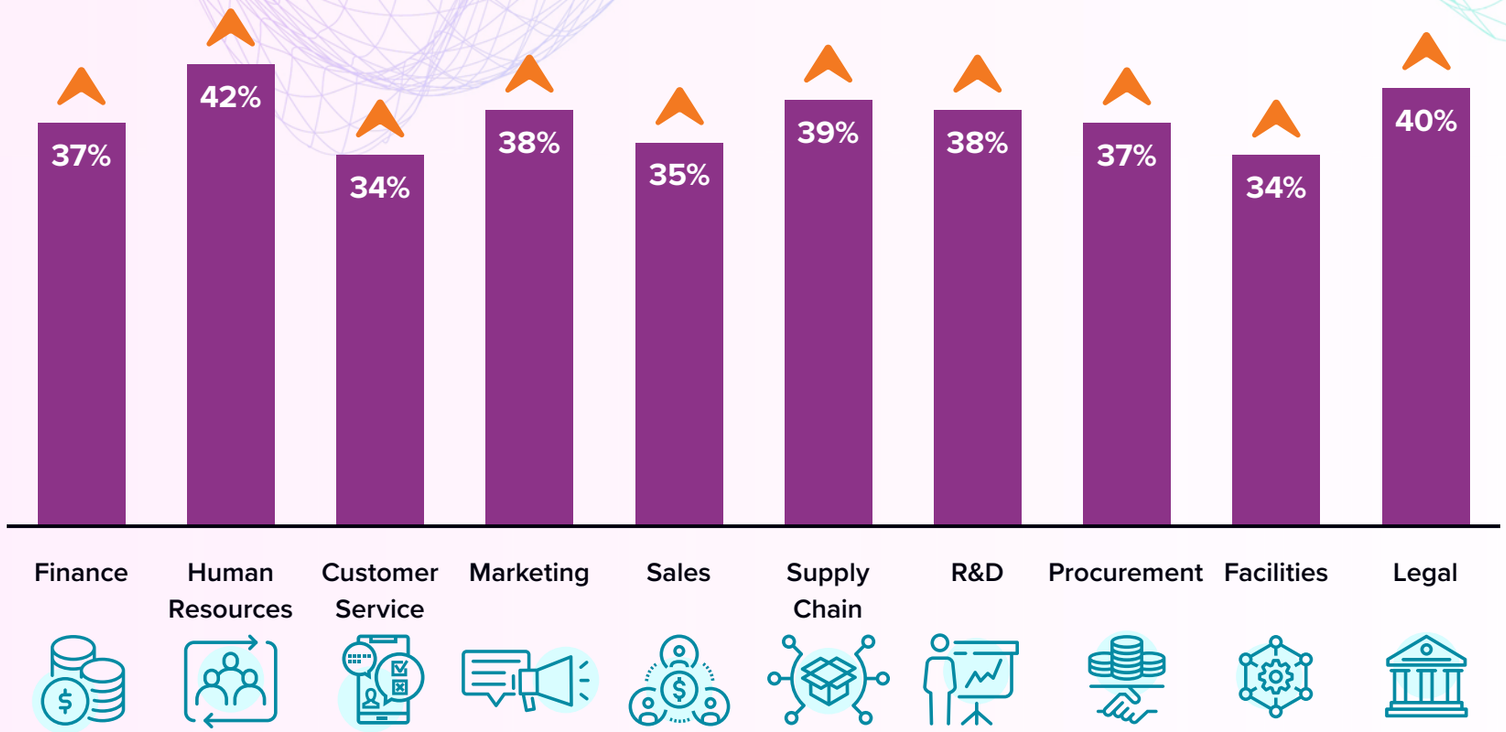
Continued on the next page ▶

n = 3,130; Source: IDC's *WW AI Use Case Survey,* August 2024

Enterprises Are Prioritizing Investment in HR, Legal, Supply Chain, Research and Development, Procurement, and Marketing Use Cases (continued)

Future Investment:

Enterprises plan to "double down" in the functional areas where they have already made investments to demonstrate a faster ROI.



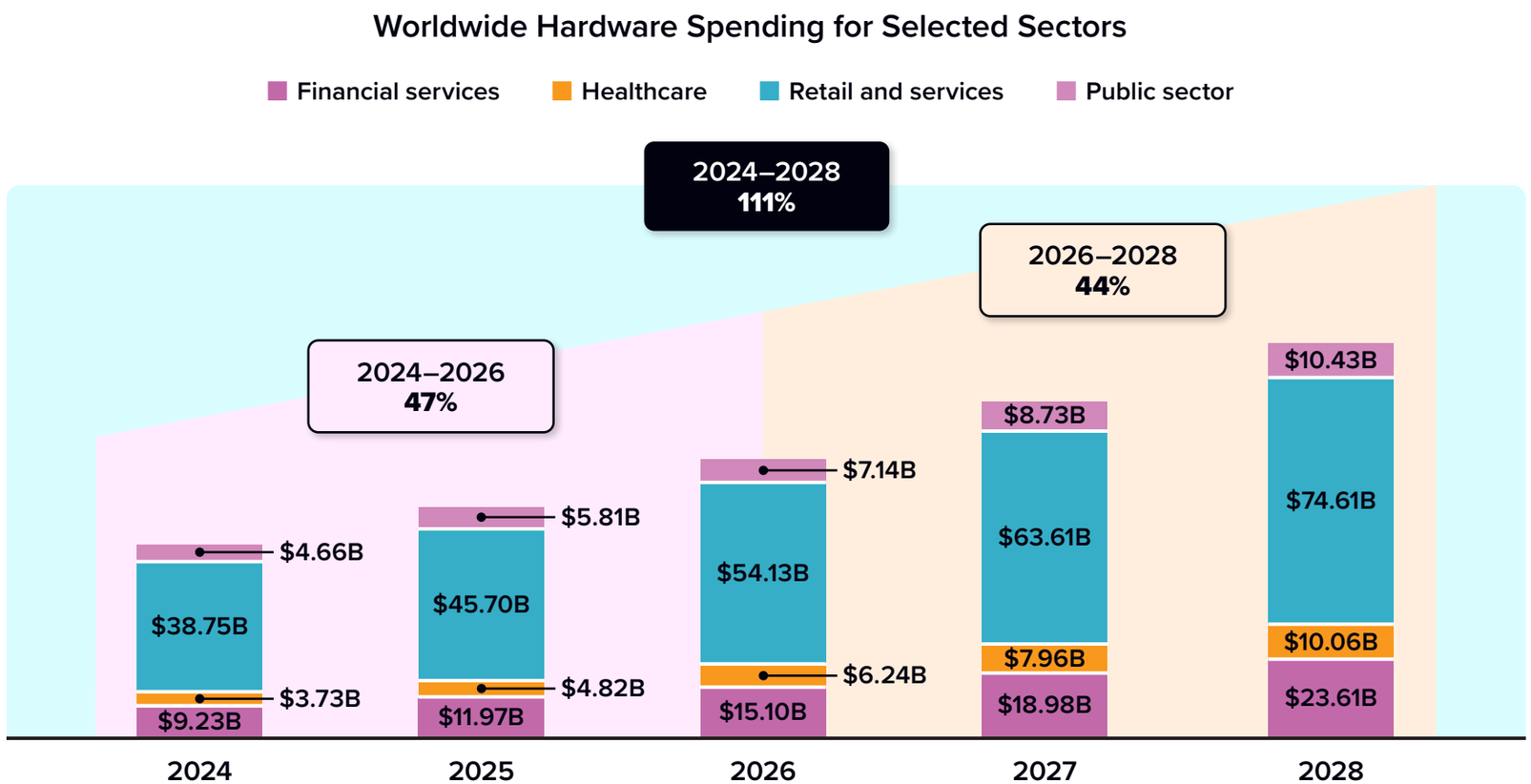
Continued from the previous page

n = 3,130; Source: IDC's *WW AI Use Case Survey,* August 2024

IT AI Hardware Spending Will Grow

Worldwide IT AI hardware spending will increase by 21.3% annually from $72 billion in 2024 to $258 billion in 2028 to support AI use cases.

CIOs must consider a datacenter portfolio that can **accommodate the immediate need** for scale while remaining **flexible to future proof operations** as the industry experiences massive innovation and growth.



Source: IDC's *AI Spending Guide*, November 2024 | For an accessible version of the data on this page, see Supplemental Data in the Appendix.

AI Is the Next Major Workload

Management of AI Workloads

84% of tech leaders believe that GenAI is a major new corporate workload, similar to ERP, requiring an incremental increase in new technology spending in the next several years

85% of tech leaders believe that supporting GenAI as a strategic workload requires a dedicated strategy across infrastructure, data, the cloud, and services

Impact of Treating AI as a Major Workload

Tech leaders cite preparing for greater AI use in the business as the top factor driving significant spending increases in 2025

Tech leaders increasingly recognize that **GenAI/AI will become a strategic workload** requiring dedicated investments and architectural choices and drive datacenter investment in 2025.

Top Areas for Significant Budget Increases in 2025

(Regardless of overall budget plans)

- **36%** Enhancing cyber-recovery/-resiliency
- **35%** Modernizing core enterprise applications
- **31%** Expanding infrastructure in edge locations
- **31%** Modernizing infra corporate datacenters

35% said updating power/cooling would be **critical.**

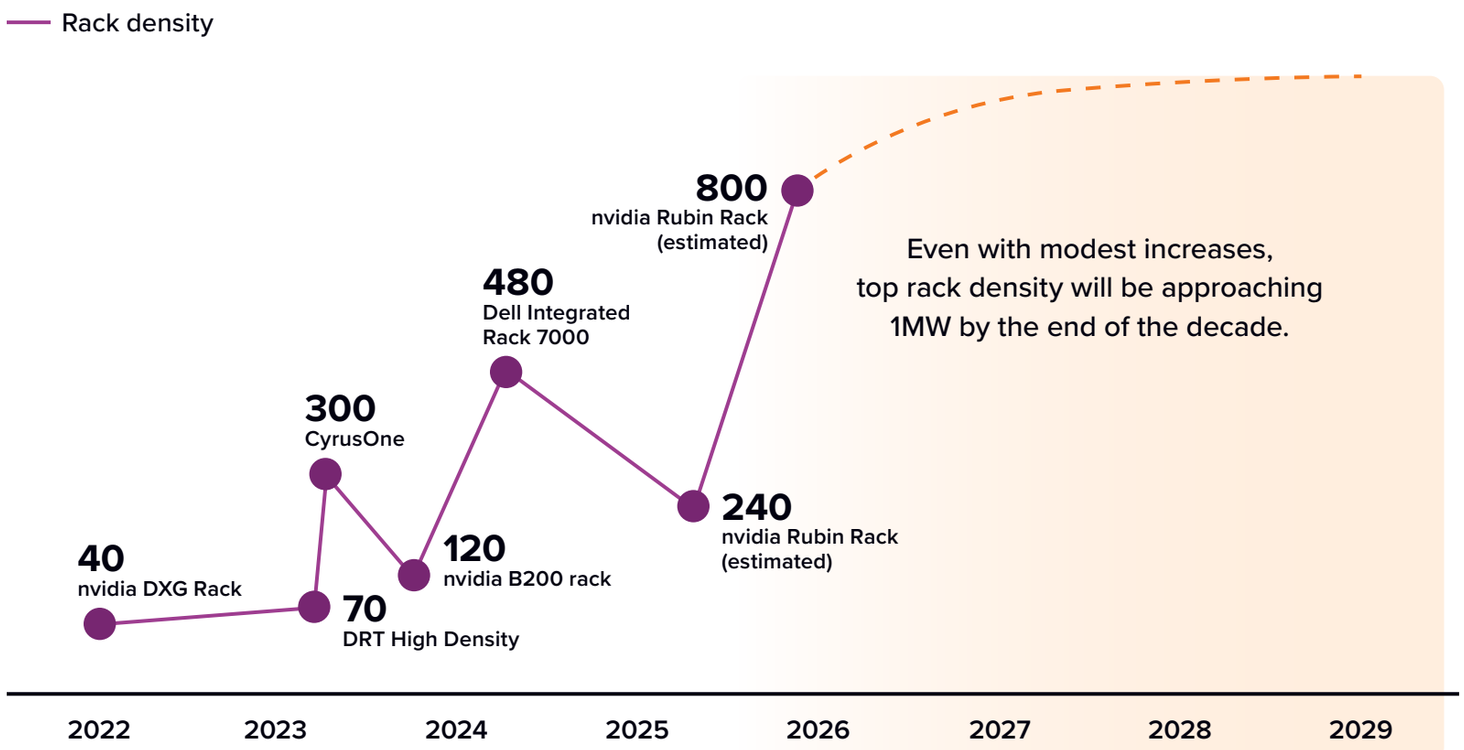
45% said it would be important.

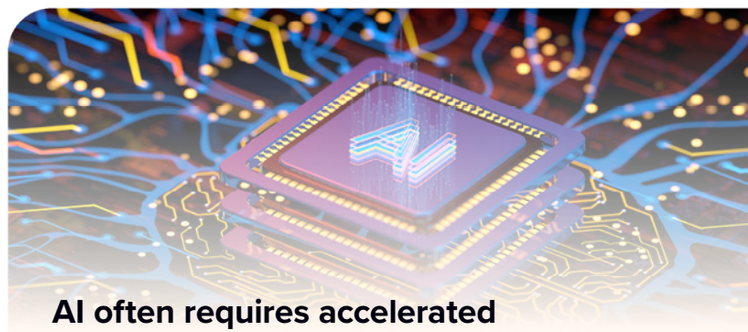
n = 889; Source: IDC's *Future Enterprise Resiliency & Spending Survey Wave 4*, April 2024. n = 891; Source: IDC's *Future Enterprise Resiliency & Spending Survey Wave 8*, September 2024

Why Updating Power Is Critical for AI

Rack densities are rising exponentially. The average rack density was less than 10kW at the beginning of 2020, but we will see racks over 1MW by the end of the decade.

Industry Examples of High AI Rack Densities





AI often requires accelerated computing, which is the ability of a computer to accelerate applications and workloads by offloading a portion of core processing functions onto adjacent silicon subsystems, such as GPUs (accelerators), resulting in racks with high power density.

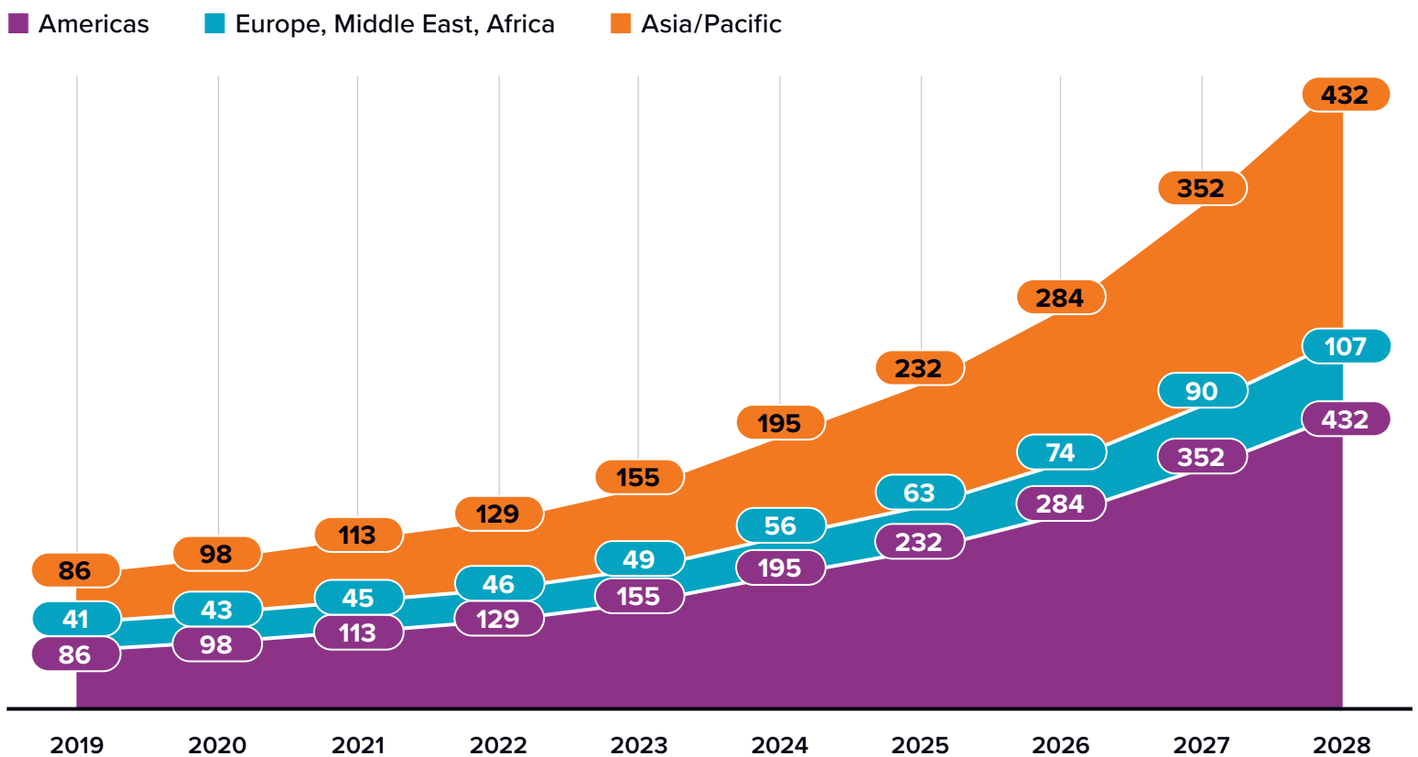
The number of GPUs in an AI rack can exceed 70.

Source: IDC's *Future Enterprise Resiliency & Spending Survey Wave 8*, September 2024 | For an accessible version of the data on this page, see Supplemental Data in the Appendix.

Forecasted Datacenter Energy and Power Growth

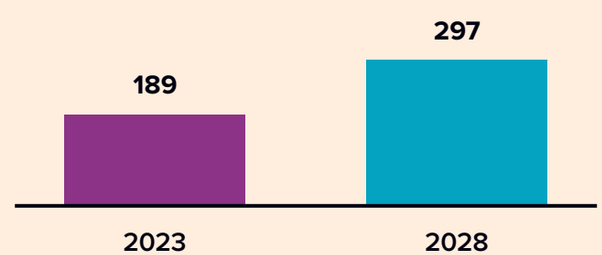
Rack density growth to support AI has industrywide implications on the need for additional energy and datacenter capacity.

Datacenter Energy Consumption (TWh)



Datacenter energy consumption will grow at a compound annual growth rate of 23.2% going from 397 terawatt-hours (TWh) in 2024 to 915 TWh in 2028, with an additional expected IT capacity of over 100GW.

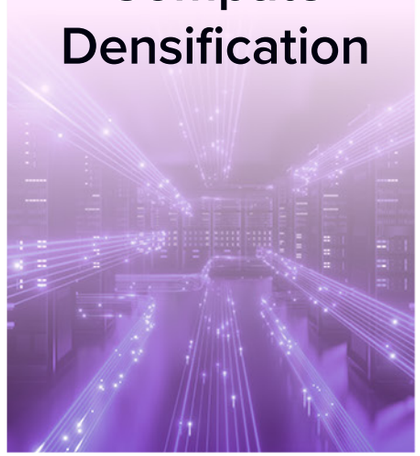
Worldwide IT Power Capacity (GW)



n = 891; Source: IDC's *Datacenter Trends: Sustainable Builds and Carbon Emissions* | For an accessible version of the data on this page, see Supplemental Data in the Appendix.

Impacts of Power Density on Datacenter Design

High power densities and rapid innovation create new requirements for supporting AI in enterprise datacenters.

AI Power Demands Creating Ripple Effects in Datacenter Design and Deployment	Trends and Requirements	Enterprise Needs and Mandates
Compute Densification 	**Power Availability and Permitting:** Energy demand outpacing new supply both globally and in key markets	**Sustainability:** Meeting corporate sustainability goals without sacrificing competitiveness
	Emissions: Governments addressing climate challenges, limiting and delaying new development	**Improved Reliability and Operability:** Strengthening fault tolerance and simplifying management
	Accelerating Deployment Cycles: IT changing every three to four years, while datacenters last 15–20 years	**Skill and Labor Shortages:** Lack of skilled workers and insufficient manpower
	Grid Stability: Aging electrical grids raising concerns about their stability and reliability	**Power Availability:** Energy demand outpacing new supply both globally and in key markets

Supporting AI in the Datacenter

Liquid Cooling: Energy-efficient heat dissipation for high-density compute

Mechanical, Electrical, and Plumbing: Support for new tech and densities

Modular Datacenters: Efficient, flexible, scalable capacity

Heat Reuse: Excess heat conversion for productive use (i.e., district heating)

Microgrid: Localized power grid with energy storage

Energy Management Systems: Software that controls energy consumption

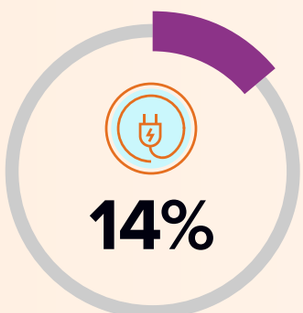
Power Purchase Agreements: Contract for buying electricity, often from renewables

Services: Design, installation, and ongoing services

Power Scarcity

Most organizations are unprepared for predicted power shortages, competing with other industries for power and facing unique problems due to the energy intensity and geographic concentration.

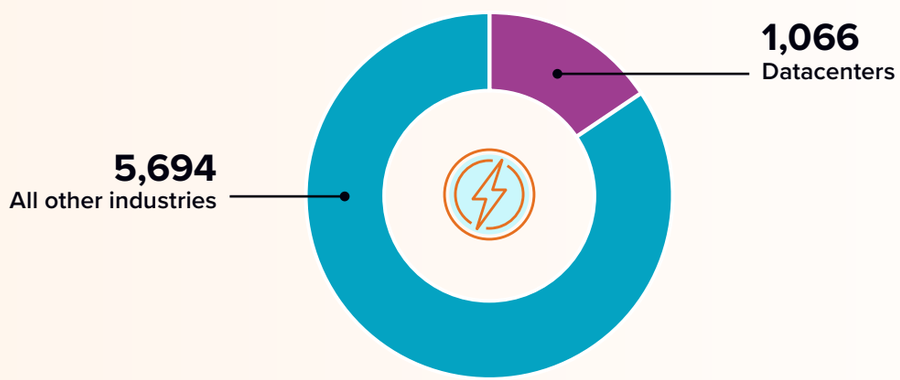
Organizations citing that addressing power scarcity is a top priority



Enterprise datacenter operators ranked addressing **power scarcity 10th** (out of 10) as a current priority, with only 14% citing it as the top priority.

Electricity Demand Growth, 2023–2030

(6,760TWh)



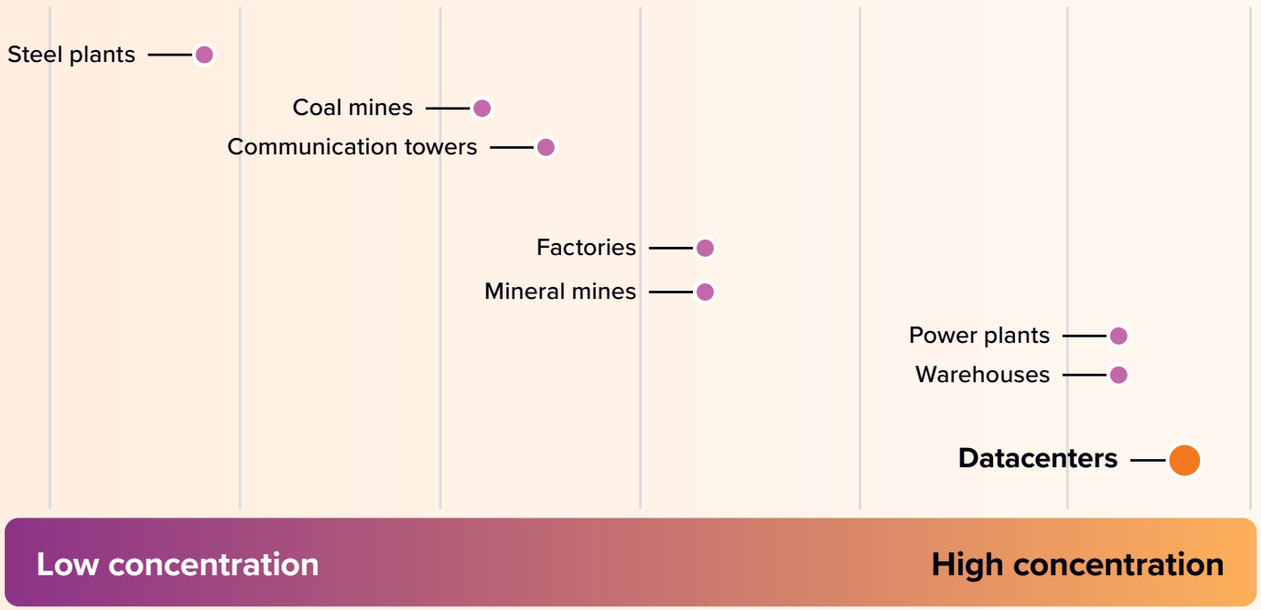
With rapid growth and increasing electrification and digitalization, datacenters will need to compete with other industries for power capacity.

Continued on the next page ▶

n = 766; Source: IDC's *Accelerated Racks, 2025: Projections and the Competition for Energy, Sean Graham, #US53232525*, March 2025

Power Scarcity (continued)

Spatial Concentration of Selected Types of Facilities, United States



Datacenters' exceptionally high spatial concentration has significant implications for local power grids because of their substantial power requirements.

Datacenters cluster to improve connectivity, share infrastructure, access skilled labor, benefit from government incentives, and reduce latency. However, in addition to competing with other industries, they will have to compete with one another for available power, a scarce resource, in areas of high datacenter concentration, such as Northern Virginia; New Albany, Ohio; Dublin; and Singapore. This will lead to prices increasing above historical rates and delayed access to power.

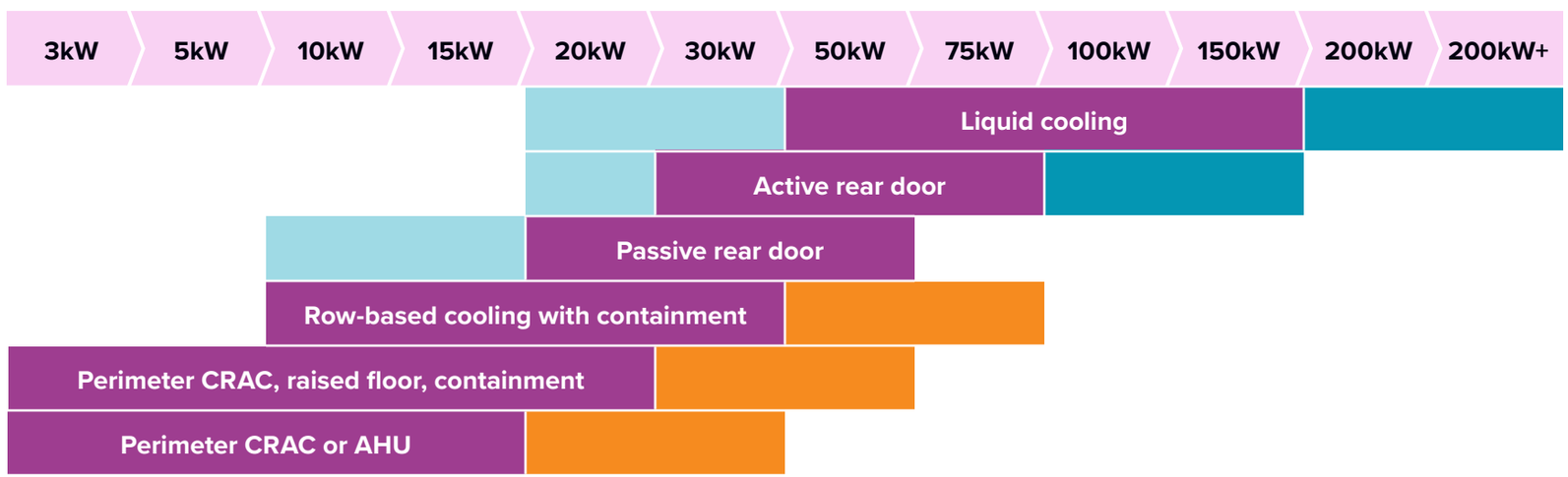
Continued from the previous page

n = 766; Source: IDC's *Accelerated Racks, 2025: Projections and the Competition for Energy, Sean Graham, #US53232525*, March 2025

GenAI's Higher-Density Racks Often Require Different Cooling Methods and Higher Cooling Efficiency

Matching the cooling method to rack density can optimize performance.

- Lower boundaries for efficiency or use case
- Combining multiple technologies for a hybrid solution
- Upper boundaries for extreme densities, with rack sizes typically increasing in height and width



High-density AI computing uses liquid cooling due to its superior thermal performance.

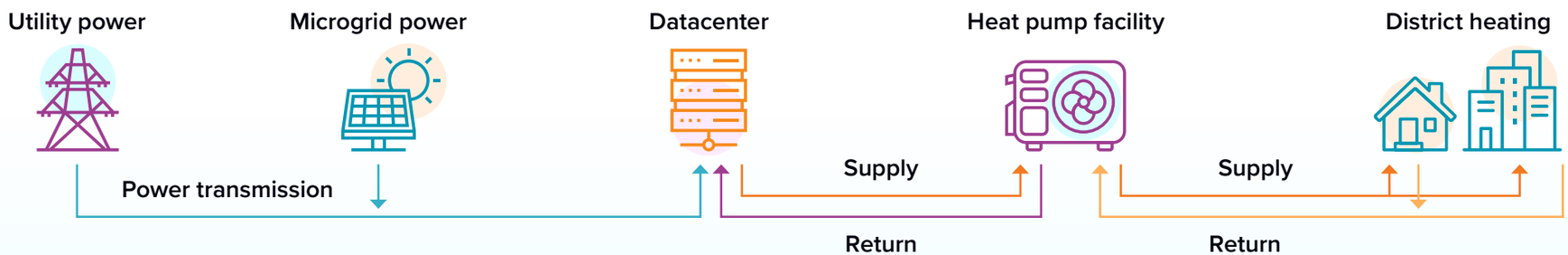
The highest power densities require hybrid cooling (liquid and other methods, such as rear door heat exchangers).

Source: IDC's *Datacenters: Best Practices to Mitigate the Environmental Impact of GenAI*

Benefits of Liquid Cooling

Liquid cooling enables the reuse of datacenter waste heat for other purposes, supporting sustainability and circularity and addressing industry power scarcity.

Up to 90% of heat from datacenters is reusable.



Benefits

Reduced carbon footprint: By reusing waste heat for other purposes, datacenters can significantly reduce their carbon footprint and greenhouse gas emissions.

Sustainable practices: Embracing waste heat reuse aligns with broader sustainability goals and contributes to a more environmentally responsible approach to datacenter operations and circular economy goals.

Cost savings: Reusing waste heat can reduce energy costs by decreasing the need for traditional heating systems and other energy-intensive processes.

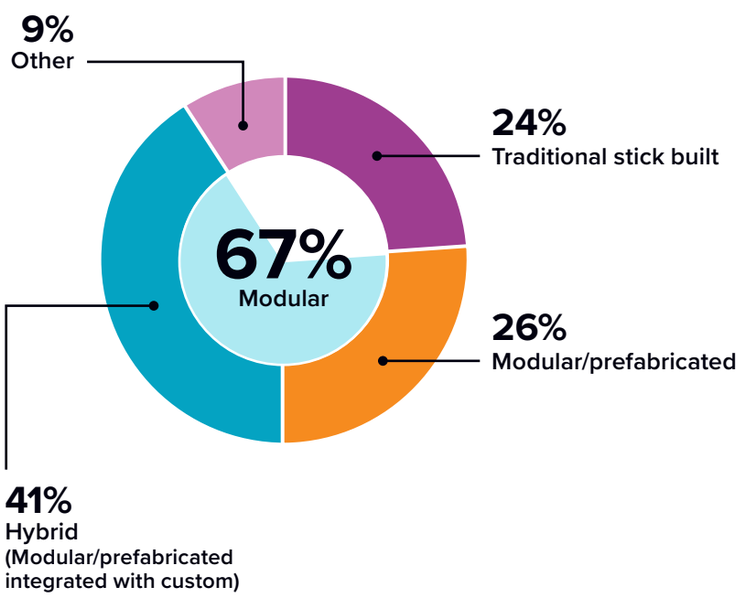
Increased revenue potential: In some cases, datacenters can generate additional revenue by selling excess heat to nearby businesses or communities.

Community impact: Reusing waste heat can benefit local communities by providing heating for homes and businesses or supporting local industries.

Modular Approach

Modular datacenters (all-in-ones, modules, or data halls) have become the preferred construction method, allowing quick and flexible scaling.

What percentage of each of the predominant construction methods did you use to construct your current datacenter?



n = 766; Source: IDC's *Datacenter Operations and Sustainability Survey*

The modular and prefabricated datacenter advantage:

Expansion

- Built with sustainable materials and low waste since designed and managed offsite
- Easier deployment for new technology near existing datacenters, business users, or data
- Faster deployment while streamlining supply chain challenges
- Optimized power and cooling for high-density racks for energy and water-efficient solutions

1 Environment Friendly
Prefabricated modular data centers minimize construction waste and eco-impact with streamlined, factory-based production, enhancing green building practices.

2 GenAI and NextGen Ready
With their scalable and swiftly deployable infrastructure, these datacenters ensure AI technologies and applications have the computing resources they need, when they need them.

3 Energy-efficient
Optimized design and advanced cooling systems drastically reduce energy consumption and operational costs, promoting environmentally sustainable computing.

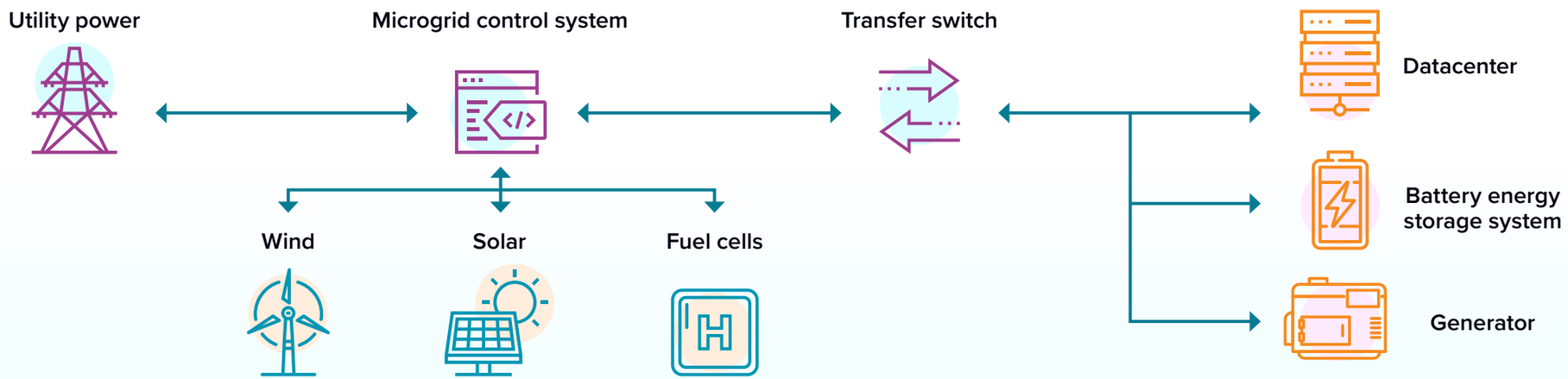
4 Location Flexibility
Requiring minimal site preparation, these solutions easily deploy in diverse environments, expanding data processing capabilities across any enterprise location.

5 Cost
Accelerated deployment delivers significant cost saving through reduced construction, operational expenses, and onsite labor needs.

Note: Modular datacenters are the most prominent example, but modular approaches extend to other areas, such as overhead busways with tap boxes, to easily accommodate growing power densities as GPU technology advances. | Source: IDC's *Datacenters: Best Practices to Mitigate the Environmental Impact of GenAI*

Microgrids

A microgrid is a self-contained power system that can operate independently or with the main power grid to meet power demands and sustainability goals.



Microgrid benefits:

Power scarcity: Microgrids can provide temporary power during utility delays and augment utility power.

Enhanced resiliency: Microgrids can provide a higher level of energy resiliency by allowing datacenters to operate independently off the main grid during outages or disruptions and spiking AI power loads.

Decarbonization/sustainability: Microgrids enable datacenters to incorporate onsite renewable energy generation, such as solar or wind power, and use battery storage to offset their carbon footprint.

Cost: By generating their own power and managing energy consumption, datacenters can reduce their reliance on the main grid, taking advantage of demand–response programs and time-of-day pricing.

Source: IDC's *Datacenters: Best Practices to Mitigate the Environmental Impact of GenAI*

Power Purchase Agreements

A power purchase agreement is a long-term contract between a datacenter operator and an energy provider that defines the sourcing and the pricing of the energy the datacenter will consume. The source is often renewable energy.

Key Terms:

- Length (usually 5–15 years)
- Volume (minimal threshold)
- Price (per kWh)
- Renewable energy attributes

Improving environmental sustainability is the number 2 priority for datacenter operators.

Power purchase agreements allow enterprises to **secure power for their massive energy consumption increase so they can support AI** without sacrificing sustainability goals.

n = 766; Source: IDC's *Datacenter Operations and Sustainability Survey*

Organizations Cite a Lack of GenAI Skills or Expertise as the Top Inhibitor to AI Adoption

28%

lack GenAI skills or expertise within the organization.

n = 889; Source: IDC's *Future Enterprise Resiliency & Spending Survey Wave 4*, April 2024

To address specific skill gaps, organizations **should partner with vendors** with proven specialized knowledge and expertise and a strong track record of delivering results.

Enterprises should opt for vendors with a comprehensive portfolio that can help them bridge skill gaps by providing access to **specialized expertise, training, end-to-end solutions, and scalable services.**

Essential Guidance

1. **Leverage CIO leadership:** CIO organizations are in a unique position at the intersection of business and technology. They can enable AI for competitive differentiation by working with the C-suite to create a strategic approach that addresses challenges across people, processes, and technology while building a framework for decision-making and technology investments.

2. **Plan workloads:** As organizations move from the AI Scramble to the AI Pivot, they will focus on their AI strategy, governance, and cost structure. This process should include AI workload forecasts, strategic decisions on where they will run, and the capital investment in ongoing operating expenses for AI workloads.

3. **Use expertise:** A significant portion of AI workloads will be on premises to support data sovereignty, flexibility, and integration with existing systems, but these on-premises datacenters will be drastically different from what organizations are familiar with. Enterprises should use third-party expertise to design, build, and operate AI datacenters to compensate for labor and skill shortages.

4. **Use modular approaches where possible:** Using modular datacenters or components (i.e., busways) will future proof datacenters against the rate of AI infrastructure change while improving maintainability, increasing reusability to maximize investments, achieving sustainability goals, and enhancing scalability and flexibility.

Appendix: Supplemental Data

The table in this appendix provides an accessible version of the data for the complex figure in this document. Click “Return to original figure” below this table to get back to the original data figure.

SUPPLEMENTAL DATA FROM PAGE 7

Worldwide Hardware Spending in $B for Selected Sectors

	2024	2025	2026	2027	2028
Financial services	$9.23B	$11.97B	$15.10B	$18.98B	$23.61B
Healthcare	$3.73B	$4.82B	$6.24B	$7.96B	$10.06B
Retail and services	$38.75B	$45.70B	$54.13B	$63.61B	$74.61B
Public sector	$4.66B	$5.81B	$7.14B	$8.73B	$10.43B

Source: IDC’s *AI Spending Guide*, November 2024

Return to original figure

Appendix: Supplemental Data (continued)

SUPPLEMENTAL DATA FROM PAGE 9

Industry Examples of High AI Rack Densities

Date	Rack Density	
22/3/2022	40	nvidia DXG SuperPod
7/8/2023	70	DRT High Density
30/8/2023	300	CyrusOne
18/3/2024	120	DGX Superpod
15/10/2024	480	Dell Integrated Rack 7000
1/1/2026	240	nvidia Rubin Superpod
1/9/2026	800	nvidia Rubin Ultra
31/12/2029	1,000	Even with modest increases, top rack density will be approaching 1MW by the end of the decade.

Source: IDC's *Future Enterprise Resiliency & Spending Survey Wave 8*, September 2024

Return to original figure

SUPPLEMENTAL DATA FROM PAGE 10

Datacenter Energy Consumption (TWh)

	Americas	EMEA	Asia/Pacific
2019	97	45	79
2020	113	49	90
2021	126	52	101
2022	142	54	113
2023	163	59	130
2024	194	68	155
2025	225	78	182
2026	270	93	220
2027	337	114	278
2028	395	132	330

n = 891; Source: IDC's *Datacenter Trends: Sustainable Builds and Carbon Emissions*

Return to original figure

About the IDC Analyst

Sean Graham
Research Director,
Cloud to Edge Datacenter Trends, IDC

Sean Graham is a research director, Cloud to Edge Datacenter Trends at IDC. He focuses on providing insights and analysis to IT infrastructure vendors, datacenter and colocation providers, cloud service providers, and datacenter services firms. Hardware areas of coverage include generators, UPS, CRAC, cabling, LAN/WAN, storage, racks, and servers. Software covered includes DCIM, building automation, artificial intelligence and machine learning, and predictive analytics. Services include datacenter design, construction, and running and operating of datacenters. All the above coverage areas will have an overarching theme of sustainability and trust. Sean draws on 25 years of industry experience to provide insights and actionable advice to assist vendors in developing, marketing, and delivering datacenters.

More about Sean Graham

Message from the Sponsor

Meet the Challenges Brought by AI

Adoption of accelerated IT and HPC architectures is growing at a fast pace to meet the demands of AI workloads.

Power consumption and heat generation from AI chips are orders of magnitude higher than conventional IT equipment. Power and cooling infrastructure designs are being transformed as traditional cooling technologies might not measure up to the needs of AI workloads.

Don't let your infrastructure slow down your AI deployments or throttle your AI workloads. Vertiv can help you rethink your critical infrastructure in the age of AI.

Vertiv's AI Hub has key insights and advice to help you learn, design, deploy and engage with AI.

Read more

IDC Custom Solutions

IDC Research, Inc.
140 Kendrick Street, Building B, Needham, MA 02494, USA
T +1 508 872 8200

idc.com @idc @idc