

A person is standing in the center of a long, narrow aisle in a server room. The aisle is flanked by rows of server racks on both sides. The floor is a light-colored grid pattern. The ceiling has several rectangular light fixtures. The overall lighting is dim, with a strong blue/cyan glow from the server racks and ceiling lights. The person is looking towards a server rack in the distance.

Present and Future HPC Trends, Forecast, and Implications Within Data Centers

Matt Vincent, Editor-in-Chief, Data Center Frontier

Hyperion Research HPC Forum | September 2025

Present and Future HPC Trends, Forecast, and Implications Within Data Centers

Matt Vincent, Editor-in-Chief, Data Center Frontier

Hyperion Research HPC Forum | September 2025

Welcome to this comprehensive analysis of where high-performance computing infrastructure is heading, how it's reshaping modern data center design, and what operators need to prepare for in the coming years. Today we'll explore the evolving landscape where traditional HPC workloads increasingly share infrastructure with AI computation, creating new challenges and opportunities for thermal management, power delivery, and facility design.



Agenda

01

Setting the Stage

Establishing context for today's HPC transformation

02

The State of HPC in the Data Center (2025)

Current deployment patterns and density profiles

03

Compute Trends: Chips, Accelerators, AI

Hardware evolution driving infrastructure changes

04

Cooling Innovation

Thermal management solutions for ultra-high density

05

Power & Energy Infrastructure

Meeting growing demand amid grid constraints

01

Regional Dynamics

Geographic trends in HPC deployment

02

AI-HPC Convergence

Architectural implications of mixed workloads

03

Modular, Edge, Sustainable HPC

Emerging deployment models and environmental factors

04

Strategic Implications

Planning considerations for operators and executives

05

Final Takeaways & Q&A

Key insights and interactive discussion

Today's presentation will methodically explore each of these critical areas, providing both strategic context and tactical implementation details. We'll examine how leading organizations are navigating these challenges and identify emerging best practices that will shape facility planning for the next generation of computational infrastructure.

What's Next for HPC in Data Centers: The Shape of Things to Come

As high-performance computing continues to evolve at an unprecedented pace, data centers face new challenges and opportunities that will fundamentally reshape their design, deployment, and operation. The convergence of traditional HPC workloads with AI is creating new demands that require forward-thinking approaches to infrastructure planning.

The next generation of HPC installations will be characterized by modularity, edge capabilities, and sustainability innovations that weren't possible just a few years ago. These developments are occurring against a backdrop of increasing collaboration between hyperscalers and national laboratories, creating a rich ecosystem for innovation and standardization.

Let's explore the key trends that will define HPC data centers through the end of this decade and examine how technical decision-makers can position their organizations for success in this rapidly evolving landscape.

HPC and AI: Now in the Same Rack

The industry is witnessing a profound shift as data centers increasingly host AI and HPC workloads side by side, creating new infrastructure challenges and opportunities.

Traditional HPC

Precision scientific modeling, research-driven applications requiring extreme computational accuracy

AI Workloads

Large datasets, dense thermal profiles, specialized accelerators, and fluctuating power demands



Combined workloads are fundamentally reshaping infrastructure planning: power delivery systems, cooling architecture, and rack density calculations must all evolve.

HPC and AI: Now in the Same Rack

Traditional HPC Characteristics

- *Precision-focused scientific modeling and simulation*
- *Highly structured data and predictable compute patterns*
- *MPI-based parallel processing across distributed nodes*
- *Research-driven with predetermined computational goals*
- *Typically CPU-heavy with specialized interconnects*
- *Steady thermal profiles with predictable power draw*

AI Workload Characteristics

- *Training on massive unstructured datasets*
- *GPU/TPU-intensive with dense accelerator configurations*
- *Bursty power consumption with extreme peak demands*
- *Hot spots reaching 500-600W per accelerator*
- *Heavy memory bandwidth requirements*
- *Often containerized with dynamic resource allocation*

i **Combined Workload Implications:** *As these disparate computational paradigms converge in shared infrastructure, data center operators must rethink fundamental aspects of facility design. Power distribution systems must handle both steady HPC loads and bursty AI training peaks. Cooling strategies must address heterogeneous thermal profiles. Network fabric must support both low-latency MPI traffic and high-bandwidth data movement for AI.*

This convergence represents perhaps the most significant shift in computational infrastructure since the advent of cloud computing, forcing operators to develop more flexible, resilient facility designs capable of supporting widely varying workload characteristics within the same physical environment.

The State of HPC in the Data Center (2025)

70kW+

Rack Density

Common in modern HPC deployments, pushing cooling and power infrastructure to new limits

2.5EF

Exascale Computing

Hyperscalers and national labs pushing beyond initial exascale achievements

Market Segmentation

Academic Clusters

- *Research-focused*
- *Grant/public funded*
- *Open science emphasis*

Enterprise AI-HPC

- *Revenue-generating*
- *Proprietary workflows*
- *Time-to-insight critical*

Notable examples: NERSC Perlmutter, Meta AI Research Cluster (RSC), Azure Quantum

The State of HPC in the Data Center (2025)

Extreme Rack Densities Now Standard

Rack densities exceeding 70kW are now commonplace in new HPC deployments, with cutting-edge facilities regularly supporting 100-120kW per rack. This represents a 3-4x increase from typical 2020 deployments and has driven wholesale redesign of power distribution and cooling systems.

Exascale Pushes Beyond National Labs

While DOE facilities pioneered exascale computing, hyperscale providers now operate multiple exascale-class systems for internal AI research. These systems blend traditional HPC architectures with massive GPU/TPU acceleration, creating hybrid infrastructure that supports both simulation and deep learning workloads.

Academic vs. Enterprise Deployment Patterns

Academic clusters remain focused on multi-tenant scientific computing with diverse workloads, while enterprise AI-HPC deployments increasingly optimize for specific production ML pipelines, often with custom silicon and specialized cooling solutions tuned for particular workload characteristics.

Notable 2025 Deployments



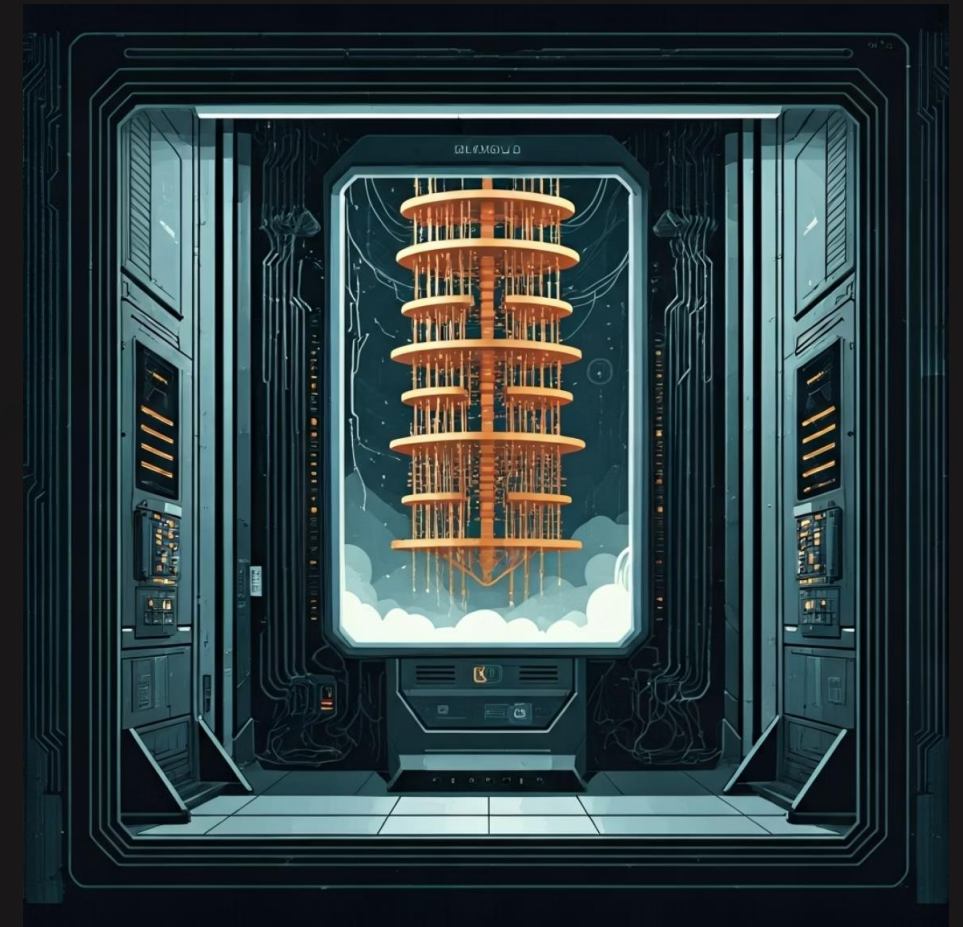
NERSC Perlmutter+

The upgraded Perlmutter system combines traditional CPU nodes with the latest NVIDIA H200 accelerators, delivering 10.5 exaflops of mixed-precision performance while maintaining compatibility with legacy scientific codes.



Meta AI Research SuperCluster

Meta's fifth-generation AI cluster integrates 35,000 NVIDIA GPUs with custom network fabric, consuming 32MW of power and requiring integrated two-phase immersion cooling throughout the facility.



Azure Quantum

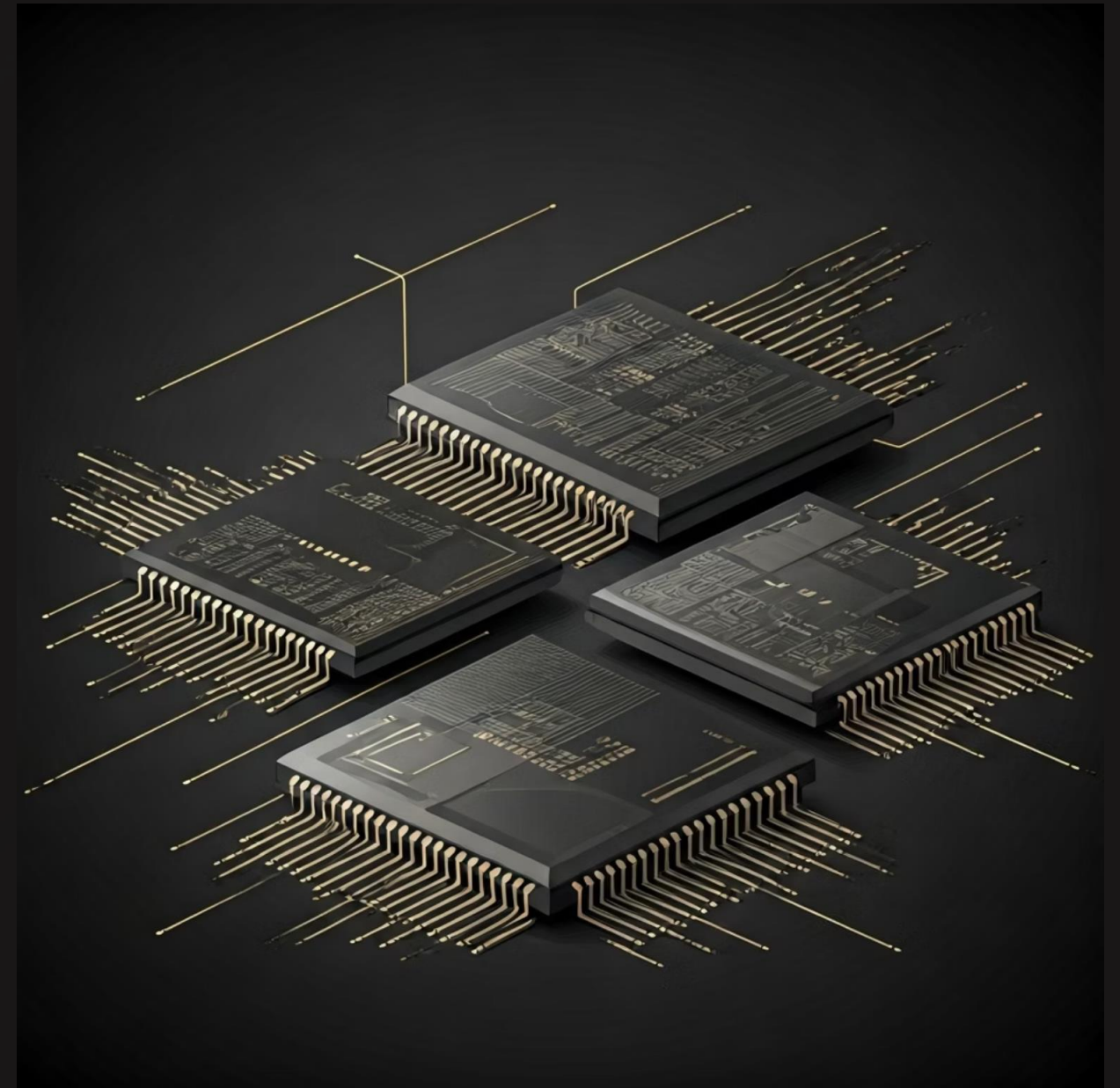
Microsoft's hybrid quantum-classical computing center pairs traditional HPC infrastructure with 128 superconducting qubit processors, establishing a new paradigm for integrated quantum-accelerated simulation.

Compute Trends: Chips, Accelerators & AI

Architectural Evolution

- *Multi-die and chiplet architectures enabling unprecedented integration*
- *Specialized accelerators dominating HPC/AI compute landscape*
- *Memory bandwidth bottlenecks driving new interconnect technologies*
- *Custom silicon adoption accelerating among hyperscalers*

i **Forecast:** By 2028, rack composition will be 60% accelerators, 25% CPUs, 15% specialized processors



Market Leaders (2025):

- *Nvidia H200 GPUs*
- *AMD MI300X APUs*

Compute Trends: Chips, Accelerators & AI



Multi-die & Chiplet Architectures

The transition to disaggregated silicon with specialized dies for compute, memory, I/O, and AI acceleration has enabled unprecedented performance scaling while working within thermal constraints. Current-generation HPC processors commonly integrate 4-8 compute chiplets with dedicated interconnect dies and HBM stacks.



Accelerator Evolution

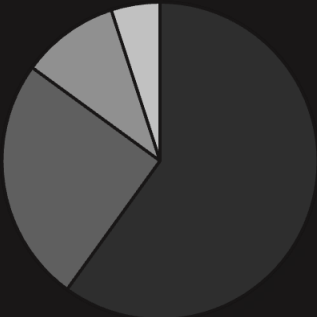
NVIDIA's H200 (successor to the H100) now delivers 30 petaflops of FP8 performance per rack, while AMD's MI300X offers competitive performance with integrated HBM. Intel's Gaudi 3 has gained market share in power-constrained environments, while Cerebras WSE-3 wafer-scale systems occupy a niche for specialized workloads.



Rise of Custom Silicon

Following the lead of hyperscalers, enterprise and national lab customers are increasingly deploying workload-specific ASICs designed for particular algorithms. These purpose-built chips offer 5-15x better performance per watt than general-purpose GPUs for specific workloads, dramatically improving data center efficiency.

2028 Rack Composition Forecast

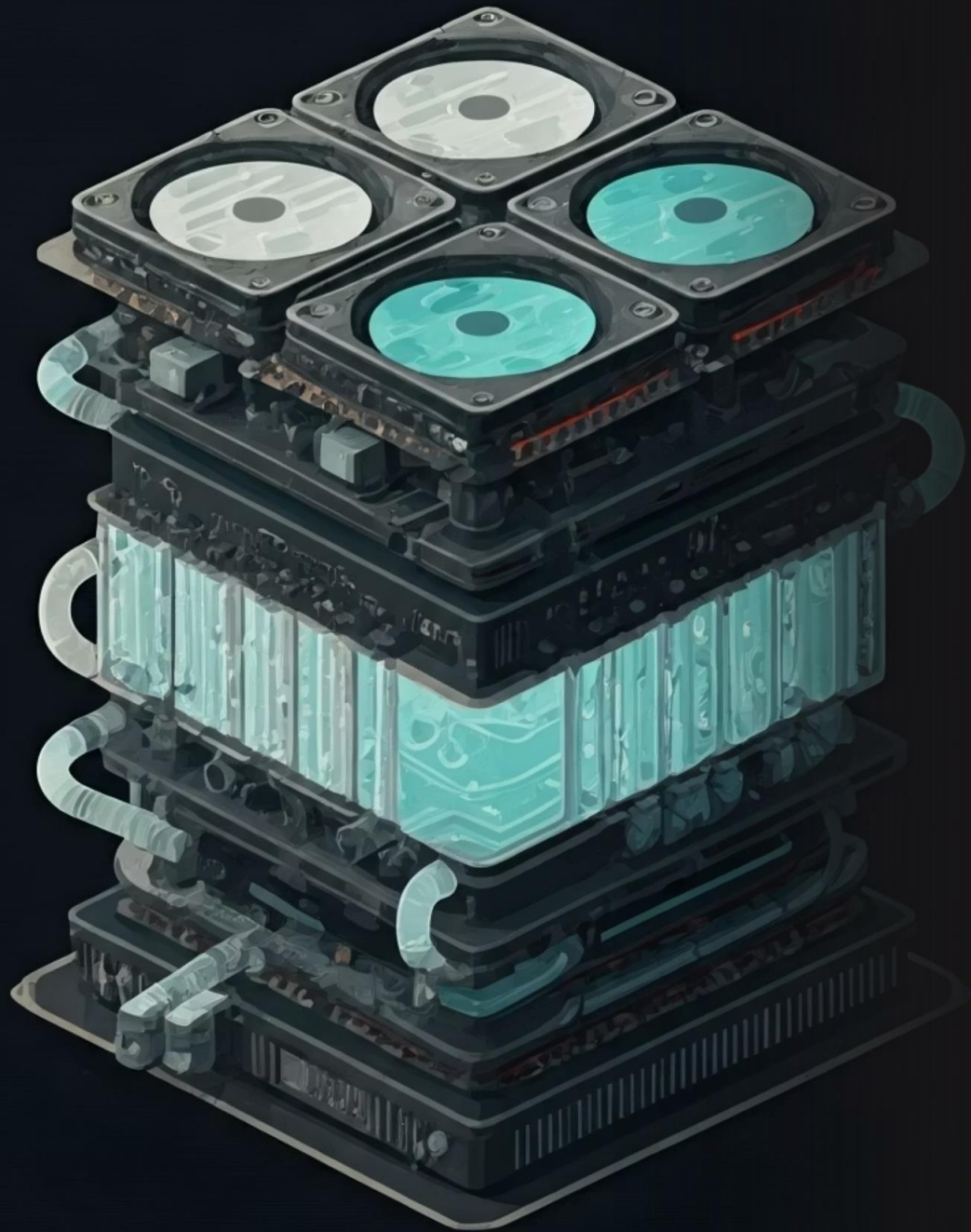


- GPU/TPU Accelerators
- CPU Compute
- Custom ASICs
- FPGA/Reconfigurable

Based on current deployment trends and vendor roadmaps, accelerators will dominate rack composition by 2028, with CPUs providing orchestration and handling portions of applications not suited to parallel processing. Custom silicon continues to grow in importance as workloads become more specialized.

Infrastructure Impact: This shift toward accelerator-dominated computing fundamentally changes data center power and cooling requirements. Accelerators draw power in concentrated areas, creating intense thermal hotspots that traditional air cooling struggles to address. Power delivery systems must handle more variability and higher peak loads.

Cooling Innovations & Thermal Strategy



Direct-to-Chip Liquid Cooling

Now mainstream for racks exceeding 50kW, with simplified manifold designs reducing deployment complexity



Single-Phase Immersion

Gaining traction for ultra-dense AI clusters, with standardized tank designs emerging



Rear-Door Heat Exchangers

Popular retrofit solution, bridging legacy infrastructure to higher densities up to 45kW

Water scarcity and emissions regulations increasingly influence cooling strategy selection, with closed-loop systems and heat reuse becoming competitive advantages.

Cooling Innovations & Thermal Strategy



Liquid Cooling Now Mainstream

Direct-to-chip (cold plate) cooling has become the default for new HPC deployments, with 85% of high-density racks (>50kW) now utilizing some form of liquid cooling. Single-phase solutions dominate due to operational simplicity, while two-phase immersion remains reserved for the most demanding applications exceeding 100kW per rack.



Retrofit Solutions

Rear-door heat exchangers (RDHx) have emerged as the preferred solution for retrofitting existing facilities, allowing gradual transition to higher densities without wholesale infrastructure replacement. Modern RDHx units can now handle up to 75kW per rack, though with less efficiency than direct liquid cooling approaches.



Thermal Density Driving Design

Component-level thermal density now exceeds 500W per GPU in latest-generation accelerators, forcing radical rethinking of chassis and rack design. Cold plates now integrate directly with GPU packages, and some vendors have moved to integrated cooling manifolds built into server backplanes.

Regulatory and Sustainability Pressures

Water Usage Regulations

New water usage restrictions in drought-prone regions have accelerated adoption of closed-loop cooling systems with dry coolers or forced-air radiators rather than evaporative cooling towers. While less energy-efficient, these approaches reduce water consumption by up to 95% compared to traditional cooling tower implementations.

Emissions Requirements

Carbon emissions regulations in Europe and increasingly in North America have shifted cooling system design toward heat recovery and reuse. Modern HPC facilities now commonly incorporate heat export capabilities, with waste heat directed to district heating systems, adjacent office spaces, or industrial processes where feasible.

The cooling paradigm has fundamentally shifted from "how do we remove heat" to "how do we efficiently capture and potentially reuse heat" while minimizing environmental impact. This represents both a technical and regulatory challenge for operators designing facilities with 20+ year lifespans.

Power is the New Battleground

Dense & Bursty Loads

AI training creates power demand spikes of 30-40%, requiring robust delivery systems

New Metrics

MW per petaflop and ERE (Energy Reuse Effectiveness) replacing PUE as key metrics



Grid Constraints

Utility capacity limitations delaying major HPC builds by 18-36 months in key markets

On-site Generation

Gas turbines, fuel cells, and micro-nuclear solutions gaining traction

 *Power availability, not land or capital, is now the primary constraint on HPC/AI infrastructure growth.*

Power is the New Battleground

Dense, Predictable, Bursty Loads

Modern HPC and AI workloads create unique power challenges, combining high baseline utilization (80%+) with extreme transient spikes during training phases. Power systems must handle steady 70-80kW per rack with headroom for 120-140kW bursts without triggering overload protection.

Grid Constraints Delaying Builds

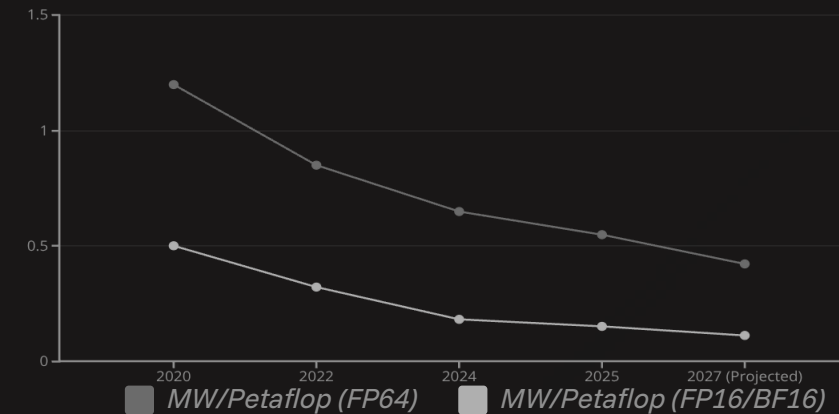
Power availability has become the primary limiting factor for new HPC deployments, with 55% of planned projects now facing delays of 18+ months due to grid capacity constraints. This has created a competitive landscape where power allocation, rather than physical space, determines data center growth potential.

On-site Generation Rising

Faced with grid limitations, 38% of new HPC facilities now incorporate significant on-site generation. Natural gas turbines remain most common, but small modular nuclear reactors (SMRs) have moved from theoretical to practical, with three facilities now powered by on-site nuclear generation in the 20-60MW range.

i **Energy-Reuse Effectiveness (ERE) Replacing PUE:** The industry is pivoting from Power Usage Effectiveness (PUE) to Energy-Reuse Effectiveness (ERE) as the primary efficiency metric. ERE accounts for heat recovery and beneficial reuse, providing a more holistic view of environmental impact. Best-in-class facilities now achieve ERE ratings below 0.8 through integrated heat recovery systems.

MW per Petaflop Trends



While overall power efficiency continues to improve, the accelerating demand for computational capacity means absolute power requirements continue to grow. Mixed-precision workloads offer dramatically better power efficiency, driving the AI-HPC convergence trend.

Regional Dynamics

Key U.S. Regions

- *Northern Virginia: Largest concentration but power-constrained*
- *Hillsboro, OR: Renewable focus with hydropower advantages*
- *Dallas-Fort Worth: Emerging AI hub with ERCOT challenges*
- *Quebec: Hydro resources attracting Canadian expansion*
- *Chicago: Financial HPC cluster with nuclear power access*

Global Growth Centers

- *Japan: Advanced cooling tech leadership*
- *Germany: Research-led HPC expansion*
- *UAE: Massive sovereign investment in AI infrastructure*
- *Singapore: Sustainable tropical HPC innovations*
- *China: Domestic accelerator ecosystem developing*

Policy factors like the CHIPS Act, DOE lab investments, and state-level tax credits are significantly reshaping the competitive landscape for HPC site selection.

Regional Dynamics

Key North American Regions

Northern Virginia

Still dominant with 35% of North American HPC capacity, but facing severe power constraints. New builds require 24-30 month lead times for grid connections exceeding 20MW. Land prices approaching \$3M per acre in prime locations.

Hillsboro, Oregon

Emerged as the premier location for hyperscale HPC due to favorable climate, renewable energy availability, and tax incentives. Now hosts 22% of North American AI compute capacity.

Dallas-Fort Worth

Benefiting from grid independence and favorable regulatory environment. Home to growing concentration of financial services HPC and regional AI training facilities.

Quebec & Chicago

Quebec leveraging abundant hydropower and cold climate for efficient operations. Chicago capitalizing on robust fiber infrastructure and central location despite higher energy costs.

Global Growth Centers

Japan: Computational Renaissance

Japan has emerged as a global leader in efficient HPC, with next-generation facilities in Kobe and Tsukuba achieving world-leading power efficiency while supporting the country's semiconductor resurgence.

Germany: Industrial AI Hub

Frankfurt and Munich regions now host 65% of European industrial AI/HPC capacity, with heavy focus on manufacturing optimization and automotive simulation.

UAE & Singapore: Strategic Investment

Both regions making massive sovereign investments in HPC infrastructure, positioning themselves as regional AI hubs with state-backed compute resources.

China: Self-Sufficient Ecosystem

Despite continued technology restrictions, China has established a parallel HPC ecosystem with domestic technologies, focused heavily on indigenous accelerator development.

Policy & Incentive Landscape

The CHIPS Act has dramatically reshaped North American HPC deployment, with \$11B in direct funding for DOE laboratory systems and another \$8.5B in tax incentives for commercial HPC facilities supporting semiconductor research and manufacturing. State-level incentives increasingly focus on power infrastructure rather than direct tax benefits, with Virginia, Texas, and Illinois establishing dedicated grid capacity allocation programs for HPC/AI facilities.

AI and HPC Convergence



Shared Infrastructure

Common cooling plants, power systems, and high-bandwidth interconnects serving both workload types



Specialized Compute

HPC remains simulation-heavy while AI focuses on data throughput, yet hardware requirements increasingly overlap



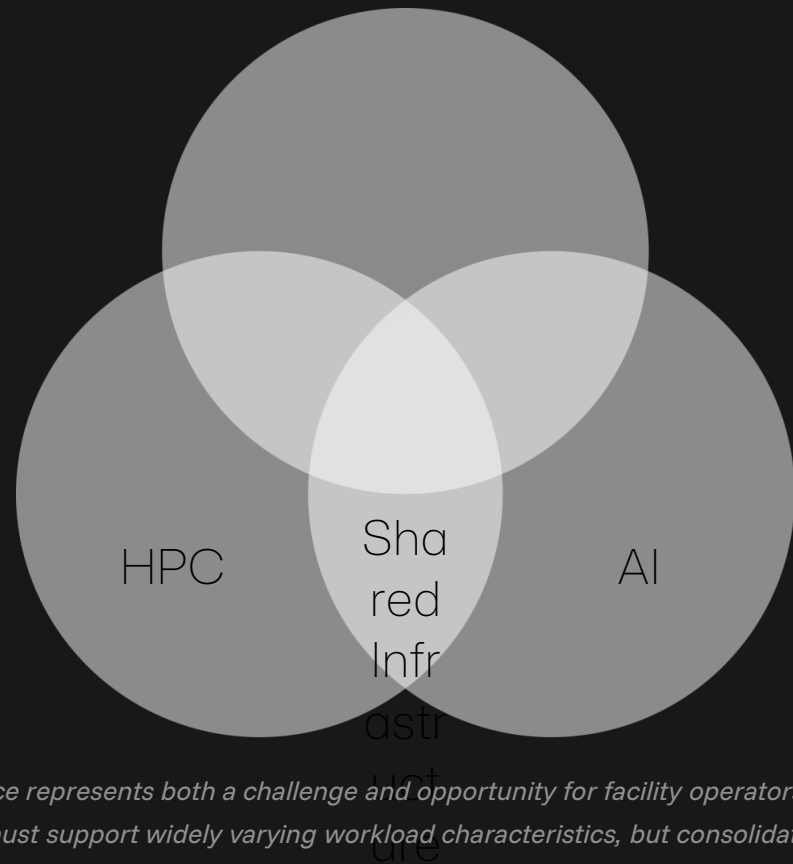
Unified Orchestration

Emerging platforms manage resources across both domains, optimizing utilization and performance

Forward-thinking organizations are now designing infrastructure explicitly for this overlap, enabling more flexible resource allocation and improved asset utilization.



AI and HPC Convergence



This convergence represents both a challenge and opportunity for facility operators. Shared infrastructure must support widely varying workload characteristics, but consolidation creates efficiency opportunities through higher utilization and better resource management.

Technical Integration Challenges

Workload Orchestration

Resource schedulers must balance traditional MPI-based HPC jobs with containerized AI workloads, each with different scaling characteristics and priority models. Leading-edge facilities now implement AI-powered workload schedulers that predict resource needs and optimize placement based on thermal and power conditions.

Infrastructure Flexibility

Physical infrastructure must adapt to changing workload mixes. Modern designs incorporate reconfigurable cooling (adjustable between air and liquid modes) and flexible power distribution that can shift capacity between rack rows based on utilization patterns.

Converged Systems

Hardware vendors now offer "AI-HPC hybrid" systems explicitly designed for mixed workloads, featuring modular node configurations where CPU-only, GPU-accelerated, and storage nodes share common power, cooling, and interconnect infrastructure within a single rack architecture.

Case Study: Los Alamos National Laboratory's Trinity-Next System

The Trinity-Next system at LANL exemplifies this convergence trend, combining traditional HPC capabilities with massive AI acceleration in a unified architecture. The system incorporates 15,000 traditional CPU nodes for simulation alongside 2,500 GPU-accelerated nodes for ML training and inference. A unified RDMA-capable network fabric allows simulation results to flow directly into AI models and vice versa, enabling new "simulation-informed ML" workflows that dramatically accelerate materials science research.

The facility infrastructure was designed specifically for this hybrid approach, with a cooling plant that can dynamically allocate cooling capacity between CPU and GPU sections based on workload characteristics, and a power distribution system that accommodates both the steady draw of simulation and the bursty load of AI training.



What's Next for HPC in Data Centers

As we look toward the future of High Performance Computing, several transformative trends are reshaping how we design, deploy, and optimize data center infrastructure. The landscape is evolving rapidly, with new approaches to modularity, edge computing, and sustainability becoming central to strategic planning.

What's Next for HPC in Data Centers

Modular HPC

Pre-engineered clusters with standardized power, cooling, and compute enabling rapid deployment

Edge HPC

Distributed computing nodes bringing simulation capabilities closer to data sources

Sustainable Design

Liquid cooling heat reuse, carbon-aware workload scheduling, and renewable integration



What's Next for HPC in Data Centers



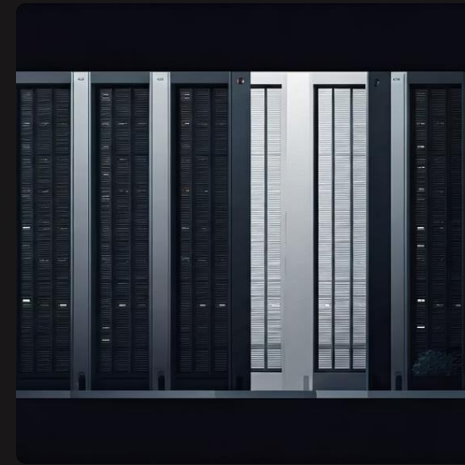
Modular HPC Clusters

Factory-built, fully integrated HPC modules with capacities from 250kW to 5MW are rapidly gaining market share, reducing deployment time by 60% compared to traditional construction. These units integrate power, cooling, and compute in weatherproof enclosures that can be deployed in non-traditional locations, opening new possibilities for distributed HPC/AI capacity.



Edge HPC Deployment

The growing need for low-latency AI inference and real-time simulation is driving HPC capabilities toward the network edge. Compact, ruggedized HPC systems capable of 100+ teraflops in under 20kW are now being deployed at manufacturing sites, healthcare facilities, and transportation hubs to enable real-time processing without cloud dependency.



Sustainable Infrastructure Design

Environmental considerations now drive HPC facility design from inception rather than as an afterthought. Leading-edge facilities implement closed-loop liquid cooling with zero water consumption, waste heat recovery for adjacent uses, and carbon-aware workload scheduling that shifts non-time-critical computation to periods of renewable energy availability.



Hyperscaler-Lab Collaboration

The lines between academic/government HPC and commercial AI infrastructure continue to blur, with joint operating models emerging. These partnerships leverage public funding for basic research while accessing commercial expertise in facility operation, creating shared infrastructure that advances both scientific computing and commercial AI development.

Final Thoughts: Strategic Planning for the Next Decade

As we look toward 2030 and beyond, HPC facility planning must incorporate unprecedented flexibility and sustainability. The most successful organizations will build infrastructure that can adapt to rapidly evolving compute paradigms while addressing growing energy and environmental constraints. This requires a fundamental rethinking of the traditional data center model, moving toward distributed, highly efficient compute resources that can be dynamically reconfigured to support diverse and changing workloads.

The convergence of HPC and AI represents not just a technical challenge but a strategic opportunity to reimagine computational infrastructure for the coming decades. Those who successfully navigate this transition will establish competitive advantages in both scientific discovery and commercial AI development.



Strategic Implications for Builders & Planners

Thermal Planning

Thermal strategy must be integrated from the earliest site selection phase, not as an afterthought. Next-gen densities demand comprehensive cooling solutions designed into the facility fabric.

Power Opportunity Mapping

Locations with stranded or underutilized power capacity represent strategic assets in an increasingly power-constrained world. Map these opportunities in your portfolio planning.

Talent Pipeline Development

Critical shortages exist in specialized roles: cooling technicians, HPC-focused electricians, and orchestration engineers. Begin targeted recruitment and training programs now.

Capital Planning Alignment

CapEx cycles must synchronize with silicon innovation timelines. Misalignment creates infrastructure that's outdated before it's operational.

Strategic Implications for Builders & Planners



The shift toward higher density, liquid-cooled HPC infrastructure necessitates a complete rethinking of traditional data center planning approaches. Organizations that fail to adapt these strategic considerations risk costly retrofits or, worse, infrastructure that cannot support next-generation computing demands.

Organizations that proactively address these strategic implications will be positioned to deploy, scale, and optimize HPC resources that support both traditional simulation workloads and emerging AI applications. Those that delay may find themselves constrained by infrastructure limitations precisely when computational capabilities are becoming the defining competitive advantage in their industries.

Thermal Planning Revolution

Thermal considerations must now be incorporated at the earliest stages of site selection and building design. With rack densities regularly exceeding 100kW, traditional raised-floor cooling is no longer sufficient. Direct-to-chip liquid cooling requires dedicated infrastructure planning, including:

- Primary and secondary cooling loops with N+1 redundancy
- Heat reuse systems for campus heating or adjacent facilities
- Temperature monitoring at unprecedented granularity

Power Opportunity Identification

Locations with stranded or underutilized power capacity represent strategic assets in the HPC landscape. Organizations should:

- Identify decommissioned industrial sites with robust electrical infrastructure
- Explore co-location with renewable energy generation
- Develop partnerships with utilities for preferential capacity allocation

Talent Pipeline Development

The specialized skills required for next-generation HPC environments extend beyond traditional IT roles:

- Liquid cooling technicians with both IT and HVAC expertise
- High-voltage electricians familiar with data center environments
- Orchestration specialists capable of workload optimization across heterogeneous systems

Capital Planning Alignment

CapEx cycles must be synchronized with silicon development roadmaps to maximize investment efficiency:

- Infrastructure planning horizon: 10-15 years
- Building systems: 7-10 years
- Compute refresh cycles: 3-4 years
- Accelerator upgrades: 18-24 months

Key Takeaways

Data Centers as Scientific Engines

The modern data center has evolved beyond hosting to become the foundational infrastructure powering scientific discovery.

Design Principles

Future-proof facilities must prioritize density, data locality, and adaptable infrastructure that can evolve with technology.

Convergent Evolution

AI and HPC workloads are following increasingly convergent paths, creating shared infrastructure needs despite distinct operational pressures.



Is your infrastructure ready for 2028?

"Exascale is here. AI scale is next. Will your data center keep up?"

Key Takeaways

The Data Center as Scientific Engine



Modern scientific discovery is increasingly defined by computational capability. From drug discovery to materials science, the data center has become the primary laboratory for breakthrough innovation. Organizations must recognize infrastructure as a scientific instrument, not merely IT equipment.

AI + HPC = Shared Destiny



While AI and traditional HPC workloads have different characteristics, their infrastructure requirements are converging. Both demand high memory bandwidth, accelerator optimization, and efficient data movement. Yet they maintain distinct pressure points around precision requirements, I/O patterns, and scaling approaches.

Build for Density, Locality, and Flexibility



Tomorrow's successful HPC environments will support unprecedented density (250kW+ per rack), minimize data movement through careful topology design, and provide the flexibility to adapt as computational paradigms continue to evolve.

"Exascale is here. AI scale is next. Will your data center keep up?"



⚠️ Is Your Infrastructure Ready for 2028?

The infrastructure decisions made today will determine computational capabilities for the next decade. Critical questions to assess readiness:

- *Can your facility support liquid cooling at scale?*
- *Is your power infrastructure sufficient for 2-3x density increases?*