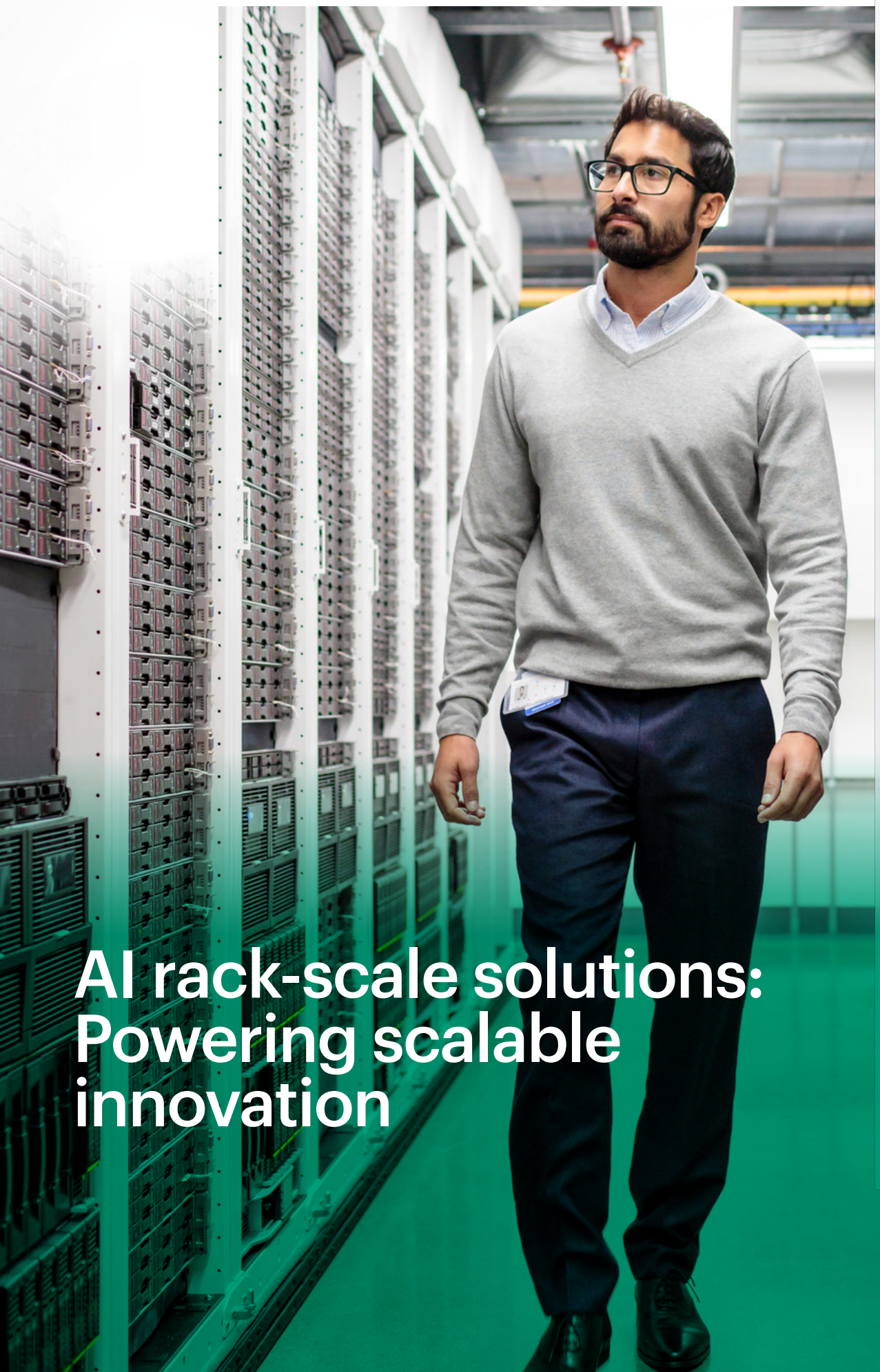
Brochure



AI rack-scale solutions: Powering scalable innovation

The challenges of scaling AI workloads

Artificial intelligence (AI) is transforming industries, unlocking innovation, and redefining what's possible. However, scaling AI presents immense challenges. As global demand for AI workloads surges, organizations face mounting pressure to deliver more compute power, manage workloads efficiently, and comply with ever-changing regulations—all while controlling costs and energy consumption.

AI workloads demand significantly higher compute power than traditional IT applications, and energy consumption is soaring. At the same time, compliance and governance also remain critical hurdles, especially in regulated industries.

HPE recognizes these challenges. That's why we offer a comprehensive AI rack-scale portfolio designed to address the demands of scaling AI. With HPE, organizations can scale responsibly, efficiently and securely—without compromising on innovation.

70%

By 2030, global data center capacity demand is expected to nearly triple, with AI workloads accounting for 70% of the growth.[1]

Significantly higher power

AI workloads require significantly higher computational power.[2]

2x or 3x

Energy demand is expected to double or triple by 2028, driven by AI growth.[3]

54%

of organizations prioritize compliance in scaling AI, but 38% cite governance as the biggest blocker.[4]

[1] McKinsey, The Cost of Compute: A $7 Trillion Race to Scale Data Centers, April 2025.

[2] McKinsey, AI power: Expanding data center capacity to meet growing demand, October 2024.

[3] Department of Energy, DOE Releases New Report Evaluating Increase in Electricity Demand from Data Centers, December 20, 2024.

[4] CIO, New study shows IT is primed to lead AI innovation, September 2025.

AI rack-scale solutions

Scaling AI effectively requires more than adding compute power—it demands a strategic approach that balances three critical pillars: performance, efficiency, security and compliance. These areas are interdependent—as a shortfall in one can undermine progress in the others.

- **Performance** fuels AI innovation, but unchecked growth can lead to unsustainable costs and energy demands.
- **Efficiency** optimizes resources, yet focusing too narrowly on efficiency can limit scalability.
- **Security and compliance** protect data, but overly rigid controls can slow deployment and innovation.

A successful solution balances these areas to drive progress. The HPE AI rack-scale solution portfolio delivers the power, efficiency, and security needed to scale AI responsibly.

Performance

High-performance CPUs, GPUs, networking and storage in modular rack-scale architectures for rapid scaling

Efficiency

Space and energy optimized with direct liquid cooling options for efficient resource utilization

Security and compliance

On-premises, air-gapped solutions and compliance-ready architecture with visibility and control

Scaling with confidence: Rack-scale solutions from HPE

The AI rack-scale portfolio from HPE enables organizations to scale AI responsibly, efficiently, and securely.

By integrating advanced technologies in servers, networking, storage and software, these solutions deliver industry-leading performance—essential for driving innovation.

With a focus on space and energy efficiency, these high-density modular rack-scale solutions come with options for air or liquid cooling, monitoring and unified management software.

From air-gapped on-premises, to multi-tenancy or multicloud, these AI rack scale solutions are available with the security and control required for compliance.

Powering intensive AI training and real-time inferencing at scale, these solutions are optimized for every stage of the AI lifecycle for maximum impact.

HPE's AI rack-scale solutions portfolio

- AMD "Helios" AI Rack
- Sovereign AI factory
- AI factory at scale
- NVIDIA® GB200 NVL72 by HPE
- NVIDIA GB300 NVL72 by HPE
- NVIDIA GB200 NVL4 by HPE

The AI rack-scale portfolio from HPE is built on six key offerings, each addressing critical challenges in scaling AI. Together, they provide the performance, efficiency, and security needed to support advanced AI workloads.

AMD "Helios" AI Rack: High-performance compute

An open, efficient and scalable foundation designed for AI workloads

- Open Compute Project (OCP) wide rack (ORW) modular design supports interoperable expansion
- Integrated liquid cooling enhances energy efficiency with high performance
- Powered by AMD Instinct™ MI450 Series GPUs, AMD EPYC™ CPUs, and AMD Pensando™ DPUs, optimized for training and inference at scale

Sovereign AI factory: Secure, compliant AI

A secure, compliant AI environment tailored for government and regulated industries

- Air-gapped infrastructure for data sovereignty and privacy
- Compliant with global regulations like GDPR and HIPAA
- Multi-tenancy visibility and control isolates environments for secure collaboration

AI factory at scale: Centralized operations

A centralized platform for managing AI workflows at scale with unified management, governance and cost optimization

- Modular, composable solutions optimized for large-scale operations
- Includes monitoring tools, which provide real-time visibility into the quality and performance of AI models to prevent performance decline and help ensure governance
- Provides multi-tenancy and lifecycle automation for streamlined operations

NVIDIA GB200 NVL72: Extreme model performance

A high-performance solution for large-scale AI training and inference

- Features NVIDIA Blackwell GPUs optimized for massive AI models
- Liquid cooling ensures maximum efficiency under heavy workloads
- Scalable design supports expanding compute capacity as AI demands grow

NVIDIA GB300 NVL72: Next-generation AI compute

Next-generation AI performance with enhanced compute density and scalability

- Advanced NVIDIA Blackwell Ultra GPUs with improved compute density for demanding workloads
- Liquid cooling system reduces energy consumption while maintaining performance
- Designed for future-proof scalability and compatibility with evolving AI frameworks

NVIDIA GB200 NVL4: Compact, agile AI power

A high-density compute solution for AI and HPC workloads

- Up to 34 1U compute trays in a 48U MGX rack, supporting 136 GPUs per rack
- Direct liquid cooling increases efficiency and reduces operational costs
- Integrated HPE Performance Cluster Manager (HPCM) simplifies deployment and management

The **NVIDIA GB200 NVL4** by **HPE** delivers up to **136 GPUs** per rack, achieving more than **2x** the density of 8-GPU servers.[5]

[5] NVIDIA GB200 NVL4 delivers up to 136 GPUs per rack, 8-GPU servers deliver 64 GPUs per rack at best

Scalable AI solutions for growing demands

Designed to address the most critical challenges in AI adoption, rack-scale solutions from HPE deliver tangible benefits in performance, efficiency and security.

Unmatched performance and scalability

Engineered for demanding AI workloads, solutions from HPE feature advanced GPUs, innovative liquid cooling, and high-density compute options. These technologies speed scaling with efficiency while providing the ability to accelerate complex AI models extracting insights from vast data sets.

Energy efficiency and sustainability

By integrating cutting-edge cooling and optimizing resource allocation, the AI rack-scale solutions portfolio from HPE minimizes energy consumption while maintaining top-tier performance. This enables organizations to use less energy without sacrificing the computational power required for large-scale AI model training and inference.

Centralized management and automation

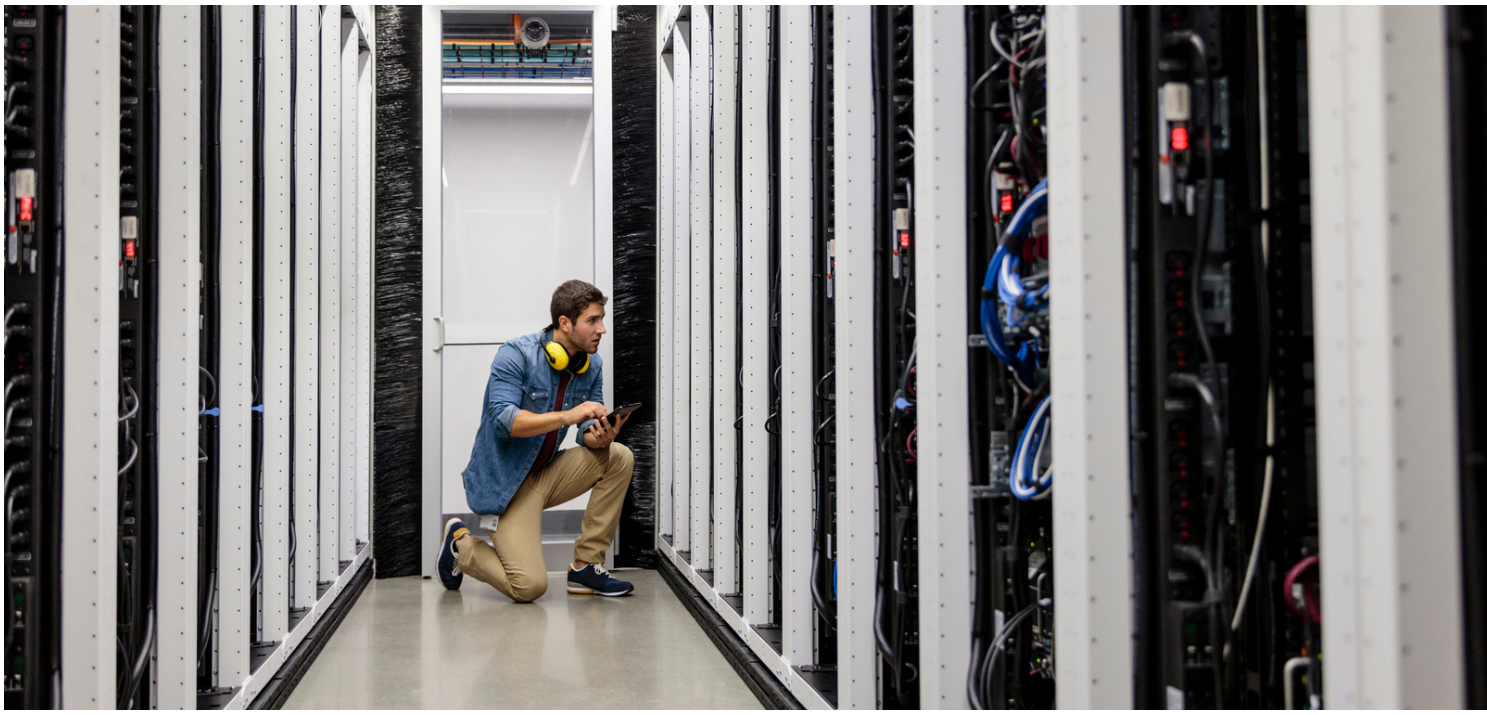
Multi-tenant platforms from HPE offer dynamic resource allocation, real-time observability, and automated workflows. These features simplify AI operations, reduce management overhead, and enable a more agile and efficient IT environment.

Robust security and compliance

With data sovereignty and compliance built into their core, solutions from HPE meet strict security standards. The portfolio provides a secure foundation for mission-critical workloads, making it well-suited for regulated industries and government entities.

Future-ready design

Built to adapt to rapid AI innovation, the portfolio from HPE supports a wide range of workloads, from training massive generative models to running complex scientific simulations. This delivers long-term value while providing a competitive edge.



What sets HPE apart

World-class services

HPE delivers comprehensive services that guide organizations through every stage of their AI transformation. This includes expert consulting, deployment and integration, dedicated support, and flexible consumption options, all tailored to de-risk complex projects and accelerate outcomes.

Edge-to-cloud expertise

By leveraging the power of HPE GreenLake and the proven Edge-to-Cloud Adoption Framework, HPE enables organizations to modernize their IT infrastructure. This holistic approach optimizes hybrid cloud environments and scales AI operations with exceptional agility and cost efficiency.

Financing and asset management services

HPE Financial Services offers a comprehensive portfolio of flexible financing solutions designed to help organizations with predictable and easy monthly payments, as well as IT asset lifecycle services that help organizations transform and repurpose IT resources into new IT capabilities and innovations.

Global network of customer innovation centers

The worldwide network from HPE, including the new AI Factory Lab in Grenoble, provides an immersive environment for collaboration and discovery. These centers enable customers to test solutions, engage with leading experts, and explore cutting-edge technologies in real-world scenarios—transforming bold ambitions into reality.

Scaling AI with confidence

Scaling AI is no longer just a technical challenge—it's a strategic imperative. AI rack-scale solutions from HPE are designed to meet the demands of today's most ambitious workloads, delivering the compute power, efficiency, and security organizations need to grow AI responsibly. By simplifying complexity and optimizing performance, HPE enables businesses to overcome scalability barriers and unlock new opportunities. With decades of expertise and a focus on real-world results, HPE offers the tools and expertise to help you scale AI effectively

Take the next step

Visit HPE.com/ai/rackscale to explore how our AI rack-scale solutions can transform your operations. Connect with our experts to discuss your unique challenges, schedule a consultation, or experience these solutions firsthand at an HPE Customer Innovation Center. Let's scale AI together.

Visit HPE.com

Learn more at

HPE.com/ai/rackscale

Chat now

a00155510enw

HEWLETT PACKARD ENTERPRISE

hpe.com