DELL Technologies

Rethinking AI Deployment with Dell Integrated Rack Scalable Systems

Contents

AI Deployment Challenges

The era of AI-driven computing has arrived, and it's transforming the way we design and deploy data centers. The explosive demand for AI requires extraordinary speed to market and immediate value, driving a need for simple and fast delivery and support of AI systems. AI infrastructure deployments, however, present unique challenges and complexities that many organizations struggle to address:

As AI workloads grow, so do power consumption and TDP, making liquid cooling and power management no longer optional, but essential.

With component sizes increasing, space constraints within servers and data centers require infrastructure that can provide more compute and acceleration in the same space. The quickly evolving components demand higher density infrastructure that is future-proof and modular.

With GPU lead times uncertain, the need for diverse silicon options is more pressing than ever.

The challenges of engineering, integrating, configuring, deploying, and installing infrastructure are exacerbated by the complexity of AI architectures, requiring more resources.

New technologies are ordered before they're available. Longer lead times mean infrastructure must be ready to produce value upon arrival. The time to deployment is longer for AI infrastructure.

AI skills and talent strategy are a major challenge for adopting AI.

AI infrastructure can be expensive, with high acquisition costs of advanced technologies, energy consumption, and heat generation, and a more complex deployment and design than typical infrastructure solutions.

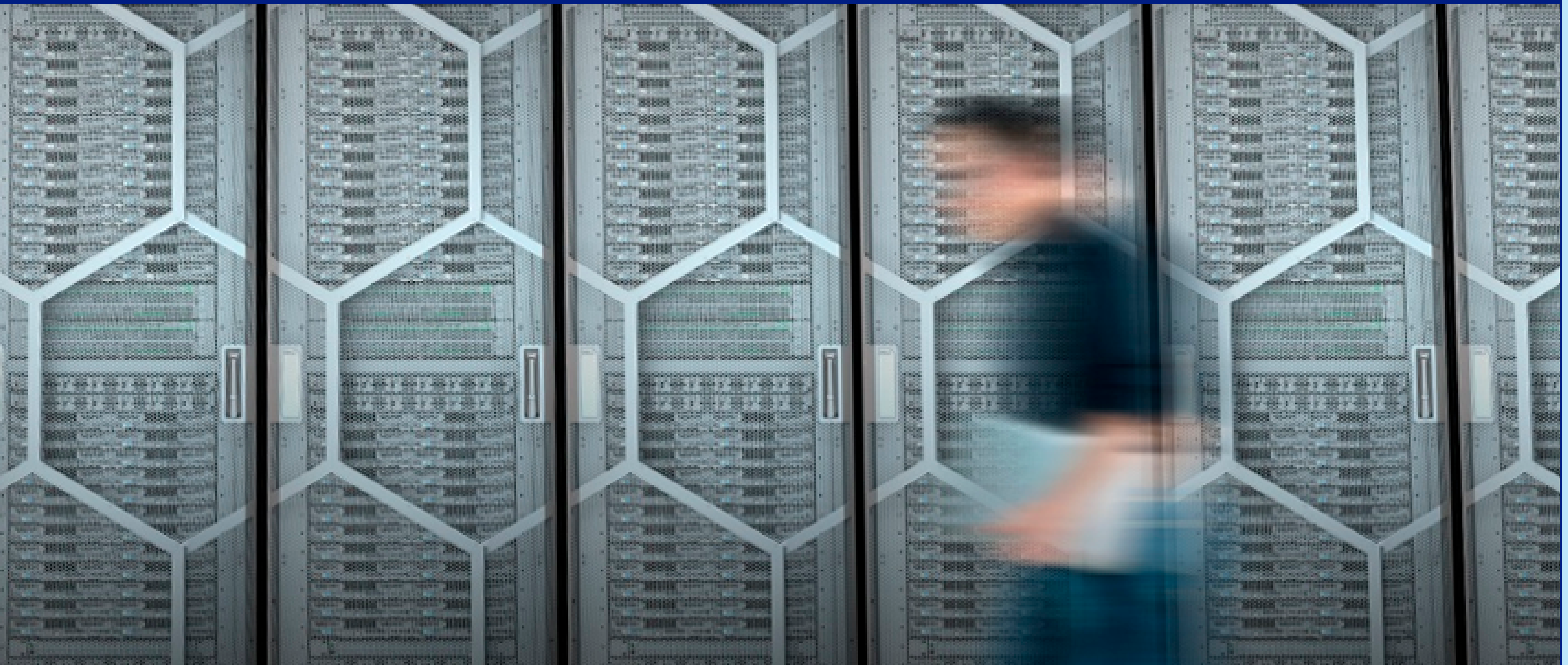
The timeline for deployment has changed. Organizations must overcome these obstacles to deploy AI infrastructure and at the same time, it must be possible to scale quickly and cost-effectively. The answer lies in implementing services that provide fully integrated, plug-and-go rack infrastructure.

Dell Integrated Rack Scalable Systems Overview

Dell's Integrated Rack Scalable Systems (IRSS) program provides a comprehensive, turnkey solution for deploying fully integrated, plug-and-play rack-scale systems. With IRSS, data centers can seamlessly integrate advanced technology to meet the demands of modern AI applications.

IRSS solutions are designed for rack-scale AI and HPC deployments, maximizing space, energy efficiency, reducing costs and are delivered as fully built and tested sets of racks. This allows for deployment in days rather than weeks, with a 3x reduction in time to build, deliver, and deploy.

Every IRSS solution undergoes comprehensive QA testing at the factory and deployment to ensure quality, and include Ethernet networking from Dell and Nvidia, Infiniband from Nvidia, and integration with 3rd party switches.



Dell's IRSS program offers a turnkey solution for deploying energy-efficient AI infrastructure. By combining expert assessment with L11/L12 factory integration & deployment services, and Dell Smart Cooling technologies, IRSS simplifies the process of deploying AI infrastructure and reduces deployment costs. IRSS solutions empower businesses to scale seamlessly, making them an attractive option for organizations looking to simplify and accelerate their AI infrastructure deployments.

With a single point of contact for service and support, IRSS enables customers to focus on their core business while Dell handles the complexities of engineering, integrating, configuring, and installing the infrastructure.

IRSS offers customers a choice of rack form factor, power and thermal management, and delivers the highest GPU density available in standard racks. The program features air and liquid cooling options, as well as advanced management capabilities, and is engineered to provide a reliable and efficient infrastructure tailored to meet specific needs.

Simplifying the Complexity of AI Infrastructure

Key benefits of Integrated Rack Scalable Systems (IRSS) include power and cooling solutions optimized for rack scale, one-call support for the entire rack, customized rack layout engineering, physical integration, validation at the rack level, and logistics support. These services reduce deployment time and free up internal resources for other value-added tasks, ultimately delivering a working solution rather than just hardware components. With expertise in engineering, integration, and installation, Dell simplifies the complexities of deploying advanced server and networking solutions, providing a seamless rack-scale procurement process that unlocks the full potential of AI.

- **Energy Efficiency:** IRSS utilizes advanced cooling technologies, including direct liquid cooling, to reduce energy consumption and optimize performance.
- **High Density:** Designed to maximize space, IRSS supports high GPU density within standard racks, featuring configurations that accommodate up to 96 GPUs.
- **Scalability:** Easily expand your infrastructure as your computing needs grow. The modular design allows for easy expansion, accommodating future growth in computing needs.
- **Future Ready:** Built with modular designs to accommodate future technology advancements. IRSS supports diverse silicon options, including processors and accelerators from Intel, AMD, and NVIDIA, readying your infrastructure for the next generation of technology.
- **Speed to Value:** Rapid deployment ensures immediate operational benefits.

Extraordinary Speed to Value

Dell’s IRSS program stands out by delivering extraordinary speed to value, transforming complex deployment challenges into seamless operations. IRSS offers a unique blend of technical innovation and operational efficiency, featuring support for multiple silicon options and sophisticated cooling technologies to meet the demands of modern AI applications.

Expert Guidance

With Dell’s technical expertise, our specialists conduct thorough data center assessments and provide tailored solutions that align with your technical goals, ensuring scalable growth as your needs evolve.

Custom Engineering

Dell customizes each solution to meet your specific demands, ensuring a perfect fit for your operational requirements. Dell’s engineering team customizes each solution to meet specific technical requirements, employing rigorous design processes to ensure optimal system performance.

Comprehensive Validation

Our validation processes include extensive testing and retesting of each component to guarantee system reliability and peak performance.

One-Call Service and Support

Dell offers comprehensive technical support for the entire rack and its components, ensuring seamless operation and quick resolution of issues.

How It Works

All components including servers, switches, power and cooling are assembled and tested at the rack level to ensure high quality, well organized infrastructure that delivers reliable operations and optimized performance. Full QA Checks at the rack level confirm your system operates at peak efficiency, reducing potential issues and ensuring smooth performance. **The steps in the Dell Integrated Rack Scalable Systems process include:**

Assess

- Data Center Readiness
- Assessment Services for GenAI
- In this optional step, you can request an expert assessment of your environment, ensuring there's enough space, power, and cooling to support optimal performance and prevent issues like overcrowding or overheating

Design

By designing a rack layout along with compute and network infrastructure, we ensure efficient organization, superior performance, and seamless connectivity. Rack designs are based on a number of factors, including:

- Compute density
- Networking
- Power capacity
- Data center cooling (Air & Water temperatures & flow rates)
- Watts & BTU per rack
- Dynamic & Static weight limits
- Data center logistics such as door height and loading dock access

Build

Servers, switches and other devices are built to exact customer specifications and tested individually to ensure a quality product

Integrate

- Integration includes:
  - Physical assembly
  - Connect servers to power, cooling and in-rack networking
  - Label all power and network cables
  - Logical configuration
  - Validate at the rack level
  - Full QA Checks
  - Asset data capture
  - Build Documentation
    - Elevations
    - Cable Maps
    - Labels
    - Integrator Instructions
  - Packaging & shipment

Deploy

With dedicated white-glove service, we handle cabling, setup, and on-site validation to guarantee smooth installation and optimal system functionality. Deployment includes:

- Shipment
  - White Glove
  - Scheduled Delivery
  - E2E Schedule
- Onsite
  - Roll in racks
  - Unload and unpack racks
  - Set and level racks
  - Connect networking, power, and water
  - Full rack & cluster test
- Customer Acceptance
  - Final Signoff

Support

We offer a single point of contact for ongoing support to ensure streamlined communication and prompt issue resolution.

Integrated Rack Options

Integrated Rack 7000 (IR7000) series

The IR7000 series to rack is a next-generation infrastructure that reimagines data centers for the new era of high-density computing. Designed for scalability, sustainability, and advanced cooling technologies, the IR7000 series to rack is optimized for large, rack-scale AI workloads with high core density and direct liquid cooling. As an Open Compute Project (OCP) Foundation standards-based infrastructure to rack, the IR7000 scales As an Open Compute Project (OCP) Foundation standards-based rack, the IR7000 scales to support multigeneration and heterogeneous technology with diverse rack compute options.

The IR7000 series features a 21” Orv3-based rack infrastructure with dense compute and liquid cooling for high TDP GPUs and CPUs. With integrated power busbars and DLC manifolds, the IR7000 series simplifies back-of-rack serviceability and cabling, paving the way for standardization. The series also includes power shelves that eliminate PSU/PDU cabling complexity and supports both in-rack and in-row CDUs.

Customer Benefits

- Unparalleled simplicity: Cable-free liquid and power delivery
- Scalable: Grow as your compute demand grows
- Rapid deployment: Your entire HPC cluster at-scale with a white glove experience
- Future-ready design: Support up to 480kW in each rack
- Efficient: Integrated DLC for energy efficiency
- Dense: Highest CPU and GPU density
- Standardized: OCP standard-based infrastructure for large-scale AI & dense compute
- Flexibility Designed for multigeneration & heterogenous technology (CPU, GPU & CPU+GPU)
- Turnkey Rack Level Deployment: Unmatched service and support from datacenter assessment to rack scale integration

Available in IR7000

PowerEdge M7725

The Future of High-Performance Dense Compute

The PowerEdge M7725 is a game-changing dense compute solution designed for the IR7000 rack. This 10U sled features 2x 2S server nodes, offering up to 72 nodes per rack and up to 27,000 cores per rack. With the latest AMD EPYC 5th Gen CPUs, the M7725 delivers uncompromised performance and energy efficiency.

Product Details

- Dense form factor: 10U with 2x 2S server nodes, up to 72 nodes per rack
- Uncompromised performance: Latest AMD EPYC 5th Gen CPUs
- Energy efficient: Hybrid cooling with air + liquid for optimized power utilization
- Easy to deploy and manage: Cold aisle serviceability, quick disconnects for cable-free liquid connectivity, and front I/O cabling
- Scalable: Up to 72 nodes per rack, delivering up to 27,000 cores per rack
- High-speed IO: 2x IO slots (PCIe Gen5x16) per node, enabling best-in-class data transfer rates (Ethernet and InfiniBand options supported)

Customer Benefits

- Uncompromised performance without constraints
- Energy efficiency with hybrid cooling and optimized power utilization
- Easy deployment and management with cold aisle serviceability and front I/O cabling
- High-speed IO for best-in-class data transfer rates

Available in the IR7000

PowerEdge XE9712 with GB200 NVL72

The Future of High-Performance Dense Acceleration

The PowerEdge XE9712 with GB200 NVL72 is a cutting-edge solution for real-time inference and generative AI workloads. With up to 30x faster performance compared to previous generations, this system is designed to accelerate large language models (LLMs) and other demanding AI workloads.

Product Details

- Lightning-fast connectivity: 72 connected GPUs acting as one with NVLink technology for seamless data transfer
- Energy efficient: Liquid cooled to maximize datacenter power utilization and reduce operating costs
- Rapid deployment: White glove deployment services ensure your entire AI cluster is up and running at-scale with a single point of contact
- Scalability: The GB200 NVL72 is designed for massive scale, with up to 72 GPUs in a single rack
- Advanced acceleration: NVIDIA’s GB200 NVL72 features 72 GB of HBM2 memory and 4320 CUDA cores

Customer Benefits

- Unparalleled performance: Up to 30x faster performance for real-time inference and generative AI workloads
- Scalability: Supports massive scale deployments with up to 72 GPUs in a single rack
- Energy efficiency: Liquid cooling reduces operating costs and environmental impact
- Rapid deployment: White glove deployment services ensure minimal downtime and maximum productivity

Product image is a represntation. Actual product appearance may vary.

Dell Integrated Rack 5000 (IR5000) series

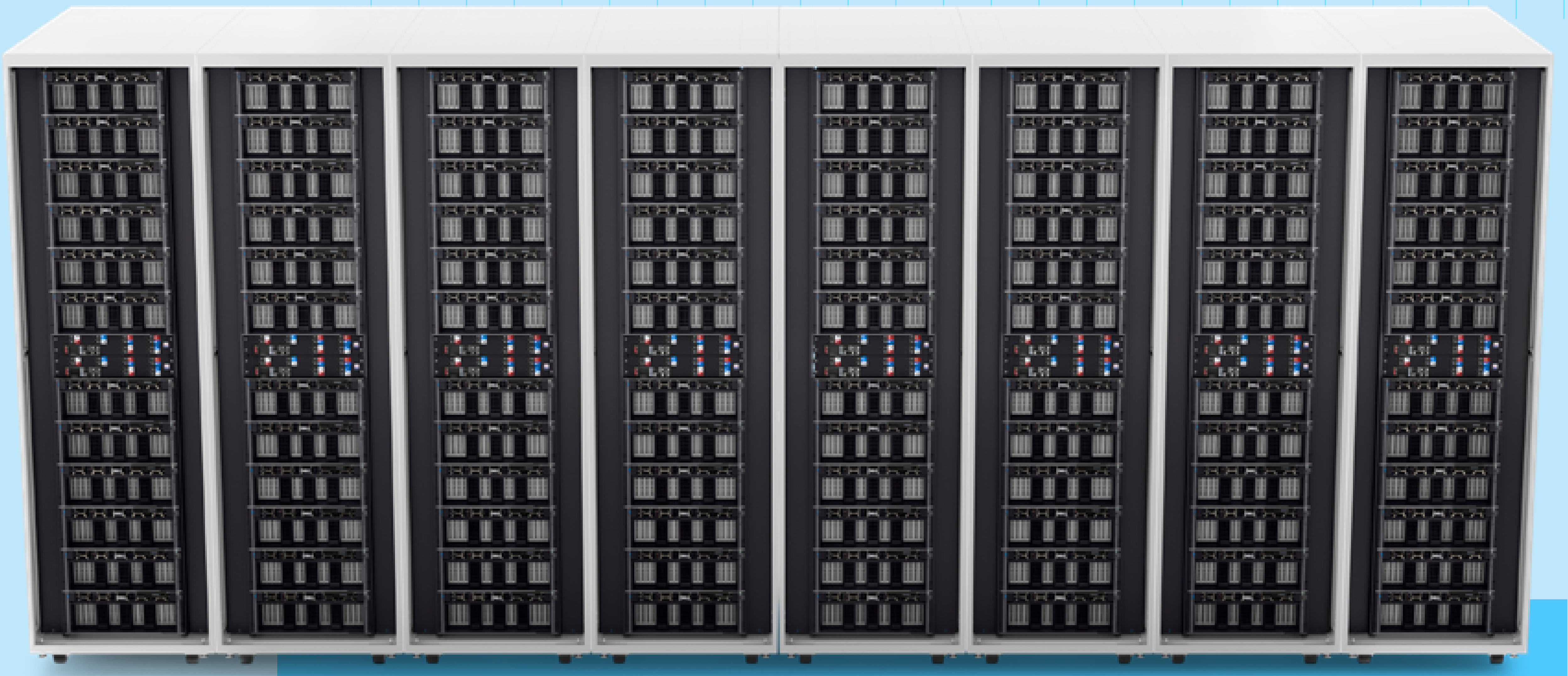
The Dell Integrated Rack 5000 (IR5000) is a standard 19-inch rack that delivers the highest available GPU density while ensuring energy efficiency through advanced thermal management. This rack offers both air and liquid cooling options, providing flexibility and scalability for your AI infrastructure.

Product Details

- Highest available GPU density in standard racks
- Advanced thermal management for energy efficiency
- Air and liquid cooling options for flexibility and scalability
- Silicon diversity for various processor choices

Customer Benefits

- Exceptional performance in a standard rack footprint
- Energy efficiency through advanced thermal management
- Flexibility and scalability with air and liquid cooling options
- Cost-effective deployment and management with turnkey services



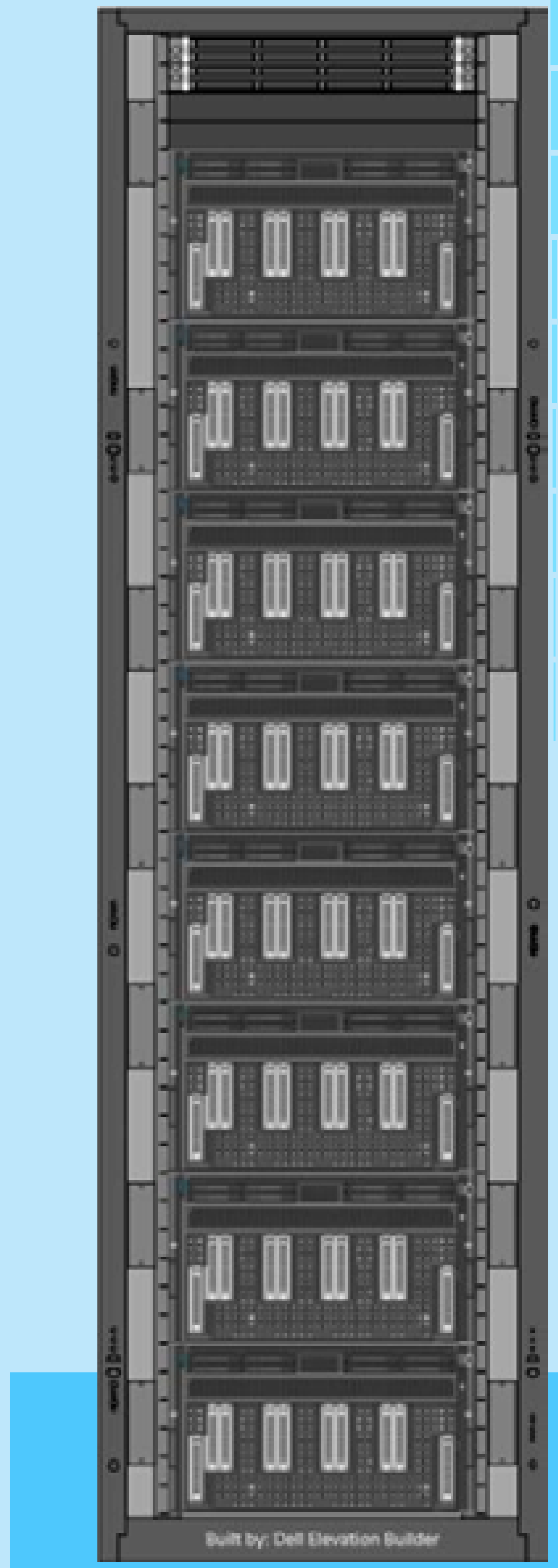
Available in IR5000

Dell PowerEdge XE9680

The Dell XE9680 air cooled 6U server was Dell's first 8x GPU platform, and it is now also available through Dell IRSS. This top-selling AI server is engineered to significantly enhance application performance by driving the most complex GenAI, Machine Learning, Deep Learning (ML/DL) and High Performance Computing workloads (HPC). This server features up to two Intel Xeon Scalable processors with up to 64 cores per processor and offers the highest GPU memory capacity and bandwidth currently available, making it capable of managing extremely large and complex models and datasets.

Product Details

- Offers up to two 5th Generation Intel Xeon Scalable processors and either eight NVIDIA H100 or eight H200 SXM5 700W GPUs fully interconnected with NVLink, eight AMD Instinct MI300X 750W OAM GPUs fully interconnected with Infinity Fabric links or eight Intel Gaudi 3 900W OAM accelerators with ethernet connectivity w/ embedded RoCE ports.
- Improve generative AI training performance with GPU-GPU communication and up to 1.5TB shared coherent GPU memory integrated into these offers.
- Deploy latest generation technologies including DDR5, PCIe Gen 5.0, and NVMe SSDs to push the boundaries of data flow and computing possibilities.
- Up to 10 front-facing PCIe Gen 5 slots and up to 16 drives enable optimal expansion for high-performance realtime AI operations.



Available in IR5000

Dell PowerEdge XE7745

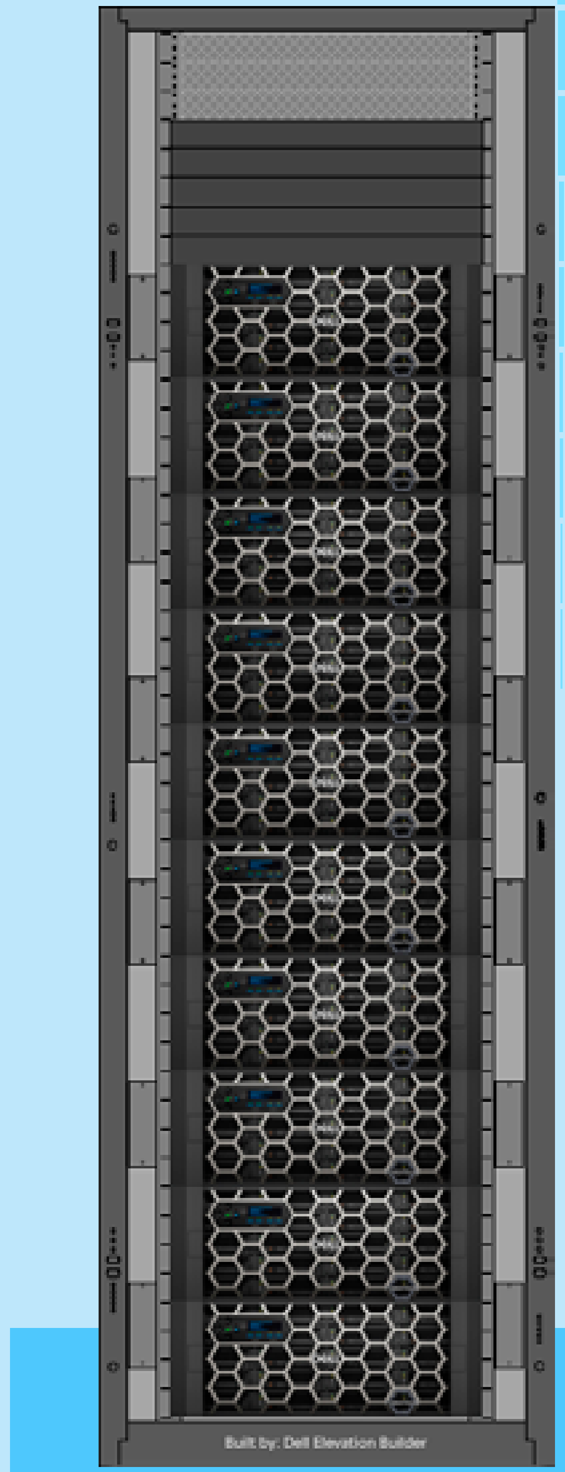
The Dell PowerEdge XE7745 is a 4U air-cooled server designed for enterprise AI workloads, available for individual purchase or in a rack-scale design in the IR5000. It provides high-performance acceleration for tasks such for tasks such as inferencing, model fine-tuning, and high-performance computing (HPC). With dual AMD EPYC processors and up to 16 internal PCIe GPUs and 8 front-facing PCIe network adapters, this server is ideal for AI inferencing, model fine-tuning, and HPC applications.

Product Details

- Supports dual 5th Gen AMD EPYC™ processors and up to 8 double-width or 16 single-width PCIe GPUs for dense AI acceleration
- 4U air-cooled chassis for efficient heat management
- Flexible GPU and AI fabric options with up to 192 CPU cores
- Scalable support for AI and HPC applications with up to 8x E3.S storage bays and 24x DDR5 DIMMs
- Quantum decryption-resistant 256-bit AES encryption and TPS processors for security
- AMD Infinity Guard and dedicated platform security co-processor for robust cybersecurity

Customer Benefits

- Designed for enterprise AI workloads, this server provides a scalable architecture for future-proofing that is adaptable to evolving AI demands
- High-performance acceleration for AI inferencing, model fine-tuning, and HPC applications
- Flexible GPU and AI fabric options for scalable and adaptable architecture
- Supports a range of AI and HPC applications, making it a versatile solution for diverse workloads



Available in IR5000

Dell PowerEdge XE7740

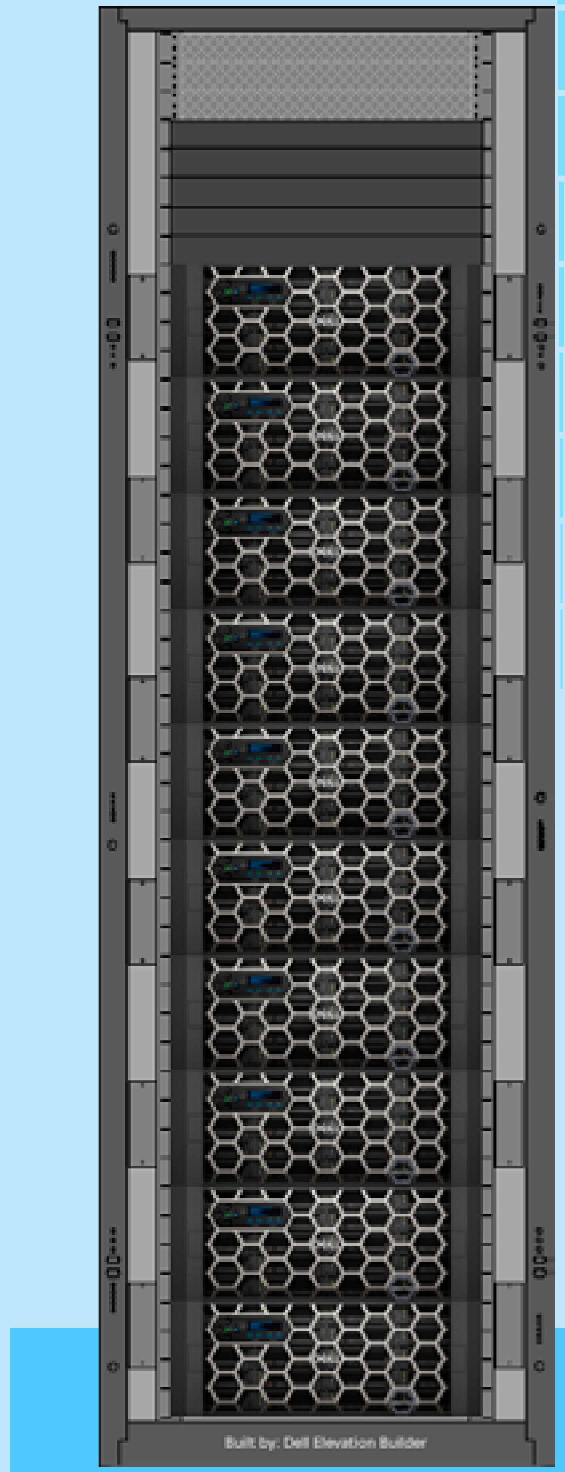
The Dell PowerEdge XE7740 is a 4U air-cooled server designed for enterprise AI workloads, available for individual purchase or in a rack-scale design in the IR5000. It features dual Intel Xeon 6 CPUs and choice of up to 8 double-wide accelerators, including Intel Gaudi® 3 PCIe accelerators or NVIDIA H200 NVL GPUs, or up to 16 single-wide accelerators, such as the NVIDIA L4. This flexibility allows enterprises to right-size their server configuration for the workload at hand, from fine-tuning or inferencing generative AI models, to extracting value from large datasets.

Product Details

- Supports dual Intel Xeon 6 processors and up to 8 double-width or 16 single-width PCIe GPUs for dense AI acceleration
- 4U air-cooled chassis for efficient heat management
- Flexible GPU and AI fabric options with up to 384 CPU cores
- Scalable support for AI and HPC applications with up to 8x E3.S storage bays and 24x DDR5 DIMMs
- Reliable cooling for high-performance GPUs with air-cooled 4U chassis

Customer Benefits

- Designed for enterprise AI workloads, this server provides the performance and flexibility needed to support complex tasks
- High-performance acceleration for AI inferencing, model fine-tuning, and HPC applications
- Scalable architecture for future-proofing and adaptable to evolving AI demands
- Flexible GPU and AI fabric options for scalable and adaptable architecture
- Offers a reliable cooling system for high-performance GPUs
- Easy deployment in existing datacenters without complex changes



Available in IR5000

Dell PowerEdge XE9680L

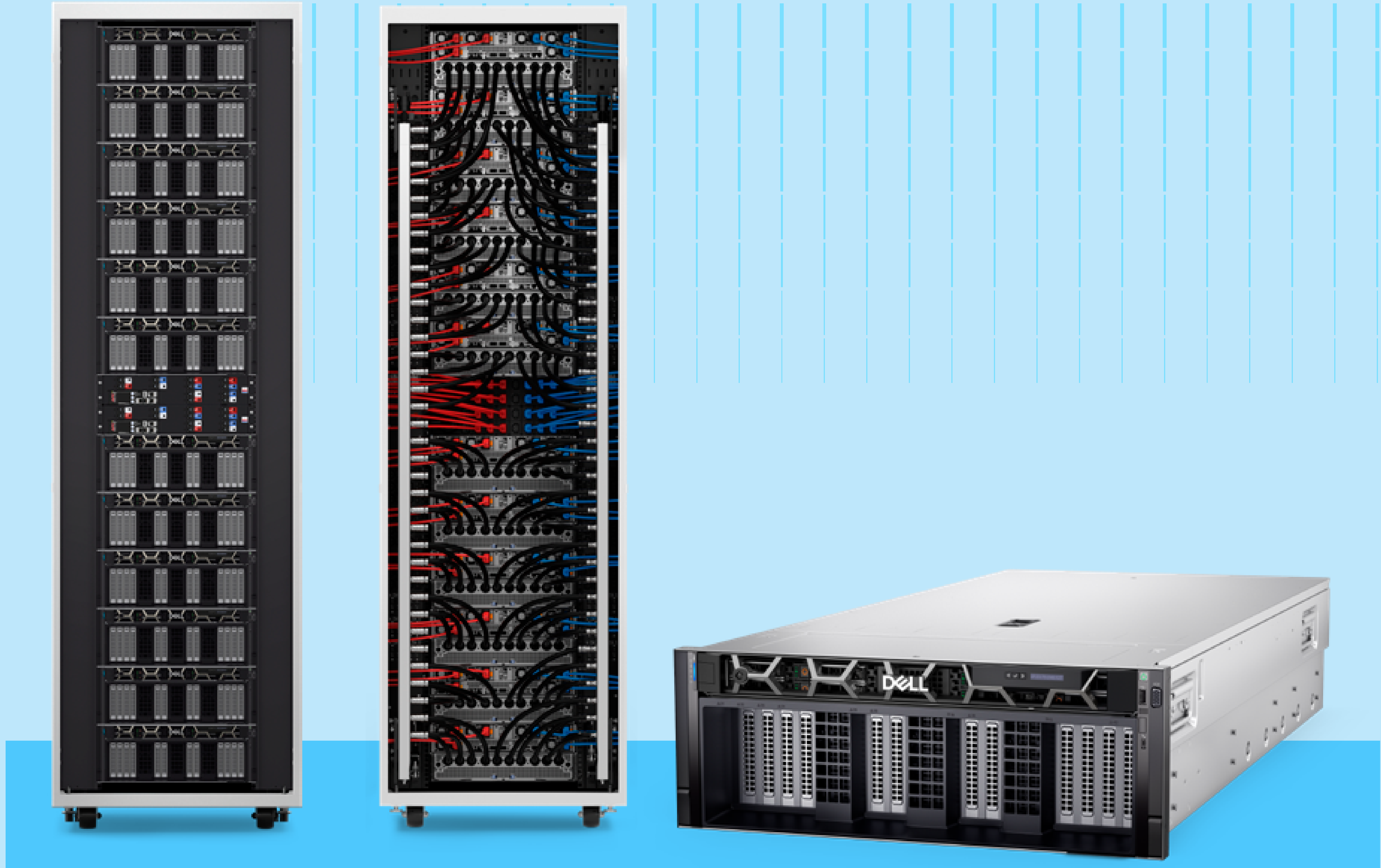
The Dell PowerEdge XE9680L is a liquid cooled, rack-scale server designed for enterprise AI workloads, only available in the IR5000 rack. It features an Intel processor and a modular architecture that simplifies infrastructure design and serviceability requirements. With optional factory integration of pre-validated networking, power distribution, and cooling options, this server is optimized for AI workloads and simplifies the process of scaling your workflow.

Product Details

- Intel processor for high-performance computing
- Modular architecture for simplified infrastructure design and serviceability
- Optional factory integration of pre-validated networking, power distribution, and cooling options
- Holistic approach to solution design, sizing, testing, optimization, and performance tuning
- Full rack solution for seamless integration and scalability
- Growing suite of services for AI, including deployment and management services

Customer Benefits

- Scalable architecture for enterprise AI workloads
- Simplifies infrastructure design and serviceability requirements with a modular architecture
- Offers liquid cooled high-performance computing with Intel processor
- Optimized for AI workloads with pre-validated networking, power distribution, and cooling options
- Holistic approach to solution design and testing for optimal performance and scalability
- Growing suite of services for AI, including deployment and management services



Available in IR5000

Dell PowerEdge XE9685L

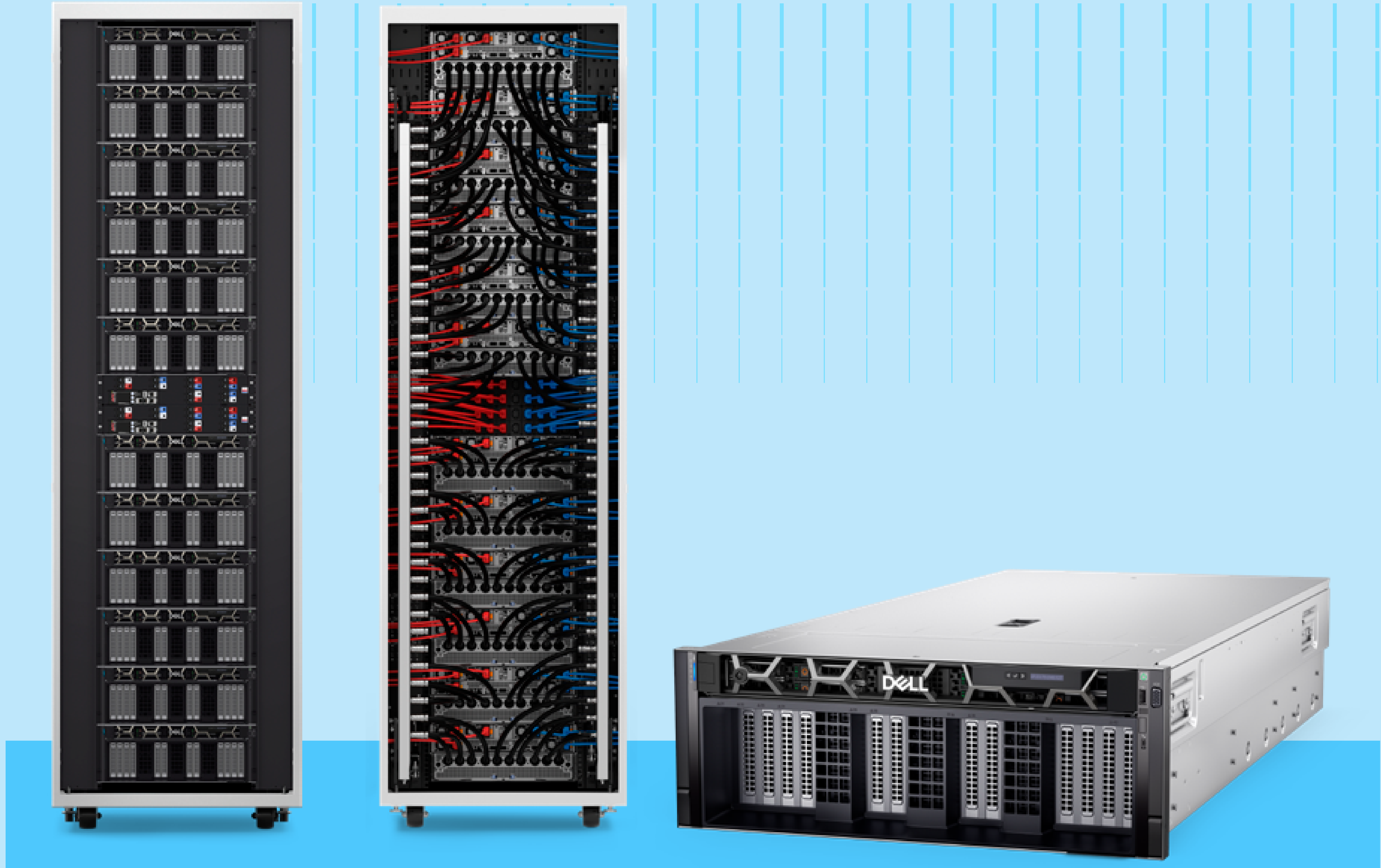
The Dell PowerEdge XE9685L is a dense, 4U liquid-cooled server that is only available in the IR5000 through Dell IRSS. It is designed for AI, machine learning, high performance computing and other data-intensive workloads. The dual AMD EPYC™ CPUs paired with NVIDIA Tensor Core GPUs, along with up to 12 PCIe gen 5.0 slots, offer customizable configurations to meet specific compute needs, optimized storage connectivity and maximum IO throughput for demanding workloads.

Product Details

- Dual AMD EPYC CPUs with 50% more cores
- 8 NVIDIA Tensor Core GPUs for maximum parallel processing
- 24 DDR5 DIMM slots with up to 6TB of memory
- 8 NVMe drives for NVIDIA GPU Direct Storage
- 12 front serviceable PCIe Gen 5.0 slots
- 4x 525W redundant power supplies for maximum performance and reliability
- Optimized for 64/72 x Tensor Core GPUs in a 48RU or 52U rack

Customer Benefits

- Designed for high-performance computing, this server provides the power and flexibility needed to support complex AI and HPC workloads
- Scalable architecture for easy expansion and upgrade of dense, liquid-cooled systems
- Optimized for dense, 4U liquid-cooled systems, this server offers unmatched performance, scalability, and efficiency
- Easy serviceability for DLC liquid hoses and hot-swap NVMe drives
- Fast deployment with pre-validated networking, power distribution, and cooling options



Expertise in Power and Cooling

Dell leads the industry in power and cooling innovations, providing cutting-edge solutions that enhance data center efficiency and sustainability. Our technologies ensure effective heat dissipation and energy management, critical for high-performance operations. For over 30 years, we've been at the forefront of developing server solutions that enhance power and cooling efficiency. Our extensive portfolio of innovations related to power and thermal management is a testament to our dedication to this innovation.

We continually innovate our broad range of air and liquid cooling technologies to align with your needs, whichever path you're taking. Our commitment to delivering reliable and efficient systems has remained steadfast, ensuring that our customers always have access to the best technology and tools available.

Power and cooling operate in tandem, which is why it's important to address both. We've made improvements in the hardware design of next-generation Dell PowerEdge servers to optimize power utilization and performance. At the same time, our software advancements help you manage and balance power resources more effectively.

With our Integrated Dell Remote Access Controller (iDRAC) embedded in Dell PowerEdge servers, we enhance your visibility and control over power consumption at the rack level. Our OME Power Manager represents the next generation of server power management, empowering you to view, measure, and monitor power consumption while fine tuning your infrastructure's performance.

Our mastery of thermal management and broad portfolio of integrated rack options allow us to offer sustainable deployments of AI infrastructure

Award-winning Services Expertise

With over 15 years of success in services, Dell has earned over 430 awards for excellence in services. Our proven track record demonstrates our commitment to delivering technically advanced, reliable solutions that transform data center operations.

Dell’s expertise in rack integration services ensures that customers receive a seamless and efficient deployment experience. With a focus on pre-validated networking, power distribution, and cooling options, Dell’s rack integration services simplify infrastructure design and serviceability, reducing the complexity and risk associated with deploying AI workloads. By choosing Dell, customers can trust that their AI solutions will be deployed quickly and efficiently, minimizing downtime and ensuring optimal performance.

Dell Services has received **over 430 industry awards** during our 15 years of service success

49 TSIA Awards from the Technology & Services Industry Association for commitment to outstanding innovation, leadership and excellence in the industry. The TSIA (Technology Services Industry Association) STAR Awards are one of the highest honors in the technology and services space. Dell is a one of only two companies in their top tier Hall of Fame (30+ awards).

Forbes World’s Best Management Consulting Firms: After surveying clients and senior executives at consulting organizations in 15 countries, Forbes named Dell Technologies among the 2024 World’s Best Management Consulting Firms. We are a strategic advisor and trusted partner, and our services expertise is leveraged far beyond hardware technologies into people, processes, operating models and more.

About Dell Services

250M Assets supported

650+ parts distribution centers

83 technical support sites

55+ languages

60,000+ Dell & partner professionals

2,000 service centers

Dell Integrated Rack Scalable Systems Portfolio

Any rack, Any cooling, Any platform

✓ Choice of rack style ✓ Choice of thermal management ✓ Choice of processor

OCP Standards-based 21” Rack (IR7000 series)

LIQUID COOLED

XE9712

NVIDIA GB200 NVL72

Dense Acceleration

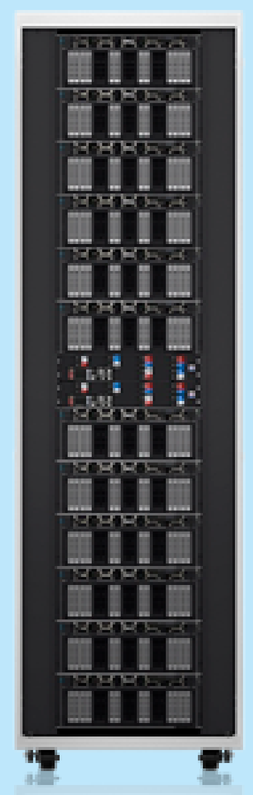
M7725

AMD Processor

Dense Compute

Traditional 19” Rack (IR5000 series)

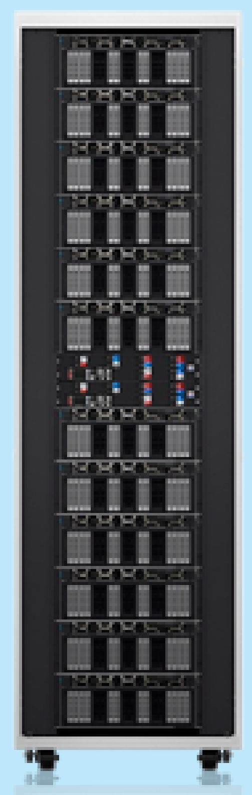
LIQUID COOLED



XE9680L

Intel Processor

Up to 96 GPUs

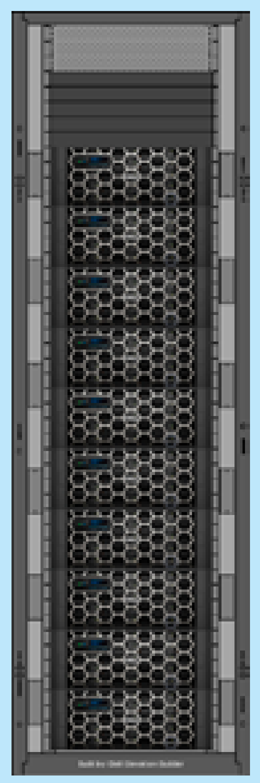


XE9685L

AMD Processor

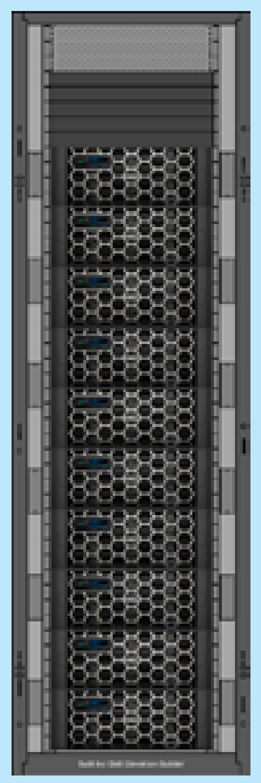
Up to 96 GPUs

AIR COOLED



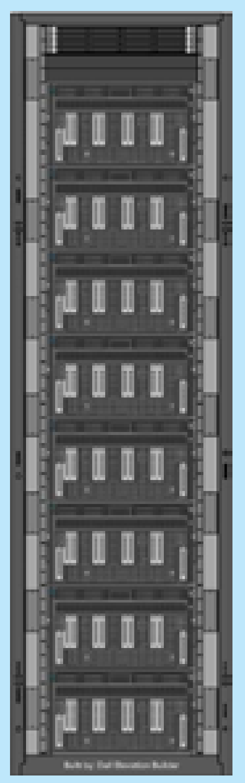
XE7740

Intel Processor Up to 160 SW GPUs or 80 DW GPUs



XE7745

AMD Processor Up to 160 SW GPUs or 80 DW GPUs



XE9680

Intel processor

Up to 64 GPUs

Delivered by expanded rack scale integration services for Integrated Rack Scalable Systems

Datacenter Assessment + Custom Rack Integration + One-call Support for Entire Rack

Visit our website to learn more: [Dell Integrated Rack Scalable Systems]

Dell Technologies believes the information in this document is accurate as of its publication date. The information is subject to change without notice.