



RIGHT-SIZING ENTERPRISE AI SCALE-OUT

What the Dell Poweredge AI Portfolio Means for Enterprise



Executive Summary

Every AI infrastructure conversation eventually collides with the same problem: workload profiles are in constant flux. What meets today's AI requirements may not meet tomorrow's. As models become larger and more demanding, organizations often move from PCIe-based GPU servers to high-density, NVLink-connected platforms designed for large-scale AI training and inference. Buying for the workload you have now rather than the portfolio of workloads you will actually run over in the next three to five years, is the one of the most common and most expensive mistakes IT organizations make when they enter AI at scale.

This makes the case that AI infrastructure decisions should start from workload diversity. We use two platforms from the Dell AI server lineup; the PowerEdge XE7745 and the XE9685L, as reference points on opposite ends of a deliberately broad spectrum, and focus on how to match the right solution to the workload, without overbuying or underbuilding.

The Inflection Point: AI Workloads Are Not One Workload

Enterprise AI adoption has moved past the pilot phase. Finance, healthcare, manufacturing, and retail organizations are no longer asking whether to deploy generative AI and machine learning capability, they are asking how to operate it economically, securely, and at a pace that keeps up with the business. That shift changes the infrastructure conversation.

Early AI infrastructure purchases were often reactive: a single dense GPU server acquired to support one training project or one proof of concept. As AI programs mature, the workload mix diversifies rapidly across an organization: retrieval-augmented generation and RAG-style inferencing for customer-facing applications, fine-tuning of open-weight models on proprietary data, batch inferencing for internal analytics, and for some, full-scale foundation model training. Each of these has a different GPU, memory, interconnect, and power profile. Standardizing on one server class to serve all of them almost always means overpaying for the lightweight workloads or underperforming on the heavy ones.

This is precisely why Dell built a portfolio with platforms designed for flexible inferencing and fine-tuning for dense multi-GPU training, all managed through a common iDRAC and OpenManage operational layer.

Two Reference Points on the PowerEdge XE Spectrum

To make the workload-first argument concrete, it helps to compare two systems that sit at genuinely different points on the spectrum: the PowerEdge XE7745, an air-cooled, PCIe-accelerator platform built for flexibility, and the PowerEdge XE9685L, a liquid-cooled, SXM-based platform built for density. Both share the same 5th Generation AMD EPYC 9005 foundation, the same iDRAC10/iDRAC9 and OpenManage management stack, and the same security architecture but solve different problems.

PowerEdge XE7745 - Flexibility as a Strategy

The XE7745 is built around a simple idea: not every AI workload needs or can justify a fully interconnected 8-GPU training pod. Its 4U air-cooled chassis supports up to eight double-wide, 600W PCIe accelerators such as the NVIDIA RTX Pro 6000 Blackwell Server Edition, or up to sixteen single-wide, 75W accelerators when the priority is inferencing density rather than raw per-GPU throughput. That range lets a single chassis design serve fine-tuning workloads on one end and high-concurrency, cost-optimized inferencing on the other, simply by changing the accelerator configuration.

Attribute	Configuration A (density focused)	Configuration B (flexibility focused)
Processors	2x AMD EPYC 9005 Series, up to 192 cores/socket	Same
GPU option	8x double-wide PCIe accelerators, up to 600W each	16x single-wide PCIe accelerators, up to 75W each
Memory	24 DDR5 DIMM slots, up to 3TB, 6400 MT/s	Same
Front storage	Up to 8x EDSFF E3.S Gen5 NVMe, 122.88 TB max	Same
Cooling	Air-cooled, standard data center friendly	Same
Networking	Up to 8 front-serviceable PCIe slots plus OCP 3.0 for cluster backend fabric	Same

Source: Dell PowerEdge XE7745 Specification Sheet, July 2025.

PowerEdge XE9685L - Density for Training & Fine-Tuning at Scale

The XE9685L represents the other end of the spectrum: eight NVIDIA HGX B200 GPUs (180GB each) fully interconnected via NVLink in a single liquid-cooled 4U chassis, paired with the same AMD EPYC 9005 Series CPU foundation. Where the XE7745 optimizes for configuration flexibility, the XE9685L optimizes for the raw GPU-to-GPU bandwidth that large-scale model training and heavy fine-tuning jobs depend on.

Attribute	PowerEdge XE9685L
Processors	2x AMD EPYC 9005 Series, up to 192 cores/socket
GPUs	8x NVIDIA HGX B200, 180GB, 1000W SXM6, fully NVLink-interconnected
Memory	24 DDR5 DIMM slots, up to 3TB, 6400 MT/s
Cooling	Direct liquid cooling for CPUs, GPUs, and NVLink switches
Storage	Up to 8x 2.5" NVMe/SAS/SATA drives, 122.88 TB max

Power	3000W Titanium redundant, hot-swap supplies
-------	---

Source: Dell PowerEdge XE series Specification Sheet and Dell.com product page, PowerEdge XE9685L.

It is a system ideally for workload that has moved from experimentation to a recurring, business-critical process and it is where liquid cooling, power delivery, and rack integration planning become as important a conversation as the server specification itself.

Beyond These Two: The Broader Dell AI Server Portfolio

The PowerEdge XE7745 and XE9685L are two points on a continuum. Dell's XE series also includes the Grace-based XE9712 and XE8712 for coherent CPU-GPU memory workloads, the AMD Instinct MI355X-based XE9785 and XE9785L for organizations standardizing on AMD's accelerator roadmap, and Intel Xeon-based XE9780/XE9780L variants for customers with existing Intel platform commitments. Having this breadth under one management framework rather than stitching together servers from multiple vendors with incompatible tooling is itself a risk-reduction argument.

Why This Matters at the Decision-Maker Level

Choosing the right AI infrastructure starts with answering three critical questions.

- Am I buying capacity for the workload I have today, or the workload portfolio I will have in 18 months? A GPU-diverse platform like the XE7745 protects against having to re-architect when the workload mix shifts from inferencing-heavy to fine-tuning-heavy.
- What does this do for my infrastructure? Liquid-cooled, SXM-class systems like the XE9685L deliver density and training performance that PCIe-based systems cannot match, but they also require power and cooling planning most enterprise data centers have not historically done. Air-cooled options remove that dependency where facility readiness is the constraint.
- Can I manage this without building a specialized AI operations team? Every system referenced in this paper shares the same iDRAC and OpenManage foundation, including Silicon Root of Trust, Secure Boot, cryptographically signed firmware, and System Lockdown. Security and lifecycle management practices transfer across the fleet rather than requiring a separate operating model per platform.

None of these questions have a universal answer. They depend on an organization's current AI maturity, its facility constraints, and its budget cycle.

Operational and Sustainability Considerations

AI infrastructure decisions increasingly get evaluated on more than raw performance. Dell reports meaningful power efficiency gains across the PowerEdge XE line, along with EPEAT Climate+ eco-label achievement across the majority of PowerEdge servers and multi-decade reductions in energy intensity per server generation. For decision makers who have to defend AI infrastructure to a board or a sustainability committee, not just an engineering team, this data matters as much as GPU throughput.

It's also an area where the operational simplicity of a single management plane (iDRAC10 plus OpenManage Enterprise, including AIOps-driven observability) pays off: power and cooling optimization is something the platform does continuously, not something an operations team has to hand-tune per server.

Why This Is Not Just a Product Conversation

Selecting the right solution is not just a technical decision. It touches facility readiness, power procurement, financing structure, and the realistic pace at which an organization's AI initiatives will actually mature, none of which show up on a spec sheet.

As an authorized Dell Technologies Titanium partner, ASA Computers works with organizations through each of those dimensions:

- Workload assessment first: mapping current and near-term AI initiatives against the PowerEdge XE lineup before recommending a specific configuration, rather than leading with a SKU.
- Configuration and integration: right-sizing GPU type and count, memory, storage, and networking to the actual workload, not the maximum available configuration.
- Facility and deployment planning: helping customers evaluate whether an air-cooled platform like the XE7745 fits existing infrastructure, or whether a liquid-cooled, rack-scale deployment via Dell's Integrated Rack Scalable Systems program is the better long-term investment.
- Financing flexibility: structuring acquisitions through Dell APEX Flex on Demand or traditional procurement, so infrastructure spend can track actual AI adoption rather than requiring a large up-front commitment.
- Lifecycle support: pairing every deployment with the appropriate tier of Dell ProSupport or ProSupport Plus, backed by ASA Computers' own local deployment and support expertise.

Closing Perspective

Organizations that get the most value from AI infrastructure investment are rarely the ones that buy the biggest system first. They are the ones that matched the platform to the workload, left room to grow into denser configurations as their AI programs matured, and built that infrastructure on a management and security foundation that scales with them. The Dell PowerEdge XE portfolio, from the flexible, air-cooled XE7745 through the dense, liquid-cooled XE9685L and beyond, is built to support exactly that trajectory.

ASA Computers is ready to help IT decision makers work through what that trajectory looks like for their organization specifically: which workloads justify which platform, what facility investments are and are not required, and how to structure the acquisition in a way that matches the pace of actual AI adoption rather than the pace of vendor announcements.

Schedule A Free Infrastructure Consultation

To discuss a configuration built around your business' AI roadmap, reach out to an ASA Technical Advisor today.

ASA Computers · 1.800.732.5727 · sales@asacomputers.com · www.asacomputers.com

About ASA Computers

ASA Computers is a leading IT solutions provider specializing in high-performance computing, cloud, and data center infrastructure for world-class institutions including NASA, Caltech, Stanford, MIT, Harvard and more. By combining deep engineering expertise with global deployment capabilities, ASA helps organizations eliminate operational bottlenecks, reduce total cost of ownership, and turn complex infrastructure challenges into production-ready realities. If institutions with the most demanding data requirements in the world trust ASA to build their infrastructure, your AI workloads are in good hands.

Note

Dell PowerEdge XE7745 Specification Sheet; Dell PowerEdge XE Series Specification Sheet; Dell.com PowerEdge XE9685L product page. Specifications are subject to change; confirm final configuration details with your ASA Computers representative.