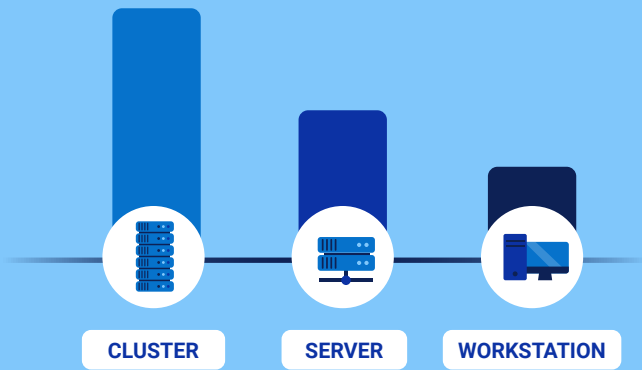


Accelerate AI innovation by right-sizing your IT infrastructure

Dell Technologies and NVIDIA help you deliver optimal efficiency and performance to maximize your ROI for AI.

Your infrastructure choices for AI depend on many factors. The smart approach **supports different model sizes, usage patterns, and user bases** while avoiding under- or over-provisioning.



Making the right infrastructure choices

Research shows that **32%** of organizations view high implementation costs as a barrier to implementing AI, and **20%** view the need to modernize infrastructure as another major hurdle.¹ How will you achieve AI success while staying on-budget?

Many factors will dictate whether you should run AI on workstations, in data centers, or in public clouds. Did you know that inferencing on-premises with Dell and NVIDIA AI solutions can be up to **4x more** cost-effective than in the public cloud?²

You can realize these savings by exploring the five key considerations for right-sizing your infrastructure.

1

How will you use AI?

While some use cases can be deployed with pre-built AI models, others require AI model creation or fine-tuning, requiring additional infrastructure, capacity, and performance.



2

How big does the AI model need to be?

The size of a model can impact speed and accuracy. Smaller models can fit on more modest hardware. Start with the smallest model you can to achieve the desired results.



3

How important is performance?

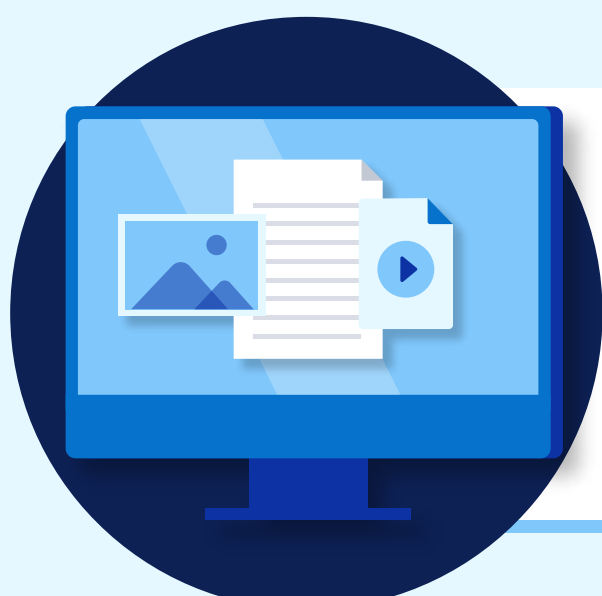
Find the right balance between cost and performance. Certain use cases, such as customer-facing digital assistants, require immediate responses and rapid scalability.



4

What is your AI model output?

Model outputs, such as text, images, and video, have very different capacity and performance requirements for the underlying infrastructure.



5

How many users are you supporting?

Understand where your AI will run and how many concurrent users will be accessing it, now and in the future, to avoid service outages and performance issues.



Accelerate AI innovation with Dell Technologies and NVIDIA

Dell Technologies and NVIDIA can help you leverage AI to drive innovation and achieve your business goals. The Dell AI Factory with NVIDIA is the industry's first and only end-to-end enterprise AI solutions,³ designed to speed AI adoption by delivering integrated Dell and NVIDIA capabilities to accelerate your AI-powered use cases, integrate your data and workflows, and enable you to design your own AI journey for repeatable, scalable outcomes.

Take the first step with Dell's half day, fee-waived AI Accelerator Workshop, and let our experts help you align priorities and achieve clarity for your AI vision.

[Review the Accelerator Workshop for GenAI](#)



[Explore the Dell AI Factory with NVIDIA](#)

1. Enterprise Strategy Group, a division of TechTarget, Inc. Research Report, Navigating Build-versus-buy Dynamics for Enterprise-ready AI, October 2024.
2. Enterprise Strategy Group, a division of TechTarget, Inc. White Paper commissioned by Dell, "Understanding the total cost of inferencing large language models," April 2024. Analyzed models show a 70B parameter LLM leveraging RAG for an organization of 50k users being up to 4x more cost effective over 3 years. Actual results may vary.
3. Based on Dell analysis, July 2024. Dell offers solutions with NVIDIA hardware and software engineered to support AI workloads from PCs with AI-powered features and workstations to Servers for High-performance Computing, Data Storage, Cloud Native Software-Defined Infrastructure, Networking Switches, Data Protection, HCI and Services.