The case for on-premises AI

Exploring the advantages of on-premises Dell PowerEdge servers with AMD EPYC processors vs. the cloud for small to medium businesses' AI workloads

Rely on predictable costs

Keep critical data fully within your control

Prioritize your performance needs

Give your IT team flexibility and control

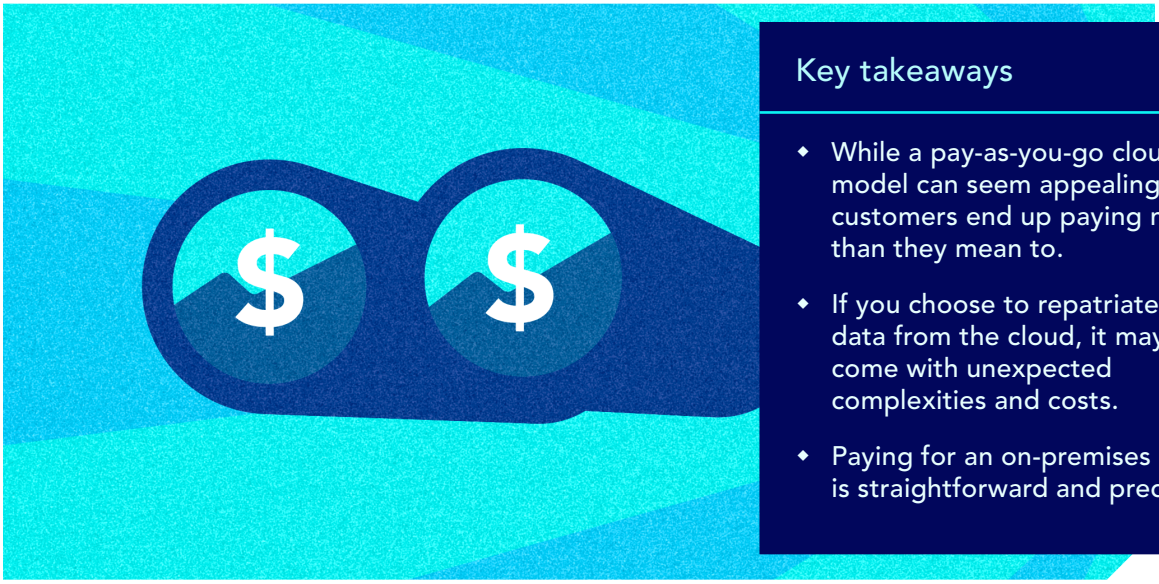
As artificial intelligence (AI) continues to dominate headlines and discussions of IT spending, organizations everywhere are considering their AI strategies. This is not a simple area. According to a recent Gartner survey, less than half of AI initiatives make it to production.[1] Ensuring your AI project doesn't fall by the wayside means making strategic, well-informed decisions.

In addition to figuring out what AI models and workloads will bring the most value for your specific use cases, you also need to select the right infrastructure to host your applications. Whether your workloads will run well with only CPUs or they demand plenty of GPUs, a plethora of cloud and on-premises options are available. The same is true for your storage needs: The choices can seem almost limitless. And particularly for small to medium businesses (SMBs) with limited budgets, it's important to make sure you're spending wisely. While purchasing hardware may require an investment up front, monthly cloud costs can add up quickly and often end up costing more over time.

We researched publicly available information to help you make this critical infrastructure decision. We focused on public cloud options generally and one specific on-premises approach: latest-generation Dell™ PowerEdge™ servers powered by AMD® EPYC™ processors. We found that compared to the public cloud, Dell PowerEdge servers can offer advantages in security, flexibility, and cost predictability, among other areas.

If your organization is mapping out your AI strategy for the next several years, read on to learn why AMD EPYC processor-powered Dell PowerEdge servers might be a good fit.

Rely on predictable costs



Key takeaways

- While a pay-as-you-go cloud model can seem appealing, many customers end up paying more than they mean to.
- If you choose to repatriate data from the cloud, it may come with unexpected complexities and costs.
- Paying for an on-premises solution is straightforward and predictable.

Any new initiative—AI or otherwise—comes at a cost. When you consider infrastructure options for your AI deployment, you need something that will both fit your budget needs and deliver the greatest impact for each dollar you spend. On the most basic level, choosing the public cloud entails ongoing costs that scale up or down as your usage increases or decreases, while investing in an on-premises solution means paying for the solution—either up front or via a different model, such as Dell APEX™ Subscriptions offerings[2]—then paying ongoing costs such as maintenance, power, and cooling.

The lower initial cost of the cloud may look appealing. But a 2023 IDC survey revealed that, due in part to new GenAI workloads, almost half of cloud buyers spent more than they planned on their cloud deployments in 2023, and nearly 60 percent anticipated similar overages the following year.[3] Conversely, an on-premises approach is predictable. You only pay for the hardware once, and your rent, power and cooling bills, and IT team's salaries are likely relatively stable.

Cost predictability is one of several factors driving a new trend: cloud repatriation, or migrating data from the cloud to on-premises solutions. A 2024 CIO article notes that between 70 and 83 percent of enterprises are planning repatriation of some workloads.[4] And one Forbes article lists cost as a major reason for cloud repatriation, explaining that "as [a small business'] data and user base grow, the cost of cloud services can quickly become expensive."[5]

Unfortunately, cloud repatriation isn't always simple—or free. If you migrate from one on-premises solution to another, there's no inherent cost in moving the data. If you've already made a cloud investment and decide to repatriate, some cloud service providers (CSPs) might require egress fees or data transfer fees, charging you for removing your data from their service. Others may have review requirements that create delays. For example:

- Microsoft Azure offers the first 100 GB per month of data egress for free, then charges 8 cents per gigabyte and up for the next 10 TB, with varying costs for more data after that.[6]

- AWS also offers 100 GB of free data egress per month, but if you need to migrate more, you must go through a review process with their support team.[7]
- Google Cloud allows you to transfer all your data off their cloud without fees, but you must first submit a request and terminate your account within 60 days of approval.[8]

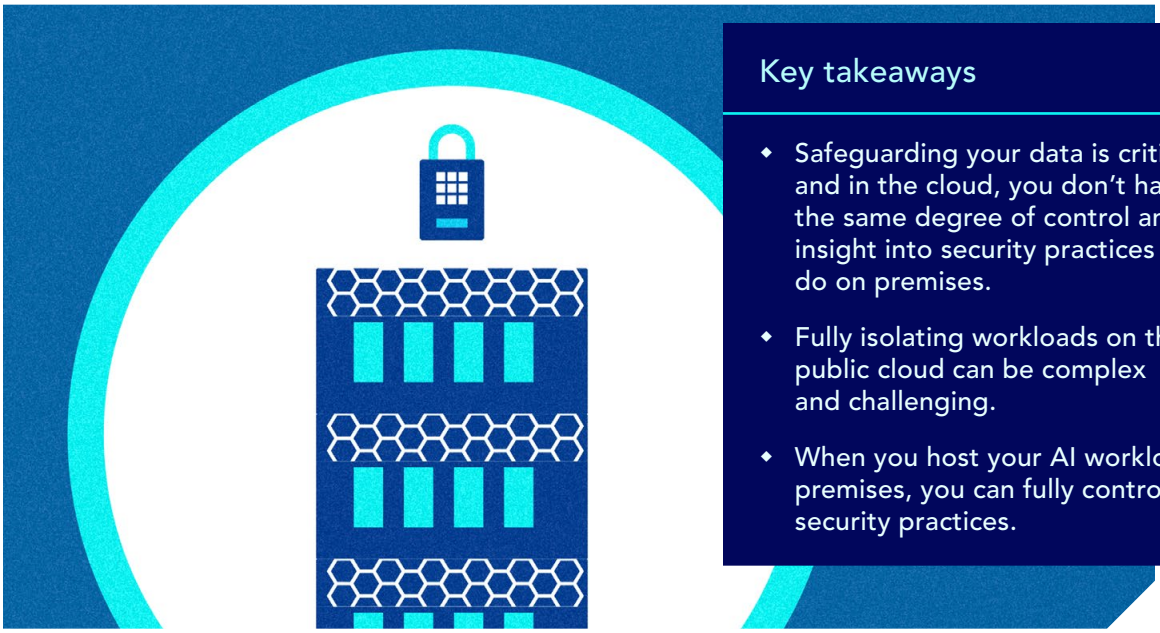
One 2025 PT study illustrates some of the cost challenges the cloud can present for deploying and managing GenAI workloads. We saw that over the course of four years, even using the most budget-friendly plans on the CSPs, Dell AI Factory could save companies up to 71 percent over comparable public cloud solutions.[9] While this study targets enterprises with much larger AI needs and budgets than those of most SMBs, it illustrates how much more companies can end up paying to host their AI workloads in the cloud versus in their own data centers.

> **Almost half of cloud buyers spent more than they planned on their cloud deployments in 2023, and nearly 60 percent anticipated similar overages the following year.**

As you can see when looking at CSP cost calculators or pricing pages, CSPs add charges across many aspects of each AI resource. Looking at Amazon SageMaker pricing, for example, you see tables for charges of compute units, API requests, storage, and more.[10] If you look at Azure Machine Learning in the Azure cost calculator, you see costs for APIs, Gateways, function executions, storage, model tokens, and much more.[11] Every one of these charge points is one your company has to keep careful track of to ensure that you stay in budget. In creating this paper, we considered incorporating screenshots of the AWS and Azure cost calculators to illustrate this point—but they were so complex and lengthy that they wouldn't fit! (See the endnotes to explore them yourself.)

By contrast, it's much more straightforward to calculate the TCO of an on-premises solution—beyond the initial CapEX investment, you simply have to calculate power and cooling, data center costs, and employee/administrator salaries. And while choosing an on-premises solution does entail up-front expense, over time, you could see cost savings versus hosting in the cloud. And you'd be in good company: In a 2025 CIO article, Chris Wolf of Broadcom says that for their customers, "running their AI services on-premises has turned out to be anywhere from a third to one-fifth of the cost of cloud-based options."[12]

Keep critical data fully within your control



Key takeaways

- Safeguarding your data is critical, and in the cloud, you don't have the same degree of control and insight into security practices as you do on premises.
- Fully isolating workloads on the public cloud can be complex and challenging.
- When you host your AI workloads on premises, you can fully control your security practices.

The top cybersecurity trend of 2025 is GenAI driving data security, per Gartner.[13] As the use of GenAI explodes, data breaches are also increasingly widespread. According to a 2024 IBM report, the average cost of a data breach for a business was $4.88 million—a 10 percent increase compared to the previous year.[14] If organizations pass these costs onto customers, as over half of those surveyed did, they could lose business in already-competitive markets.[15]

The level of security you require for your AI workloads will differ depending on the workload, the data the workload uses, your location, and the nature of your work. If you're using a great deal of your organization's private data in your AI deployment, security is even more vital. As just one example, large language models (LLMs) are a popular AI use case, and many organizations are building LLMs that draw on their own in-house data. But LLMs bring a number of potential security issues, including poor tenant isolation, shared memory across GPUs, and simple user error in failure to follow security best practices.[16]

For SMBs choosing where to deploy their AI workloads, an on-premises approach offers potential security advantages over the cloud:

- **Complete insight:** Keeping data on-premises gives you visibility and control over which users and applications access data. In contrast, with data in the cloud, you don't have the same degree of insight into what truly happens with your information.[17]
- **Reduced risk:** Simply introducing the added internet layer can expose data to risk: Cloud APIs that you may use to manage and monitor cloud resources are internet-accessible, unlike those for on-premises computing, and could give bad actors a larger window to access sensitive information.

- **Fewer issues with tenant isolation:** The cloud can exacerbate some of the security issues we mentioned surrounding LLMs such as tenant isolation. While CSPs offer security services and best practices guides, completely isolating workloads on the public cloud isn't always easy or straightforward. Cloud users must be conversant in identity and access management (IAM) policies—which include settings across multiple application layers, from storage buckets to VM settings and beyond—in order to ensure that their application is isolated and secure from top to bottom.[18]

In a traditional data center, security falls squarely on the shoulders of your organization, meaning admins can take all the steps necessary to safeguard against breaches. But resolving security issues in the cloud can require a lot of knowledge and responsibility from your IT team, which is ironic, considering that many cloud services promise to reduce work on IT. If your data isn't properly protected, it could leak out through shared memory on the host servers or other vulnerabilities in your cloud configuration.

"Responsibility for mitigating the risks that result from these software vulnerabilities is shared between the CSP and the cloud consumer," notes a blog post from Carnegie Mellon University.[19] In the cloud, you have to remain hyperaware of your particular plan's shared responsibilities, especially if the CSP changes policies.

> According to a 2024 IBM report, the average cost of a data breach for a business was $4.88 million.

For those in heavily regulated industries, such as healthcare, government, and financial services, the disadvantages of the cloud in this area are clear. Data sovereignty and compliance laws such as HIPAA and GDPR dictate strict privacy measures that include stringent requirements for data storage and processing. While HIPAA lays out a robust set of rules and agreements for healthcare organizations using CSPs,[20] cloud usage is still, tellingly, lower in heavily regulated industries. These companies store only 47 percent of sensitive data in the public cloud on average—much lower than the 61 percent usage rate across other industries.[21]" A cloud, while offering scalability, is a shared resource. For businesses with highly proprietary information or strict compliance needs, the potential risks of not having end-to-end control over the storage of that information far exceed the benefits," says Timothy E. Bates, a current professor and the former CTO of Lenovo.[22] Even if your data isn't quite as sensitive as healthcare records, you can gain more control by keeping your most privacy-critical AI workloads on premises.

In a fraught security and compliance landscape, running AI workloads in-house offers promising benefits. 56 percent of respondents to a Gartner Peer Community poll said that the biggest advantage of on-premises computing is the greater control and security from keeping data and apps on their own servers.[23] Others prefer the traditional data center "for maximum control and security over highly confidential data or those subject to strict regulations."[24] By keeping data on premises, you can enjoy visibility and control over who accesses data and when, potentially minimizing risks by maintaining a greater degree of agency over your resources.

The Dell and AMD approach to security

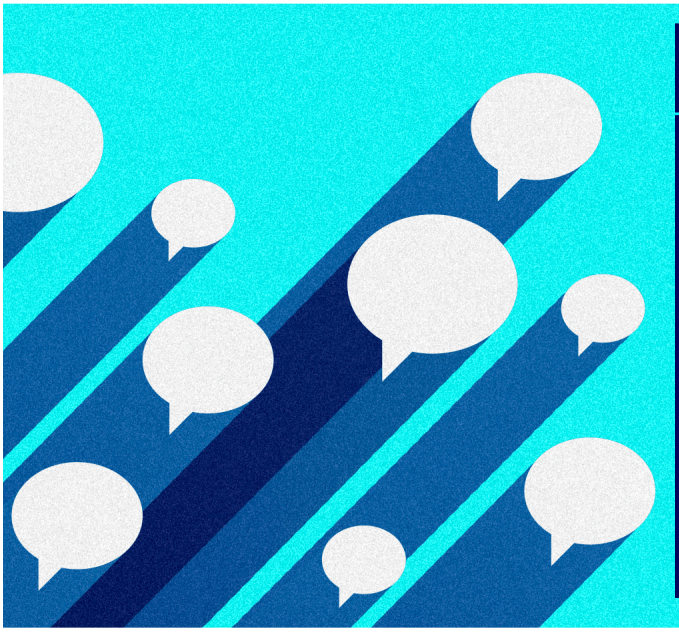
When it comes to AI, Dell understands the distinct importance of data security. According to a Dell whitepaper, "AI systems typically rely on massive amounts of data, which may include sensitive and confidential information such as personal details, financial data, or proprietary information. Safeguarding this data is critical to prevent unauthorized access or data theft, as well as to ensure the precision, dependability, and consistency of AI models and predictions."[25]

To bolster organizations' security efforts, Dell PowerEdge servers with AMD processors offer AMD Infinity Guard[26] security features built into the AMD EPYC processors, including Secure Encrypted Virtualization (SEV) and AMD SEV-Secure Nested Paging (SEV-SNP).[27] These features come free in the latest AMD processors and, per AMD, add only a small overhead to the processors.[28] In fact, in a 2024 study, we found that when we enabled AMD SEV and another AMD security feature, SEV-ES, on a Dell PowerEdge R7625 server, performance did not decrease at all.[29]

Dell PowerEdge servers can also leverage federated learning, a method for training an AI model across decentralized devices to maintain data security. With the Federated AI platform from Dell, algorithms run at the network edge and share only models, metadata, and results to other edge devices, "enabling the near real-time extraction of actionable insights from large, distributed datasets without revealing the data and any intellectual property."[30]

To learn more, visit https://www.delltechnologies.com/asset/en-us/products/servers/industry-market/dell-and-amd-empowering-enterprises-with-ai-whitepaper.pdf.

Prioritize your performance needs



Key takeaways

- Latency is a significant concern for AI workloads, and the cloud can introduce network-related inefficiencies that affect latency.
- Performance is complicated and dependent on many factors—you can't necessarily count on the performance you'd expect based on components you choose.
- We've consistently seen strong AI performance from AMD EPYC processor-powered Dell PowerEdge servers.

When you're designing a new AI project and mapping out the infrastructure to support it, one of the most important considerations is performance. If your new chatbot can't keep up with the real person it's talking to, or if your AI-enhanced search takes too long to return results, users could abandon these services. Poorly designed or implemented AI-based applications could drive away business or slow employee productivity.

Latency—or the length of time for an application to respond across its various components—is especially important for AI apps. Derreck Van Gelderen, head of AI strategy at PA Consulting, says that latency matters "particularly for applications requiring real-time or low-latency responses, such as autonomous systems or edge-based solutions. Delays introduced by transmitting data to and from cloud servers can be a limiting factor."[31] AI developers must consider the latency between where the data resides and where data computations take place. Key latency considerations include:

- **What your AI application requires:** If you're implementing a business-critical AI app that many users will rely on, low latency is critical. If you're still experimenting and don't need optimal performance, higher latencies may be acceptable.
- **Where your data is stored:** With the public cloud, your data likely lives in cloud storage separate from the instance on which your AI app runs, so the app must access the data across a network. This makes performance and latency dependent on that network, which can introduce network-related inefficiencies.[32] In contrast, if you choose on-premises hardware, you can store your data on disks attached directly to your server, cutting out the need for a potentially limiting network connection. Instead of moving your data to your compute in the cloud, you can move your compute resources to your data on site.

- **What drive options you choose:** You could configure the AMD EPYC processor-powered Dell PowerEdge R7725, for example, with more affordable SATA disks for less-sensitive applications or extremely fast NVMe drives to serve real-time applications that require the greatest responsiveness.[33]

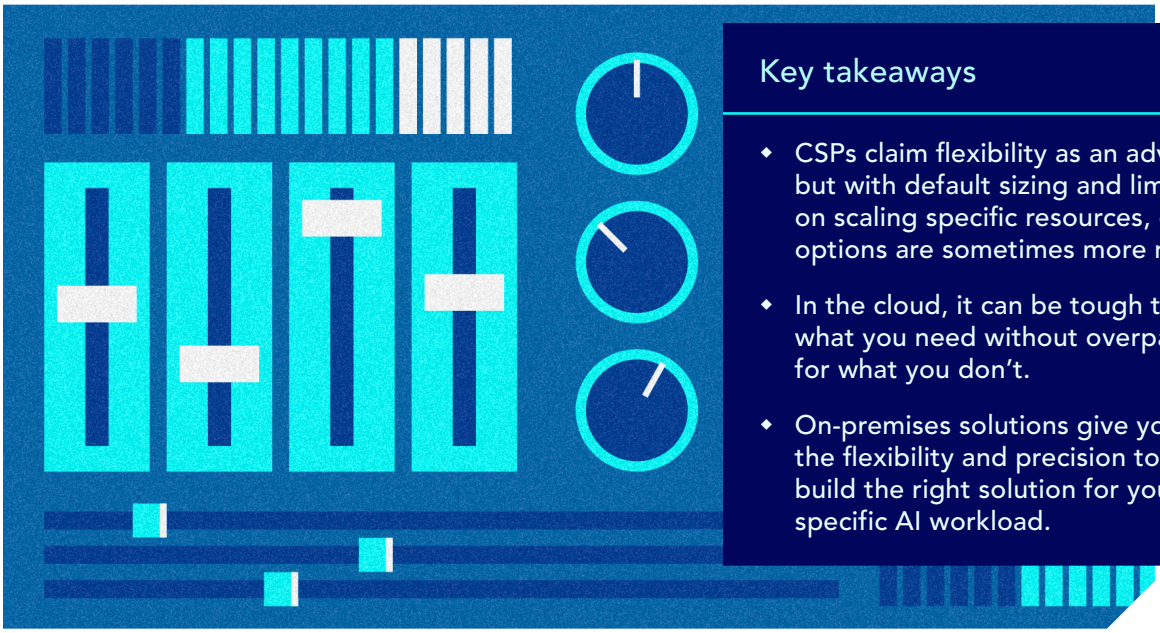
Latency isn't the only performance concern to keep in mind as you're planning your AI strategy. It's possible, if you opt for shared resources on the cloud, that you'll get varied performance over time due to usage of the shared resources by other users.[34] Even if you ignore the potential "noisy neighbor" problems, though, performance isn't necessarily completely predictable.

> Even if you choose CSP resources using the same hardware as your on-premises resources, you may not get the performance you're used to.

In our experience testing performance in both on-prem and cloud environments, identical hardware resources can perform differently in various environments. Most hardware vendors publicly note what performance you can expect from their product (i.e., disk manufacturers might post IOPS numbers). But the performance you actually get from those components depends on many different factors, which can lead to different outcomes across workload environments. That means that even if you choose CSP resources using the same hardware as your on-premises resources, you may not get the performance you're used to.

Still not sure what kind of performance you can count on? A section later in this paper, *"Our experience with AMD processor-powered Dell PowerEdge servers"* covers several recent PT studies that have highlighted the strong, real-world AI performance of Dell PowerEdge servers with AMD EPYC processors.

Give your IT team flexibility and control



Key takeaways

- CSPs claim flexibility as an advantage, but with default sizing and limits on scaling specific resources, cloud options are sometimes more rigid.
- In the cloud, it can be tough to get what you need without overpaying for what you don't.
- On-premises solutions give you the flexibility and precision to build the right solution for your specific AI workload.

Many cloud service providers tout flexibility as one of their advantages. The myriad of CSP offerings may seem to offer infinite customization and flexibility, and the ability to upgrade instances' vCPUs or memory on the fly can come in handy as a workload's needs change over time. But CSPs can be more limiting than they first appear, and in some ways, you might find the cloud rather rigid.

Let's say you have a workload that needs a lot of compute power and memory but only one GPU. If you ran this workload on the cloud, you'd likely find that like clothing manufacturers who use standard sizing, CSPs often make whole instances larger or smaller, which is the opposite of flexibility. An instance that offers more CPUs and more memory might be available only with more GPUs or other resources you don't need—resources that increase the cost of the instance. As just one example, take Amazon Web Services (AWS) G4 multi-GPU instances, which don't allow you to scale CPUs without scaling GPUs. If you want four GPUs and 96 vCPUs, you're out of luck.[35]

In the cloud, it can be challenging to ensure you're getting all of what you need, particularly for heavy AI workloads, without overpaying for what you don't. CSP instances often come with less-than-obvious limitations, such as how much bandwidth they receive or how many storage disks you can connect. Some even limit how many minutes or hours per day you receive the maximum performance they advertise. With resource-intensive AI workloads that demand very high performance, this can be a significant problem.

As you browse the data on instance options, you have to pay close attention to all the caveats and footnotes to make sure you understand what you're getting. For example, only certain instances on AWS come with Elastic Fabric Adapters (EFAs), and if you want certain processor features, they may be available on only some sizes or types of instances.[36] Storage can also come with limits and caveats; for example, the Remote Storage tab for the Azure Dsv6 series shows options and limits changing at every instance size.[37]

> Investing in an on-premises solution gives you the flexibility to design the specific environment your AI workload needs.

In contrast, investing in an on-premises solution gives you the flexibility to design the specific environment your AI workload needs with greater precision than the cloud can deliver. For example, you may wish to invest in CPU cores for servers handling growing AI pre-processing needs or add a GPU to your model training server without changing anything else. This is possible when you're building your own environment on premises. You can also more easily swap out or upgrade hardware as needed—such as by purchasing a few disks to add to your storage pool—without having to absorb larger monthly payments.[38]

With the ability to customize your hardware for the exact performance you need by pairing the right number of CPU cores or GPU cards with the right drive capacity and type for your AI workload, you can simply configure your server with those resources and get to work. On-premises environments allow you to spend your money on only the resources your workloads need, avoiding paying for any extra resources they don't need.

Taking a hybrid approach

While we're comparing public cloud and on-premises approaches in this paper, you don't necessarily have to choose just one. Many companies opt for hybrid solutions that combine the strengths of both cloud and on-premises options.[39] Some organizations may choose to put workloads with less sensitive data in the cloud, while keeping the most critical workloads—which may include AI workloads—on premises. According to Dell, PowerEdge servers are built to run in any environment without compromising control or security.[40]

How Dell PowerEdge servers with AMD EPYC processors can meet your AI needs

We've made the argument for why using on-premises infrastructure for your AI workloads instead of the public cloud might be right for you. Now, let's consider the advantages of choosing AMD EPYC processor-powered Dell PowerEdge servers:

Multiple high-performance options for AI

Dell offers a wide array of one-socket and two-socket servers with AMD EPYC processors, many of which Dell notes are ideal for AI workloads.[41] For AI workloads that require more power than CPUs, for example, these servers can support up to six single-wide or three double-wide GPUs.[42]

Dell AI Factory

When you buy from Dell, you also get the backing of the powerful Dell AI portfolio. Dell AI Factory offers a range of AI-specific solutions and services—including data services, infrastructure, software, use-case reference architectures, and more—to help you get the most out of your on-premises solution.

Full portfolio of products

Dell offers all the hardware you'll need for your AI implementation, from workstations to storage to networking and even data management and protection offerings.

Robust partner ecosystem

By choosing Dell, you also gain the advantages of the large Dell AI partner ecosystem, including offerings from AMD, to leverage AI tools such as accelerators, models, and more.[43]

Professional services and support

If you're feeling lost in the shifting AI landscape, Dell offers many professional services geared toward helping companies strategize, build, implement, and operate AI workloads—an especially valuable option for SMBs that may not have in-house knowledge across the AI spectrum.[44]

The Dell and AMD partnership strengthens the PowerEdge solution: AMD EPYC processors present an excellent choice for optimizing your AI workload. According to AMD, you can boost data analysis speeds, save money, and improve power efficiency[45] with AMD EPYC processors that "offer leadership performance and efficiency to enable material workload consolidation, allowing more space and energy to support new AI workloads."[46] AMD customers also benefit from the company's partnerships with AI organizations, such as Hugging Face.[47]

According to Dell, "What Dell and AMD's collaboration offers is a unified ecosystem of hardware and software, designed to allow developers to create end-to-end AI solutions that incorporate transfer learning, fine-tuning, and inferencing easily and efficiently."[48] If you go with an on-premises solution instead of the public cloud, leveraging Dell and AMD AI expertise, services, performance, and partnerships can help make your next AI project a success.

Our experience with AMD processor-powered Dell PowerEdge servers

PT has almost two decades of experience assessing Dell PowerEdge servers across a variety of workloads, and recently, much of our testing has focused on AI. In recent studies, we found that PowerEdge servers featuring AMD EPYC processors offer a compelling value proposition for SMBs planning their AI deployments.


Run your in-house AI chatbot on an AMD EPYC 9534 processor-powered Dell PowerEdge R6615 server

Deploy an in-house chatbot on a server with only a CPU

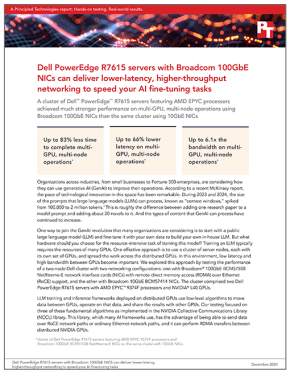
Support up to 9 simultaneous chatbot users

Support more users with a clear and easy upgrade path

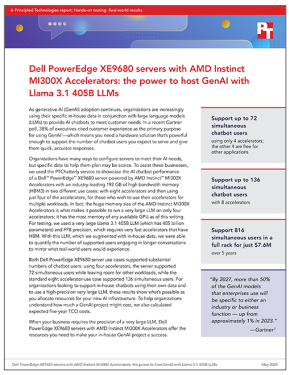
In a February 2025 study, we highlighted the Dell PowerEdge R6615 server, powered by a 64-core AMD EPYC 9534 processor, as an excellent solution for SMBs looking to implement in-house AI chatbots.

Our testing demonstrated that the server could:

- Support up to nine simultaneous users with a median response time of under five seconds, making it a cost- effective entry point for organizations exploring GenAI
- Accommodate up to 23 simultaneous users when we added an NVIDIA® L4 GPU, providing a clear upgrade path as your needs grow


Dell PowerEdge R7615 servers with Broadcom 100GbE NICs can deliver lower-latency, higher-throughput networking to speed your AI fine-tuning tasks

A December 2024 study focused on the performance of new PowerEdge R7615 servers for multi-GPU, multi-node operations such as you might see in algorithms used for training large language models (LLMs). A cluster of these servers, featuring AMD EPYC 9374F processors along with GPUs, completed multi-GPU, multi-node operations faster, with higher bandwidth and with lower latency when we equipped them with 100GbE Broadcom NICs than they did with 10GbE NICs. If you're building your own in-house LLM, it's worth considering this powerful compute and networking combination.

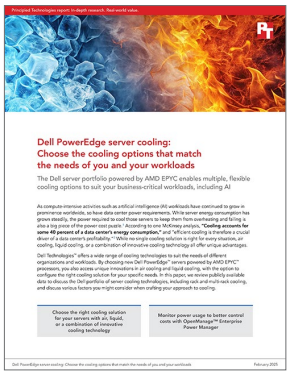

Dell PowerEdge XE9680 servers with AMD Instinct MI300X Accelerators: the power to host GenAI with Llama 3.1 405B LLMs

Support up to 72 simultaneous chatbot users

Support up to 136 simultaneous chatbot users

Support 816 simultaneous users in a full rack for just $7.6M

A similar study in May 2025 tested the AI chatbot performance of Dell PowerEdge XE9680 servers with AMD Instinct™ MI300X Accelerators and AMD EPYC processors. Unlike the previous study, this testing focused on enterprises and featured the very large Llama 3.1 405B LLM. As we saw in February, however, the Dell and AMD solution delivered very strong performance for this AI use case, supporting up to 136 simultaneous chatbot users.


Dell PowerEdge server cooling: Choose the cooling options that match the needs of you and your workloads

While these studies assessed performance, we've also recently explored other features of AMD EPYC processor-powered Dell PowerEdge servers. Our February 2025 research report examined the cooling and power management options these servers offer and laid out considerations for choosing the right cooling path for your needs. Because AI workloads are so resource-intensive, cooling and power are critical factors to consider as you map out infrastructure for your AI deployment.

Conclusion

AI initiatives can bring tremendous value to your business, but you need to support your new AI workloads effectively. That means choosing the best possible infrastructure for your needs—and many companies are finding that the cloud isn't right for them. According to a recent Rackspace survey of IT executives, 69 percent of companies have moved some of their applications on-premises from the cloud, with half of those citing security and compliance as the reason and 44 percent citing cost.[49]

On-premises solutions provide a number of advantages. With full control over your security infrastructure, you can be certain that all compliance requirements remain firmly in the hands of your IT team. Opting for on-premises also gives you the ability to design your infrastructure to the precise needs of that team and your new AI workloads. Depending on the workload, you may also see performance benefits, along with more predictable costs. As you start to build your next AI initiative, consider an on-premises solution utilizing AMD EPYC processor-powered Dell PowerEdge servers.

---

1. "Gartner Survey Finds Generative AI Is Now the Most Frequently Deployed AI Solution in Organizations," accessed March 26, 2026, https://www.gartner.com/en/newsroom/press-releases/2024-05-07-gartner-survey-finds-generative-ai-is-now-the-most-frequently-deployed-ai-solution-in-organizations.
2. Dell Technologies, "Simplified cloud experiences with Dell APEX Subscriptions," accessed April 3, 2025, https://www.dell.com/en-us/dt/apex/subscriptions.htm.
3. Daniel Saroff, "Storm Clouds Ahead: Missed Expectations in Cloud Computing," accessed March 26, 2025, https://blogs.idc.com/2024/10/28/storm-clouds-ahead-missed-expectations-in-cloud-computing/.
4. Lightedge, "The future is hybrid," accessed March 31, 2025, https://www.cio.com/article/3541342/the-future-is-hybrid.html.
5. Marcin Zgola, "The Rise Of Cloud Repatriation: Why Companies Are Bringing Data In-House," accessed March 31, 2025, https://www.forbes.com/councils/forbestechcouncil/2023/04/18/the-rise-of-cloud-repatriation-why-companies-are-bringing-data-in-house/.
6. Microsoft Azure, "Bandwidth pricing," accessed March 31, 2025, https://azure.microsoft.com/en-us/pricing/details/bandwidth/.
7. Sébastien Stormacq, "Free data transfer out to internet when moving out of AWS," accessed April 3, 2025, https://aws.amazon.com/blogs/aws/free-data-transfer-out-to-internet-when-moving-out-of-aws/.
8. Google Cloud, "Applying for free data transfer when exiting Google Cloud," accessed April 3, 2025, https://cloud.google.com/exit-cloud?hl=en.
9. Principled Technologies, "Investing in GenAI: Cost-benefit analysis of Dell on-premises deployments vs. similar AWS and Azure deployments," accessed April 1, 2025, https://www.principledtechnologies.com/Dell/PowerEdge-on-prem-GenAI-0524-v2.pdf.
10. "Amazon SageMaker pricing," accessed April 25, 2025, https://aws.amazon.com/sagemaker/pricing/
11. "Pricing calculator," accessed April 25, 2025, https://azure.microsoft.com/en-us/pricing/calculator/
12. Chris Wolf, "Why AI on-premises means big bottom-line advantages in the long-run," accessed April 1, 2025, https://www.cio.com/article/3830651/why-ai-on-premises-means-big-bottom-line-advantages-in-the-long-run.html.
13. "Cybersecurity Trends: Resilience Through Transformation," accessed May 14, 2025, https://www.gartner.com/en/cybersecurity/topics/cybersecurity-trends.
14. IBM, "Cost of a Data Breach Report 2024," accessed March 26, 2025, https://www.ibm.com/downloads/documents/us-en/107a02e94948f4ec.
15. IBM, "Cost of a Data Breach Report 2024."
16. Nahla Davies, "GPU Hosting, LLMs, and the Unseen Backdoor," accessed April 25, 2025, https://www.secureworld.io/industry-news/gpu-hosting-llms-unseen-backdoor
17. Timothy Morrow, "12 Risks, Threads, & Vulnerabilities in Moving to the Cloud," accessed March 26, 2025, https://insights.sei.cmu.edu/blog/12-risks-threats-vulnerabilities-in-moving-to-the-cloud/.

18. "17 Security Risks of Cloud Computing in 2025," accessed April 25, 2025, https://www.sentinelone.com/cybersecurity-101/cloud-security/security-risks-of-cloud-computing/

19. Timothy Morrow, "12 Risks, Threads, & Vulnerabilities in Moving to the Cloud," accessed March 26, 2025, https://insights.sei.cmu.edu/blog/12-risks-threats-vulnerabilities-in-moving-to-the-cloud/.

20. US Department of Health and Human Services, "Guidance on HIPAA & Cloud Computing," accessed March 26, 2025, https://www.hhs.gov/hipaa/for-professionals/special-topics/health-information-technology/cloud-computing/index.html.

21. Skyhigh Security, "Skyhigh Security Cloud Adoption and Risk Report: Healthcare Edition," accessed March 26, 2025, https://www.skyhighsecurity.com/wp-content/uploads/2023/09/SkyhighSecurity_DataDilemmaReport_HC_2023.pdf.

22. Joe McKendrick, "Why some companies are backing away from the public cloud," accessed April 1, 2025, https://www.zdnet.com/article/why-some-companies-are-backing-away-from-the-public-cloud/.

23. Gartner Peer Community, "What is the biggest advantage of on-premises computing?" accessed March 26, 2025, https://www.gartner.com/peer-community/poll/biggest-advantage-premises-computing.

24. Aaron Tan, "Top AI infrastructure considerations," accessed March 26, 2025, https://www.computerweekly.com/feature/Top-AI-infrastructure-considerations.

25. Dell Technologies, "Empower Enterprises with AI: Entering the Era of Choice," accessed March 27, 2025, https://www.delltechnologies.com/asset/en-us/products/servers/industry-market/dell-and-amd-empowering-enterprises-with-ai-whitepaper.pdf.

26. "AMD Infinity Guard," accessed April 25, 2025, https://www.amd.com/en/products/processors/server/epyc/infinity-guard.html

27. Dell Technologies, "Empower Enterprises with AI: Entering the Era of Choice," accessed March 27, 2025, https://www.delltechnologies.com/asset/en-us/products/servers/industry-market/dell-and-amd-empowering-enterprises-with-ai-whitepaper.pdf.

28. Raghu Nambier, "AMD EPYC™ Processors Deliver Confidential Computing for Public and Private Cloud Environments," accessed April 24, 2025, https://community.amd.com/t5/server-processors/amd-epyc-processors-deliver-confidential-computing-for-public/ba-p/544955

29. "Enable security features with no impact to OLTP performance with Dell PowerEdge R7625 servers powered by 4th Gen AMD EPYC 9274F processors," accessed April 24, 2025, https://www.principledtechnologies.com/clients/reports/Dell/PowerEdge-R7625-AMD-EPYC-9274F-security-0524/index.php.

30. Dell Technologies, "Empower Enterprises with AI: Entering the Era of Choice," accessed March 27, 2025, https://www.delltechnologies.com/asset/en-us/products/servers/industry-market/dell-and-amd-empowering-enterprises-with-ai-whitepaper.pdf.

31. Stephen Pritchard, "Why run AI on-premise?" accessed March 31, 2025, https://www.computerweekly.com/feature/Why-run-AI-on-premise.

32. Conor Bronsdon, "Understanding Latency in AI: What It Is and How It Works," accessed March 31, 2025 https://www.galileo.ai/blog/understanding-latency-in-ai-what-it-is-and-how-it-works.

33. Dell Technologies, "New PowerEdge R7725 Rack Server," accessed March 31, 2025, https://www.dell.com/en-us/shop/cty/pdp/spd/poweredge-r7725/pe_r7725_tm_vi_vp_sb.

34. Christopher Tozzi, "Cloud vs. On-Prem AI Accelerators: Choosing the Best Fit for Your AI Workloads," accessed April 25, 2025, https://www.itprotoday.com/cloud-computing/cloud-vs-on-prem-ai-accelerators-choosing-the-best-fit-for-your-ai-workloads

35. "Amazon EC2 G4 Instances," accessed April 25, 2025, https://aws.amazon.com/ec2/instance-types/g4/.

36. "Amazon EC2 Instance types," accessed April 25, 2025, https://aws.amazon.com/ec2/instance-types/.

37. "Dsv6 sizes series," accessed April 25, 2025, https://learn.microsoft.com/en-us/azure/virtual-machines/sizes/general-purpose/dsv6-series?tabs=sizestorageremote

38. Gaurab Acharya, "Cloud or On-Prem? The AI/ML Dilemma for Small Businesses and the Path to Efficiency," accessed March 31, 2025, https://www.techtimes.com/articles/309834/20250331/cloud-prem-ai-ml-dilemma-small-businesses-path-efficiency.htm.

39. Ken Kaplan, "The Amalgamation of AI and Hybrid Cloud," accessed March 31, 2025, https://www.nutanix.com/theforecastbynutanix/technology/where-ai-meets-the-hybrid-cloud.

40. Rod Mercado, "Empowering AI-Driven Innovation with Windows Server 2025 and Dell PowerEdge Servers," accessed March 31, 2025, https://www.dell.com/en-us/blog/empowering-ai-driven-innovation-with-windows-server-2025-and-dell-poweredge-servers/.

41. Dell Technologies, "PowerEdge Servers with AMD," accessed March 26, 2025, https://www.dell.com/en-us/dt/servers/amd.htm.

42. "PowerEdge servers with AMD," accessed April 25, 2025, https://www.dell.com/en-us/dt/servers/amd.htm#tab0=0&tab1=0&accordion0

43. Dell Technologies, "Dell AI Solutions," accessed April 1, 2025, https://www.dell.com/en-us/shop/dell-ai-solutions/sc/artificial-intelligence#ai-factory-dell.

44. Dell Technologies, "AI Services," accessed April 1, 2025, https://www.dell.com/en-us/lp/dt/artificial-intelligence-services.

45. AMD, "Accelerate Queries & AI Inference to Transform Data into Actionable Insights," accessed April 1, 2025, https://www.amd.com/content/dam/amd/en/documents/epyc-business-docs/infographics/amd-epyc-analytics-infographic.pdf.

46. AMD, "Advance Data Center AI with Servers Powered by AMD EPYC Processors," accessed April 1, 2025, https://www.amd.com/en/products/processors/server/epyc/ai.html.

47. Julien Simon, "Hugging Face and AMD partner on accelerating state-of-the-art models for CPU and GPU platforms," accessed April 1, 2025, https://huggingface.co/blog/huggingface-and-amd.

48. Dell Technologies, "Empowering Enterprises with AI: Entering the Era of Choice," accessed April 1, 2025, https://www.delltechnologies.com/asset/en-us/products/servers/industry-market/dell-and-amd-empowering-enterprises-with-ai-whitepaper.pdf.

49. Joe McKendrick, "Why some companies are backing away from the public cloud," accessed April 1, 2025, https://www.zdnet.com/article/why-some-companies-are-backing-away-from-the-public-cloud/.

This project was commissioned by Dell Technologies.

Facts matter.®