

AUGUST 2024

Accelerate AI Initiatives on Dell VxRail

Scott Sinclair, Practice Director; and Monya Keane, Senior Research Analyst

Abstract: The rapid evolution of AI in enterprise settings is driving demand for infrastructure that supports generative AI workloads on premises. Dell Technologies VxRail is a hyperconverged infrastructure (HCI) solution offering scalability, automation, lowered costs, and a consistent experience across data center, edge, and cloud environments. Organizations can start with light AI workloads on VxRail and scale up to more demanding tasks with validated GPUs. Dell's holistic approach also delivers robust security, ensuring protection across AI attack vectors. Leveraging VxRail's capabilities, businesses can position themselves well at any stage of their AI maturity, maximizing return on investment (ROI) and accelerating time to value for their AI initiatives.

The Rapid Pace of AI in the Enterprise

The quick rise of AI initiatives that we are now witnessing is reflective of the massive excitement around generative AI. According to research by TechTarget's Enterprise Strategy Group, 54% of organizations expect to have generative AI workloads in production within the next 12 months.¹

While discussions about suitable infrastructure to support generative AI often touch on expensive, high-performance compute infrastructure and/or public cloud services, modern generative AI-related techniques and approaches have greatly reduced the burden associated with kicking off AI-based initiatives in house. This is a good development because data locality, security, and governance requirements along with cost considerations are keeping more generative AI deployments on premises. According to Enterprise Strategy Group research, 78% of organizations identified that their organization prefers to run AI applications on premises.²

Dell Technologies is keeping pace with these shifts. It is a leader in enterprise IT storage infrastructure, servers/compute, and HCI. As a key innovation partner with Intel®, Dell is working to simplify organizations' access to infrastructure that can support AI. Recently, Dell has announced multiple AI-centric solutions, along with a collaboration with Hugging Face to better support the deployment of generative AI models on premises.

Of the organizations with a planned hyperconverged deployment in the next six months, 41% expect all or part of that new HCI to support generative AI.

Simplify Infrastructure to Accelerate AI-related Time to Value

Enterprise Strategy Group research highlights both the rise of AI initiatives and the importance of HCI environments in relation to IT modernization. The research shows that 71% of organizations currently have a new data center HCI

¹ Source: Enterprise Strategy Group Complete Survey Results, *Beyond the GenAI Hype: Real-world Investments, Use Cases, and Concerns*, August 2023.

² Source: Enterprise Strategy Group Complete Survey Results, *Understanding Workload, Application, and Data Deployment and Migration Decision-making*, July 2024.

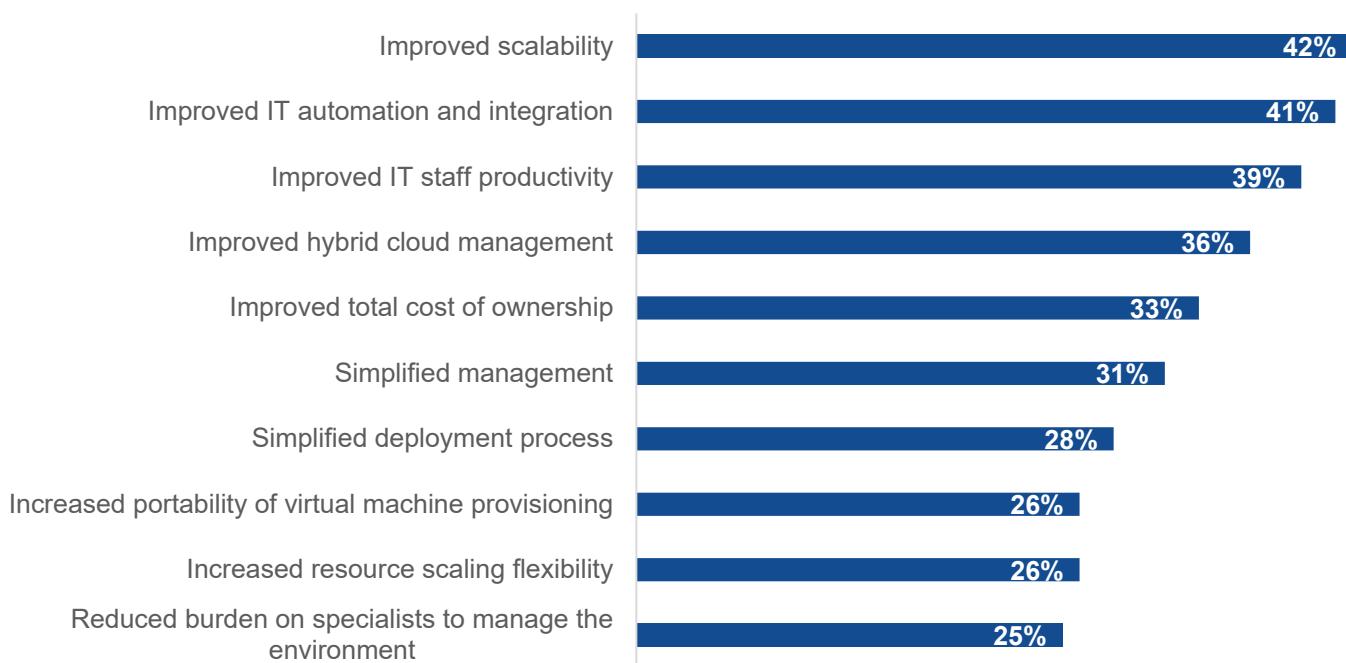
project planned, and among those organizations, 41% expect either part or all of that new HCI environment to support generative AI.³

Organizations often select HCI over more traditional three-tier (server, network, and storage) infrastructure as a means to simplify both the deployment and the management of the application and infrastructure environment. HCI also offers flexible scalability, with the option to start with a smaller initial deployment to reduce the upfront investment and then scale as application demands increase. HCI offers benefits related to supporting generative AI initiatives. Specifically, HCI enables organizations to scale their environments faster and more efficiently. Among survey respondents, 82% reported that HCI is core to their data center modernization plans.⁴

Some of the benefits IT decision-makers reported seeing from their HCI environments include improved scalability, improved automation, and enhanced IT staff productivity. HCI is also making hybrid cloud management easier for them by simplifying the deployment, hosting, and management of both traditional and cloud-native workloads across data center, edge, and cloud environments, while lowering the total cost of ownership (see Figure 1).⁵

Figure 1. Top 10 Realized Benefits From Using a Hyperconverged Infrastructure

Which of the following benefits, if any, has your organization experienced from its use of HCI? (Percent of respondents, N=170, multiple responses accepted)



Source: Enterprise Strategy Group, a division of TechTarget, Inc.

The popularity and usefulness of HCI also raises the question of what organizations should be looking for in terms of infrastructure intended for AI training and inference. Organizations are at various stages of AI maturity. A critical first step for many of them will center on identifying the most suitable use cases for AI implementation. To achieve maximum ROI with an AI initiative, careful planning and infrastructure selection are essential.

³ Source: Enterprise Strategy Group Research Report, [Navigating the Cloud and AI Revolution: The State of Enterprise Storage and HCI](#), March 2024.

⁴ Ibid.

⁵ Ibid.

Because generative AI supports such a wide variety of use cases and model types—and because it can work within such a wide variety of environment sizes—these organizations will gain an advantage by working with a partner such as Dell Technologies. Dell can help them identify the infrastructure best capable of advancing their unique AI-related goals.

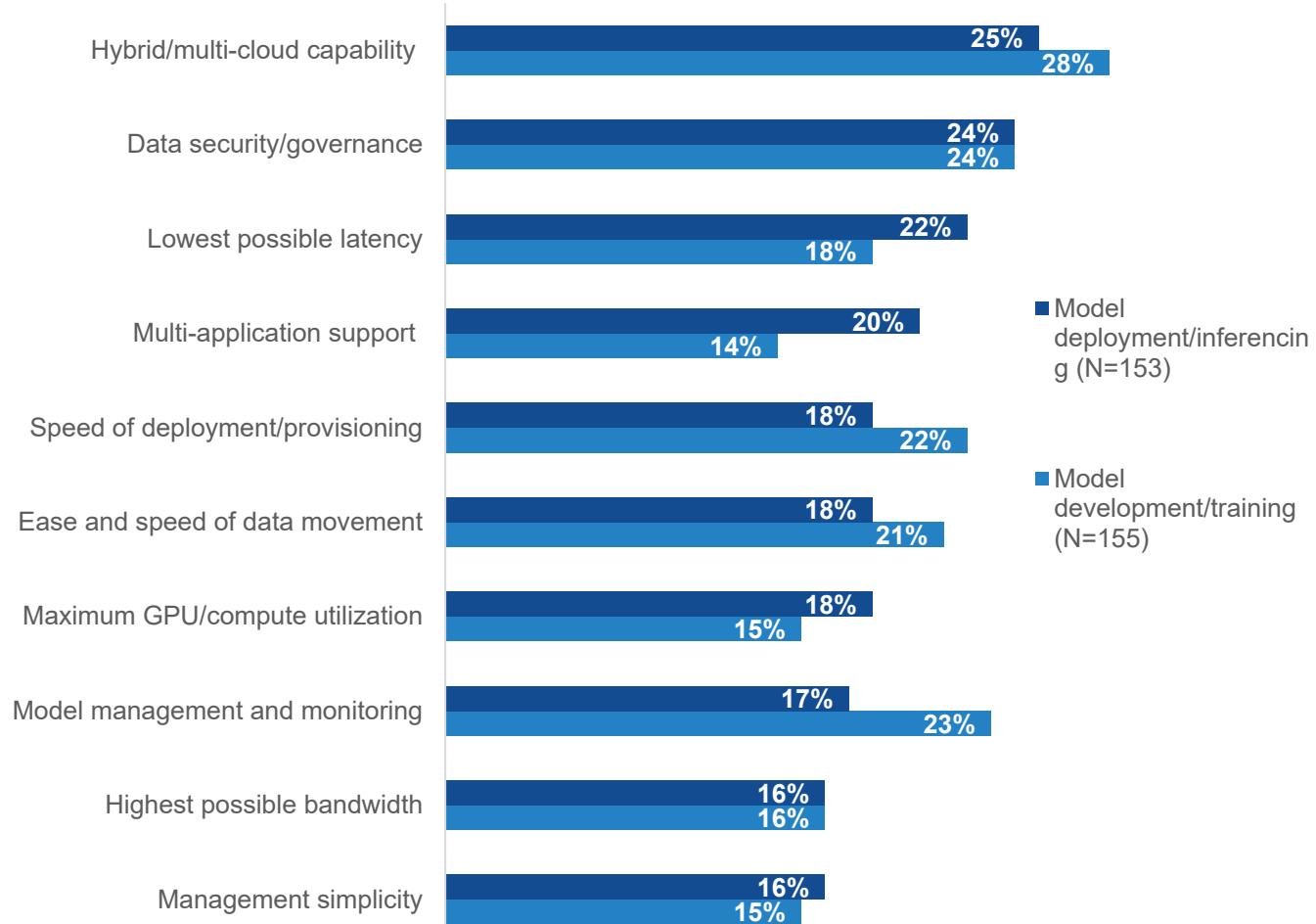
In regard to generative AI workloads, often, a smaller model can be more cost-effective and accessible than might be expected. And keeping that HCI infrastructure on premises is an excellent way to better control the cost and scale of the infrastructure.

Figure 2 highlights criteria that now influence organizations' AI-related infrastructure purchases.⁶ The importance of hybrid cloud capability was the most commonly cited response for supporting both AI inference and AI training.

Similarly, data security and governance are critical considerations for any AI deployment, as is having an ability to speed deployment and provisioning. These types of needs align with the simplicity and control offered by an on-premises hyperconverged solution.

Figure 2. Top Ten Criteria Influencing Data Center Infrastructure Purchases for AI Workloads

Which of the following buying criteria will have the greatest influence on your organization's data center storage infrastructure purchases to support the following AI workloads commonly found in the AI lifecycle?



Source: Enterprise Strategy Group, a division of TechTarget, Inc.

⁶ Ibid.

AI on Dell VxRail

VxRail is an HCI appliance jointly engineered with VMware to optimize support for VMware environments. It is built on Dell PowerEdge server technology, powered by the latest Intel processors. This seamless, integrated platform designed for a turnkey VMware experience provides robust, automated, consistent operations through VxRail HCI system software with thoroughly tested and validated full-stack upgrades.

For AI use cases, VxRail offers versatility to support both CPU accelerator-based and GPU-based AI scenarios. VxRail simplifies AI with built-in Intel AMX accelerators designed for handling matrix operations, which are crucial for AI workloads. The latest-generation Intel CPUs support the Intel AMX accelerators, which are integrated into the CPU and provide accelerated matrix operations out of the box, delivering a simplified approach for supporting environments where AI is part of the workload. VxRail also offers optional validated GPUs for more advanced and more demanding generative AI inference and testing. This flexibility enables organizations to tune VxRail to meet their specific needs, improving simplicity and efficiency, while reducing cost and making AI more accessible.

As a result, VxRail is highly optimized to perform the matrix operations common to heavy-duty AI workloads, while also being optimized to run lighter AI workloads. VxRail also supports multiple AI frameworks, including TensorFlow, PyTorch, and OpenVINO. AI on VxRail is designed for training/inference of lighter-weight AI-infused workloads that do not require a dedicated GPU. Suitable use case examples include image recognition, natural language processing, recommendation systems, media processing, and machine translation.

Dell recommends between 1 and 12 billion parameters to help control or reduce the cost of the AI infrastructure, and it offers validated, add-on accelerator/GPU options for larger model environments that require a larger number of parameters.

The benefits of opting for AI on Dell VxRail include:

- Accelerated time to value and a reduced complexity burden thanks to the preconfigured integration of VxRail's built-in Intel AMX accelerators.
- Optimized/reduced AI infrastructure costs by enabling AI usage without needing add-on accelerators or GPUs.
- The option to purchase validated and integrated NVIDIA GPUs to support heavier AI workloads as necessary.
- Reduced complexity and risks related to AI initiatives as a result of Dell's AI partner and services ecosystem. Dell offers a rich set of AI services and partners to help address the complexities of AI initiatives. The help from these services and partners extends beyond the infrastructure—for instance, they will help organizations identify the right AI use cases and provide assistance with preparing and cleaning the data for training and inference.

Infrastructure Security

VxRail's approach to security follows the zero-trust approach (i.e., never trust, and always verify) from cradle to grave. A zero-trust architecture is complex and requires multiple layers of security that start and stop at different points within the lifecycle, each attending to specific security concerns. According to Dell, this approach provides a formidable level of protection against bad actors, and it is especially important in an AI environment to protect against all AI attack vectors.

Conclusion

With the market exuberance often associated with AI initiatives, too many organizations are overspending on the initial infrastructure. They are justifying those expenses by citing that time to value outweighs optimization. However, for a wide variety of enterprise AI use cases, initial success can be achieved with more reasonable

investments. Right-sizing AI infrastructure is often essential to ensuring a positive ROI. And achieving a positive return on early investments in AI is essential to finding long-term success.

Infrastructure solutions such as VxRail offer an opportunity to reduce the upfront infrastructure spending for AI initiatives and right-size the investment to help ensure a stronger return, while offering the simplicity to accelerate time to value for AI initiatives. Often, the secret to long-term success with AI is starting with smaller projects and seeing positive value early, then building on that success with subsequent additional projects.

Why AI on VxRail? Organizations that are using the VxRail portfolio already will find it a simple process to get started achieving value from AI. They'll have flexibility wherever they are in their AI journey, with validated GPUs that they can use later, as needed, to widen the scope of AI training and inference projects. There are also the extremely important security benefits to consider.

Running AI on VxRail is part of a larger AI-centric shift at Dell Technologies overall. Given Dell's breadth of portfolio and expertise helping enterprise organizations identify AI use cases, right-size their infrastructures, and achieve success with AI, the time is right to talk to Dell as you begin your AI journey.

©TechTarget, Inc. or its subsidiaries. All rights reserved. TechTarget, and the TechTarget logo, are trademarks or registered trademarks of TechTarget, Inc. and are registered in jurisdictions worldwide. Other product and service names and logos, including for BrightTALK, Xtelligent, and the Enterprise Strategy Group might be trademarks of TechTarget or its subsidiaries. All other trademarks, logos and brand names are the property of their respective owners.

Information contained in this publication has been obtained by sources TechTarget considers to be reliable but is not warranted by TechTarget. This publication may contain opinions of TechTarget, which are subject to change. This publication may include forecasts, projections, and other predictive statements that represent TechTarget's assumptions and expectations in light of currently available information. These forecasts are based on industry trends and involve variables and uncertainties. Consequently, TechTarget makes no warranty as to the accuracy of specific forecasts, projections or predictive statements contained herein.

Any reproduction or redistribution of this publication, in whole or in part, whether in hard-copy format, electronically, or otherwise to persons not authorized to receive it, without the express consent of TechTarget, is in violation of U.S. copyright law and will be subject to an action for civil damages and, if applicable, criminal prosecution. Should you have any questions, please contact Client Relations at cr@esg-global.com.

About Enterprise Strategy Group

TechTarget's Enterprise Strategy Group provides focused and actionable market intelligence, demand-side research, analyst advisory services, GTM strategy guidance, solution validations, and custom content supporting enterprise technology buying and selling.

 contact@esg-global.com

 www.esg-global.com