ESG SHOWCASE

Modernize Compute for an AI-driven Future with Dell Servers and NVIDIA

Date: March 2022 **Author:** Mike Leone, Senior Analyst

ABSTRACT: The continued evolution and adoption of AI applications is putting pressure on IT. They recognize that force-fitting legacy hardware with an advanced component like a powerful GPU is simply not enough to satisfy all workload and stakeholder requirements. And this scenario is pushing IT to look to a modern compute platform that can better enable organizations to democratize and scale AI.

AI Adoption Trends

Whether turning to AI to provide more predictive insights into future scenarios/outcomes or developing AI-based products and services to capture new revenue opportunities, businesses are continuing to emphasize the importance of AI adoption as a game-changer for modern businesses. As organizations continue to rely on specialized infrastructure to power their AI workloads, rising costs and longer time-to-value are driving organizations to look for a holistic AI strategy based on standardized components like optimized servers and GPUs. Simply put, organizations want smarter, faster, and more cost-effective ways to gain value from their data with AI. Today nearly half of organizations (45%) currently have AI in production running on specialized infrastructure to handle their AI initiatives. Additionally, more than half are either currently in the piloting stage with AI projects (21%), in the proof-of-concept stage with AI projects (18%), or planning to develop AI initiatives within the next 12 months (16%).[1]

62% of organizations plan to increase their year-over-year spend on AI, including investments in people, process, and technology.

Regardless of where organizations are in their adoption of AI, expected spend continues to rise as organizations look to prioritize a holistic AI strategy with a modern compute stack. Whether just getting started or looking to scale AI, 62% of organizations plan to increase their YoY spend on AI.[2] And the investments are being made to help organizations overcome skills gaps, infrastructure shortcomings, and improve time-to-value while ensuring AI stakeholder requirements are met. As an example, with a modern compute platform, data scientists and AI researchers could spend less time integrating, troubleshooting, and supporting hardware and software, and more time confidently utilizing right-sized resources collaboratively across their end-to-end workflows, from development to training at scale.

[1] Source: ESG Complete Survey Results, *Supporting AI/ML Initiatives with a Modern Infrastructure Stack*, May 2021.
[2] Source: ESG Complete Survey Results: *2022 Technology Spending Intentions*, Nov 2021.

AI Transformation Driving Infrastructure Modernization

As businesses look to rapidly adopt AI and transform by democratizing AI at scale with modern compute platforms and integrated AI frameworks, infrastructure change is all but imperative. Simply put, legacy infrastructure cannot effectively satisfy the unique demands of common workloads found throughout the AI lifecycle. Components like traditional CPUs and even commodity GPUs are not optimized to deliver the performance required for a diverse set of AI workloads. In fact, ESG research shows that nearly 1 in 3 organizations (29%) state that one of their greatest barriers to AI success is the need for better IT infrastructure capabilities.[3] Between inadequate processing power, storage capacity, networking capabilities, and an inability to properly manage resource allocation, infrastructure readiness is proving to be a significant issue in keeping up with the performance and concurrency demands of diverse AI workloads that include data analysis and experimentation, feature engineering, model training, model serving, and inference within a deployed application. This is a major reason why 86% of organizations identified at least one of the following areas as a weak link in their AI infrastructure stack: GPU processing, CPU processing, data storage, networking, resource sharing, or integrated development environments.[4]

86% of organizations identified at least one of the following areas as a weak link in their AI infrastructure stack: GPU processing, CPU processing, data storage, networking, resource sharing, or integrated development environments.

AI Success through Compute Platform Standardization

As businesses look for a fast AI onramp that balances simplicity, performance, scale, reliability, and price, IT, as influenced by the needs of key stakeholders such as data scientists, developers, and line-of-business leaders, is in search of a compute platform that can enable organizations to effectively and optimally scale their AI environments. This is especially important as AI use cases become more complex over time, whether enabling self-driving cars, smart personal assistants, and smarter Web services or shaping innovation across industries with fraud detection and supply chain modernization. As model complexity grows, the last thing organizations want is to make a cost tradeoff sacrificing robust compute, timely results, and improved accuracy due to rising costs. Organizations need dense computational power to support the massively parallel architecture of neural networks, and high-performance storage with ultra-low latency networking to keep compute clusters fed with the right data. New compute offerings are enabling the business to create a new IT infrastructure standard through a viable AI compute platform that supports all AI workloads, from the temporary needs of AI experimentation to the power required to train massive data sets as quickly as possible.

Transforming AI Compute with Dell and NVIDIA

As enterprises race to find the right infrastructure that meets the unique needs of their business while making management and maintenance easy for IT, businesses need to partner with market leaders who not only deliver standard solutions but holistically enable the fastest AI outcomes. Dell and NVIDIA are at the forefront of providing a wide portfolio of foundational solutions, along with strong complementing portfolios and software. To meet the accelerated computing needs of modern AI applications and workloads, NVIDIA-Certified Dell EMC PowerEdge servers deliver the technology and solutions needed to scale AI. And paired with NVIDIA technology such as powerful GPUs and the NVIDIA AI Enterprise software suite, organizations are set up for success with an accelerated computing platform for enterprise and high-performance compute deployments.

[3] Source: ESG Complete Survey Results, *Artificial Intelligence and Machine Learning: Gauging the Value of Infrastructure*, March 2019.
[4] Source: ESG Complete Survey Results, *Supporting AI/ML Initiatives with a Modern Infrastructure Stack*, May 2021.

Dell PowerEdge servers feature PCIe Gen 4.0, which delivers double the throughput performance over previous generations, and up to six accelerators per server to support the most challenging, data-intensive workloads. With PCIe Gen 4.0, these servers can host more NVIDIA GPUs (than previous generation servers) and high-speed NVIDIA networking to allow for greater compute and network acceleration in the same form factor. PCIe Gen 4.0 also enables networking speeds of 200Gb/s, such as the NVIDIA Mellanox ConnectX family of HDR 200Gb/s InfiniBand adapters and 200Gb/s Ethernet NICs as well as the forthcoming NDR 400Gb/s InfiniBand adapter technology. In addition, data transfer rates within the system match the native speed of NVIDIA Ampere architecture GPUs such as the NVIDIA A100 Tensor Core GPU. These platform enhancements enable enterprises to run accelerated applications with even better performance and at data center scale.

Dell's NVIDIA-Certified Systems support the NVIDIA AI Enterprise suite of AI and data analytics software that is optimized, certified, and supported by NVIDIA to run on VMware vSphere. The software suite includes several technologies and software from NVIDIA that simplify the AI lifecycle, from rapid deployment and management to scaling AI workloads in a hybrid AI architecture. The Dell EMC PowerEdge lineup features several servers that have passed this certification and are perfectly suited for enterprises that wish to run this new software along with their existing enterprise applications. Dell also offers NVIDIA accelerator-optimized servers that are purpose-built for AI and power the latest Dell Technologies Validated Design for AI and Data Analytics, making it easier to run AI, analytics, and advanced computing workloads on one system while significantly increasing HPC and machine learning performance.

Between the latest CPU technologies from AMD and Intel, along with the power of NVIDIA GPUs, PowerEdge servers deliver a modern compute approach needed for customers' most critical AI workloads and applications. Enterprises gain peace of mind knowing they are enabled to modernize their data centers and scale the use of AI throughout the business with a next-generation computing architecture built for the future.

The Bigger Truth

Between distributed data sets, data gravity, operational silos, and the need for access to cost-effective infrastructure with dense computational power, IT needs help in addressing current and future demands from key AI stakeholders (i.e., data scientists, developers, line-of-business leaders, etc.). As IT teams look for solutions to overcome barriers to adopting AI at scale, they recognize the shortcomings of legacy infrastructure being unable to deliver the right level of performance and scale to satisfy diverse AI workloads. Looking to the future, IT must ensure the entire organization has access to powerful and reliable infrastructure that is tightly integrated with next-generation components to rapidly scale the use of AI throughout the business.

Together, Dell and NVIDIA are aligned with meeting customers where they are in their AI journeys. Whether an organization is just getting started and simply looking for a fast AI on-ramp with proven technology and integrations to experiment or is in need of a robust solution that delivers unprecedented performance and ultra-low latency for complex, deep-learning model training, Dell and NVIDIA drive innovation and solutions that customers can trust for their current and future AI initiatives.

Enterprise Strategy Group is an integrated technology analysis, research, and strategy firm that provides market intelligence, actionable insight, and go-to-market content services to the global IT community.

www.esg-global.com contact@esg-global.com 508.482.0188