DELL Technologies

Technical Whitepaper

The Buzz of Artificial Intelligence — AI Now and in the Future

Executive summary

It all started with the question, “Can machines think?” The idea goes back to the very first computer builders and programmers. The term *artificial intelligence* (AI) was coined in 1956 at Dartmouth College in New Hampshire.

AI is an extensive branch of computing related to building smart machines capable of executing tasks that generally require human intelligence. While AI’s journey of progress has been slow and unpredictable, recent advances are thanks to increased data volumes, algorithmic advances and quantum leaps in computing power and storage.

AI systems have created a multi-billion-dollar industry with sustained growth projected, but this has also created the need to make AI systems trustworthy and secure. AI and its associated models —such as machine learning (ML) and deep learning (DL) — have the potential to create a paradigm shift in virtually every sector of the tech industry and for substantial innovation and economic growth.

Organizations embrace AI for different use cases:

- Human resources uses AI for talent acquisition.
- IT uses AI for cybersecurity.
- Manufacturing uses AI for smart operations.
- Healthcare uses AI for clinical diagnosis and AI-assisted robotic surgery.

This widespread adoption of AI across industries requires seamless integration of AI capabilities into data center operations.

Introduction

AI and ML are among the top trending technologies forecasted by many research groups that will revolutionize the way organizations work. Technology has helped organizations in data capture and storage for analytics. In the digital transformation era, success is based on knowing your data and using analytics to discover insights that these massive volumes of data can provide and making data-centric decisions. As we have accumulated more and more of this ever-growing data from varied sources, the fundamental decision-making tasks or intelligence has mainly been left with humans. AI and ML are all about letting machines make most of these decisions for us.

Definition

AI is a wide-ranging field of computer science concerned with building smart machines capable of performing tasks that typically require human intelligence, such as decision making, problem solving and learning. Simply put, AI is intelligence exhibited by machines, for example, getting a machine to sort of learn, and to understand what it learned to make predictions.

ML is the field of study that gives computers the capability to learn from data and automatically improve through experience. An associated computing model of AI, ML deals with the creation of intelligent machines that can simulate human thinking, capability, and behavior and learn and develop their own programs without being explicitly programmed.

AI can do lot of things, but it can't do everything.

AI is intelligence programmed into machines or computers that includes ML, DL, neural networks and other associated computing models and works by using algorithms, historical data along with constraints and other aspects that help create a propensity model that channel toward thinking, perception and action.

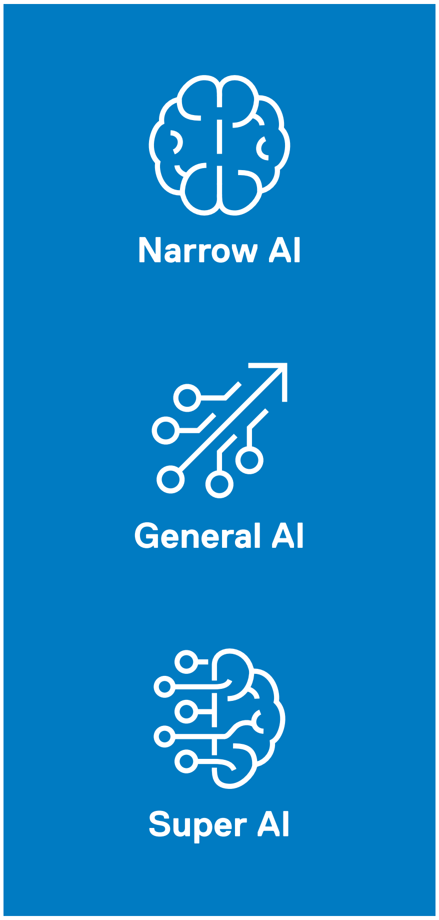


Figure 1 — Three categories of machine learning

AI works by combining vast volumes of data using iterative processing and superior algorithms that make it possible for systems to learn automatically from patterns or features in the data. The earlier approaches to creating intelligent machines were more human-centric, where a human expert created precise rules for the system to follow step by step to decide how to respond to a given situation. The rules were expressed in simple algorithms, such as "do… while" and "if… then… else" formats, and said to keep the human in the loop because the decision-making process is closely aligned to how human experts make decisions.

While these types of machines can perform tasks autonomously, they can do only what they are told and can improve only with direct human intervention. With recent advances in technology and a tremendous increase in the volume and quality of data and intelligent algorithms, ML helps improve learning and performance without humans directly encoding their expertise. These types of data-driven learning based on algorithms find their own ways of identifying patterns and implement what they learn to make statements about data. There are different approaches to ML that are suited for different tasks and scenarios with different implications. These approaches can be categorized into the following concepts.

- **Narrow AI**, also known as basic AI, specializes in one area and performs singular tasks by replicating human intelligence.
- **General AI**, also referred to as strong AI, a computer that is as smart and performs much like a human being, it can apply that intelligence to solve any problem.
- **Super AI** exceeds human intelligence and is far more sophisticated than any other AI or even a human brain in practically every field.

AI works with large amounts of data which are first combined with fast, iterative processing and smart algorithms to allow the system to learn from the patterns within the data. This way, the system can deliver accurate or close to accurate outputs.

AI is a vast subject that involves advanced and complex processes, and the field includes many theories, methods and technologies. To understand how AI actually works, one needs to dive deep into the various subdomains of AI and understand how those subdomains could be applied into the various fields of the industry.

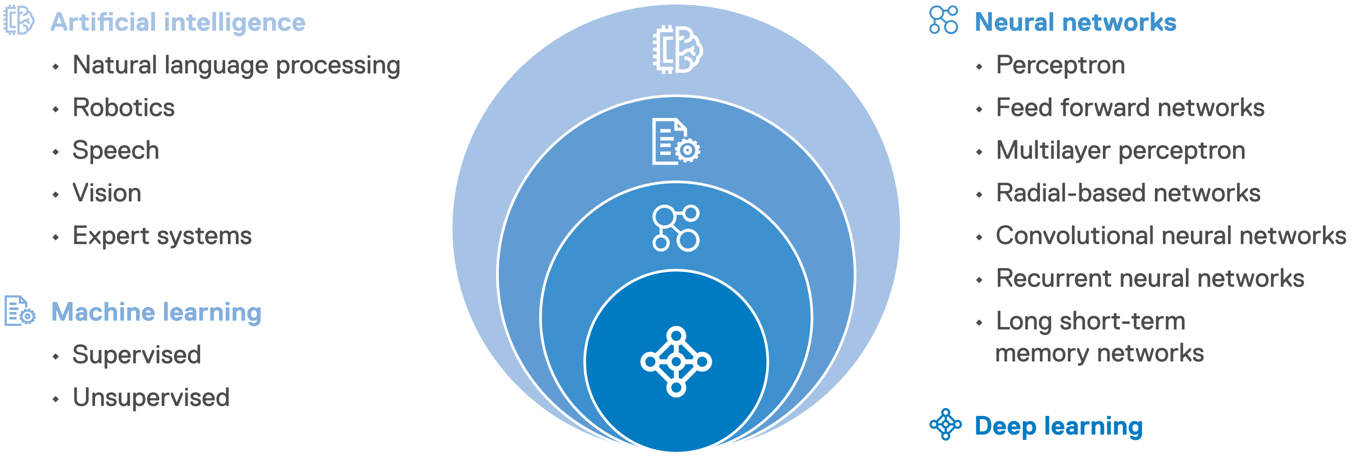


Figure 2 — AI and its subdomains

Machine learning: ML teaches a machine how to make inferences and decisions based on its own examples and on previous experience. It identifies patterns and analyzes historical data to infer the meaning of these data points to reach a possible conclusion without having to involve a human expert. Reaching conclusions driven by data saves time and helps make better decisions.

ML is further distinguished by three types of learning: supervised learning, unsupervised learning and reinforced learning. In supervised learning, the AI system already knows the correct answers and must only adapt the algorithms so that the answers can be derived as precisely as possible from the existing data set. In unsupervised learning, the AI system does not have predefined target values and must recognize similarities and thus patterns in the data independently. Finally, reinforced learning uses the process of trial and error. Reward and punishment signal positive and negative behavior, and the goal is to find a suitable action model that maximizes the total cumulative reward of the agent. This technique has many positives, such as solving various complicated problems which cannot be solved with conventional techniques and lessening the potential for repeat mistakes.

Deep learning: DL is an ML technique, where a large amount of data is analyzed and the algorithm tends to perform the tasks repeatedly, editing and modifying a little to improve the outcome. It teaches a machine to classify, infer and predict results by processing inputs through multiple layers.

Neural networks: Neural networks work on principles similar to biological neural (brain) cells. This series of algorithms captures the relationship between various underlying variables and process the data as a human brain does. Neural networks are one of the most important tools in ML to find patterns within the data, which are far too complex for a human to figure out and teach to the machine.

Natural language processing: Natural language processing (NLP) is the science of reading, understanding and interpreting a language by a machine. It means developing methods that help humans communicate with machines using natural languages, like English. Once a machine understands what the user intends to communicate, it responds accordingly.

Computer vision: Computer vision algorithms allow computers to see, recognize and process images in the same way human vision does. This helps the machine understand what it sees, analyze, classify and learn from a set of images, to make a better output decision based on previous observations.

Cognitive computing: The goal of cognitive computing is to emulate the human thought process by analyzing text, speech, images and objects like a human does and try to give the desired output. Using self-learning algorithms, pattern recognition, neural networks and NLP, a machine can mimic the human way of thinking and simulate the human cognition process.

Benefits of AI

Though many day-to-day tasks are automated and use some kind of intelligence, AI still sounds like the stuff of far-off future. But AI has come a long way from being science-fiction or future technology to becoming a reality. Thanks to rapid advancements in technology, machines powered by AI specialize in precise, repetitive tasks such as basic analysis, subtle judgements, data management and problem solving. AI holds the potential to create a world of endless possibilities.

Some areas that use AI are as follows:

- Data management
- Improving manufacturing
- Aiding cybersecurity
- Accurately diagnosing diseases
- Improving education
- Reducing job hazards at the workplace
- Predicting natural disasters
- Preserving environmental resources
- Aiding the criminal justice system

The benefits of AI in these areas are many because AI machines are highly efficient, execute quickly, save time and money by automating routine processes and tasks, increase revenue by identifying and maximizing sales opportunities, and make faster business decisions based on cognitive technology outputs. AI machines are available 24x7, with zero scope for error, better forecasts, increased productivity, operational efficiencies and optimal distribution of resources.

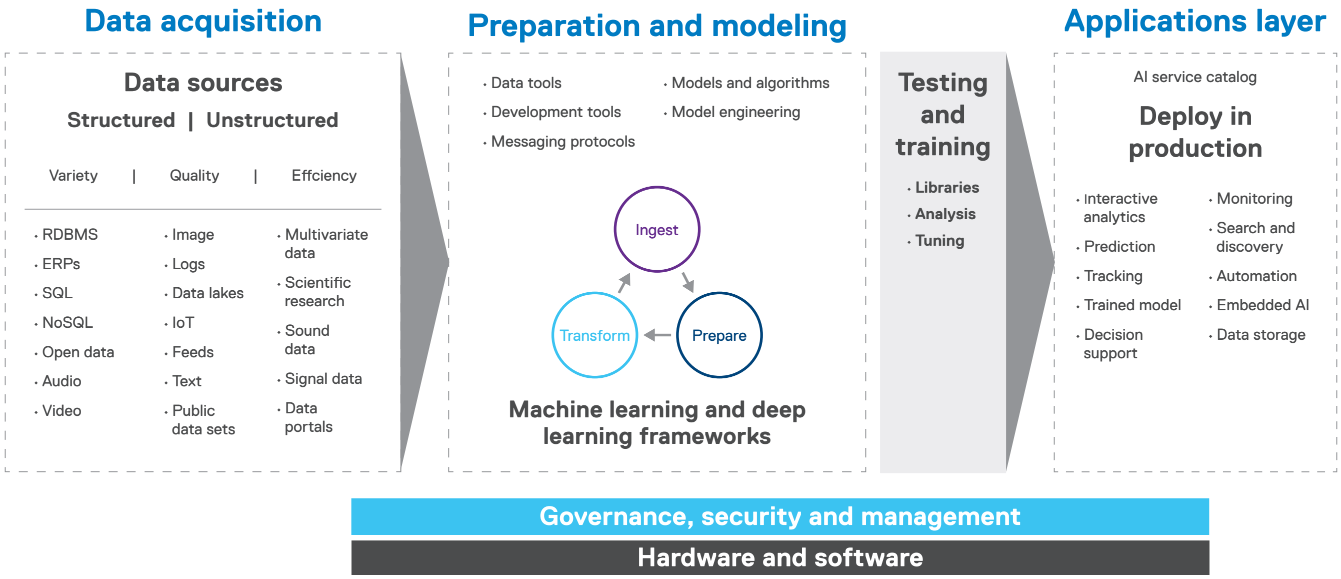


Figure 3 — Reference framework

Hardware and software for AI

AI has moved beyond experimental and preproduction phases toward real life adoption with tremendous growth in recent years. Though most AI projects begin with standard hardware, a challenge hindering its potential is the demand for high performance computing (HPC) resources. The need for significant computing and storage hardware that supports the complex production level functions goes beyond the standard hardware resources available today. Choosing the right hardware requires an extensive understanding on the outcome expected, what its actual hardware requirements are and how it differs from general computing hardware.

Nearly all digital computers are based on the Von Neumann architecture, consisting of a central processing unit (CPU), primary memory and input/output (I/O) devices. This architecture provides significant benefits in terms of modularity, individual components upgrade and use advanced peripherals. However, all these components are interfaced with a bus to transfer data between the CPU, memory and I/O devices. No matter how fast the bus performs its task, a bottleneck that reduces speed is always a possibility.

To overcome this challenge there are many types of hardware accelerators being built, using graphics processing units (GPUs), vision processing units (VPUs), application-specific integrated circuits (ASICs), and field programming gate arrays (FPGAs) to accelerate computationally intensive tasks. These specialized accelerators provide high performance and the required accuracy to accelerate AI apps and its associated computing models, such as ML, DL, neural networks and other data-intensive tasks.

Regular computing systems can handle basic AI workloads to a large extent, but when we dive deeper, involving multiple large data sets and scalable neural networks, the computing capacity and density grows. For that, a CPU-based computing may not be sufficient. This is where hardware accelerators come into the picture, for example, a GPU with thousands of cores and high memory bandwidth can process multiple computations with almost 100% accuracy and efficiency simultaneously. It is for this reason the GPUs are well suited for high performance computation of AI and its associated computing models.

Servers with hardware accelerators

Servers with hardware accelerators or GPUs transform the way computing is done for AI, ML and DL. Varied GPU configuration support helps organizations choose the right GPU and server combination to best enable breakthrough performance for the intended workload.

GPUs are built differently than CPUs, to process fast and accurate floating-point arithmetic. Though GPUs have slower core speeds than CPUs, they compensate with thousands of cores running in parallel. The number of calculations involved are extremely high and not suitable for the algorithms used in most methods of AI, ML and DL — especially DL, which uses neural networks. The algorithms are mostly offloaded from the CPUs to GPUs and are easier to process in parallel with GPUs than traditional CPUs, while the CPU handles the main sequential processes. Such GPU-to-CPU strategies are critical to delivering better services that cater to accelerated performances.

AI and its associated computing models, ML and DL, generally require fast computational resources and specialized computing environments to run more processes and algorithms simultaneously. This is where hardware accelerators enter the game with their thousands of cores that help AI analyze large sets of data repeatedly, learn, gain insights and participate in decision making at speed.

Currently, GPUs and FPGAs are the best bet for cost-effective hardware accelerators compared to other technologies that help in training the system for ML or DL. Compared to CPUs, GPUs are cheaper and offer higher performance, and today we have servers that support from one to tens of single-width or double-width GPUs and FPGAs. Many aspects need to be considered when choosing the right GPU/FPGA — for example, number of cores, compute power, single root vs. dual root, memory size and bandwidth, and PCIe bus lanes — for the type of application at hand.

$

The global AI market has an approximate market value of USD $40 billion and is expected to grow.

Market value — Why AI?

The global AI market has an approximate market value of USD $40 billion and is expected to grow at a compound annual growth rate (CAGR) of 38–45% from 2020 to 2027 according to most of the research groups. Studies also indicate that 35–42% of organizations have implemented AI in some form and that 90% of customer interactions will be powered by AI. The continuous research and innovation by the technology giants are driving the adoption of advanced technologies in industry verticals — the driving factors being rapid increase in the number of connected devices, growing adoption of the Internet of Things (IoT), augmenting customer experience, rapid increase in requirements for cloud-based applications and services, the growing need to analyze and interpret large amounts of data, and the increase in the adoption of intelligent virtual applications.

AI has dramatically changed the business landscape, and its applications are endless. AI technology and applications have significantly evolved over the past few years and are applied in many different sectors and industries to maximize output on the operational front.

Some of the applications and use cases of AI are as follows:

- Agriculture
- Artificial creativity
- Automobiles
- Banking and finance
- Chatbots
- Customer service
- Enhanced images
- Gaming
- Healthcare
- Intelligence and law
- Logistics and supply chain
- Manufacturing
- Marketing
- Navigation
- Personalized online shopping
- Research and development
- Retail
- Smart cars
- Social media
- Space exploration
- Surveillance
- Virtual assistance
- Workplace communication

As listed above, AI and its subdomains are prevalent in every aspect of the business and technology, thereby revolutionizing industries with its applications and helping solve complex problems.

The below figure shows a pattern of how industries are adopting AI, ML and DL.

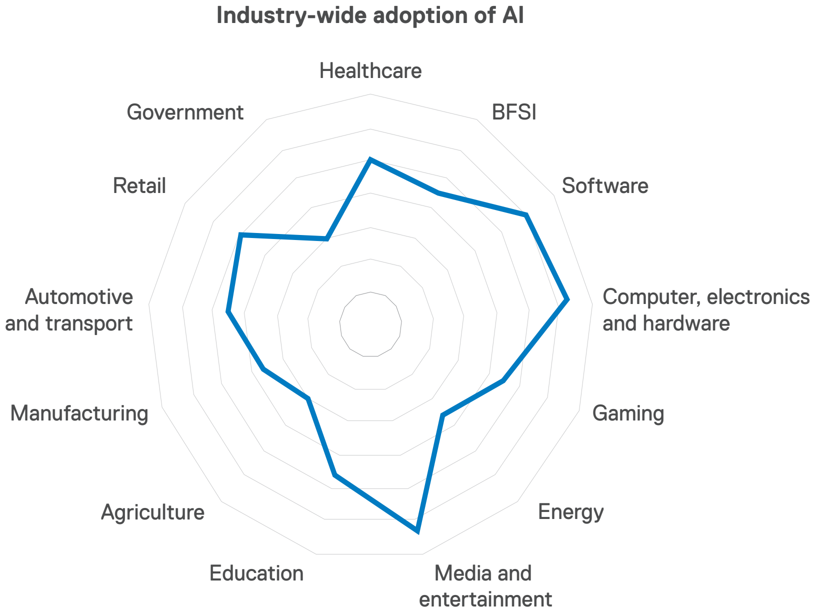


Figure 4 — Industry usage probability radar chart

Conclusion

AI is the future and is attracting growing amounts of corporate investment. As the technology develops and the potential increases, AI is being adopted by public and private organizations alike, is already impacting the way enterprises engage with customers, and will fundamentally redefine how we work. This paper discussed the concepts that make up one of the most powerful technologies invented, offering the ability to learn autonomously and to exceed human capabilities. What most of us think of AI goes back to the original question: Can machines think? The answer depends on training and inference through ML, DL and the other associated models. To help determine the right solution for an organization, it is helpful to have a good understanding of AI, ML, DL and how they use the volume and type of data.

Dell Technologies and AI, ML and DL

Unlike the science-fiction dystopias that originally made AI famous, today's AI is generally adopted as a supporting tool adept at processing and analyzing vast amounts of data for learning insights and actions. As organizations pursue business solutions based on AI and its associated computing models (ML and DL), Dell Technologies is at the forefront, with a portfolio for AI providing the cutting-edge technologies that make tomorrow possible, today.

Dell Technologies has a vast portfolio of solutions to help organizations get their AI applications and projects up and running without delay. From accelerators such as GPUs, FPGAs and intelligence processing units (IPUs) that boost AI-driven workloads and analytics on Dell PowerEdge servers, to Validated Designs that deliver business results across industry segments, Dell offers one of the industry's richest portfolio for AI solutions, including IoT gateways, workstations, servers and storage.

Dell Technologies Validated Designs for AI

Dell Technologies is at the forefront of AI, providing the technology that makes tomorrow possible, today. Dell Technologies has invested to create a portfolio of Validated Designs for AI, simplifying the IT infrastructure to provide faster, deeper insights, no matter the industry.

Validated Designs are available for many industries, including:

- Healthcare
- Financial services
- Government
- Media and entertainment
- Energy
- Transportation
- Manufacturing
- Retail

Deep dive into AI for healthcare

AI has the potential to transform how healthcare is delivered. From perennial diseases, such as diabetes, heart disease, tuberculosis, coronavirus and cancer, to risk assessment, there are nearly endless opportunities to leverage technology to deploy more precise, efficient and impactful interventions at exactly the right moment in a patient's care. As the volume of patient data grows, researchers, doctors and scientists infuse this data into the AI systems. The AI algorithms can review, interpret and even suggest solutions to complex medical problems that help doctors make better decisions, create personalized medicine plans from complex data sets. This data can be further analyzed and used for information management, medical diagnostics, chronic care management, drug discovery, clinical trials, personalized medicine, clinical decision support, improved patient outcomes and so on.

Some of the use cases for AI in healthcare are as follows:

- Preliminary diagnosis
- Precision medicine
- AI image analysis and diagnosis
- Dosage error reduction
- Robot-assisted surgery
- Virtual nursing assistants
- Clinical trials
- Security

Though there are many benefits to AI and its associated models in the healthcare sector, it has drawbacks in utilizing these technologies. Some of them are lack of personal involvement, rising unemployment among healthcare workers, possible defective diagnosis, data privacy concerns, compliance to regulations and cybersecurity are some of the hindrances to utilizing these technologies.

Despite these challenges and risks, AI in healthcare has tremendous benefits that can boost AI-driven healthcare ecosystem adoption worldwide.

Dell Technologies Validated Design for AI in Healthcare

Dell Technologies Validated Designs for AI include everything an organization needs to accelerate their AI initiatives with pre-designed and pre-validated hardware and software stacks optimized to shorten the time to architect a new solution.

Validated Designs for AI with VMware and NVIDIA deliver unprecedented performance at scale with Dell PowerEdge servers; NVIDIA® accelerators; AMD®, Intel®, Graphcore® and Xilinx® processors; Dell scale-up, scale-out, all-flash, file, object, and software-defined storage; and Dell Technologies edge, core and cloud networking solutions — combined with VMware® vSphere® end-to-end AI/ML solutions.

Dell Technologies is the first to deliver the latest virtualization capabilities for AI and advanced computing in VMware environments. With the ability to virtualize AI solutions with VMware, IT can quickly provision hardware as needed, speed up initial deployment and save time with simpler centralized management and security.

Dell Technologies is the first to deliver the latest virtualization capabilities for AI and advanced computing in VMware environments.

Dell Technologies powers AI

Dell Technologies innovative solutions for AI help customers with the powerful and secure Dell PowerEdge server portfolio. PowerEdge servers provide the power needed to gain and act on real-time insights from data wherever it resides, offers greater IT efficiency, embraces AI and address the demands of IT from the core data centers to public clouds and edge locations.

The PowerEdge server portfolio provides next-level performance and capabilities to help customers get the most out of their valuable data. A diverse portfolio allows them to bring the compute closer to the data, perform a deeper analysis at a faster pace and empower their AI journey.

PowerEdge servers now feature PCIe Gen 4.0, the fourth major iteration of the PCIe standard. PCIe 4 generates double the throughput per lane at 16GT/s, double that of PCIe Gen 3. These technological enhancements, coupled with the autonomous intelligence of PowerEdge servers, make this the most AI-enabled PowerEdge portfolio to date, empowering organizations to anticipate and more quickly respond to their needs.

Accelerators such as GPUs, FPGAs and IPUs complement and accelerate CPUs, using parallel processing to crunch large volumes of data faster. Some of the use cases for accelerators are in ML, DL, predictive analytics, visualization, modeling and simulation, financial modeling and signal processing.

Organizations across industry segments adopting AI and its associated computing models need accelerated servers to enhance productivity and application performance, optimize operations with fast and powerful analytics. Accelerated data centers can also deliver better economics, providing breakthrough performance with fewer servers, resulting in faster insights and lower costs.

Dell Technologies offers a range of PowerEdge servers with accelerators for AI, and a few are highlighted below:

- Dell PowerEdge C6525, R7525, R6525, R7515 and R6515 Servers with 3rd Generation AMD EPYC™ processors
- Dell DSS 8440 with 2nd Generation Intel Xeon® Scalable processors
- Dell PowerEdge XE8545 servers with 3rd Generation AMD EPYC processors
- Dell PowerEdge C6520, MX750c, R750, R750xa, R650 servers with 3rd Generation Intel Xeon Scalable processors
- Dell PowerEdge R750xs, R650xs, R550, R450 and the ruggedized PowerEdge XR11 and XR12

R750xa

R750xa is optimized to tackle GPU workloads and deliver outstanding performance for demanding and emerging applications.

- Maximizes performance
- Front-to-back air-cooled design
- Supports all GPU cards

R940xa

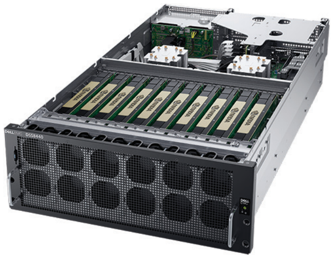
R940xa is optimized to tackle workloads that are compute-intensive, combining up to 4 CPUs, up to 112 cores, with four GPUs in a powerful 1:1 ratio to drive AI, ML and DL workloads.

- Accelerate applications
- Scale dynamically
- Streamline IT operations

DSS 8440

DSS 8440 is specifically designed to reduce time to insight for ML training by providing substantially increased horsepower.

- Scale up GPU density as needed
- Provision GPU virtualization to workloads and users
- Support multi-tenant environments



XE8545

XE8545 delivers optimized CPU and GPU performance for AI and ML training and inferencing by pairing the maximum core count AMD EPYC processors.

- Supercharged AI/ML and HPC performance
- Interconnected 4-way NVLink architecture
- GPU virtualization

vSAN Ready Nodes and VxRail

Many organizations need their AI and analytics infrastructure on-premises to access existing data sets. These data sets are too large and spread out within multiple organizations to be conveniently or cost-effectively transferred to and from the cloud. Additionally, some organizations have the capability to acquire and operate compute resources at much cheaper rates than cloud providers offer.

A recent Moor Insights publication reports on the benefits and limits of cloud hosting for AI and HPC. It states, “Starting with the cloud may make a lot of sense if an organization wants to experiment with AI” due to the services that are available and the ease of acquiring the required compute resources.

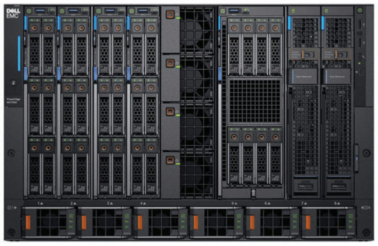
The report continues, “However, many organizations will eventually need significant computing infrastructure for AI and HPC as their applications begin to run at scale. This, along with data transfer and throughput fees, begins to tip the cost balance in favor of building on-premises infrastructure as the organization matures in AI.”

In addition to cost, other important factors that organizations must consider are data gravity and security/privacy concerns. AI workloads need substantial amounts of data from different sources both within and outside the organization. It usually makes more sense to keep the application co-resident with the data sets. When it comes to security, AI projects that involve massive proprietary data sets and security concerns can override the potential ease of use, especially in financial and healthcare markets.

The PowerEdge server portfolio offers flexible designs for optimized application performance. The single-socket server portfolio provides balanced performance and storage capacity for future growth. The two-socket server portfolio brings a mix of features to maximize performance, scale to meet future demands, and adapt to virtually any workload with an optimum balance of compute and memory. The four-socket server portfolio fills the top end with the highest performance and extensive scalability for your applications from in-memory database workloads and HPC to data analytics, AI and GPU database acceleration.

Going a step beyond, the Dell Technologies hyperconverged infrastructure (HCI) portfolio, built on top of PowerEdge servers, offers scalable infrastructure platforms to simplify IT operations and lower risk. Dell VxRail and Dell vSAN Ready Nodes are integrated systems designed with VMware, for VMware, and to enhance VMware-enabled organizations through intelligent lifecycle management. Designed for today’s mission-critical workloads, they have multiple compute, memory, storage, network, and graphics options to cover a wide variety of applications, including AI and ML.

	PowerEdge R740xd	PowerEdge R740	PowerEdge R640
Profiles	All-flash, All-NVMe™, hybrid	All-flash, hybrid	All-flash, All-NVMe, hybrid
Processor	Up to two Intel Xeon Scalable processors, up to 28 cores per processor	Up to two Intel Xeon Scalable processors, up to 28 cores per processor	Up to two Intel Xeon Scalable processors, up to 28 cores per processor
Memory	Up to 24 DIMM slots, supports RDIMM/LRDIMM, speeds up to 3,200MT/s	Up to 24 DIMM slots, supports RDIMM/LRDIMM, speeds up to 3,200MT/s	Up to 24 DIMM slots, supports RDIMM/LRDIMM, speeds up to 3,200MT/s
Cache	SAS/SATA SSDs, NVMe	SAS/SATA SSDs	SAS/SATA SSDs, NVMe
Capacity	SAS HDDs, SAS/SATA SSDs	SAS HDDs, SAS/SATA SSDs	SAS HDDs, SAS/SATA SSDs
Controller	HBA330	HBA330	HBA330
Network	10GB or more network card	10GB or more network card	10GB or more network card
Boot device	Boot Optimized Storage Subsystem: 2x M.2 SSDs 120GB or 240GB	Boot Optimized Storage Subsystem: 2x M.2 SSDs 120GB or 240GB	Boot Optimized Storage Subsystem: 2x M.2 SSDs 120GB or 240GB



	PowerEdge R440	PowerEdge MX740c	PowerEdge C6420
Profiles	All-flash, hybrid	All-flash, All-NVMe, hybrid	All-flash, hybrid
Processor	Up to two Intel Xeon Scalable processors, up to 28 cores per processor	Up to two Intel Xeon Scalable processors, up to 28 cores per processor	Up to two Intel Xeon Scalable processors, up to 28 cores per processor
Memory	Up to 24 DIMM slots, supports RDIMM/LRDIMM, speeds up to 3,200MT/s	Up to 24 DIMM slots, supports RDIMM/LRDIMM, speeds up to 3,200MT/s	Up to 16 DIMM slots, supports RDIMM/LRDIMM, speeds up to 3,200MT/s
Cache	SAS/SATA SSDs	SAS/SATA SSDs	SAS/SATA SSDs
Capacity	SAS HDDs, SAS/SATA SSDs	SAS HDDs, SAS/SATA SSDs	SAS HDDs, SAS/SATA SSDs
Controller	HBA330	HBA330	HBA330
Network	10GB or more network card	10GB or more network card	10GB or more network card
Boot device	Boot Optimized Storage Subsystem: 2x M.2 SSDs 120GB or 240GB	Boot Optimized Storage Subsystem: 2x M.2 SSDs 120GB or 240GB	Boot Optimized Storage Subsystem: 2x M.2 SSDs 120GB or 240GB

Figure 5 — Dell vSAN Ready Nodes portfolio

Intel offers libraries and frameworks optimized for Intel Xeon Scalable processors. These include TensorFlow™, MXNet, PaddlePaddle, Caffe and PyTorch®. They enhance DL performance using software optimizations. The Intel distribution for Python®, for example, accelerates AI-related Python libraries such as NumPy, SciPy, and scikit-learn with integrated Intel performance libraries such as Intel MKL to deliver faster AI training and inferencing. These libraries and frameworks enable enterprises to quickly deploy a reliable, powerful, comprehensive AI development software platform.

VMware in the AI era

VMware introduced the world to virtualization on the x86 platform. This allowed IT to make efficient use of server capacity by creating multiple independent virtual servers on a single physical server. Today we have both virtual machines (VMs) and containers running side by side, and VMware is still making it easier to develop and deploy AI and its supporting computing models on their virtual platform.

For organizations on the path to gain business value through AI, ML and DL, one of the most important things is access to the best hardware that supports complex functions and extensive computations, with access to a variety of CPUs, GPUs, IPUs, FPGAs and ASICs. Hence it becomes a little complex and confusing when considering hardware infrastructure for these projects.

With VMware, AI can be easily managed with the same VMware flexibility as other applications, making it easier than ever to develop, deploy and manage diverse workloads anywhere. VMware and NVIDIA have partnered to unlock the power of AI for every business by delivering an end-to-end enterprise platform optimized for AI workloads. VMware vSphere delivers enterprise virtualization for traditional and emerging workloads, including GPU-powered AI with powerful support for the latest NVIDIA A100 and A30 Tensor Core GPUs.

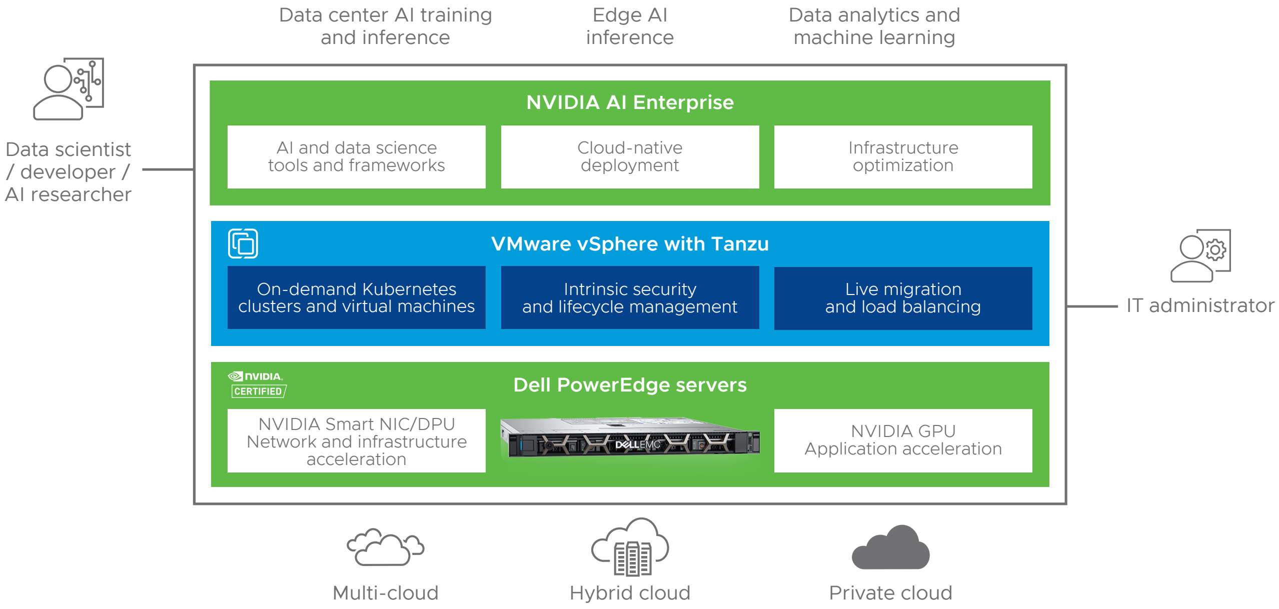

Data center AI training and inference
Edge AI inference
Data analytics and machine learning
Data scientist / developer / AI researcher
NVIDIA AI Enterprise
AI and data science tools and frameworks
Cloud-native deployment
Infrastructure optimization
VMware vSphere with Tanzu
On-demand Kubernetes clusters and virtual machines
Intrinsic security and lifecycle management
Live migration and load balancing
IT administrator
NVIDIA CERTIFIED
Dell PowerEdge servers
NVIDIA Smart NIC/DPU Network and infrastructure acceleration
DELL EMC
NVIDIA GPU Application acceleration
Multi-cloud
Hybrid cloud
Private cloud

Figure 6 — End-to-end enterprise platform for AI workloads

VMware has pioneered compute, storage and network virtualization, reshaping yesterday's bare-metal data centers into modern software-defined data centers (SDDC). Despite this platform availability, many ML workloads are still run on bare-metal systems. DL workloads are so compute-intensive that they require compute accelerators like NVIDIA GPUs and software optimized for AI; however, many accelerators are not yet fully virtualized. Deploying unvirtualized accelerators makes such systems difficult to manage when deployed at scale in data centers. VMware's collaboration with NVIDIA brings virtualized GPUs to data centers, allowing data center operators to leverage the many benefits of virtualization.

VMware vSphere

VMware vSphere's hypervisor platform delivers the rich features of virtualization for GPU-powered AI workloads that is the need of organizations who depend on GPU-powered workloads. vSphere optimizes performance, increases availability, tightens security, and creates an agile, efficient, resilient and intrinsically secure infrastructure platform that supports AI workloads.

VMware vSphere supports the following:

- Support for the latest generation of GPUs from NVIDIA, based on the NVIDIA Ampere architecture, including the NVIDIA A100 Tensor Core GPU.
- Support for address translation services (ATS), for enhanced peer-to-peer performance between NVIDIA NICs/GPUs.
- Support for the latest spatial partitioning-based NVIDIA multi-instance GPUs (MIGs):
  - VMware vSphere is the only virtualization platform that enables live migration (using vMotion®) for NVIDIA MIG vGPU-powered VMs, simplifying infrastructure maintenance such as consolidation, expansion, or upgrades, and enabling non-disruptive operations.
  - With the VMware vSphere Distributed Resource Scheduler™ (DRS), vSphere provides automatic initial workload placement for AI infrastructure at scale for optimal resource consumption and avoiding performance bottlenecks.

VMware vSphere with Tanzu

vSphere with Tanzu™ provides a platform for running Kubernetes workloads natively on the VMware ESXi™ hosts, side by side with virtual machines. vSphere with Tanzu supports deployment of compute-intensive workloads such as AI and ML applications that require the use of GPU accelerators, on Tanzu Kubernetes clusters provisioned by the Tanzu Kubernetes Grid service.

NVIDIA vGPU

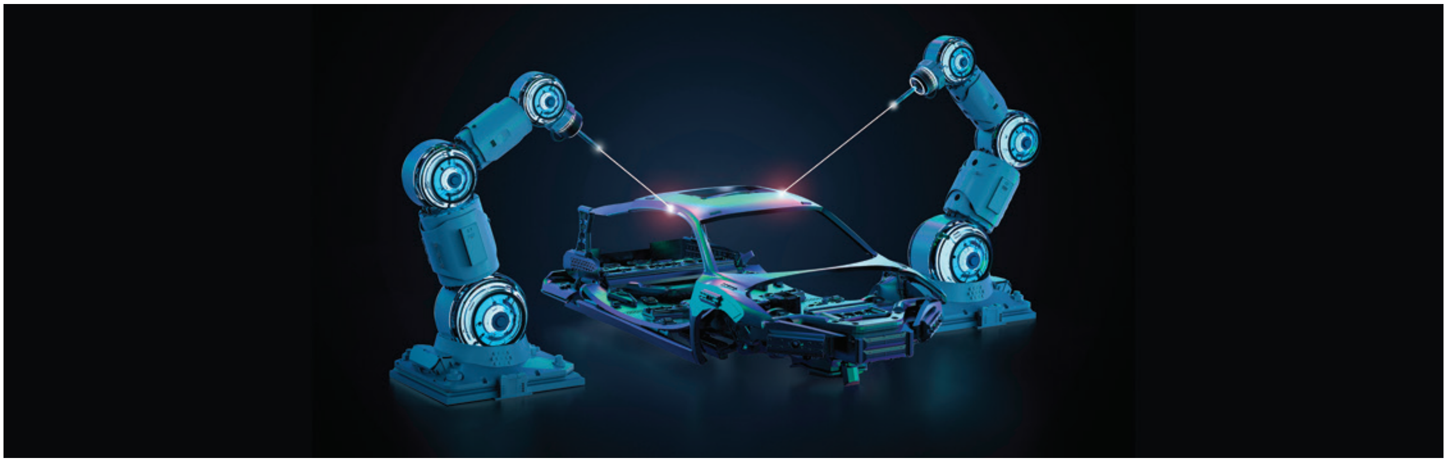
NVIDIA vGPU enables multiple VMs to have simultaneous, direct access to a single physical GPU, or GPUs can be aggregated within a single VM. By doing so, NVIDIA vGPU provides VMs with high-performance compute, application compatibility, cost-effectiveness, and scalability since multiple VMs can be customized to specific tasks that may demand more or less GPU compute or memory.

NVIDIA AI Enterprise Suite

NVIDIA AI Enterprise suite is an end-to-end, cloud-native suite of AI and data science applications and frameworks, optimized and exclusively certified by NVIDIA to run on VMware vSphere with NVIDIA-Certified Systems. It includes key enabling technologies and software from NVIDIA for rapid deployment, management and scaling of AI workloads in the modern hybrid cloud. NVIDIA AI Enterprise is licensed and supported by NVIDIA.

VMware vSphere Bitfusion

vSphere Bitfusion decouples physical resources from servers within an environment. The platform can share GPUs in a virtualized infrastructure, to provide a pool of shared network-accessible resources capable of supporting resource-intensive AI and ML workloads. Bitfusion works across AI frameworks, clouds, networks and in environments such as VMs, containers and notebooks.



Dell PowerEdge and VMware — Unlocking the power of AI

Our strength in partnership allows for:

- Global reach with best-of-breed solutions
- Time savings and investment protection thanks to hours of joint testing and validation
- Streamlined experience with a massive supply chain and distribution
- Peace of mind with best-in-class security and cyber-recovery solutions

Dell Technologies and VMware Together

The combination of technologies from Dell Technologies and VMware with end-to-end hardware and software makes it easier to access accelerated parallel-computing applications, AI frameworks, models and software development kits (SDKs). It gives AI researchers, data scientists and developers the software they need to deliver successful AI projects while arming IT professionals with the ability to support AI using the tools they're most familiar with for managing data centers and hybrid cloud environments.

This AI-ready platform accelerates the speed at which developers can build AI and high-performance data analytics (HPDA) for their business, enables organizations to scale modern workloads on the same VMware vSphere infrastructure they've already invested in, and delivers enterprise-class manageability, security and availability.

Dell Technologies and VMware's strategy is to make AI real and scalable, supporting workloads such as scientific simulations, weather forecasting, predictive analytics, visualization, modeling and simulation, financial modeling and signal processing. Collaborative designs in the AI space are designed to move the enterprise conversation from theoretical to reality, where intelligence and computing technologies make a significant business impact.

Learn more about PowerEdge servers.

Learn more about our systems management solutions.

Contact a Dell Technologies Expert for Sales or Support.

Search our resource library.

Follow PowerEdge servers on Twitter.

Dell Technologies believes the information in this document is accurate as of its publication date. The information is subject to change without notice.