Generative AI in the Enterprise

A Scalable and Modular Production Infrastructure for Artificial Intelligence Large Language Models

May 2023

H19611

White Paper

Abstract

This white paper presents an overview of Generative AI, and introduces Project Helix, a collaboration between Dell Technologies and NVIDIA to enable high performance, scalable, and modular full-stack generative AI solutions for large language models in the enterprise.

Dell Technologies Solutions

Copyright

The information in this publication is provided as is. Dell Inc. makes no representations or warranties of any kind with respect to the information in this publication, and specifically disclaims implied warranties of merchantability or fitness for a particular purpose.

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

Contents

Introduction

Executive summary

The growth of artificial intelligence (AI) applications and use cases is astounding, with impacts across nearly all facets of business and personal lives. Generative AI, the branch of AI that is designed to generate new data, images, code, or other types of content that humans do not explicitly program, is becoming particularly impactful and influential.

According to one analyst, the global generative AI market size was already estimated at USD 10.79 billion in 2022. It is projected to approach USD 118 billion by 2032, growing at a compound annual growth rate (CAGR) of 27 percent from 2023 to 2032[1].

As well as myriads of other applications, use cases include:

- Conversational agents and chatbots for customer service
- Audio and visual content creation
- Software programming
- Security, fraud detection, and threat intelligence
- Natural language interaction and translation

There are few areas of business and society that are not impacted in some way by this technology.

While public generative AI models such as ChatGPT, Google Bard AI, DALL-E, and other and more specialized offerings are intriguing, there are valid concerns about their use in the enterprise. These concerns include ownership of output, which encompasses issues of accuracy, truthfulness, and source attribution.

Therefore, there is a compelling need for enterprises to develop their own Large Language Models (LLMs) that are trained on proprietary datasets or developed and fine-tuned from known pretrained models.

Dell Technologies and NVIDIA

Dell Technologies and NVIDIA have been leading the way by delivering joint innovations for AI and high-performance computing. We are actively collaborating in this new space to enable customers to create and operate generative AI models for the enterprise.

- Dell Technologies has industry-leading infrastructure that includes powerful servers with NVIDIA graphical processing unit (GPU) acceleration, data storage systems, networking, systems management, reference designs, and years of experience helping enterprises with their AI initiatives.

- NVIDIA has the leading GPU acceleration, end-to-end networking solutions, cluster management software, NVIDIA AI Enterprise software, state-of-the-art,

[1] Presence Research (https://www.precedenceresearch.com/generative-ai-market)

pretrained foundation models including the NeMo framework, and the expertise to build, customize, and run generative AI.

We are now partnering on a new generative AI project called Project Helix, a joint initiative between Dell Technologies and NVIDIA, to bring generative AI to the world's enterprise data centers. Project Helix is a full-stack solution that enables enterprises to create and run custom AI models, built with the knowledge of their business. We have designed a scalable, modular, and high-performance infrastructure that enables enterprises everywhere to create a wave of generative AI solutions that will reinvent their industries and give them competitive advantage.

Generative AI is one of the most exciting and rapidly evolving fields in AI today. It is a transformative technology and the combination of powerful infrastructure and software from Dell Technologies together with accelerators, AI software, and AI expertise from NVIDIA is second to none.

About this document

In this whitepaper, readers can gain a comprehensive overview of generative AI, including its underlying principles, benefits, architectures, and techniques. They can also learn about the various types of generative AI models, and how they are used in real-world applications.

This white paper also explores the challenges and limitations of generative AI, such as the difficulty in training large-scale models, the potential for bias and ethical concerns, and the trade-off between generating realistic outputs and maintaining data privacy.

This white paper also provides guidance about how to develop and deploy generative AI models effectively. It includes considerations about hardware and software infrastructure from Dell Technologies and NVIDIA, data management, and evaluation metrics – all leading to a scalable, high-performance, production architecture for generative AI in the enterprise.

Audience

This white paper is intended for business leaders Chief Technology Officers (CTOs), Chief Information Officers (CIOs), IT infrastructure managers, and systems architects who are interested in, involved with, or considering implementation of generative AI.

Generative AI background and concepts

Background

AI has gone through several phases of development since its inception in the mid-20th century. The major phases of AI development, along with approximate timeframes, are:

1. **Rule-based systems (1950s-1960s)**—The first phase of AI development addressed the creation of rule-based systems, in which experts encoded their knowledge into a set of rules for the computer to follow. These systems were limited in their ability to learn from new data or adapt to new situations.

2. **Machine learning (1960s-1990s)**—The next phase of AI development addressed the use of machine learning algorithms to train computers to recognize patterns in data and make predictions or decisions based on those patterns. This phase saw the development of algorithms such as decision trees, logistic regression, and neural networks.

3. **Deep learning (2010s-present)**—The next phase of AI addressed deep learning. Deep learning is a subset of machine learning that uses neural networks with multiple layers to recognize complex patterns in data. This phase has been effective at processing images, videos, and natural language data.

4. **Generative AI (present)**—The present phase addresses generative AI. Generative AI uses deep learning algorithms to generate content such as images, videos, music, and even text that closely resembles the patterns of the original data. This phase has enormous potential for creating new types of content and generating new insights and predictions based on large amounts of data.

While these phases are not strictly defined or mutually exclusive, they represent major milestones in the development of AI and demonstrate the increasing complexity and sophistication of AI algorithms and applications over time.

Definition and overview

Generative AI is a branch of artificial intelligence that builds models that can generate content (such as images, text, or audio) that is not explicitly programmed by humans and that is similar in style and structure to existing examples. Generative AI techniques use deep learning algorithms to learn from large datasets of examples, learn patterns, and generate new content that is similar to the original data.

One of the significant aspects of generative AI is its ability to create content that is indistinguishable from content created by humans, which has numerous applications in industries such as entertainment, design, and marketing. For example, generative AI can create realistic images of products that do not exist yet, generate music that mimics the style of a particular artist, or even generate text that is indistinguishable from content written by humans.

An important area of generative AI is natural language generation (NLG), which is a subset of natural language processing (NLP) and involves generating natural language text that is coherent, fluent, and similar in style to existing or human-produced text. NLG has been used for various applications, including chatbots, language translation, and content generation.

Overall, generative AI has the potential to transform the way we create and consume content. It has the potential to generate new knowledge and insights in various fields, making it an exciting area of development in AI.

Evolution

Advances in deep learning algorithms and the availability of large datasets of natural language text have driven the evolution of NLG to generative AI. Early NLG systems relied on rule-based or template-based approaches, which were limited in their ability to generate diverse and creative content. However, with the rise of deep learning techniques such as recurrent neural networks (RNNs) and transformers, it has become possible to build models that can learn from large datasets of natural language text and generate new text that is more diverse and creative.

An important milestone in the evolution of generative AI was the development of the Generative Pretrained Transformer (GPT) series of models by OpenAI. The original GPT model, released in 2018, was a transformer-based model trained on a large corpus of text data. The model was able to generate coherent and fluent text that was similar in style to the original data. Subsequent versions of the model, including GPT-2 and GPT-3, have pushed the boundaries of what is possible with NLG, generating text that is increasingly diverse, creative, and even human-like in some cases.

Today, generative AI techniques are used in a wide range of applications, including content generation, chatbots, language translation, and more. As the field continues to evolve, we can expect to see more sophisticated generative AI models that can generate even more creative and diverse content.

Transformer models

Transformer models are a type of deep learning model that are commonly used in NLP and other applications of generative AI. Transformers were introduced in a seminal paper by Vaswani and others in 2017. They have since become a key building block for many state-of-the-art NLP models.

At a high level, transformer models are designed to learn contextual relationships between words in a sentence or text sequence. They achieve this learning by using a mechanism called self-attention, which allows the model to weigh the importance of different words in a sequence based on their context. This method is in contrast to traditional recurrent neural network (RNN) models, which process input sequences sequentially and do not have a global view of the sequence.

A key advantage of transformer models is their ability to process input sequences in parallel, which makes them faster than RNNs for many NLP tasks. They have also been shown to be highly effective for a range of NLP tasks, including language modeling, text classification, question answering, and machine translation.

The success of transformer models has led to the development of large-scale, pretrained language models, referred to as generative pretraining transformers (GPTs), such as OpenAI's GPT series and Google's Bidirectional Encoder Representations from Transformers (BERT) model. These pretrained models can be fine-tuned for specific NLP tasks with relatively little additional training data, making them highly effective for a wide range of NLP applications.

Overall, transformer models have revolutionized the field of NLP and have become a key building block for many state-of-the-art generative AI models. Their ability to learn contextual relationships between words in a text sequence has offered new possibilities for language generation, text understanding, and other NLP tasks.

Workload characteristics

Generative AI workloads can be broadly categorized into two types: training and inferencing. Training uses a large dataset of examples to train a generative AI model, while inference uses a trained model to generate new content based on an input. Data preparation before training can also be a significant task in creating custom models. All these workloads have characteristics that must be considered in the design of solutions and their infrastructure.

The characteristics of a generative AI workload can vary depending on the specific application and the type of model being used. However, some common characteristics include:

- **Compute intensity**—Generative AI workloads can be computationally intensive, requiring significant amounts of processing power to train or generate new content. This scenario particularly applies to large-scale models such as GPT-3, which can require specialized hardware such as GPUs to train efficiently.

- **Memory requirements**—Generative AI models require significant amounts of memory to store the model parameters and intermediate representations. This scenario particularly applies to transformer-based models such as GPT-3, which have many layers and can require hundreds of millions or even billions of parameters. Therefore, having sufficient GPU memory capacity is key.
- **Data dependencies**—Generative AI models are highly dependent on the quality and quantity of training data, which can greatly affect the performance of the model. Data preparation and cleaning are important parts of a solution as tapping into large, high-quality datasets is key to creating custom models.
- **Latency requirements**—Inference workloads might have strict latency requirements, particularly in real-time applications such as chatbots or voice assistants. Models must be optimized for inference speed, which can involve techniques such as model quantization or pruning. Latency considerations also favor on-premises or hybrid solutions, as opposed to purely cloud-based solutions, to train and infer from models closest to the source of the data.
- **Model accuracy**—The accuracy and quality of the generated content is a critical outcome for many generative AI applications, and is typically evaluated using metrics such as perplexity, bilingual evaluation understudy (BLEU) score, or human evaluation.

Overall, generative AI workloads can be highly complex and challenging, requiring specialized hardware, software, and expertise to achieve optimal outcomes. However, with the right tools and techniques, they can enable a wide range of exciting and innovative applications in fields such as NLP, computer vision, and creative arts.

Types of workloads

There are several specific types of generative AI workloads; each has different requirements. The system configurations described later in this white paper reflect these requirements.

Inferencing

Inferencing is the process of using a generative AI model to generate new predictive content based on input. A pretrained model is trained on a large dataset, and when new data is fed into the model, it makes predictions based on what it has learned during training. This training involves feeding an input sequence or image into the model and receiving an output sequence or image as the result. Inferencing is typically faster and less computationally intensive than training because it does not involve updating the model parameters.

Model customization

Pretrained model customization is the process of retraining an existing generative AI model for task-specific or domain-specific use cases. For large models, it is more efficient to customize than to train the model on a new dataset. Customization techniques in use today include fine-tuning, instruction tuning, prompt learning (including prompt tuning and P-tuning), reinforcement learning with human feedback, transfer learning, and use of adapters (or adaptable transformers).

The most useful types of customizations are fine-tuning, prompt learning, and transfer learning.

Fine-tuning

Fine-tuning retrains a pretrained model on a specific task or dataset, adapting its parameters to improve performance and make it more specialized. This traditional method of customization either freezes all but one layer and adjusts weights and biases on a new dataset or adds another layer to the neural network and re-recalculates weights and biases on a new dataset.

Prompt learning

Prompt learning is a strategy that allows pretrained language models to be repurposed for different tasks without adding new parameters or fine-tuning with labeled data. These techniques can also be used on large generative AI image models.

Prompt leaning can be further categorized into two broader techniques: prompt tuning and P-tuning.

- **Prompt tuning** is the process of retraining a pretrained generative AI model for task-specific or domain-specific use cases. It uses tailored datasets to improve its performance on a specific domain, use case, or task, or to incorporate additional knowledge into the model. This process allows the model to adapt to the specific characteristics of the new dataset and can improve its accuracy and performance on the task.
- **P-tuning**, or parameter tuning, focuses on adjusting prompts or instructions during inference to shape the model's output without modifying its underlying weights. Both techniques play a role in customizing and optimizing large language models for specific use cases.

Transfer learning

Transfer learning is a traditional technique for using pretrained generative AI models to accelerate training on new datasets. This technique starts with a pretrained model that has already learned useful features from a large dataset, and then adapts it to a new dataset with a smaller amount of training data. It can be much faster and more effective than training a model initially on the new dataset because the pretrained model already understands the underlying features of the data. Transfer learning is useful when there is limited training data available for a new task or domain. Transfer learning is not typically used for generative AI LLMs but is effective with general AI models.

In this solution design, the configurations related to customization are optimized for fine-tuning and P-tuning. However, the scalability and overall architecture design considerations still apply to other customization techniques and for datasets other than text.

Training

Training is the process of using a dataset to train a generative AI model initially. Training feeds the model examples from the dataset and adjusts the model parameters to improve its performance on the task. Training can be a computationally intensive process, particularly for large-scale models like GPT-3.

In an end-to-end workflow for generative AI, the exact sequence of these steps depends on the specific application and requirements. For example, a common workflow for LLMs might involve:

- Preprocessing and cleaning the training data
- Training a generative AI model on the data
- Evaluating the performance of the trained model
- Fine-tuning the model on a specific task or dataset
- Evaluating the performance of the fine-tuned model
- Deploying the model for inferencing in a production environment

Transfer learning can also be used at various points in this workflow to accelerate the training process or improve the performance of the model. Overall, the key is to select the appropriate techniques and tools for each step of the workflow and to optimize the process for the specific requirements and constraints of the application.

Types of outputs

The type of data used and the generative AI outcome varies depending on the type of data being analyzed. While the focus of this project is on LLMs, other types of generative AI models can produce other types of output.

- **Text**—LLMs can be used to generate new text based on a specific prompt or to compile long sections of text into shorter summaries. For example, ChatGPT can generate news articles or product descriptions from a few key details.
- **Image**—Generative AI models for images can be used to create realistic images of people, objects, or environments that do not exist. For example, StyleGAN2 can generate realistic portraits of nonexistent people.
- **Audio**—Generative AI models for audio can be used to generate new sounds or music based on existing audio samples or to create realistic voice simulations. For example, Tacotron 2 can generate speech that sounds like a specific person, even if that person never spoke the words.
- **Video**—Generative AI models for video can be used to create videos based on existing footage or to generate realistic animations of people or objects. For example, DALL-E can generate images of objects that do not exist, and these images can be combined to create animated videos.

In each case, the generative AI model must be trained on large datasets of the appropriate datatype. The training process is tailored to the requirements of the datatype and to the specific datatype because different input and output formats are required for each type of data. Recent advancements are now capable of integrating differing datatypes, for example, using a text entry to generate an image.

Business and technical challenges

There are both business and technical challenges to consider when using generative AI models, particularly those models in the public domain that have not been developed and controlled from within the enterprise.

The following examples show the challenges that businesses might face when implementing generative AI models, and potential solutions for addressing those challenges. It is important to approach each challenge on a case-by-case basis and work with experts in the field to develop the best possible solutions.

Ownership of content

There are valid concerns in the enterprise about ownership of output and intellectual property when using some generative AI models. These concerns include issues of accuracy, truthfulness, and source attribution. Data used for training public models, while extensive, might be based on incomplete or outdated knowledge or lead to the inability to verify facts or access real-time information.

Data quality

One of the biggest challenges with any machine learning model is ensuring that the training data is of high quality. This need is especially important for generative AI models, which might require large amounts of training data to generate accurate results. To address this challenge, businesses must ensure that their data is clean, well-labeled, and representative of the problem they are trying to solve.

Model complexity

Generative AI models can be complex and require significant computational resources to train and run. This requirement can be a challenge for businesses that do not have access to powerful hardware or that are working with large datasets.

Ethical considerations

Generative AI models can have ethical implications, especially if they are used to create content or make decisions that affect people's lives. To address this challenge, businesses must carefully consider the potential ethical implications of their generative AI models and work to ensure that they cause no harm.

Sustainability

Large-scale generative AI models require significant computational resources and power to operate. Training and inference processes for such models can consume substantial amounts of energy, contributing to increased carbon emissions, cooling demands, and environmental impact.

Regulatory compliance

Depending on the industry and application, there might be regulatory requirements that businesses must meet when implementing generative AI models. For example, in healthcare, there might be regulations for patient privacy and data security. To address this challenge, businesses must work closely with legal and compliance teams to ensure that their generative AI models meet all regulatory requirements.

Benefits

Generative AI benefits

Generative AI can provide numerous benefits to an organization across multiple dimensions. These benefits include:

- **Improved productivity**—To automate repetitive and time-consuming tasks, allowing employees to focus on more high-level tasks and increasing overall productivity
- **Enhanced customer experience**—To develop conversational interfaces and chatbots that can improve customer engagement and satisfaction by providing personalized and timely responses
- **Better decision-making**—To generate insights and recommendations from data that can help inform business decisions and improve overall business performance
- **Cost savings**—To help reduce operational costs by automating tasks and improving process efficiency, ultimately resulting in cost savings
- **Increased innovation**—To generate new ideas and solutions that can help drive innovation and create new revenue streams
- **Competitive advantage**—To help enterprises stay ahead of the competition by enabling faster and more efficient processes, better customer engagement, and improved decision making

Dell and NVIDIA advantages

The advantages that Dell Technologies and NVIDIA provide are significant, as together we:

- Deliver full-stack generative AI solutions built on the best of Dell infrastructure and software, with the latest NVIDIA accelerators, NVIDIA AI software, and AI expertise
- Deliver validated designs that reduce the time and effort to design and specify AI solutions, accelerating the time to value
- Provide sizing and scaling guidance so that your infrastructure is efficiently tailored to your needs but can also grow as those needs expand
- Enable enterprises to build, customize, and run purpose-built generative AI on-premises to solve specific business challenges, and use the same accelerated computing platform to create leading models
- Assist enterprises with the entire generative AI life cycle, from infrastructure provisioning, large model training, pretrained model fine-tuning, multisite model deployment, and large model inferencing
- Enable custom generative AI models that focus on the wanted operating domain, have the up-to-date knowledge of your business, have the necessary skills, and can continuously improve in production
- Include state-of-the-art, pretrained foundation models to rapidly accelerate the creation of custom generative AI models

- Ensure security and privacy of sensitive and proprietary company data, as well as compliance with government regulations
- Powerful yet performance-optimized server and storage hardware designs coupled with GPU acceleration, plus system management software that includes advanced power management, thermal optimization, and overall energy utilization monitoring
- Include the ability to develop safer and more trustworthy AI with known models and datasets – a fundamental requirement of enterprises today

Use cases

Generative AI models have the potential to address a wide range of use cases and solve numerous business challenges across different industries. Generative AI models can be used for:

- **Customer service**—To improve chatbot intent identification, summarize conversations, answer customer questions, and direct customers to appropriate resources.
- **Content creation**—To create content such as product descriptions, social media posts, news articles, and even books. This ability can help businesses save time and money by automating the content creation process.
- **Sales and marketing—**To create personalized experiences for customers, such as customized product recommendations or personalized marketing messages.
- **Product design**—To design new products or improve existing products. For example, a generative AI model can be trained on images of existing products to generate new designs that meet specific criteria.
- **Education**—To create personal learning experiences, similar to tutors, and generate learning plans and custom learning material.
- **Fraud detection—**To detect and prevent fraud in financial transactions or other contexts. For example, a generative AI model can be trained to recognize patterns of fraudulent behavior and flag suspicious transactions.
- **Healthcare**—To analyze medical images or patient data to aid in diagnosis or treatment. For example, a generative AI model can be trained to analyze medical images to identify cancerous cells or analyze protein structures for new drug discovery.
- **Gaming—**To create more realistic and engaging gaming experiences. For example, a generative AI model can be trained to create more realistic animations or to generate new game levels.
- **Software development—**To write code from human language, convert code from one programming language to another, correct erroneous code, or explain code.

These examples show the many business challenges that generative AI models can help solve. The key is to identify the specific challenges that are most pressing for a specific business or industry, and then to determine how generative AI models can be used to address those challenges.

Dell and NVIDIA solution architecture

High-level architecture

Dell Technologies and NVIDIA have been leading the way in delivering joint innovations for AI and high-performance computing for years. With this project, we have jointly designed a full-stack workflow-centric solution that enables enterprises to create and run generative AI models at any scale—from AI experimentation to AI production.

The architecture is modular, scalable, and balances performance with efficiency. The modularity enables the architecture to support numerous different AI workflows, as explained in the following sections.

Spirit of modularity

The cornerstone of this joint architecture is modularity, offering a flexible design that caters to a multitude of use cases, sectors, and computational requirements. A truly modular AI infrastructure is designed to be adaptable and future-proof, with components that can be mixed and matched based on specific project requirements. The Dell-NVIDIA solution uses this approach, enabling businesses to focus on certain aspects of generative AI workloads when building their infrastructure. This modular approach is accomplished through specific use-case designs for training, model tuning, and inference that make efficient use of each compute type. Each design starts with the minimum unit for each use case, with options to expand.

A modular software stack is also critical to allow AI researchers, data scientists, data engineers, and other users to design their infrastructure quickly and achieve rapid time to value. The Dell-NVIDIA solution uses the best of NVIDIA AI software, with partner solutions to build an AI platform that is adaptable and supported at each layer—from the operating system to the scheduler to multiple AI Operations (AIOps) and Machine Learning Operations (MLOps) solutions.

The following figure shows a high-level view of the solution architecture, with emphasis on the software stack, from the infrastructure layer up through the AI application software:

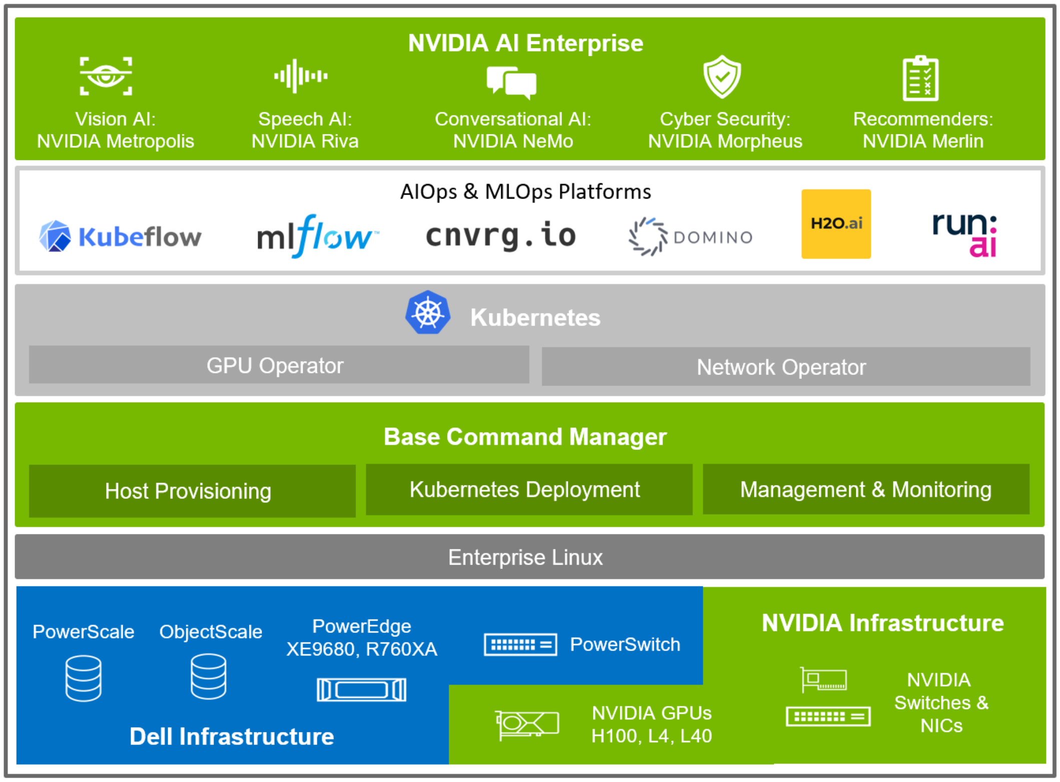


Figure 1. Solution architecture and software stack

At a high level, the solution architecture starts with the base hardware components from Dell Technologies and NVIDIA, which are combined in permutations that focus on specific AI workloads, such as training, fine-tuning, and inferencing. This white paper describes the individual hardware components in a later section.

Each control plane or compute element supports either Red Hat Enterprise Linux or Ubuntu as the operating system, which is preloaded with NVIDIA GPU drivers and Compute Unified Device Architecture (CUDA) for bare metal use.

NVIDIA Base Command Manager (BCM) serves as the cluster manager by installing software on the host systems in the cluster, deploying Kubernetes, and monitoring the cluster state. Host provisioning is core to a well-functioning cluster, with the ability to load the operating system, driver, firmware, and other critical software on each host system. Kubernetes deployment includes GPU Operator and Network Operator installation, a critical part of GPU and network fabric enablement. NVIDIA BCM supports both stateful and stateless host management, tracking each system, its health, and collecting metrics that administrators can view in real time or can be rolled up into reports.

At the top layer of the solution, is the NVIDIA AI Enterprise software that accelerates the data science pipeline and streamlines development and deployment of production AI including generative AI, computer vision, speech AI, and more. Whether developing a new AI model initially or using one of the reference AI workflows as a template to get started, NVIDIA AI Enterprise offers secure, stable, end-to-end software that is rapidly growing and fully supported by NVIDIA.

With Kubernetes deployed in the solution, there are several different MLOps solutions that can be installed, whether open-source solutions like Kubeflow and MLFlow, or featured supported solutions such as cnvrg.io, Domino, H2O.ai, Run:ai, and more. Each of these solutions can be deployed to work in a multicluster and hybrid cloud scenario.

Architectural modules

The generative AI solution architecture addresses three primary workflows:

- Large model inferencing
- Large model customization (fine-tuning and P-tuning)
- Large model training

Each of these workflows has distinct compute, storage, network, and software requirements. The solution design is modular and each of the components can be independently scaled depending on the customer's workflow and application requirements. Also, some modules are optional or swappable with equivalent existing solutions in an organization's AI infrastructure such as their preferred MLOps and Data Prep module or their preferred data module. The following table shows the functional modules in the solution architecture:

Table 1. Functional architecture modules for generative AI solution

Module	Description
Training	Module for AI optimized servers for training, powered by PowerEdge XE9680 and XE8640 servers with NVIDIA H100 GPUs
Inferencing	Module for AI optimized servers for inferencing, powered by PowerEdge XE9680 servers with NVIDIA H100 or R760xa servers with NVIDIA L40 or L4 GPUs
Management	Module for system and cluster management, including a head node for NVIDIA BCM, powered by PowerEdge R660 servers.
MLOps and Data Prep	Module for machine learning operations and data preparation for running MLOps software, database, and other CPU-based tasks for data preparation, powered by PowerEdge R660 servers
Data	Module for high-throughput, scale-out Network Attached Storage (NAS) powered by Dell PowerScale, plus high-throughput scale out object storage powered by Dell ECS and ObjectScale
InfiniBand	Module for very low latency, high-bandwidth GPU-to-GPU communication, powered by NVIDIA QM9700 InfiniBand switches
Ethernet	Module for high throughput and high-bandwidth communication between other modules in the solution powered by Dell PowerSwitch Z9432F-ON

Scalability

In the solution architecture, the functional modules can be scaled according to the use cases and the capacity requirements. For example, the minimum training module unit for large model training consists of eight PowerEdge XE9680 servers with 64 NVIDIA H100 GPUs.

As a theoretical example, the training module with an InfiniBand module could train a 175B parameter model in 112 days. To illustrate the scalability, six copies of these modules could train the same model in 19 days. As another example, if you are training a 40B parameter model, then two copies of the training module are sufficient to train the model in 14 days.

There is a similar scalability concept for the InfiniBand module. For example, one module with two QM9700 switches can support up to 24 PowerEdge XE9680 servers. If you double the InfiniBand module, in a fat-tree architecture, you can scale up to 48 PowerEdge XE9680 servers. The Ethernet module and Inference modules work similarly.

The Data module is powered by scale-out storage architecture storage solutions, which can linearly scale to meet performance and capacity requirements, as you increase the number of servers and GPUs in your Training and Inference modules.

Scalability and modularity are intrinsic to the Dell and NVIDIA design for generative AI across the board.

Security

The Dell approach to security is intrinsic in nature—it is built-in, not bolted-on later, and it is integrated into every step through the Dell Secure Development Lifecycle. We strive to continuously evolve our PowerEdge security controls, features, and solutions to meet the ever-growing threat landscape, and we continue to anchor security with a Silicon Root of Trust.

Security features are built into the PowerEdge Cyber Resilient Platform, enabled by the integrated Dell Remote Access Controller (iDRAC). There are many features added to the system that span from access control to data encryption to supply chain assurance. These features include Live BIOS scanning, UEFI Secure Boot Customization, RSA Secure ID MFA, Secure Enterprise Key Management (SEKM), Secured Component Verification (SCV), enhanced System Erase, Automatic Certificate Enrollment and Renewal, Cipher-Select, and CNSA support. All features make extensive use of intelligence and automation to help you stay ahead of threats, and to enable the scaling demanded by ever-expanding usage models.

As enterprises move to production AI, maintaining a secure and stable AI platform can be challenging. This challenge is especially true for enterprises that have built their own AI platform using open source, unsupported AI libraries and frameworks. To address this concern and minimize the burden of maintaining an AI platform, the NVIDIA AI Enterprise software subscription includes continuous monitoring for security vulnerabilities, ongoing remediation, and security patches as well priority notifications of critical vulnerabilities. This monitoring frees enterprise developers to focus on building innovative AI applications instead of maintaining their AI development platform. In addition, maintaining API stability can be challenging due to the many open-source dependencies. With NVIDIA AI Enterprise, enterprises can count on API stability by using a production branch that NVIDIA AI experts maintain. Access to NVIDIA Support experts means that AI projects stay on track.

Infrastructure component considerations for AI

There are many important considerations regarding the various hardware infrastructure components for a generative AI system, including high performance computing, high-speed networking, and scalable, high-capacity, and low-latency storage to name a few.

Compute

Generative AI models require significant amounts of computational power, particularly during the training phase because they typically involve large-scale matrix multiplication and other computationally intensive operations. For training, it is common to use many powerful GPUs to accelerate the process. For inferencing, less powerful hardware can be used but a significant amount of compute power to provide fast responses is required.

Accelerators

As mentioned earlier, accelerators such as GPUs are often used to expedite the training process. These accelerators are designed specifically for parallel processing of large amounts of data, making them well suited for the matrix multiplication and other operations required by generative AI models. In addition to specialized hardware, there are also software-based acceleration techniques such as mixed precision training, which can expedite the training process by reducing the precision of some of the calculations.

Storage

Generative AI models can be sizable, with many parameters and intermediate outputs. This volume means that the models require significant amounts of storage to hold all the data. It is common to use distributed storage systems such as Hadoop or Spark to store the training data and intermediate outputs during training. For inferencing, it might be possible to store the model on a local disk, but for larger models, it might be necessary to use network-attached storage or cloud-based storage solutions. Scalable, high-capacity, and low-latency storage components for both file object and file store are essential in AI systems.

Networking

Networking is an important consideration for generative AI, particularly in distributed training scenarios. During training, data is typically distributed across multiple nodes, each with their own accelerator and storage. These nodes must communicate with each other frequently to exchange data and update the model. High-speed networking solutions such as InfiniBand or RDMA are often used to minimize the latency of these communications and significantly improve the performance of the training process.

Summary

Generative AI requires significant amounts of computational power and storage, and often involves the use of specialized accelerators such as GPUs. Also, high-speed networking solutions are important to minimize latency during distributed training. By carefully considering these requirements, businesses can build and deploy generative AI models that are fast, efficient, and accurate.

Dell infrastructure and software components

This section describes the primary Dell hardware and software components used in the generative AI solution architecture.

Dell PowerEdge servers

Dell Technologies offers a range of acceleration-optimized servers and an extensive acceleration portfolio with NVIDIA GPUs. Two Dell servers are featured in the solution for generative AI.

The PowerEdge adaptive compute approach enables servers engineered to optimize the latest technology advances for predictable profitable outcomes. The improvements in the PowerEdge portfolio include:

- **Focus on acceleration**—Support for the most complete portfolio of GPUs, delivering maximum performance for AI, machine learning, and deep learning training and inferencing, high performance computing (HPC) modeling and simulation, advanced analytics, and rich-collaboration application suites and workloads
- **Thoughtful thermal design**—New thermal solutions and designs to address dense heat-producing components, and in some cases, front-to-back, air-cooled designs
- **Dell multivector cooling**—Streamlined, advanced thermal design for airflow pathways within the server

PowerEdge XE9680 server

The PowerEdge XE9680 server is a high-performance application server made for demanding AI, machine learning, and deep learning workloads that enable you to rapidly develop, train, and deploy large machine learning models.

The PowerEdge XE9680 server is the industry's first server to ship with eight NVIDIA H100 GPUs and NVIDIA AI software. It is designed to maximize AI throughput, providing enterprises with a highly refined, systemized, and scalable platform to help them achieve breakthroughs in NLP, recommender systems, data analytics, and more.

Its 6U air-cooled design chassis supports the highest wattage next-generation technologies up to 35C ambient. It features nine times more performance and two times faster networking with NVIDIA ConnectX-7 smart network interface cards (SmartNICs), and high-speed scalability for NVIDIA SuperPOD.

PowerEdge XE8640 server

The PowerEdge XE8640 server is an air-cooled 4U performance-optimized server featuring four NVIDIA H100 Tensor Core GPUs and NVIDIA NVLink technology, along with two upcoming 4th Gen Intel Xeon Scalable processors. It is designed to help businesses develop, train, and deploy machine learning models to accelerate and automate analysis.

PowerEdge R760xa server

Optimized for PCIe GPUs, the dual-socket 2U PowerEdge R760xa server enables businesses to accelerate a wide variety of applications including AI training and inferencing, analytics, virtualization, and performance rendering applications, all within an air-cooled design. The PowerEdge R760xa server delivers outstanding performance using Intel CPUs and supports a diverse set of GPU accelerators from AMD, Intel, and NVIDIA to drive a large and powerful range of demanding processing needs. Deploy and enable demanding graphics applications and dense AI inferencing applications business-wide with powerful features and capabilities, using the latest technology.

Dell file storage

Dell PowerScale supports the most demanding AI workloads with all-flash NVMe file storage solutions that deliver massive performance and efficiency in a compact form factor.

There are several models used in the generative AI solution architecture, all powered by the PowerScale OneFS operating system and supporting inline data compression and deduplication. The minimum number of PowerScale nodes per cluster is three nodes, and the maximum cluster size is 252 nodes.

PowerScale F900

PowerScale F900 provides the maximum performance of all-NVMe drives in a cost-effective configuration to address the storage needs of demanding AI workloads. Each node is 2U in height and hosts 24 NVMe SSDs. PowerScale F900 supports TLC or QLC drives for maximum performance. It enables you to scale raw storage from 46 TB to 736 TB per node and up to 186 PB of raw capacity per cluster.

PowerScale F600

PowerScale F600 includes NVMe drives to provide larger capacity with massive performance in a cost-effective compact 1U form factor to power demanding workloads. The PowerScale F600 supports TLC or QLC drives for maximum performance. Each node allows you to scale raw storage capacity from 15.36 TB to 245 TB and up to 60 PB of raw capacity per cluster.

Dell object storage

Dell Technologies offers a choice of object-based storage products, all of which are scalable and cost-effective for high volumes of unstructured data for AI workloads.

Dell ECS

ECS enterprise object storage combines the simplicity of S3 with extreme performance at scale for modern workloads such as AI, machine learning, and real-time analytics applications. ECS EXF900 offers all-flash, NVMe performance with capacity that scales up to 5.898 PB per rack—as well as 21 times faster performance* than the previous generation. Using ECS to fuel GPU servers with throughput-optimized storage rapidly exposes training algorithms and applications to more data than ever before.

*Based on Dell Technologies internal analysis comparing the max bandwidth of the ECS EXF900 (511 MB/s) to the maximum bandwidth of the ECS EX300 (24 MB/s) for 10 KB writes, November 2020. Actual performance will vary.

Dell ObjectScale

ObjectScale is software-defined object storage that delivers performance at scale to support AI workloads. It delivers datasets at high transfer rates to the most demanding CPU and GPU servers, exposing AI training algorithms to more data without introducing the complexity of HPC storage. This storage includes fast stable support for objects as large as 30 TB. Clusters can be scaled out easily to enhance performance and capacity linearly. With the ability to deploy on NVMe-based, all-flash drives, storage performance is no longer a bottleneck. Additionally, object tagging provides inference models with richer datasets from which to make smarter predictions.

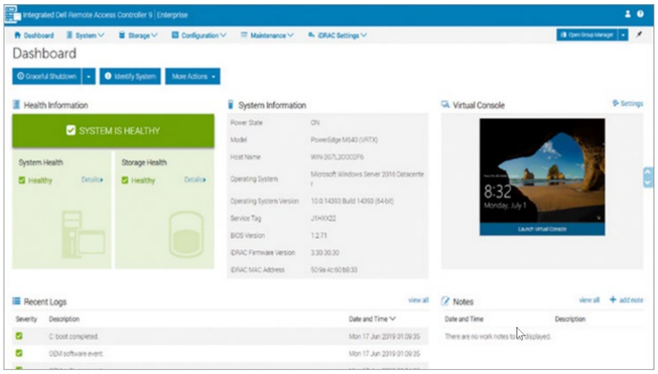
Dell PowerSwitch networking

Future-ready networking technology helps you improve network performance that lowers overall costs and network management complexity and experience flexibility to adopt new innovations.

The Dell PowerSwitch Z9432F-ON 100/400GbE fixed switch consists of Dell’s latest disaggregated hardware and software data center networking solutions, providing state-of-the-art, high-density 100/400 GbE ports and a broad range of functionality to meet the growing demands of today’s data center environment. This innovative, next-generation open networking high-density aggregation switch offers optimum flexibility and cost-effectiveness for the Web 2.0, enterprise, mid-market, and cloud service providers with demanding compute and storage traffic environments.

Dell OpenManage Enterprise

IT management is the foundation for operational success and running a large multinode system that is needed for generative AI workloads can be especially complex. OpenManage Enterprise reduces the time and effort required to manage IT implementations. It enables server life cycle management capabilities that return value through real-time efficiencies and cost-savings.



Power State	ON
Model	PowerEdge M640 (VRTX)
Host Name	WIN-2G7L2OO02F6
Operating System	Microsoft Windows Server 2016 Datacente r
Operating System Version	10.0.14393 Build 14393 (64-bit)
Service Tag	J1HXX22
BIOS Version	1.2.71
iDRAC Firmware Version	3.30.30.30
iDRAC MAC Address	50:9a:4c:60:b8:33

Severity	Description	Date and Time
✓	C boot completed.	Mon 17 Jun 2019 01:09:35
✓	OEM software event.	Mon 17 Jun 2019 01:09:35

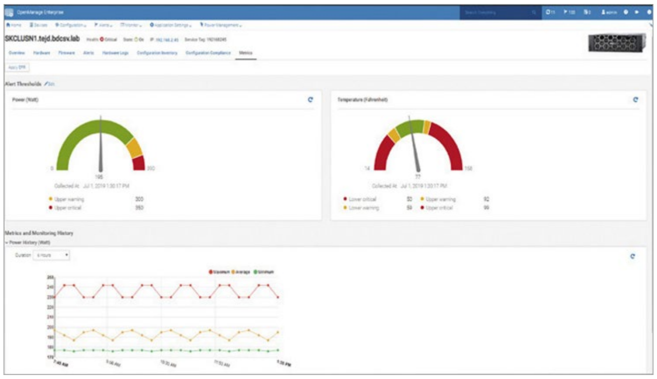
Date and Time	Description
There are no work notes to be displayed.	

OpenManage Enterprise powers innovation through predictive analysis, added insight, and extended control that improves security, enhances efficiency, and accelerates time to value. It supports up to 8,000 devices, manages Dell servers, and monitors Dell networking and storage infrastructure as well as third-party products.

With intelligent automation, OpenManage Enterprise has full-lifecycle configuration management with editable templates, and configuration management with firmware drift detection. It also has an extensible plug-in architecture and streamlined remote management capabilities.

Dell OpenManage Enterprise Power Manager

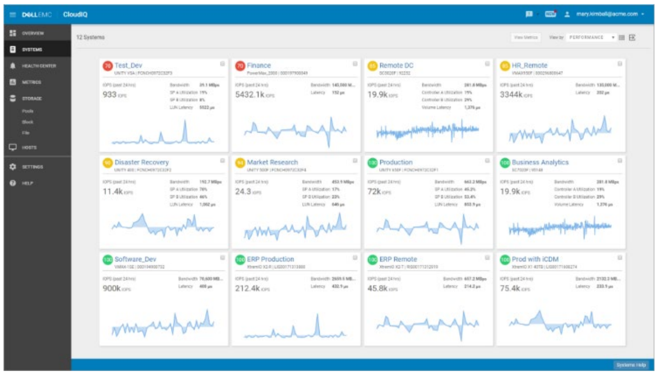
OpenManage Enterprise Power Manager allows you to maximize data center uptime, control energy use, and monitor and budget server power based on the consumption and workload needs as well as monitor thermal conditions. Because generative AI has demanding and resource-intensive workloads, managing power consumption efficiently is critical.



Through integration with the integrated Dell Remote Access Controller (iDRAC) embedded in all PowerEdge servers, you can set policy-based controls to maximize resource use and throttle-back power when performance demand ebbs. By using predefined power policies, OpenManage Enterprise Power Manager can help mitigate operational risks and ensure that your servers and their key workloads continue to operate.

Dell CloudIQ

CloudIQ is a cloud-based proactive monitoring and predictive analytics application for the Dell infrastructure portfolio. It combines the human intelligence of expert engineering and the machine intelligence of AI and machine learning to provide you with the insight to manage your IT infrastructure efficiently and proactively to meet business demand.



CloudIQ integrates data from all your OpenManage Enterprise Power Manager consoles to monitor the health, capacity, performance, and cybersecurity of Dell components across all your locations.

The CloudIQ portal displays your Dell infrastructure systems in one view to simplify monitoring across your core and secondary data centers and edge locations as well as data protection in public clouds. With CloudIQ, you can easily assure that critical business workloads get the capacity and performance that they need, spend less time monitoring and troubleshooting infrastructure, and spend more time innovating and focusing on projects that add value to your business.

Dell Services

Dell Technologies provides multiple services, linking people, processes, and technology to accelerate innovation and enable optimal business outcomes for AI solutions and all your data center needs.

Consulting Services

Consulting Services help you create a competitive advantage for your business. Our expert consultants work with companies at all stages of data analytics to help you plan, implement, and optimize solutions that enable you to unlock your data capital and support advanced techniques, such as AI, machine learning, and deep learning.

Deployment Services

Deployment Services help you streamline complexity and bring new IT investments online as quickly as possible. Use our over 30 years of experience for efficient and reliable solution deployment to accelerate adoption and return on investment (ROI) while freeing IT staff for more strategic work.

Support Services

Support Services driven by AI and deep learning will change the way you think about support with smart, groundbreaking technology backed by experts to help you maximize productivity, uptime, and convenience. Experience more than fast problem resolution – our AI engine proactively detects and prevents issues before they impact performance.

Managed Services

Managed Services can help reduce the cost, complexity, and risk of managing IT so you can focus your resources on digital innovation and transformation while our experts help optimize your IT operations and investment.

Residency Services

Residency Services provide the expertise needed to drive effective IT transformation and keep IT infrastructure running at its peak. Resident experts work tirelessly to address challenges and requirements, with the ability to adjust as priorities shift.

NVIDIA infrastructure and software components

This section describes the primary NVIDIA hardware acceleration and AI software components used in the generative AI solution architecture.

NVIDIA accelerators

The following NVIDIA GPUs are among the NVIDIA acceleration components used in this generative AI solution architecture.

NVIDIA H100 Tensor Core GPU

The NVIDIA H100 Tensor Core GPU delivers unprecedented performance, scalability, and security for every workload. With NVIDIA fourth-generation NVLInk Switch System, the NVIDIA H100 GPU accelerates AI workloads with a dedicated Transformer Engine for trillion parameter language models. The NVIDIA H100 GPU uses breakthrough innovations in the NVIDIA Hopper architecture to deliver industry-leading conversational AI, speeding up large language models by 30 times over the previous generation.

For small jobs, the NVIDIA H100 GPU can be partitioned to right-sized Multi-Instance GPU (MIG) partitions. With Hopper Confidential Computing, this scalable compute power can secure sensitive applications on shared data center infrastructure. The inclusion of the NVIDIA AI Enterprise software suite reduces time to development and simplifies deployment of AI workloads and makes NVIDIA H100 GPU the most powerful end-to-end AI and HPC data center platform.

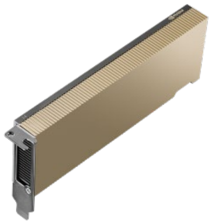
NVIDIA L40 GPU

The NVIDIA L40 GPU accelerator is a full height, full-length (FHFL), dual-slot 10.5-inch PCI Express Gen4 graphics solution based on the latest NVIDIA Ada Lovelace Architecture. The card is passively cooled and capable of 300 W maximum board power.

The NVIDIA L40 GPU supports the latest hardware-accelerated ray tracing, revolutionary AI features, advanced shading, and powerful simulation capabilities for a wide range of graphics and compute use cases in data center and edge server deployments. This support includes NVIDIA Omniverse, cloud gaming, batch rendering, virtual workstations, and deep learning training as well as inference workloads.

As part of the NVIDIA OVX server platform, the NVIDIA L40 GPU delivers the highest level of graphics, ray tracing, and simulation performance for NVIDIA Omniverse. With 48 GB of GDDR6 memory, even the most intense graphics applications run with the highest level of performance.

NVIDIA L4 Tensor Core GPU



The NVIDIA Ada Lovelace L4 Tensor Core GPU delivers universal acceleration and energy efficiency for video, AI, virtualized desktop and graphics applications in the enterprise, in the cloud, and at the edge. With NVIDIA's AI platform and full-stack approach, the NVIDIA L4 GPU is optimized for inference at scale for a broad range of AI applications, including recommendations, voice-based AI avatar assistants, generative AI, visual search, and contact center automation to deliver the best personalized experiences.

The NVIDIA L4 GPU is the most efficient NVIDIA accelerator for mainstream use. Servers equipped with the NVIDIA L4 GPU power up to 120 times higher AI video performance and 2.5 times more generative AI performance over CPU solutions, as well as over four times more graphics performance than the previous GPU generation. The NVIDIA L4 GPU's versatility and energy-efficient, single-slot, low-profile form factor make it ideal for global deployments, including edge locations.

NVIDIA NVLink and NVSwitch

NVIDIA NVLink is a fast, scalable interconnect that enables multinode, multi-GPU systems with seamless, high-speed communication between every GPU. The fourth generation of NVIDIA NVLink technology provides 1.5 times higher bandwidth and improved scalability for multi-GPU system configurations. A single NVIDIA H100 Tensor Core GPU supports up to 18 NVLink connections for a total bandwidth of 900 gigabytes per second (GB/s), over seven times the bandwidth of PCIe Gen5.

For even greater scalability, NVIDIA NVSwitch builds on the advanced communication capability of NVIDIA NVLink to deliver higher bandwidth and reduced latency for compute-intensive workloads. To enable high-speed, collective operations, each NVIDIA NVSwitch has 64 NVIDIA NVLink ports equipped with engines for NVIDIA Scalable Hierarchical Aggregation Reduction Protocol (SHARP) for in-network reductions and multicast acceleration.

NVIDIA AI Software

NVIDIA enterprise software solutions are designed to give IT admins, data scientists, architects, and designers access to the tools they need to easily manage and optimize their accelerated systems.

NVIDIA AI Enterprise

NVIDIA AI Enterprise, the software layer of the NVIDIA AI platform, accelerates the data science pipeline and streamlines development and deployment of production AI including generative AI, computer vision, speech AI and more. This secure, stable, cloud-native platform of AI software includes over 100 frameworks, pretrained models, and tools that accelerate data processing, simplify model training and optimization, and streamline deployment.

- **Data preparation**—Increase data processing time by up to 5 times while reducing operational costs by 4 times with the NVIDIA RAPIDS Accelerator for Apache Spark.
- **AI Training**: Create custom, accurate models in hours, instead of months, using NVIDIA TAO Toolkit and pretrained models.

- **Optimization for inference**—Accelerate application performance up to 40 times over CPU-only platforms during inference with NVIDIA TensorRT.
- **Deployment at scale**—Simplify and optimize the deployment of AI models at scale and in production with NVIDIA Triton Inference Server.

Available in the cloud, the data center and at the edge, NVIDIA AI Enterprise enables organizations to develop once and run anywhere. Since the full stack is maintained by NVIDIA, organizations can count on regular security reviews and patching, API stability, and access to NVIDIA AI experts and support teams to ensure business continuity and AI projects stay on track.

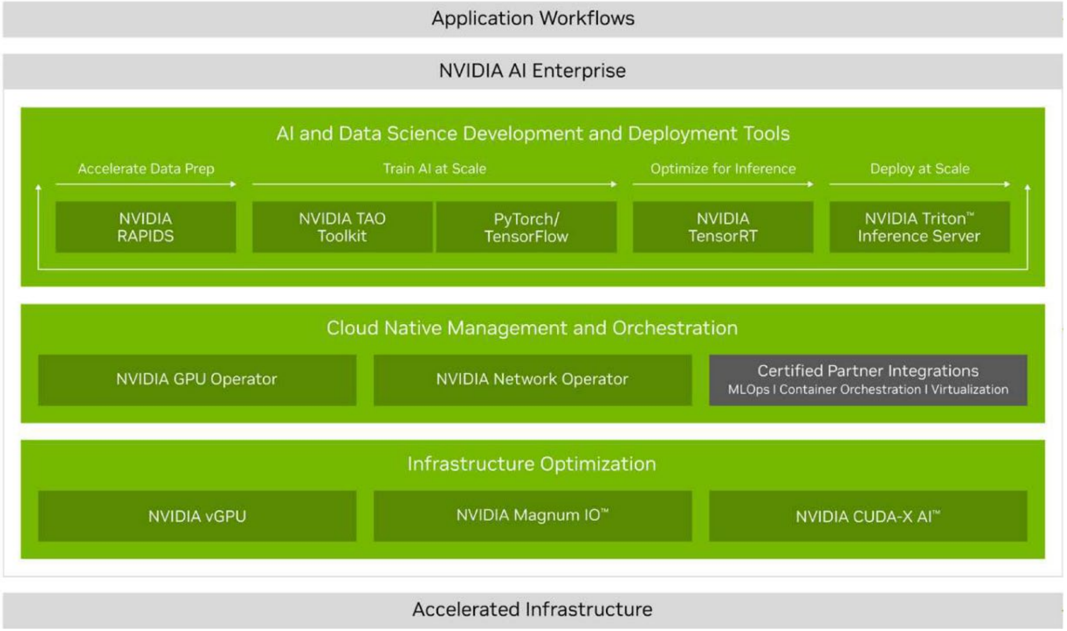


Figure 2. NVIDIA AI Enterprise

NVIDIA AI Enterprise includes NVIDIA NeMo, a framework to build, customize, and deploy generative AI models with billions of parameters. The NVIDIA NeMo framework provides an accelerated workflow for training with 3D parallelism techniques. It offers a choice of several customization techniques and is optimized for at-scale inference of large-scale models for language and image applications, with multi-GPU and multinode configurations. NVIDIA NeMo makes generative AI model development easy, cost-effective, and fast for enterprises.

NVIDIA Base Command Manager

NVIDIA BCM is NVIDIA's cluster manager for AI infrastructure. It facilitates seamless operationalization of AI development at scale by providing features like operating system provisioning, firmware upgrades, network and storage configuration, multi-GPU and multinode job scheduling, and system monitoring, thereby maximizing the utilization and performance of the underlying hardware architecture.

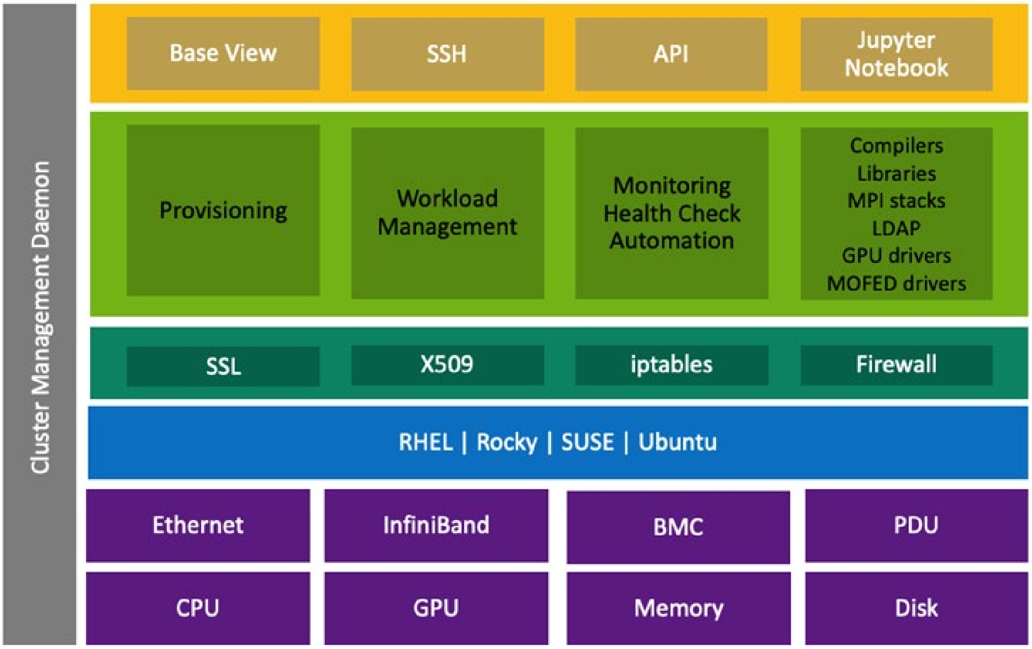

Cluster Management Daemon
Base View
SSH
API
Jupyter Notebook
Provisioning
Workload Management
Monitoring Health Check Automation
Compilers Libraries MPI stacks LDAP GPU drivers MOFED drivers
SSL
X509
iptables
Firewall
RHEL | Rocky | SUSE | Ubuntu
Ethernet
InfiniBand
BMC
PDU
CPU
GPU
Memory
Disk

Figure 3. NVIDIA Base Command Manager

NVIDIA BCM supports automatic provisioning and management of changes in nodes throughout the cluster's lifetime.

With an extensible and customizable framework, it has seamless integrations with the multiple HPC workload managers, including Slurm IBM Spectrum LSF, OpenPBS, Univa Grid Engine, and others. It offers extensive support for container technologies including Docker, Harbor, Kubernetes, and operators. It also has a robust health management framework covering metrics, health checks, and actions.

System Configurations

Based on the modular, scalable architecture for generative AI described earlier and powered by Dell and NVIDIA components, there are initially three system configurations in this family of designs, each focused on a particular use case. The three optimized system configurations are designed for inferencing, customization, and training use cases.

The following sections describe the system configurations for each area of focus at a high level. Note that the control plane, data storage, and Ethernet networking for each case is similar. Therefore, if you are building AI Infrastructure that addresses two or more cases, these core resources can be shared.

Large model inferencing

Many enterprises elect to start with a pretrained model and use it without modification or conduct some prompt engineering or P-tuning to make better use of the model for a specific function. Starting with production deployment in mind is critical in the case of LLMs because there is a heavy demand for compute power. Depending on the size of the model, many larger models require multiple 8x GPU systems to achieve second or subsecond-level throughput. The minimum configuration for inferencing pretrained models starts with a single PowerEdge R760XA server with up to four NVIDIA H100 GPUs or one PowerEdge XE9680 server with eight NVIDIA H100 GPUs based on model size and number of instances. The number of nodes can then scale out as needed for performance or capacity, though two nodes are recommended for reliability purposes.

Design considerations for inferencing large models include:

- Large models tend to have a large memory footprint. While there might not be a clear boundary that defines a large model, for the sake of simplicity, anything above 10B parameters can be considered a large model.
- When the model is split between GPUs, the communication between GPUs plays a crucial role in delivering optimum performance. Therefore, the NVIDIA Triton Inference Server software with multi-GPU deployment using fast transformer technology might be employed.
- For large models above 40B parameters, we recommend the PowerEdge XE9680 server. For model sizes less than 40B parameters, the PowerEdge R760xa server delivers excellent performance.
- The PowerSwitch Z9432F supports 32 ports of 400 (QSFP56-DD optical transceivers) or up to 128 ports of 100 GbE. Inference does not have the InfiniBand module or high throughput requirement; therefore, it scales linearly for concurrency needs up to 32 nodes.
- Throughput (inference per second) requirements require multiple GPUs to be deployed depending on the workload needs.

Large model customization

Many enterprises forgo initial training and elect to use and customize a pretrained model as the basis for their solution. Using fine-tuning and P-tuning, it is possible to apply enterprise-specific data to retrain a portion of an existing model or build a better prompt interface to it. This method requires significantly less compute power than training a model initially, with the ability to start with a similar configuration to the inference-only configuration. The key difference is the addition of InfiniBand networking between compute systems.

Design considerations for large model customization with fine-tuning or P-training using pretrained large models include the following:

- Even though this task is relatively less compute-intensive than large model training, there is a need for a tremendous amount of information exchange (for example, weights) between GPUs of different nodes. Therefore, InfiniBand is required for optimized performance and throughput with an eight-way GPU and an all-to-all NVLInk connection. In some cases, when the model sizes are less than 40 B parameters and based on the application latency requirements, the InfiniBand module can be optional.
- P-tuning uses a small trainable model before using the LLM. The small model is used to encode the text prompt and generate task-specific virtual tokens. Prompt-tuning and prefix-tuning, which only tune continuous prompts with a frozen language model, substantially reduce per-task storage and memory usage at training.
- For models less than 40B parameters, you might be able to use a PowerEdge XE8640 server. For larger models, we recommend the PowerEdgeXE9680 server.
- The Data module is optional because there are no snapshot requirements. Certain prompt-engineering techniques might require a large dataset and require a high-performance data module.

Large model training

Large model training is the most compute-demanding workload of the three use cases, with the largest models requiring data centers of large numbers of GPUs to train a model in a few months. The minimum configuration for training requires eight PowerEdge XE9680 servers with eight NVIDIA H100 GPUs each. The largest model training requires expansion to greater cluster sizes of 16-times, 32-times, or even larger configurations.

Design considerations for large model training include:

- Large generative AI models have significant compute requirements for training. According to OpenAI, for Chat GPT-3 with 175B parameters, the model size is approximately 350 GB, and it would take 355 years to train GPT-3 on a single NVIDIA Tesla V100 GPU. Alternatively, it would take 34 days to train with 1,024 NVIDIA A100 GPUs.
- The training model has a considerable memory footprint that does not fit in a single GPU; therefore, you must split the model across multiple GPUs (N-GPUs).
- The combination of model size, parallelism techniques for performance, and the size of the working dataset requires high communication throughput between

GPUs, thus benefitting from PowerEdge XE9680 servers with eight NVIDIA GPUs fully connected to each other by NVIDIA NVLink and NVIDIA NVSwitch.

- During the training phase, there is also a significant amount of information exchange (for example, weights) between GPUs of different nodes; InfiniBand is required for optimized performance and throughput.
- The QM9700 InfiniBand switch has 64 network detection and response (NDR) ports. Therefore, 24 nodes of the PowerEdge XE9680 servers in this cluster fill the ports on the QM9700 in the InfiniBand module. Add additional InfiniBand modules in a fat-tree network topology.
- As you add additional PowerEdgeXE9680 server nodes to your cluster, expand the PowerScale switches appropriately to meet the input/output performance requirements.
- Checkpointing is a standard technique used in large model training. The size of the checkpoints depends on the size and dimensions of the model and pipeline parallelism used in training.
- Four Dell PowerScale F600 Prime storage platforms deliver 8 GBS write and 40 GBS read throughput performance with linear scaling.

Summary

The information contained in this section is a high-level overview of the characteristics and key design considerations of the suggested configurations for inferencing, customization, and training of large language generative AI models. As mentioned earlier, further details about each use case will follow this white paper in a series of design guides for these Dell Validated designs for AI.

Conclusion

Generative AI advantage

This document has explored the concepts, benefits, use cases, and challenges of generative AI, and presented a scalable and modular solution architecture designed by Dell Technologies and NVIDIA.

Project Helix is a unique collaboration between Dell Technologies and NVIDIA that make the promise of generative AI real for the enterprise. Together, we deliver a full-stack solution, built on Dell infrastructure and software, and using the award-winning software stack and accelerator technology of NVIDIA. Bringing together the deep knowledge and creativity of NVIDIA with the global customer knowledge and technology expertise of Dell Technologies, Project Helix:

- Delivers full-stack generative AI solutions built on the best of Dell infrastructure and software, in combination with the latest NVIDIA accelerators, AI software, and AI expertise.
- Enables enterprises to use purpose-built generative AI on-premises to solve specific business challenges.
- Assists enterprises with the entire generative AI life cycle, from infrastructure provisioning, large model development and training, pretrained model fine-tuning, multisite model deployment and large model inferencing.
- Ensures trust, security, and privacy of sensitive and proprietary company data, as well as compliance with government regulations.

With Project Helix, Dell Technologies and NVIDIA enable organizations to automate complex processes, improve customer interactions and unlock new possibilities with better machine intelligence. Together, we are leading the way in driving the next wave of innovation in the enterprise AI landscape.

We value your feedback

Dell Technologies and the authors of this document welcome your feedback on this document. Contact the Dell Technologies Solutions team by email.

For more information about this solution, you can engage with an expert by emailing AI.Assist@dell.com.

References

These materials may provide additional information about the solutions and components presented here, as well as related offers.

Dell Technologies documentation

The following Dell Technologies documentation and resources provide additional and relevant information to that contained within this white paper.

- *Dell Technologies AI Solutions*
- *Dell Technologies Info Hub for Artificial Intelligence Solutions*
- *Dell PowerEdge XE Servers*
- *Dell PowerEdge Accelerated Servers and Accelerators (GPUs)*
- *Dell PowerScale Storage*
- *Dell ECS Enterprise Object Storage*
- *Dell ObjectScale Storage*
- *Dell PowerSwitch Z-series Switches*
- *Dell OpenManage Systems Management*

NVIDIA documentation

The following NVIDIA documentation and resources also provide additional and relevant information:

- *NVIDIA AI Enterprise NVIDIA NeMo*
- *NVIDIA Data Center GPUs*
- *NVIDIA Networking*