Delivering Edge AI with NVIDIA Fleet Command

A Dell EMC Ready Solution

April 2021

H18710

Design Guide

Abstract

This design guide describes Dell EMC PowerEdge server options for hosting AI applications on edge systems managed by the NVIDIA Fleet Command software platform. Based on a micro-services architecture, Fleet Command enables customers to provision edge systems quickly and deploy both custom AI applications and applications available from the NVIDIA NGC Catalog.

Dell Technologies Solutions

Copyright

Contents

Executive summary

The development of smart devices that can be deployed wherever there is a network connection is driving the need for innovations in distributed computer systems management. An evolving edge-computing model uses “pods” of servers, storage, and networking hardware with groups of Internet of Things (IoT) devices at many locations under the control of a single organization. The main advantages of this model are easy scalability and significant reduction of the costs for transporting large data volumes over limited-bandwidth wide-area networks (WANs). Edge computing also addresses the need for low-latency decision-making.

The complication for organizations implementing edge computing solutions is the challenge of deploying, managing, and securing distributed systems with hundreds to thousands of edge devices and servers.

In this guide, we describe an integrated and easy-to-use solution that is based on products from Dell Technologies and NVIDIA that address these management challenges. Dell Technologies Solutions engineers, with support from NVIDIA, conducted proof-of-concept testing in our engineering labs. This guide discusses the motivation for edge computing and the integration of NVIDIA® Fleet Command™, the NVIDIA NGC™ catalog, and NGC Private Registry, along with Dell hardware to quickly deploy edge analytics and artificial intelligence (AI) applications.

Audience

This guide is intended for solution architects, system administrators, and others who are interested in edge computing for AI applications.

We value your feedback

Dell Technologies and the authors of this guide welcome your feedback on the solution and the solution documentation. Contact the Dell Technologies Solutions team by email or provide your comments by completing our documentation survey.

Authors: Bala Chandrasekaran and Phillip Hummel

Note: For links to additional documentation for this solution, see the Dell Technologies Solutions Info Hub for AI and Data Analytics Workloads.

Note: This guide might contain language from third-party content that is not under Dell's control and is not consistent with the current guidelines for Dell's own content. When such third-party content is updated by the relevant third parties, this guide will be revised accordingly.

AI at the edge

Introduction

In the periods leading up to major innovations in the IT sector, there is frequently a burst of new concepts, buzzwords, and marketing hype before the changes solidify. It took many years before there was consensus on the adoption of the term “cloud computing” to describe a model for the use of automation and self-service provisioning of infrastructure and IT services. Now we frequently encounter the terms “public cloud” and “private cloud” to refer to variants of that model without the need to define what they mean.

Today, the IT industry is on the verge of an innovation at the intersection of the Internet of Things (IoT) and data analytics. Instead of having to move IoT data across a WAN for processing, a new model is emerging that is built on infrastructure systems deployed outside of traditional data-center environments. Many IT professionals are evaluating new hardware and software product options that can advance the state-of-the-art for edge computing. Typically, when innovations come to market, there are comparisons with existing operating models. For example, with momentum now building for a distributed mode of edge computing, some of the questions surfacing are “Will all analytics move to the edge?” and “Will the edge eat the cloud?”

Computer technology shifts are rarely zero-sum games. It is highly unlikely that every dollar spent on edge computing displaces a dollar that is invested in cloud computing. There are IoT use cases that are most suited to cloud-centric computing, edge-centric computing, and hybrid architectures. Investments in both edge and cloud computing technologies will grow for the foreseeable future, but the rate of increase of new distributed edge solutions will gain momentum quickly.

We have seen enough pilot projects to understand the range of potential edge-computing solutions. A “one size fits all” approach will not work for this diverse market. Most organizations need a strategy that embraces different types of edge implementations and multiple cloud options to provide the best services at the lowest cost. Organizations can move beyond the pilot phase by finding the most cost-effective architecture choice for a specific set of application requirements that consider an “edge-first” strategy for workload placement. If this approach seems overly complex, there are new multicloud applications with robust technology-management solutions coming to market. There is a lack of clear information in the industry press about finding the right approach for your use case.

In this guide, we present some recent findings from our edge computing solutions engineering proof-of-concept labs involving several specific use cases. We include the prerequisites, hardware setup, software used, and sizing details for these use cases.

Our lab setup included servers and networking from Dell Technologies and edge deployment management software from NVIDIA, known as NVIDIA Fleet Command.

Understanding the motivation for edge computing

Although the media attention being paid to edge computing has exploded in the last year, edge computing is not a recent innovation. Researchers in the early to mid-2000s were studying the potential for growth in the number and types of network-connected “smart” devices and the associated trends in worldwide Internet network use. Researches alerted the IT industry that a new model for data processing was going to be required as the IoT era developed. While the forecasts for total IoT devices deployed in this decade reached the tens of billions, the amount of Internet bandwidth available held relatively steady at

about 15 percent. The need for change was apparent. Until this boom in IoT, media streaming for consumers was the only use case that required significant Internet bandwidth. The IT industry quickly learned that following its model of sending data from centralized data centers to devices would not meet the latency, bandwidth, or time-to-response needs for this new Internet of Things. Any application designs that needed to collect and concentrate data from billions of devices for transport to relatively few centralized data centers would increase the competition for already limited Internet network bandwidth. The IoT industry needed alternative models for dealing with the large amounts of data being produced from edge devices.

Refining the edge computing model

The term “fog computing” was a buzzword that described edge computing as “a collection of numerous distributed tiny clouds deployed closer to the [IoT] devices at the edge of the network.”[1] The term did not persist for long. The term “distributed cloud computing” was also used before the industry settled on the term “edge computing.” Edge computing defines groups of devices paired with some computing and storage resources hosted at the edge of the Internet or other network.

The combination of the need to reduce bandwidth consumption into and out of centralized data centers and the huge projections for the deployment of devices at the edge drove the urgency for the development of distributed edge computing.[2] There are also benefits for applications hosted at the edge that require low roundtrip latency for any data that needs to be evaluated before making an operational or control change locally. Applications that need to access high-speed, highly available services remotely, even for small amounts of data, are complicated when separated by a WAN link. Capacity planning and equipment sizing is also less risky when applications can be deployed in edge configurations using a few building block components.

Edge computing management

The potential value of applications and services with embedded AI deployed near edge devices is nearly limitless. A major challenge that must be addressed is the management complexity that comes with distributed systems. The development of an efficient tool-set and processes to manage the ever-growing number of edge devices that organizations are deploying has never been accomplished at the scale needed for many edge computing scenarios.

The most common approach to developing tools for systems management of distributed computing systems is to separate the functions that manage the system from the functions that do the work of the system. An example of this approach is to have one system to communicate status and a separate system to make operational decisions for several interrelated components. The terms “control plane” and “data plane” are the two most commonly defined subsystems in distributed computing. Because the goal of edge computing is to restrict most of the heavy data movement to an area near the devices, the

[1] VikingPLoP '16: Proceedings of the 10th Travelling Conference on Pattern Languages of Programs April 2016 Article No.: 13 Pages 1–10 https://doi.org/10.1145/3022636.3022649

[2] https://link.springer.com/chapter/10.1007/978-3-319-99061-3_4

requirements for the data plane are less complex. The bigger challenge for the design of edge computing systems is providing a scalable, secure control plane.

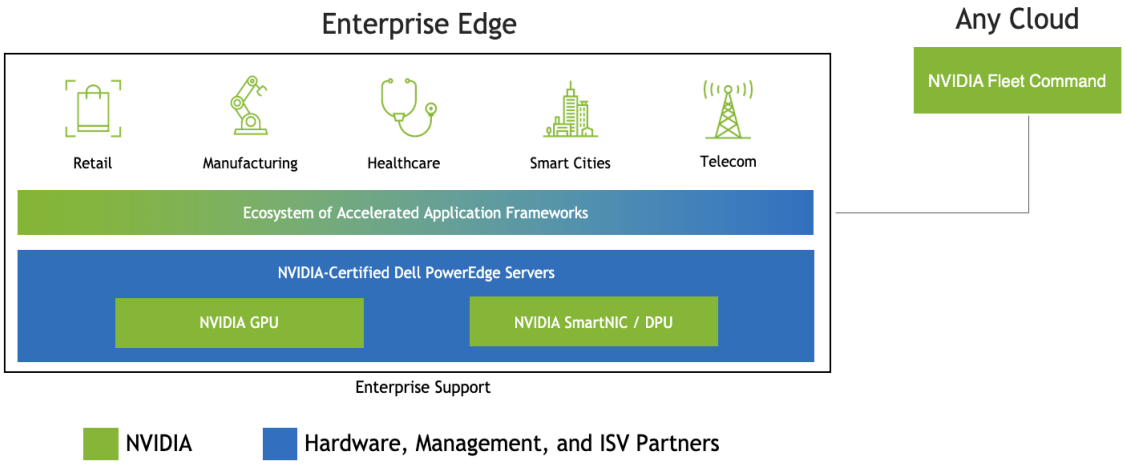

Figure 1. Managing the enterprise edge with the Fleet Command cloud console

Fleet Command enables organizations to easily deploy and manage a growing number of AI applications for many industries and public sector use cases. The Fleet Command controller can be accessed from a centralized cloud service including private clouds or a public cloud service provider.

Fleet Command streamlines automated deployments by managing fleets of devices connected to physical edge devices. Customers can easily choose to which locations the AI applications are deployed and are continuously updated.

Administrators can centrally monitor health and remotely fix systems, and with one-click updates, simplify AI operations at scale. Fleet Command is the tool for organizations to effectively make AI operational at the edge.

Fleet Command connects with NVIDIA-Certified Systems™ at edge locations for end-to-end security protocols, and it minimizes application downtime with a resilient software platform.

Technology components

Dell Technologies infrastructure solutions designed with NVIDIA hardware and software enable customers to deploy AI to their edge locations in retail, manufacturing, health care, and public services environments. The following figure shows how this Ready Solution uses a combination of hardware and software components at both the edge and in a cloud data center:

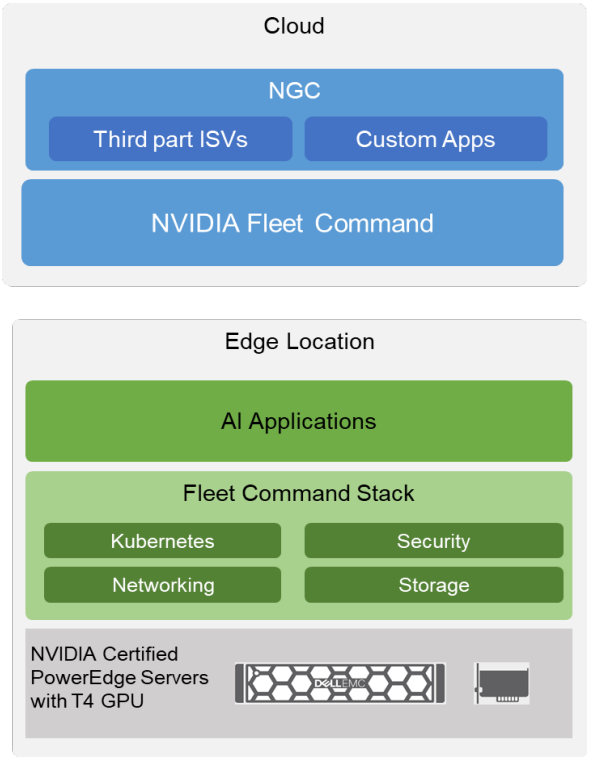

Figure 2. Dell EMC Ready Solution hardware and software components

NVIDIA NGC catalog and NGC Private Registry

The NGC catalog is a hub of GPU-optimized AI, high-performance computing (HPC), and data analytics software that simplifies and accelerates end-to-end workflows. With enterprise-grade containers, pretrained AI models, and industry-specific SDKs, data scientists and developers can accelerate their development process. Also, with Helm charts, IT and DevOps can reliably deploy their AI applications on premises, in the cloud, or at the edge.

To foster collaboration, enterprise AI teams can take advantage of the NGC Private Registry, a secure, cloud-hosted portal in the NGC catalog. AI teams can securely store and share their custom containers, models, model scripts, and Helm charts within their organization.

The NGC container registry service provides researchers, data scientists, and developers with access to a comprehensive catalog of GPU-accelerated software for AI, machine learning, and HPC. These containers take full advantage of NVIDIA GPUs whether on-premises, at the edge, or in the cloud. Each container image is fully optimized to work with NVIDIA-Certified Dell EMC PowerEdge servers. Also, the NGC Catalog hosts Kubernetes-ready Helm charts that simplify deployment of powerful third-party software.

NVIDIA Fleet Command

Fleet Command is a hybrid-cloud software platform for deploying and managing AI systems and applications at the edge. Its features enable you to:

- **Manage edge systems in multiple locations with a single control plane**—You can pair PowerEdge servers in multiple locations with Fleet Command, which allows you to deploy a complete operating environment and application software

stack on those servers. Groups of servers in a specific edge location are managed as a single Kubernetes cluster under the control of Fleet Command.

- **Deploy applications from private or public catalogs**—Fleet Command allows you to deploy applications from the public NGC catalog and from your NGC Private Registry to the edge systems. Organizations can use applications already available in the public catalog, develop and deploy in-house AI applications, and store them in their private registry. Administrators can then push these applications to edge locations as Helm charts so that organizations can take advantage of consistent, secure, and reliable environments to accelerate development-to-production cycles.
- **Connect safely to systems with remote management**—You can track system status to ensure that systems are ready to run applications. Systems that are accessed remotely are marked as having been accessed using a remote console. If the system is marked as accessed but you do not know who accessed it, you can reimage the system to alleviate security concerns.

  Fleet Command enables remote console access to edge systems for troubleshooting. It also provides centralized access to all the system logs, including Kubernetes and Linux system logs.
- **Secure deployments**—Fleet Command has integrated, end-to-end security to ensure that AI applications and data are always protected:
  - Applications are scanned for vulnerabilities and malware before they are loaded. Signed containers ensure that only authenticated software is deployed to the edge.
  - Storage in the local drives is encrypted.
  - Secure Boot and Measured Boot provide a trusted boot process.
- **Complete the life cycle (for edge systems and applications)**—Fleet Command provides an end-to-end life cycle for both the edge systems and the applications. New edge systems can be deployed and then kept current through over-the-air updates or unneeded systems removed with the Fleet Command stack. Also, applications can also be deployed and updated from the same Fleet Command console.

  Fleet Command also handles all the once-complex "day two" management tasks like updating system software over the air or monitoring location health at all your edge locations from a central control plane.
- **Provide resiliency**—Fleet Command takes advantage of Kubernetes capabilities to provide a resilient software stack that allows all systems to self-heal when applications are disrupted. Multiple nodes in a single location form a Kubernetes cluster.

For a comprehensive list of capabilities, see the NVIDIA Fleet Command documentation.

The Fleet Command software stack includes Ubuntu, Kubernetes, Helm, Tiller, and tools for deploying and managing the NVIDIA hardware assets necessary for GPU-enabled Kubernetes. With a few clicks, the software stack is automatically installed on any targeted edge servers. Fleet Command configures and manages the remote systems through an abstraction layer that shields customers from the complexity of full edge software stack management.

Dell EMC PowerEdge Server

You can choose from several Dell EMC PowerEdge servers to deploy your AI applications at the edge. Dell EMC PowerEdge XE2420 servers are purpose-built for the edge and NVIDIA-Certified for AI applications.

PowerEdge servers support NVIDIA T4 Tensor Core GPUs for AI inference. The NVIDIA T4 GPU is a single-slot, low profile, PCI Express (PCIe) Gen3 accelerator card that is based on the NVIDIA TU104 GPU. The T4 GPU is a passively cooled board that has 16 GB GDDR6 memory and a 70 W maximum power limit. NVIDIA Turing™ Tensor Cores power the T4 GPU to accelerate inference, video transcoding, and virtual desktops. The small PCIe form factor and energy efficiency of the T4 make it an ideal accelerator for inference at the edge.

The following table lists the rack server options that provide an optimum balance of computing power, memory, and GPU performance:

Table 1. Server options

Server model	Processor	Memory	Number of GPUs supported by Fleet Command	NVIDIA certification
PowerEdge R650	Dual-socket 3rd Gen Intel Xeon scalable processors	32 DIMMs	1 or 2	In progress[3]
PowerEdge R750xa	Dual-socket 3rd Gen Intel Xeon scalable processors	32 DIMMs	1, 2, or 4	In progress[3]
PowerEdge R6515	Single-socket 2nd Gen AMD EPYC processor	16 DIMMS	1	NVIDIA NGC-Ready™
PowerEdge R7515	Single-socket 2nd Gen AMD EPYC processors	16 DIMMS	1, 2, or 4	NGC-Ready
PowerEdge R6525	Dual-socket 2nd Gen AMD EPYC processors	32 DIMMS	1 or 2	NGC-Ready
PowerEdge R7525	Dual-socket 2nd Gen AMD EPYC processors	32 DIMMS	1, 2, or 4	NVIDIA-Certified System for T4
PowerEdge R940xa	Quad-socket 2nd Gen Intel Xeon scalable processors	48 DIMMS	1, 2, or 4	NGC-Ready
Dell EMC DSS 8840	Dual-socket 2nd Gen Intel Xeon scalable processors	24 DIMMS	1, 2, or 4	NGC-Ready
PowerEdge XE2420	Dual-socket 2nd Gen Intel Xeon scalable processors (maximum 150 W Gold)	16 DIMMs	1, 2, or 4	NGC-Ready
PowerEdge R640	Dual-socket 2nd Gen Intel Xeon scalable processors	24 DIMMs	1 or 2	NGC-Ready
PowerEdge R740	Dual-socket 2nd Gen Intel Xeon scalable processors	24 DIMMs	1, 2, or 4	NVIDIA-Certified System for T4
PowerEdge R740xd	Dual-socket 2nd Gen Intel Xeon scalable processors	24 DIMMs	1, 2, or 4	NVIDIA-Certified System for T4

[3] Certification is in progress. This guide will be updated when certification for these servers is complete.

NVIDIA-Certified Systems

NVIDIA-Certified Dell EMC PowerEdge Systems bring together NVIDIA GPUs and NVIDIA networking in servers from Dell Technologies in optimized configurations. These servers are validated for performance, manageability, security, and scalability and are backed by enterprise-grade support from NVIDIA and Dell Technologies. With an NVIDIA-Certified System, enterprises can confidently choose performance-optimized hardware solutions to power their accelerated computing workloads—both in smaller configurations and at scale.

The NVIDIA-Certified Systems program encompasses a wide range of enterprise GPUs. These GPUs include NVIDIA Ampere architecture-based data center GPUs and T4, as well as NVIDIA ConnectX®-6 and ConnectX-6 Dx smart network interface cards (SmartNICs) and NVIDIA BlueField®-2 data processing units (DPUs). The certification test suite is designed to exercise the performance and functionality of the configured server by running a set of software that represents a range of real-world applications. These applications include deep learning training, AI inference, data science algorithms, intelligent video analytics (IVA), HPC, and CUDA® functions and rendering. The program also covers infrastructure performance acceleration, such as network and storage offload, security features, and remote management capabilities.

The NVIDIA-Certified Systems program is a program for servers with new GPUs and NVIDIA networking for multimode tests. NGC-Ready for Edge is a validation program for existing single-node systems with NVIDIA V100 Tensor Core, NVIDIA T4 Tensor Core, and NVIDIA RTX™ 6000/8000 GPUs. NGC-Ready servers consist of NVIDIA GPUs installed in qualified enterprise-class servers that have passed an extensive suite of single-node tests that validate their ability to deliver high performance running NGC containers.

For a list of PowerEdge servers that are validated as NVIDIA-Certified Systems or NCG-Ready servers, see the NVIDIA Qualified Server Catalog.

Ordering considerations

Consider the following factors when ordering PowerEdge servers for Fleet Command:

- **NVIDIA T4 GPU**—Configure the server with at least one NVIDIA T4 GPU.
- **Trusted Platform Module (TPM)**—Order the server with a TPM module. TPM is a security device that holds computer-generated keys for encryption. It is a hardware-based solution that prevents hacking attempts to capture passwords, encryption keys, and other sensitive data. Fleet Command uses TPM and secure boot capabilities in the BIOS to securely deploy edge servers.
- **Storage adapter and hard drive considerations**—The Fleet Command software stack detects all the drives and creates a single logical volume. This volume is used for the Linux operating system, Kubernetes, and applications (both data and executable) running as pods and logs.

  Note: To provide redundancy, Dell Technologies recommends that the PowerEdge server is configured with a PERC storage controller and the RAID volume is configured with SSD.

  The BOSS controller and non-RAID storage controllers are supported but not recommended. Installing the Fleet Command software stack on the SD card is not supported.

- **Network adapter**—Configure the PowerEdge server with NVIDIA network adapters based on edge infrastructure requirements. For edge deployments that support 10 Gigabit Ethernet or more, we recommend Mellanox Connect-X network adapters as NVIDIA-Certified Systems are qualified with Connect-X adapters. For edge deployments that require 1 Gigabit Ethernet, we recommend Broadcom network adapters.
- **Dell EMC iDRAC Enterprise**—The Integrated Dell Remote Access Controller (iDRAC) is a hardware device connected to a server motherboard that allows systems administrators to update and manage Dell systems. The iDRAC is available to administrators even when the server is powered off. We recommend that the PowerEdge server is configured with iDRAC Enterprise to enable virtual console and virtual media capabilities. Systems administrators managing edge locations can use the capabilities to deploy and manage the Fleet Command stack. For more information about licensing options, see Support for Integrated Dell Remote Access Controller 9 (iDRAC9).

Dell EMC OpenManage Enterprise

OpenManage is a portfolio of solutions to help you discover, monitor, manage, update, and deploy your PowerEdge infrastructure. You can use Dell OpenManage Enterprise to manage the edge servers before or after installation of Fleet Command for server specific “day 2” operations. OpenManage Enterprise is a simple-to-use, one-to-many systems management console. It cost-effectively facilitates comprehensive life cycle management for PowerEdge servers in one console. It allows IT staff to discover, deploy, update, and monitor Dell EMC PowerEdge servers. It also enables IT administrators to view and change settings on edge servers with the same tools used with data center infrastructure when connected through a WAN.

NVIDIA Fleet Command-certified applications

You can deploy either your in-house-developed AI applications, or Fleet Command can deploy GPU-accelerated software for the top AI and data science use cases for every industry. The following table provides a sample of vendors who are certified with NVIDIA Fleet Command:

Table 2. Vendors certified with NVIDIA Fleet Command

Partner	Solution scope
Data Monsters	Data Monsters is an AI R&D lab and consulting company of scientists, engineers, product managers, and business consultants that are passionate about data, science, and creativity. Over 100+ successful projects have been completed and research groups at five university campuses around the world have been united.
Deep North	The Deep North platform enables you to architect robust, analytics-driven operations and sales strategies by turning existing video assets into real market intelligence for shopping centers, retail, commercial real estate, travel, and restaurants.
Deep Vision	Deep Vision empowers communities with computer vision to create smarter cities and safer environments. It improves the shopper experience through AI by providing computer vision for your online marketplace or eCommerce brand. Deep Vision changes the way media is organized by letting trained machines automatically organize your data.
Dematic	The Dematic Micro-Fulfillment System is the latest solution to help customers meet aggressive growth strategies for omni-channel product distribution. Dematic has been designing and building distribution centers for retail partners for decades. It is a natural extension to bring global innovation and supply chain thought leadership to microfulfillment.
Everseen	Everseen provides AI applications that transform how retailers see and solve business problems—from end-to-end. Everseen applications can be deployed across the entire business to: • Break the cyclical patterns that affect gross margin • Quantify problems at the intersection of product, people, and place • Correct flawed processes that contribute to negative financial and operational consequences

Sizing

AI applications range from image and video analytics to sensor analysis and more. They can be applicable to industries such as retail, manufacturing, health care, and smart cities. Sizing the PowerEdge server for edge deployments depends on the AI applications that you plan to run.

We use intelligent video analytics from Deep Vision as an example to size the PowerEdge server. The following table provides three configurations and a sample scenario for which the configuration might be applicable:

Table 3. Recommended configurations

Component	Small configuration	Medium configuration	Large configuration
Scenario	6 to 7 video processing streams for facial recognition	10 to 12 video processing streams for facial recognition and people counting	20 video processing streams for facial recognition, people counting, and vehicle identification
PowerEdge server model	R7515	R740	R7525
Processor	AMD EPYC 7502P 2.5 GHz, 32C/64T, 128 M Cache (180 W)	Intel Xeon Gold 6252 2.1 G, 24C/48T, 10.4 GT/s, 35.75 M Cache, Turbo, HT (150 W) DDR4-2933	AMD EPYC 7452 2.35 GHz, 32C/64T, 128 M Cache (155 W) DDR4-3200
Memory	8 x 8 GB	8 x 16 GB	16 x 16 GB
GPUs	1 NVIDIA T4 GPU	2 NVIDIA T4 GPUs	4 NVIDIA T4 GPUs
Network	Broadcom 5720 Quad Port 1 GbE BASE-T, rNDC	Broadcom 5720 Quad Port 1 GbE BASE-T, rNDC	Either of the following: • 1 GbE: Broadcom 5720 Quad Port 1 GbE BASE-T, rNDC • 25 GBE: Mellanox ConnectX-5 Dual Port 10/25GbE SFP28 Adapter, PCIe Low Profile
Storage	6 x 480 GB SAS SSDs in RAID 6	8 x 480 GB SAS SSDs in RAID 6	12 x 480 GB SAS SSDs in RAID 6
Trusted Platform Module	Trusted Platform Module 2.0	Trusted Platform Module 2.0	Trusted Platform Module 2.0
iDRAC	iDRAC Enterprise	iDRAC Enterprise	iDRAC Enterprise

These configurations are a basis and your Dell sales representative can customize the server configuration, including processing, memory, and disk, to meet your requirements. Multiple factors can drive these configurations. In video analytics, these factors include type and number of video analytics modules used, number of cameras, resolutions of cameras, and the nature of the video being analyzed. For guidelines for configuring the edge systems, see NGC-Ready Recommended Configurations.

For large deployments at a single location, choose between a scale-up server configuration and a scale-out configuration (for example, a single PowerEdge server with several T4 servers compared to multiple servers each with one or two GPUs). Cost, resiliency, and power requirements of the AI application help drive the decision.

Deployment

This section provides an overview of the steps for deploying Fleet Command on PowerEdge servers.

Setting up the NGC Private Registry and application repository

Fleet Command enables administrators to deploy applications from the NGC catalog and from their NGC Private Registry. To access applications in a private registry from Fleet Command, administrators must synchronize their private registry to Fleet Command by using an NGC API key that provides authenticated access.

When set up, administrators can then add applications from the registry for use by Fleet Command using the Fleet Command user interface. Applications are added from a Helm chart repository. We recommend that you use the NGC Private Registry as that repository.

Adding a location

A Fleet Command location defines an edge location or site such as a retail store or hospital where edge deployments will be installed. Locations are added by giving a name and a description in the Fleet Command user interface.

Edge site preparation

Before deploying the Fleet Command Stack on the PowerEdge servers and pairing them with Fleet Command, the following prerequisites must be met at the location:

- Ensure that the edge server has access to a DNS server and NTP server
- Ensure that the server has an IP address allocated by DHCP through a reservation and that the IP address does not change during operations.
- Ensure that you can make secure outbound connections using TLS 443. Ensure that both outbound and inbound UDP connections through port 31111 are allowed for access to the remote console.

Pairing a PowerEdge server with Fleet Command

To deploy the Fleet Command Stack on a PowerEdge server using iDRAC, and then pair a server with the Fleet Command interface:

1. Configure iDRAC.

   You can order the server with several iDRAC System Management options such as Static IP or DHCP with Zero Touch Configuration. When iDRAC is configured, ensure that in the system BIOS:

   a. The **Secure Boot** option is enabled.

   b. The **TPM 2.0** algorithm is set to **SHA256**.

2. Deploy the edge system image from ISO format, which you can download from the Fleet Command portal, and boot the server.

   The ISO formatted image can be attached as a virtual media in iDRAC.

3. When the server is booted with the ISO, provide the administrator password, network settings, and an activation code.

   The activation code allows the edge server to be paired with Fleet Command. Fleet Command provides the activation code when there is a request to add a

system to a location. The activation code is valid for 12 hours after it is created and a technician at the location must enter the activation code in that time.

After the activation is completed, the Fleet Command Stack is deployed on the edge system. You can now manage the system from Fleet Command.

Use the preceding steps to add multiple servers. The first server added to the location servers is a “master-worker” in a Kubernetes cluster. All subsequent servers act as Kubernetes “workers.”

Deploying applications

Fleet Command can now deploy applications from the NGC Private Registry in the cloud to the servers at the edge locations. Using the Fleet Command interface, select the applications that you want to deploy and the locations to which you want to deploy them. The deployment operates based on the Helm chart of the application and the specified configuration.

The following workflow summarizes the power and simplicity of the comprehensive Fleet Command approach to manage AI systems and applications at the edge. A single control plane handles everything from managing a catalog of applications to mapping and deploying applications to edge locations, as well as keeping applications at the edge updated and healthy.

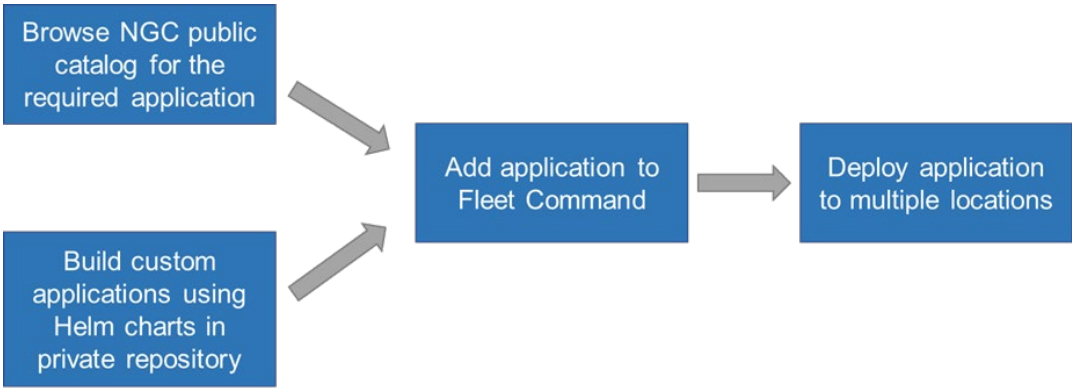

Figure 3. Application selection and deployment workflow

1. Using Fleet Command, deploy independent software vendor (ISV) applications or custom applications at the edge:
   a. For vendor applications (like Deep Vision), browse the NGC catalog for the required application and download it to the Private Repository.
   b. For custom-built applications, upload the application by providing information about the Helm chart to the NGC Private Repository.
2. From the NGC repository, add the application to Fleet Command. Fleet Command Applications define what can be deployed to the edge locations.
3. Create a deployment by providing the application from the private repository, any required application configuration information, and the locations for the application deployment. You can provide more than one location. Fleet Command deploys the application to all selected locations.

   The application is now ready to use.

4. Obtain the application's IP information from the Fleet Command interface. If required, use this IP information to log in to the application for further configuration and use.

Managing edge servers with Dell OpenManage and iDRAC

OpenManage Enterprise can be installed on the data center and can connect to edge devices using a WAN solution. OpenManage Enterprise can be used to discover and inventory the server, monitor the health of the server as well perform BIOS and firmware updates.

Also, you can use iDRAC for secure remote server management of the edge servers. In addition to using virtual media and console for deployment as explained earlier, iDRAC can also be used to update BIOS and firmware and monitor PowerEdge servers.

Conclusion

Applications that require either high bandwidth or low latency data communications and highly available access to data coming from devices deployed at the edge of the network are best engineered to run as close as possible to those devices. This form of distributed computing is often referred to as edge computing.

The requirements for a management control plane for a distributed edge computing environment are common to many types of application scenarios. This commonality of functions provides incentives to organizations to look for an existing solution instead of taking on the challenge of design, coding, and maintenance of a custom control plane. Fleet Command is a centralized control plane operating in the cloud that must be evaluated when choosing technology for a distributed edge computing environment. When paired with PowerEdge servers, NVIDIA and Dell Technologies provide a scalable edge solution for enterprises of any size.

In this guide, we have shared how to size the infrastructure for edge computing using a simple small, medium, and large hierarchy. This guide, together with the links to related resources, provides the information to start a pilot use case project with Dell servers and NVIDIA software and accelerators.