Enterprise AI

Buyer's Guide for Smart Hospitals

How to Design and Deploy AI Infrastructure for AI Chatbots, Telehealth, Medical Imaging, and more using NVIDIA and VMware

Table of Contents

Introduction

State of the Hospital Data Center

The need for smart hospitals is rapidly increasing— to improve patient care and to address numerous IT challenges. More than ever, patients are looking for healthcare services to be delivered more efficiently in a convenient and comfortable setting. For these reasons, healthcare facilities are seeking new ways to incorporate intelligent solutions that can improve their productivity, efficiency, and patient care; but unfortunately, they do not always have the resources to get started. Artificial intelligence lays the foundation for making smart hospitals a reality. In fact, 78% of health services executives believe AI will be part of their mainstream technology with the intention to increase productivity and efficiency, grow revenue, and create innovative products and services[1]. AI-powered tools and applications help healthcare facilities strengthen communication, streamline clinical workflows, and create a seamless patient experience.

While healthcare organizations know they need to invest in AI to secure their future, many struggle with finding the strategy and platform that can enable success. Unlike traditional enterprise applications, AI applications are relatively new for healthcare IT departments. They are anchored in rapidly evolving, open-source, bleeding-edge code and lack proven approaches that meet the rigors of scaled production settings in enterprises. Healthcare organizations and hospitals need infrastructure that can encompass not only their core critical applications today but also be AI-ready in the future. They need a foundational platform to support the caregivers and the patients they serve.

AI-Ready Platform for Smart Hospitals

VMware and NVIDIA have partnered to unlock the power of AI for every business by delivering an end-to-end enterprise platform optimized for AI workloads. This integrated platform delivers best-in-class AI software, the NVIDIA AI Enterprise suite, optimized and certified for the industry's leading virtualization platform, VMware vSphere.

[1] https://www.pwc.com/us/en/tech-effect/ai-analytics/ai-predictions/health-industries.html

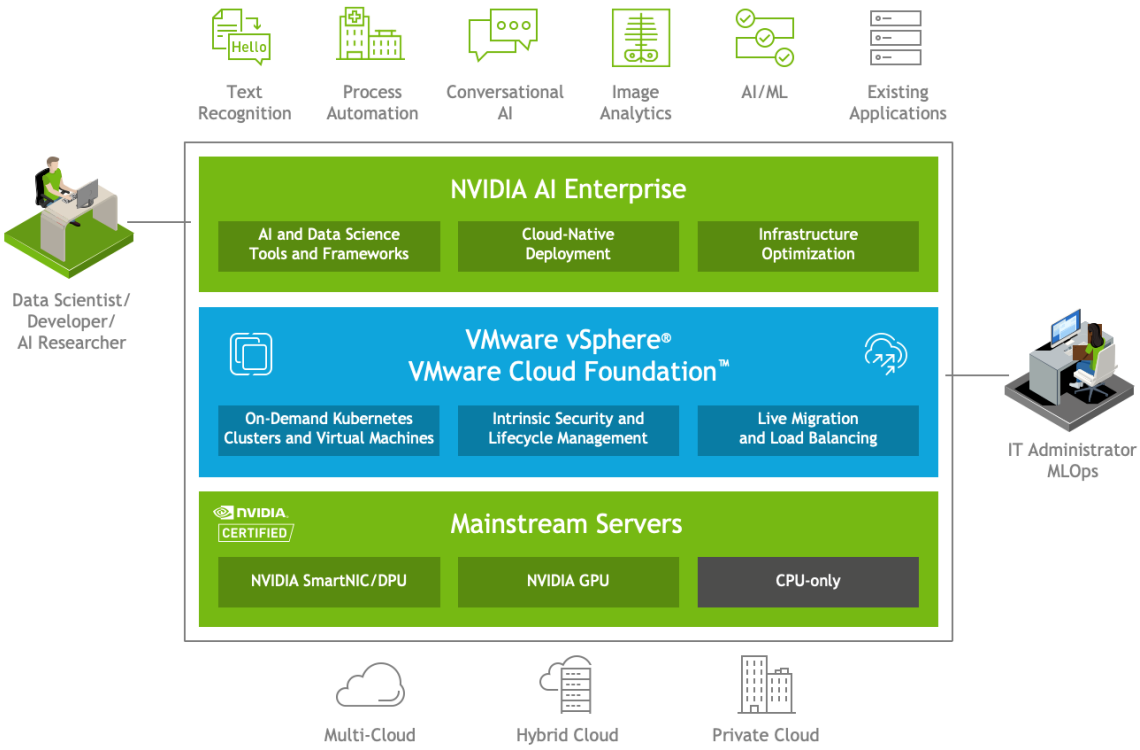

Figure 1 NVIDIA and VMware AI-Ready Platform

NVIDIA AI Enterprise is an end-to-end, cloud-native suite of AI and data science tools and frameworks optimized and certified by NVIDIA to run on VMware vSphere with NVIDIA-Certified Systems. It includes key enabling technologies and software from NVIDIA for rapid deployment, management, and scaling of AI workloads in the modern hybrid cloud. NVIDIA AI Enterprise is licensed and supported by NVIDIA. NVIDIA GPU virtualization software integrates with VMware vSphere for provisioning and managing GPUs in virtualized environments. It includes optimizations for bare-metal performance of accelerated virtual machines (VMs), supports containerized workloads, and enables GPU sharing so multiple VMs can be accelerated by a single GPU. Healthcare organizations using it are realizing significant benefits, including improved performance and increased productivity.

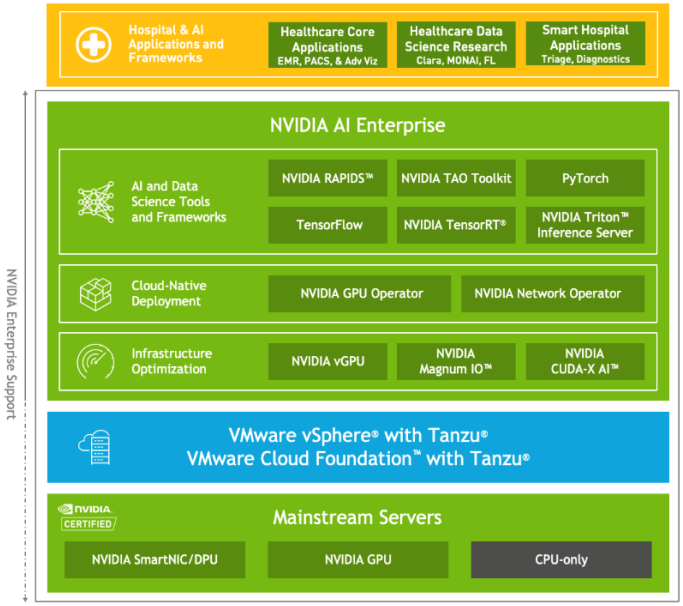

Figure 2 The NVIDIA AI Enterprise suite includes tools and frameworks used by data scientists and researchers as well as tools for cloud-native deployments and infrastructure optimization

The NVIDIA AI Enterprise software suite includes AI frameworks and containers that provide performance-optimized data science, training, and inference frameworks and tools that simplify building, sharing, and deploying AI software, so enterprises can gather insights faster and deliver business value sooner. Even organizations that lack AI expertise can adopt AI because NVIDIA AI Enterprise includes easy-to-use tools for every stage of the AI workflow, from data prep to training, inferencing, and deploying at scale.

- **NVIDIA TAO Toolkit** - gives you a faster, easier way to accelerate training and quickly create highly accurate and performant, domain-specific vision, and conversational AI models. It abstracts away the AI/deep learning framework complexity, letting you fine-tune on high-quality NVIDIA pre-trained models with only a fraction of the data compared to training from scratch. Developers can go beyond customization and optimize these models required for low-latency, high-throughput inference. This enables you to create custom, production-ready AI models in hours, rather than months, without a huge investment in AI expertise.

- **NVIDIA RAPIDS** - the first step in the end-to-end AI flow requires data prep before the neural networks can be trained. NVIDIA RAPIDS is optimized for GPU acceleration. It reduces data science processes from hours to seconds, when combined with NVIDIA A100, for up to 70x faster performance, and up to 20x more cost-effective when compared to similar CPU-only configurations.

- **PyTorch and TensorFlow** - Open-source deep learning frameworks for training and machine learning, such as PyTorch and TensorFlow, are integrated with NVIDIA RAPIDS to simplify enterprise AI development. Leveraging these tools and pre-trained models, accelerates development and deployment cycles, eliminating the need to procure, manage, certify and deploy different environments.

- **TensorRT** - based applications perform up to 40X faster than CPU-only platforms during inference. With TensorRT, you can optimize neural network models trained in all major frameworks, calibrate for lower precision with high accuracy, and deploy to hyperscale data centers, embedded, or automotive product platforms.

- **NVIDIA Triton Inference Server** - Triton Inference Server simplifies and optimizes the deployment of AI models at scale in production. It integrates with Kubernetes for orchestration and auto-scaling and allows front end client applications to submit inference requests from an AI inference cluster and can service models from an AI model repository. Triton Inference Server supports all major frameworks, such as TensorFlow, TensorRT, PyTorch, MXNet, Python, and more. Triton Inference Server also includes the RAPIDS Forest Inference Library (FIL)1 backend for GPU and CPU inference of Random Forests, GBDTs, and Decision Tree models. Triton with FIL backend delivers the best inference performance for tree-based models on GPUs, enabling simplified deployment of large tree models on GPUs with low latency and high accuracy.

Smart Hospital Workloads

AI Chatbots and Telehealth

Smart hospitals are utilizing Natural Language Processing (NLP) to develop and deploy AI chatbots to assist with appointments, gathering patient feedback, and triage of healthcare services. AI chatbots can help improve patient interactions by assisting with appointment scheduling and reminders.

Telehealth services can also utilize NLP tools to deploy Text to Speech (TTS) applications to gather patient feedback, symptoms, appointment notes, and physician recommendations and seamlessly interpret and input the appropriate data into the hospital's electronic medical record (EMR) system. NVIDIA AI Enterprise provides hospitals with AI tools and frameworks to streamline the development and deployment of NLP applications. Hospitals will be able to accelerate development with PyTorch and TensorFlow for training and TensorRT for optimization, and efficiently deploy in production with Triton Inference Server.

Medical Imaging and PACS

Radiology is a key part of the healthcare enterprise, enabling caregivers to peer into the human body to detect, diagnose, and treat the patient. Modalities such as X-Ray, ultrasound, CT, and MRI have transformed the diagnostic process, and healthcare organizations rely on GPUs to drive reconstruction and visualization of medical images. With the rise of computer vision and deep learning, hospitals are now able to use AI to augment this work to reconstruct 3D images quickly and accurately. Netherlands Cancer Institute (NKI) utilized NVIDIA AI Enterprise and NVIDIA A100 GPUs to reconstruct cone beam computed tomography (CBCT) scans in as little as 5 minutes at 1mm resolution. With near real-time reconstruction, NKI is researching the potential for more efficient and accurate radiation treatments for patients with cancer.

To develop next generation AI applications – whether it is for reconstruction, detection, classification, or segmentation, researchers and developers use deep learning frameworks to accelerate their work. MONAI (Medical Open Network for AI) is one such framework that accelerates medical imaging

research, with components like Label, Core and Deploy, that help with each step of the process. MONAI Label is an SDK that accelerates identifying ground truth in segmenting organs, tissues, and pathologies. MONAI Core is a framework for training and tuning AI models, with a shared set of libraries for data ingestion, transformations, and processing. MONAI Deploy enables developers to package up an AI model for delivery into a validation or deployment environment to connect into a medical ecosystem.

MONAI runs atop PyTorch, one of the containers used within NVIDIA AI Enterprise. NVIDIA AI Enterprise makes it easy to stand up an environment for running frameworks like MONAI, providing not only a common infrastructure, and accelerating not only the initial time -to-train - but also for getting enterprise support when it is needed.

Monitoring and Engagement

A hospital is built with the goal of caring for patients when they are sick or hurt – but at its core, it is a diverse place with many different personas. Whether it is the patient themselves, or their family or other visitors, or the doctors, nurses, technologists, specialists, surgeons, social workers, schedulers, unit coordinators, and many more – there is so much going on at any given moment. AI and computer vision can be used to augment clinical and administrative workflows, giving a boost to productivity and safety. Patient monitoring – for example, for those who are at risk of falling, or for those in the ICU – is an important activity that AI can help support and augment.

Federated Learning

In healthcare, where patient privacy is a hard requirement, it can be difficult to balance the constraints of data privacy and locality with the data requirements for building AI models. One technique that has emerged to address this need is federated learning, where models are trained at individual sites and aggregated to build a generalized global model without sharing data. NVIDIA FLARE is an open-source framework for federated learning that enables researchers and data scientists to build machine learning and deep learning applications in a federated paradigm. These applications can then be deployed securely, with privacy-preservation, across multiple hospitals and clinics.

Using NVIDIA AI Enterprise on NVIDIA-Certified systems simplifies this process by providing a consistent hardware and software stack across the set of participating institutions and ensuring that all underlying dependencies required to support the federated workflow are available and correctly configured.

Compute-as-a-service

As Smart Hospitals look to future-proof their data center, Compute-as-a-service (CaaS) becomes a solution as researchers and data scientists are defining future projects and workloads. Using NVIDIA AI Enterprise running on VMware vSphere, hospital IT can immediately provision VMs complete with necessary AI tools and frameworks as needed by researchers and data scientists.

Understanding the Smart Hospital Architecture

NVIDIA AI solutions enable healthcare organizations to support a variety of demanding workloads in both traditional data center and hybrid cloud deployments, leveraging NVIDIA AI Enterprise running on VMware vSphere with Tanzu, and NVIDIA-Certified systems. To understand what this looks like for hospitals and hospital workloads, NVIDIA has developed a full virtualized hospital stack that includes an EMR, a PACS, visualization, AI services, and other common hospital services.

Topology

The Smart Hospital data center is built to provide resilient, always-on services to doctors, nurses, and technicians, as well as clinical and operational teams. With NVIDIA AI Enterprise and NVIDIA GPUs, NVIDIA accelerates core hospital applications and frameworks including existing core hospital applications (EMR, PACS, radiology imaging), healthcare data science research, and emerging AI applications, on one common stack.

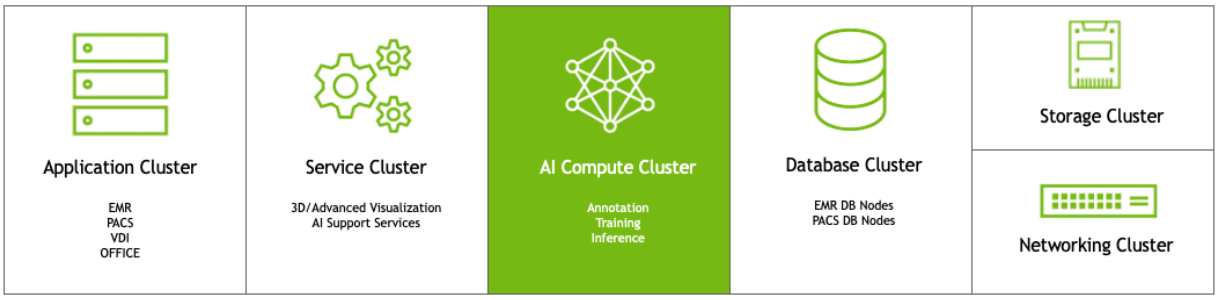

Figure 3 Smart Hospital Data Center

This is an important part of the healthcare enterprise data center roadmap: as applications continue to adopt innovative technologies through accelerating processing with GPUs, and accelerated networking and storage with DPUs, they will need a strategy to incorporate this for the enterprise. To support the thousands of applications healthcare needs in the coming years, a virtualization strategy will be paramount to success.

Sizing VMs

Selecting the correct NVIDIA Virtual GPU (vGPU) profile size for each workload will help to improve GPU utilization and workload density. Typically, machine learning (ML) training workloads are very computationally intensive and should be provided a full profile and/or multiple GPUs per virtual machine (VM) if available. Once model training is completed, inference workloads can often be deployed with a smaller fractional profile as it is less compute intensive.

Leverage the performance tools and troubleshooting section to help determine the resource demands of your deployment.

Unified Data Center

By leveraging the Smart Hospital architecture accelerated by NVIDIA AI Enterprise, healthcare organizations and hospitals can utilize a common infrastructure that supports the hospital applications of today and the AI-enhanced applications of the future.

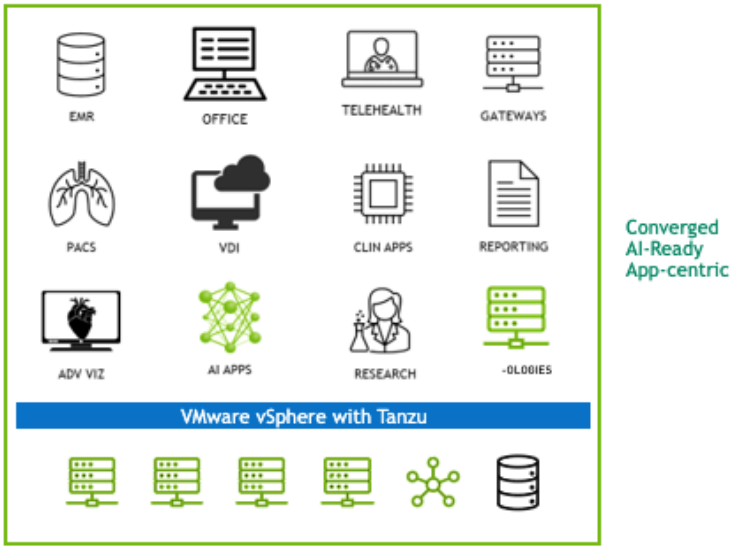

Figure 4 Unified Smart Hospital Data Center

Getting Started

Two Ways to Get Started

Depending on where you are in your AI journey, how much time you want to commit, and whether you have infrastructure in place, it is easy to get started with NVIDIA AI Enterprise. Choose the path that works for you: NVIDIA LaunchPad Hands-on Labs or On-Prem evaluation.

Take LaunchPad Hands-on Labs

Healthcare organizations can experience the difference of the NVIDIA AI-Ready Platform with free, immediate, short-term access to NVIDIA AI Enterprise on NVIDIA LaunchPad. NVIDIA LaunchPad includes a set of curated hands-on labs for AI practitioners and IT staff and healthcare organizations can speed the development and deployment of modern, data-driven applications and quickly test and prototype their entire AI workflow on the same complete stack that is available for purchase and deployment

Experience the difference of the NVIDIA AI Enterprise software suite with hands-on labs for every user. IT administrators will learn best practices for deploying and managing NVIDIA AI Enterprise. AI practitioners will learn how to optimize training and inferencing workloads using AI tools and frameworks, including NVIDIA RAPIDS™, PyTorch, NVIDIA® TensorRT™, TensorFlow, NVIDIA Triton™ Inference Server, and more.

https://www.nvidia.com/en-us/launchpad/ai/ai-enterprise/

On-Premises Eval

Getting started with NVIDIA AI Enterprise can be done by leveraging the NVIDIA Certified systems that are already in your data center. A complete list of NVIDIA AI Enterprise compatible systems can be found in the Qualified Systems Catalog. Familiar tools and frameworks like Triton, RAPIDS, PyTorch & TensorFlow are fully supported by NVIDIA on CPU based systems and can be accelerated with GPUs when necessary. Start with the CPU Only Deployment Guide to create an AI ready OS to deliver to your Data Scientists and AI Practitioners today!

Setting Up Your Smart Hospital

Every AI/ML workload will require various resource demands. Right sizing and configuring the hardware powering your AI/ML use cases is key to successful deployments. A hardware architect needs to ensure nodes are networked appropriately and each node has the optimal resources to meet demands.

For the Smart Hospital accelerated with NVIDIA AI Enterprise, a cluster of NVIDIA-Certified Systems with a minimum of four nodes is recommended. An NVIDIA-Certified System contains powerful NVIDIA GPUs and networking, which offers maximum performance per node. By adding high-performance NVIDIA Mellanox Networking, performance gains can be achieved when executing multi-node AI Enterprise workloads. The following table describes an example hardware configuration for each of the nodes within the cluster.

Table 1 Hardware Node Configuration

EGX Node Configuration	
Server Model	2U NVIDIA-Certified System
CPU	Dual Intel® Xeon® Gold 6240R 2.4G, 24C/48T
RAM	12 x 64GB RDIMM, 3200MT/s, Dual Rank
Storage	1 x 446GB SSD SATA Mix Use 6Gbps
Storage	1 x 6TB Enterprise NVMe
Network	Onboard networking
Power	Dual, Hot-plug, Redundant Power Supply (1+1), 1600W
Network	1 x NVIDIA® Mellanox® ConnectX®-6 Dx 100G
GPU	1 x NVIDIA A100 for PCIe

Connecting Nodes

This cluster size is the minimum viable size since it offers a balanced approach with NVIDIA GPUs and NVIDIA networking for various workloads. The cluster can also be expanded with additional nodes as needed and each node has identical hardware and software specifications. When creating an NVIDIA AI Enterprise cluster, leverage the Reference Architecture to guide your design decisions.

Sizing Nodes

When determining which resources to outfit your NVIDIA certified systems, leverage the Sizing Guide for a starting reference point. Three configurations are outlined, starting with an Entry level config that is designed to be dropped into your existing rack infrastructure without modifying the current power or networking. This configuration will allow an organization to support AI Enterprise workloads quickly but has lower performance throughput potential when compared to the Mainstream and best configurations. It is ideal for organizations who want to get started quickly. A Mainstream configuration builds upon the Entry configuration. Yet it provides more powerful GPUs and networking, which offers maximum performance per node while maximizing the number of nodes per rack. More throughput

between the nodes is achieved by adding high-performance NVIDIA Mellanox Networking, resulting in performance gains when executing multi-node AI Enterprise workloads.

Scaling Up Over Time

Going from proof of concepts to enterprise deployments needs effective scaling. This means making efficient use of precious GPU resources, as well as ensuring manageability and availability. The costs of provisioning siloed bare metal AI infrastructure or public cloud infrastructure can also be challenging.

With features like GPUDirect Communication integrated with vSphere, a fast path is provided between the NICs and GPU memory, resulting in boosted scale out performance enabling larger, more complex AI training and data analytics.

NVIDIA AI Enterprise can run in both bare metal and virtualized environments. The supported Bare Metal Operating Systems are RHEL and Ubuntu, while VMware vSphere is supported for virtualization. For a complete list of supported platforms refer to the NVIDIA AI Enterprise Product Support Matrix.

All the required software and drivers for running a smart hospital can be found on NGC and in the NVIDIA AI Enterprise Catalog for easy access and version control for your IT department.

Operationalizing The Smart Hospital

Resilience and Business Continuity

NVIDIA AI Enterprise also includes NVIDIA Enterprise Support, a guaranteed software license agreement (SLAs) for enterprise-grade support. Instead of waiting for a response from a community developer for example, an NVIDIA AI Enterprise customer will receive a response within four hours and can count on prioritization of their ticket. With access to NVIDIA AI experts who can provide guidance on how to optimally configure their AI infrastructure, enterprises can be more confident that their AI projects will be successful. For enterprises that require a stable, reliable environment, the constant change of open source can be challenging to manage. With NVIDIA AI Enterprise, businesses can control upgrade and maintenance schedules. Quarterly certified releases ensure developers have access to the latest software while long-term support for up to 3 years means the IT team can plan updates as needed. Additionally, NVIDIA AI Enterprise has been tested and certified with leading MLOps and industry ISV partners to ensure that your AI workloads can be easily deployed.

Enterprises deploying mission critical AI applications and requiring a higher level of support, can upgrade to Business Critical Support which provides 24x7 live agent access.

Measuring Performance and Troubleshooting

There are various tools for monitoring system health and utilizations for NVIDIA AI Enterprise virtualized deployments running on VMware vSphere. NVIDIA Management Library (NVML) and NVIDIA Data Center GPU Manager (DGCM) are tools for monitoring virtualized systems. Further details and examples of NVML and DGCM are included in the Appendix. Hospitals can also leverage VMware's monitoring solution vRealize Operations (vROPS) to provide additional insight for virtualized system.

Summary

AI is transforming the healthcare industry with the integration of AI-powered tools and applications to help hospitals improve patient experience, advance clinical research, and accelerate AI workflows. By leveraging NVIDIA and VMware's AI-Ready Platform, hospitals can unify AI and common hospital workloads in their data center with an end-to-end AI solution. NVIDIA AI Enterprise, which includes industry-leading AI tools and frameworks, such as TensorFlow, Pytorch, NVIDIA RAPIDS, and Triton Inference Server, enable hospitals to easily develop and deploy AI chatbots for healthcare services triage and appointment reminders, Telehealth and TTS applications to gather and integrate patient notes into an EMR system, AI models to quickly annotate and reconstruct high resolution images, and more. With NVIDIA LaunchPad, hospitals can kick start their AI journey with immediate, short-term access to NVIDIA AI Enterprise and gain hands-on experience to confidently develop and deploy AI workloads.

Resources

NVIDIA AI Enterprise Healthcare solution brief

NVIDIA AI Enterprise Documentation

NVIDIA LaunchPad

NVIDIA LaunchPad Documentation

Netherlands Cancer Institute Customer Success Story

Appendix

NVIDIA MANAGMENT LIBRARY (NVML)

There are two main areas in the NVIDIA AI Enterprise stack, Operating System, and Cloud Native, which can be monitored using NVIDIA's monitoring tools. Both layers leverage NVIDIA's telemetry technology built into the NVIDIA Management Library (NVML), while the cloud native layer leverages the NVIDIA Data Center GPU Manager (DCGM) which is built on top of NVML.

The NVIDIA System Management Interface (nvidia-smi) is a command line utility, based on top of the NVIDIA Management Library (NVML), intended to aid in the management and monitoring of NVIDIA GPU devices.

AI workloads can run on the OS within container but if you want to fully embrace containers, support for Kubernetes clusters is offered as well. With this in mind, NVML offers GPU monitoring within the following:

- OS
- Container (via docker)
- Container managed by Kubernetes (via kubectl)

The main advantage of NVML, is that it runs with very low overhead. This tool is ideal for a quick look at your system's current GPU telemetry and can be leveraged by running from the OS or a container runtime.

NVIDIA-SMI

Example of running nvidia-smi in an OS:

```
nvidia@cloud-native-core:~$ nvidia-smi
Mon May 23 20:22:14 2022
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 470.63.01    Driver Version: 470.63.01    CUDA Version: 11.4     |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
```

```
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|                               |                      |               MIG M. |
|===============================+======================+======================|
|   0  GRID A100-8C        On   | 00000000:02:02.0 Off |                    0 |
| N/A   N/A    P0    N/A /  N/A |    656MiB /  8187MiB |      0%      Default |
|                               |                      |             Disabled |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                                  |
|  GPU   GI   CI        PID   Type   Process name                  GPU Memory |
|        ID   ID                                                   Usage      |
|=============================================================================|
|  No running processes found                                                 |
+-----------------------------------------------------------------------------+
```

Example of running nvidia-smi in a container:

```
$ docker run --gpus all --rm -it nvcr.io/nvaie/pytorch-2-0:22.02-nvaie-2.0-py3


$ docker ps -a
CONTAINER ID   IMAGE                                          COMMAND
CREATED              STATUS                  PORTS               NAMES
ff479dfac216   nvcr.io/nvaie/pytorch-2-0:22.02-nvaie-2.0-py3   "/opt/nvidia/nvidia_…"   About
a minute ago   Up About a minute       6006/tcp, 8888/tcp   tender_shamir


$ docker exec -it ff479dfac216 nvidia-smi
Mon May 23 21:19:21 2022
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 470.63.01    Driver Version: 470.63.01    CUDA Version: 11.6     |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|                               |                      |               MIG M. |
|===============================+======================+======================|
|   0  GRID A100-8C        On   | 00000000:02:02.0 Off |                    0 |
| N/A   N/A    P0    N/A /  N/A |    656MiB /  8187MiB |      0%      Default |
|                               |                      |             Disabled |
```

```
+-----------------------------+---------------------+----------------------+


+-----------------------------------------------------------------------------+
| Processes:                                                                  |
|  GPU   GI   CI        PID   Type   Process name                  GPU Memory |
|        ID   ID                                                   Usage      |
|=============================================================================|
|  No running processes found                                                 |
+-----------------------------------------------------------------------------+
```

Example of running nvidia-smi from a Kubernetes managed container:

```
$ kubectl exec -ti tensorflow-jupyter-notebook-5796d755d7-qgrps -- bash
root@tensorflow-jupyter-notebook-5796d755d7-qgrps:/workspace# nvidia-smi
Fri May 20 18:48:08 2022
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 510.47.03    Driver Version: 510.47.03    CUDA Version: 11.6     |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|                               |                      |               MIG M. |
|===============================+======================+======================|
|   0  NVIDIA A30-24C      On   | 00000000:02:00.0 Off |                    0 |
| N/A   N/A    P0    N/A /  N/A |      0MiB / 24576MiB |      0%      Default |
|                               |                      |             Disabled |
+-------------------------------+----------------------+----------------------+


+-----------------------------------------------------------------------------+
| Processes:                                                                  |
|  GPU   GI   CI        PID   Type   Process name                  GPU Memory |
|        ID   ID                                                   Usage      |
|=============================================================================|
|  No running processes found                                                 |
+-----------------------------------------------------------------------------+
```

NVIDIA-SMI LOGGING

NVIDIA-SMI can be used to log GPU metrics to a file over time. This is accomplished via the -l & -f options. Add the option "-f <filename>" to redirect the output to a file and the -l to change the polling interval in seconds

Example:

```
$ nvidia-smi --query-gpu=timestamp,name,pci.bus_id,driver_version,pstate,pcie.link.gen.max,pcie.link.gen.current,utilization.gpu,utilization.memory,memory.total,memory.free,memory.used --format=csv -l 5 -f output.csv

$ cat output.csv

timestamp, name, pci.bus_id, driver_version, pstate, pcie.link.gen.max, pcie.link.gen.current, utilization.gpu [%], utilization.memory [%], memory.total [MiB], memory.free [MiB], memory.used [MiB]

2022/05/23 20:47:58.253, GRID A100-8C, 00000000:02:02.0, 470.63.01, P0, [N/A], [N/A], 0 %, 0 %, 8187 MiB, 7531 MiB, 656 MiB
```

Additional information on nvidia-smi can be found here.

Short-term logging

Example logging use cases:

Purpose	nvidia-smi "-l" value	interval
Fine-grain GPU behavior	5	5 seconds
General GPU behavior	60	1 minute
Broad GPU behavior	3600	1 hour

Long-term logging

Create a shell script to automate the creation of the log file with timestamp data added to the filename and query parameters

Add a custom cron job to /var/spool/cron/crontabs to call the script at the intervals required.

NVIDIA DATA CENTER GPU MANAGER (DCGM)

NVIDIA Data Center GPU Manager (DCGM) is a suite of tools for managing and monitoring NVIDIA datacenter GPUs in cluster environments. DCGM integrates into the Kubernetes ecosystem using DCGM-Exporter to provide rich GPU telemetry in containerized environments. This exposes GPU metrics to an HTTP endpoint for monitoring and reporting solutions such as Prometheus and Grafana. To get started with DGCM Explorer, refer the docs.nvidia.com

DCGM can also be used standalone by infrastructure teams and easily integrates into cluster management tools, resource scheduling and monitoring products from NVIDIA partners. It includes active health monitoring, comprehensive diagnostics, system alerts and governance policies including power and clock management.

The following graphic illustrates Kubernetes orchestration in Virtualized environment. The NVIDIA GPU Operator is highlighted which manages all Kubernetes components includes DCGM monitoring:

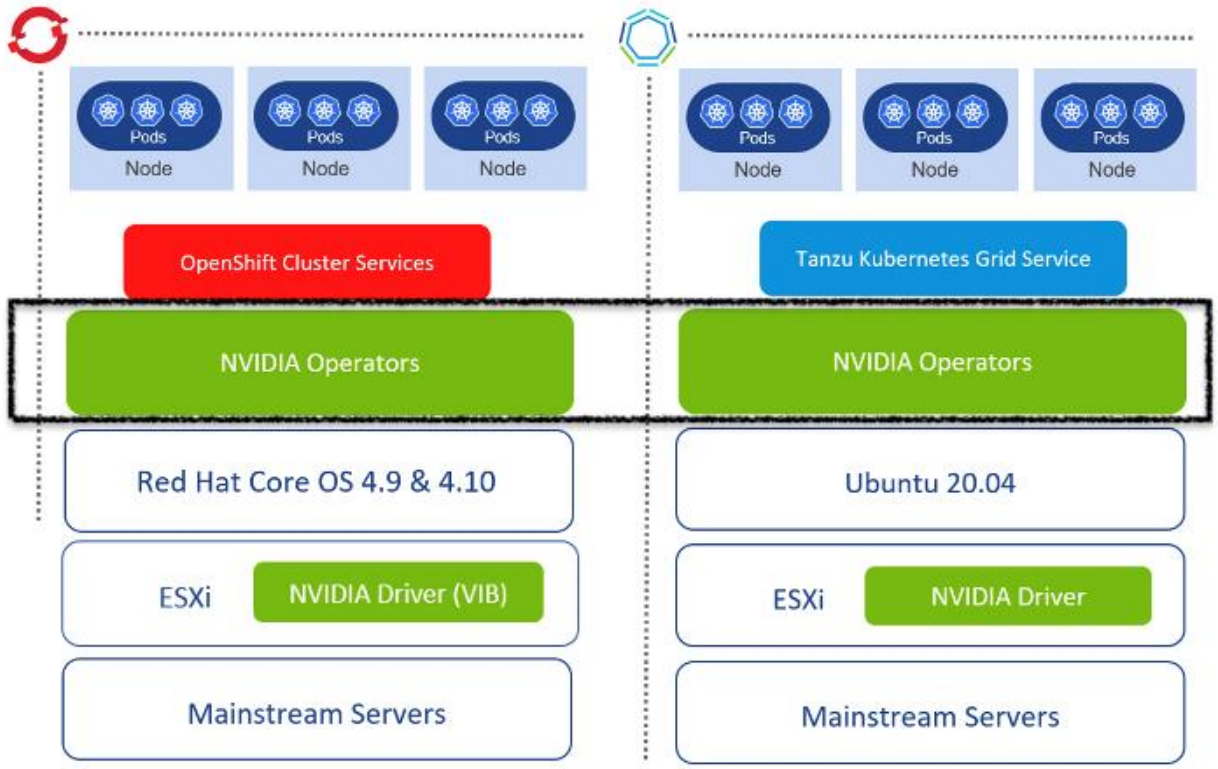

Figure 5 Kubernetes Orchestration