驅動教學、科研與場域創新：以 NVIDIA 全棧 AI 軟體平台建構前瞻校園 AI 生態系

Senior Solutions Architect, NVIDIA

2026 Aug 21

Agenda

- 校園 AI 化挑戰與藍圖
- BCM 自動化叢集維運
- Run:ai 算力池化與切分
- NIM 雲原生 AI 微服務
- Agent Toolkit & NeMo Claw
- 效益總結與交流問答

校園 AI 化挑戰與藍圖

校園 AI 化面臨的常見挑戰

算力孤島現象： 各系所獨自採購 GPU 伺服器，資源無法跨跨單位彈性共享，閒置與不足並存。

維運管理繁鎖： 傳統 AI 叢集部署耗時，IT 人員需花費大量時間維護環境與韌體升級。

自主 Agent 落地困難： 師生在部署生成式 AI 與開發校園應用時，缺乏標準化推論服務框架以及隔離防護。

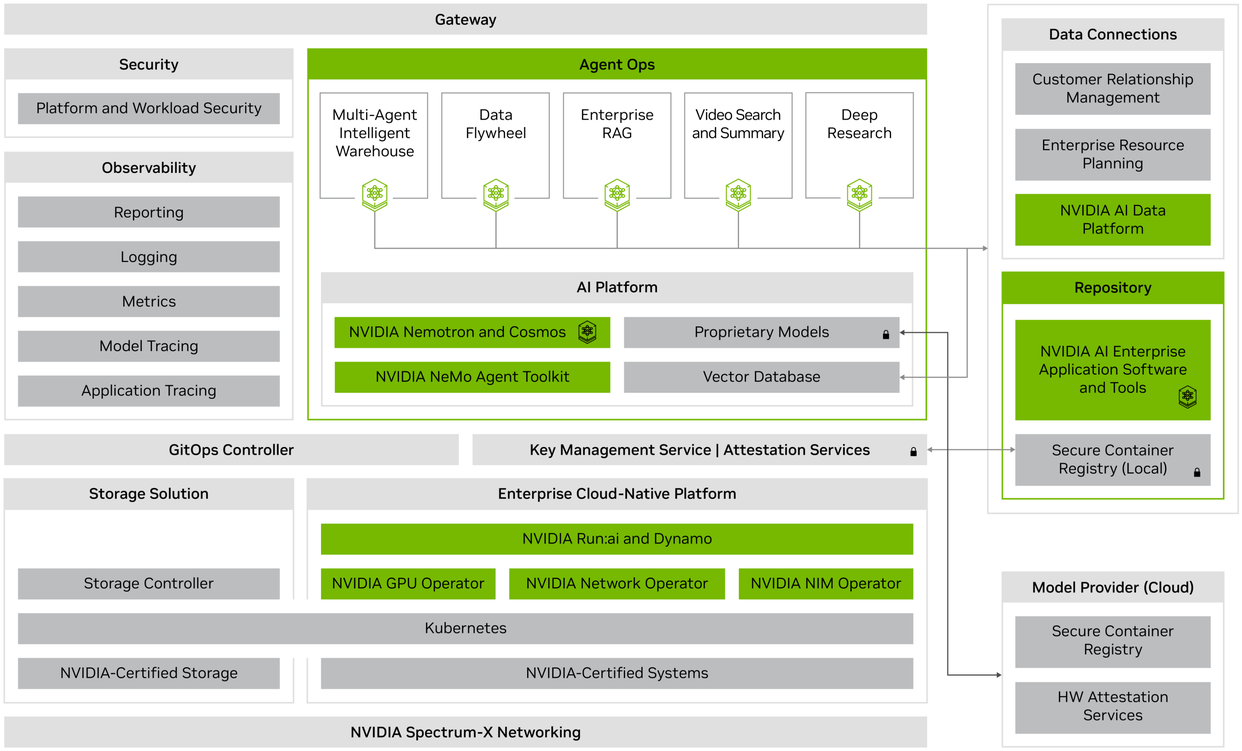
NVIDIA 全棧 AI 軟體生態藍圖

從硬體維運到自主 Agent 應用

NVIDIA 提供貫穿 AI 生命週期的完整軟體架構：

由底層營運 (BCM)、算力池化 (Run:ai)、雲原生推論 (NIM) 至 Agent 工具與沙盒安全防護 (Agent Toolkit & NemoClaw)。

此架構無縫串聯 IT 維運、教學實作、前瞻科研與自主代理應用，全面普及校園 AI 創新。無縫串聯 IT 維運、教學實作、前瞻科研與校園智慧場域，全面普及校園 AI 創新。



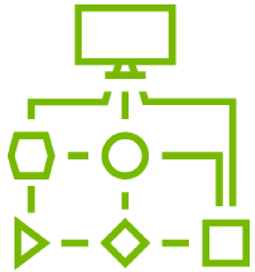
BCM 自動化叢集維運

Base Command Manager

Purpose-built for Enterprise AI Infrastructure Management

Infrastructure Provisioning

- Maintain a secure, up-to-date, and reliable AI infrastructure



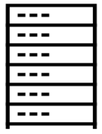
Workload Management

- Easily provide data scientists with all the tools and resources they need

Resource Monitoring

- Get detailed insights for informed decision-making

Cloud



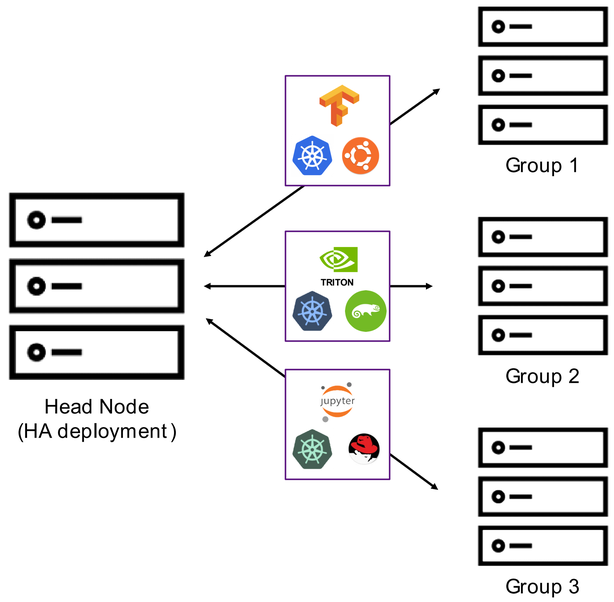
Data Center

Edge

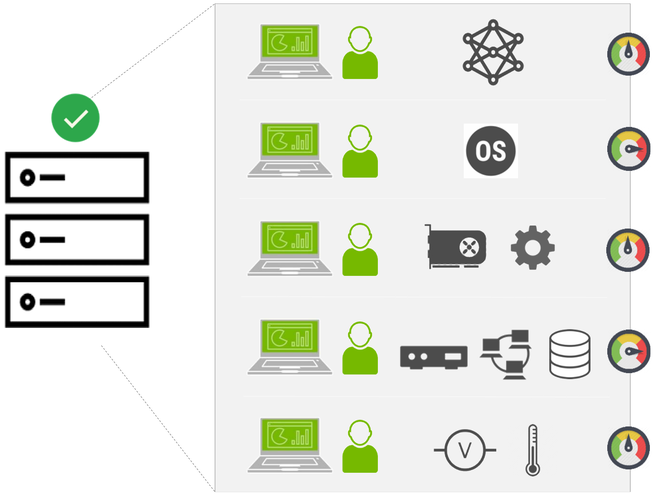
Infrastructure Management

Base Command Manager

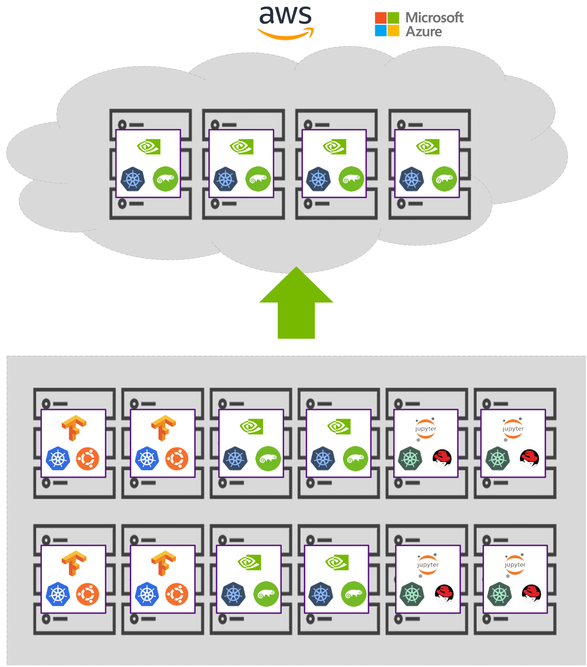
Image Management and Configuration



Utilization Metrics



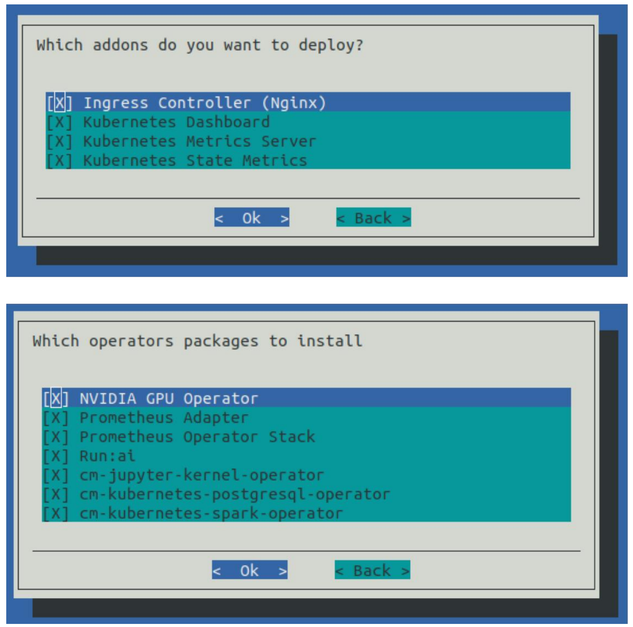
Cloud Integration



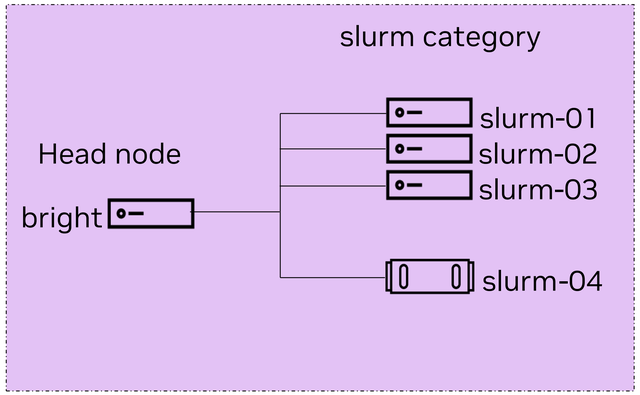
Workload Management

Base Command Manager

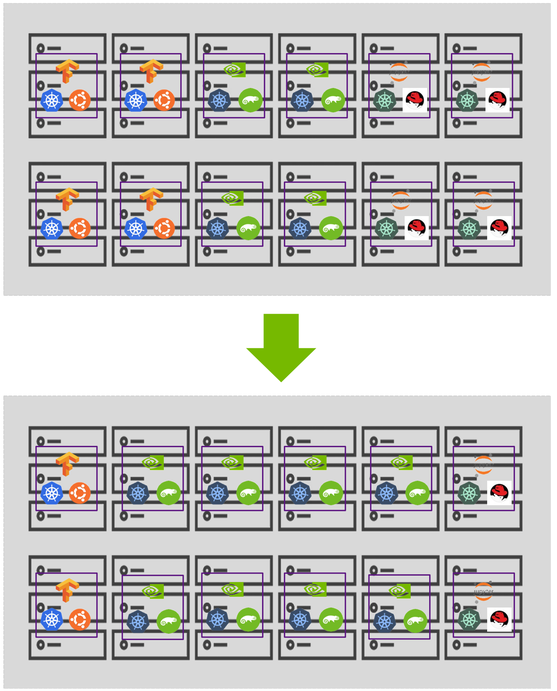
Kubernetes



Slurm



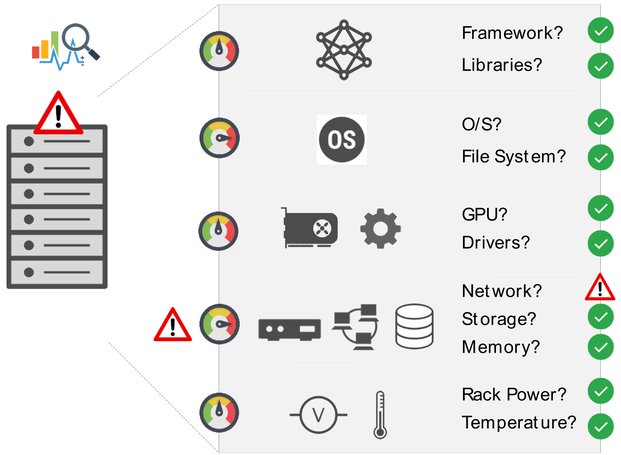
Auto Scaler



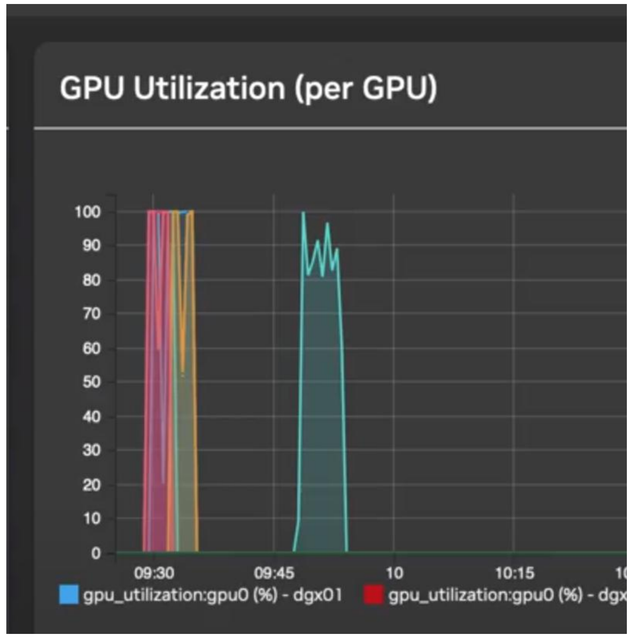
Resource Monitoring

Base Command Manager

Health Monitoring and Remediation



NVIDIA GPU Management and Monitoring



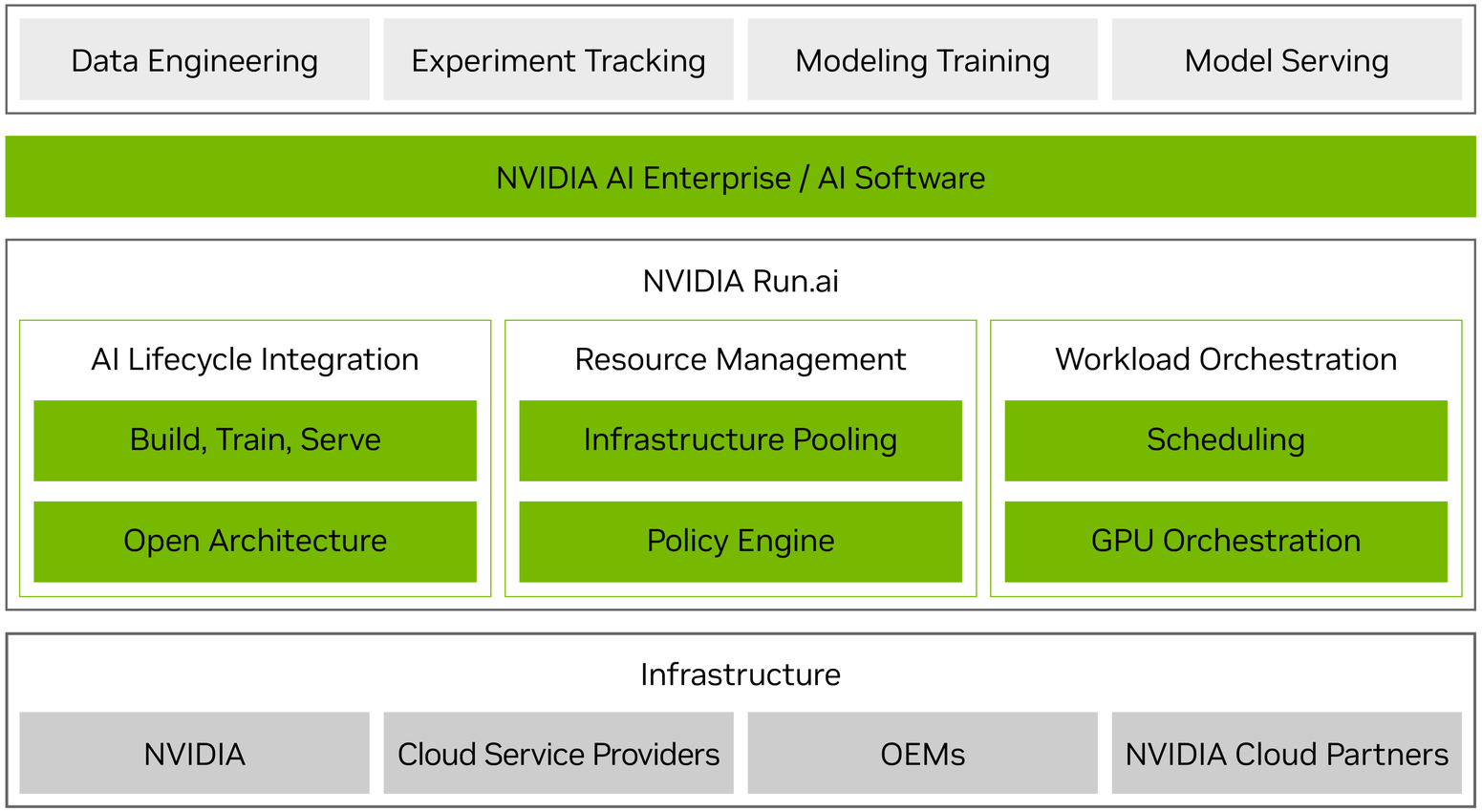
Reporting and Chargeback

User	Jobs	Runtime (s)	CPU ($)
bob	1	1804	0.3608
charline	1	1804	0.7216
dennis	2	3606	1.0818
frank	1	1803	0.1803

Run:ai 算力池化與切分

NVIDIA Run:ai Platform

AI Workload and GPU Orchestration to Build, Train and Deploy AI Workloads at Scale



GPU Availability

10X

Workloads Running

20X

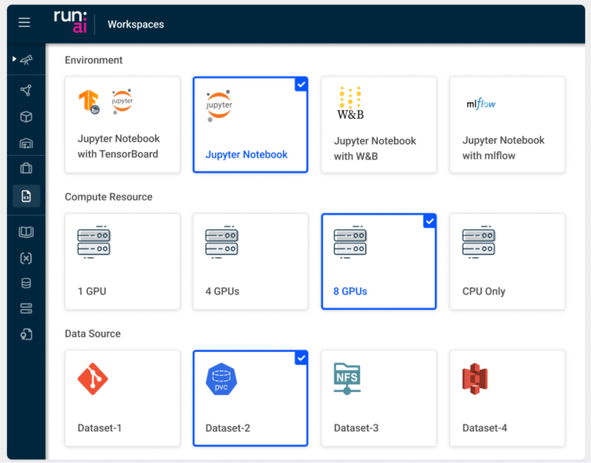
GPU Utilization

5X

AI LifeCycle Integration

Accelerate AI development with comprehensive life cycle support from concept to deployment

Spin up pre-defined environments with your tools, libraries and data in one place



WORKSPACES

Iterate faster, run more jobs and scale up your training to multiple nodes

```
$ runai list jobs --project my-project

~ showing jobs for project my-project
name           GPUs allocated     type          status
my-workspace   1                  interactive   running
my-job-0       1                  train         running
my-job-1       4                  train         running
my-job-2       1                  train         pending
my-job-3       1                  train         pending
my-job-4       1                  train         pending
my-job-5       2                  train         pending
```

EXPERIMENTS & TRAINING

Increase efficiency and move faster from the lab to production

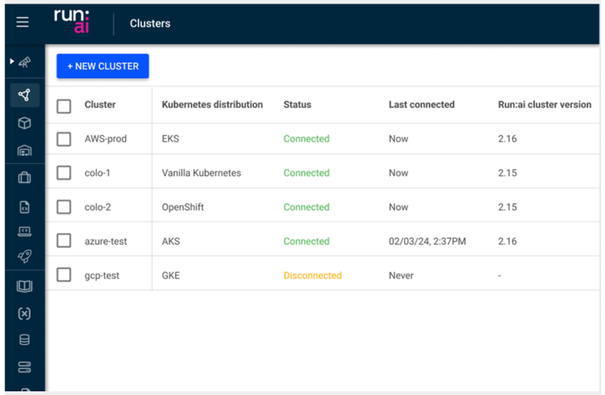
- Model Catalog
- Scale to Zero
- GPU Fractions

INFERENCE

Resource Management

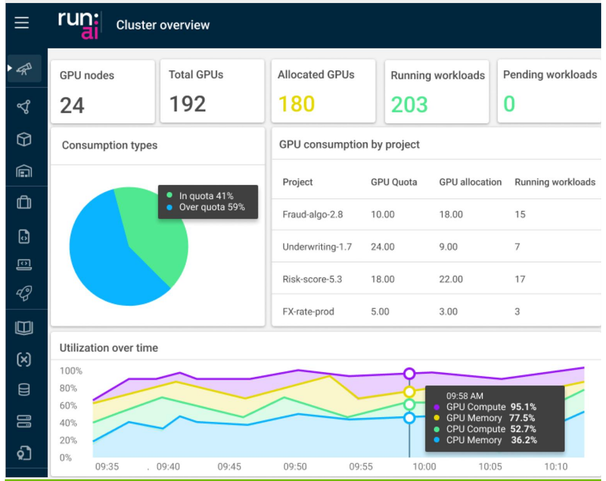
Take control and get visibility into your AI infrastructure, workloads, and users

Consolidate and manage infrastructure across public clouds and on-premises



CENTRALIZED MANAGEMENT

Improve planning and decision making with insights into historical utilization patterns



VISIBILITY

Optimize cluster utilization and workload scheduling with Run:ai's advanced policy engine

- Guaranteed Quotas
- Over Quota
- Hierarchical Quotas
- Priority Overrides

POLICY ENGINE

Workload Orchestration

Dynamically Orchestrate AI workloads and resources for optimal GPU utilization

Continuously optimize compute allocations based on policy driven priorities and quotas

- Fairshare Scheduling
- Job Queues
- Guaranteed Quotas
- Bin Packing and Consolidation
- Gang Scheduling

AI WORKLOAD SCHEDULER

Maximize efficiency and reduce costs. Perfect for Jupyter Notebook farms and Inference

```
$ runai submit my-inference -i nvcr.io/nvidia/tritonserver -g 0.5

$ runai list jobs --project my-project
name              GPUs allocated     type
my-inference      0.5                inference
my-notebook       0.2                interactive
my-job            0.3                train
```

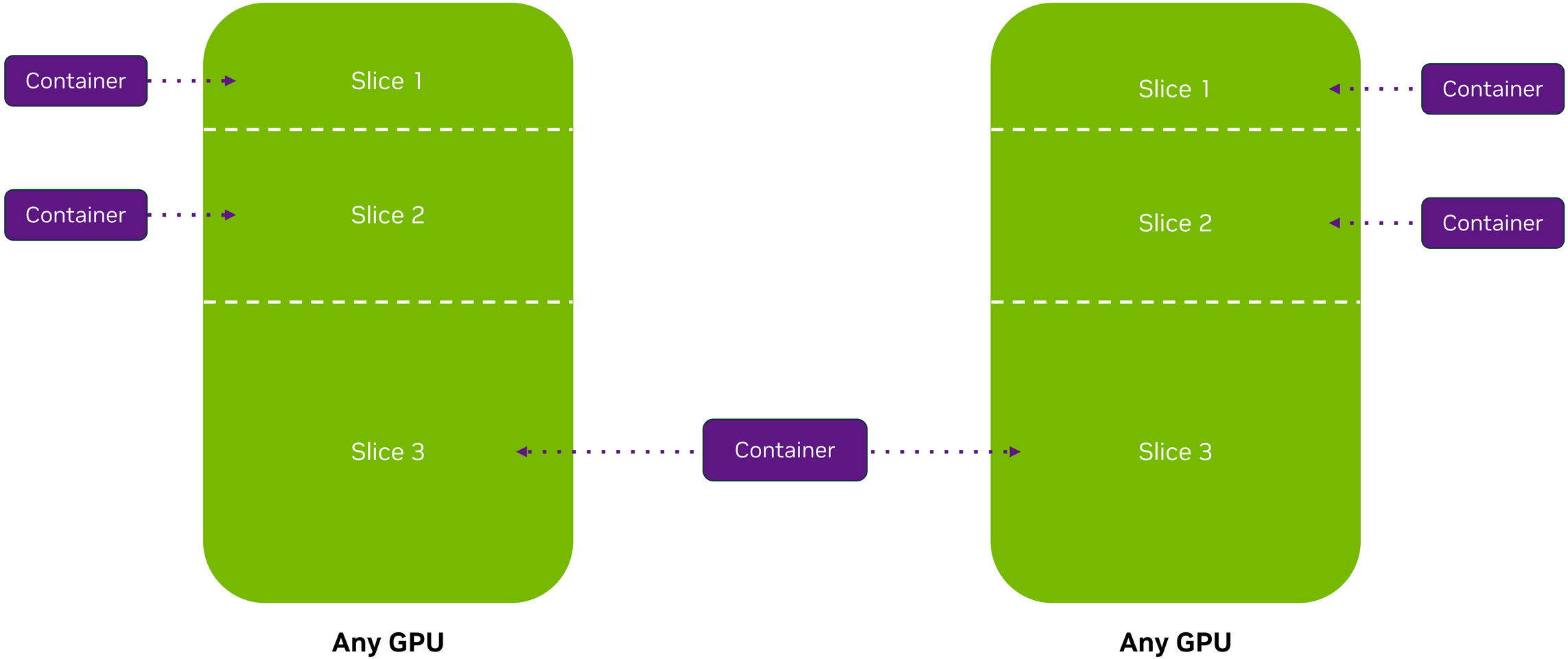
GPU FRACTIONING

Dynamically offload idle workloads to CPU, freeing GPUs for high-priority tasks, enabling multiple model replicas to share GPUs during low demand times.

Extends beyond inference to interactive workspaces

GPU MEMORY SWAP

NVIDIA Run:ai Multi-GPU Fractioning (v.2.20)



Working with NVIDIA Run:ai

Customer administrator and user workflows

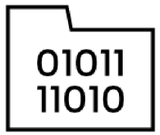
Administrator flow

Implement Core Run:ai Configurations

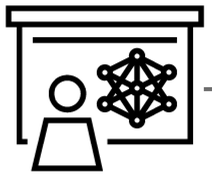
Set Resource Quotas

Onboard Users

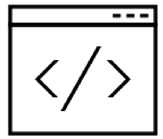
User flow



Integrate Data and Tools



Interactive Experiments



Distributed Training

Inference & Deployment

Working with NVIDIA Run:ai

Customer administrator and user workflows

Administrator flow

Implement Core Run:ai Configurations

Set Resource Quotas

Onboard Users

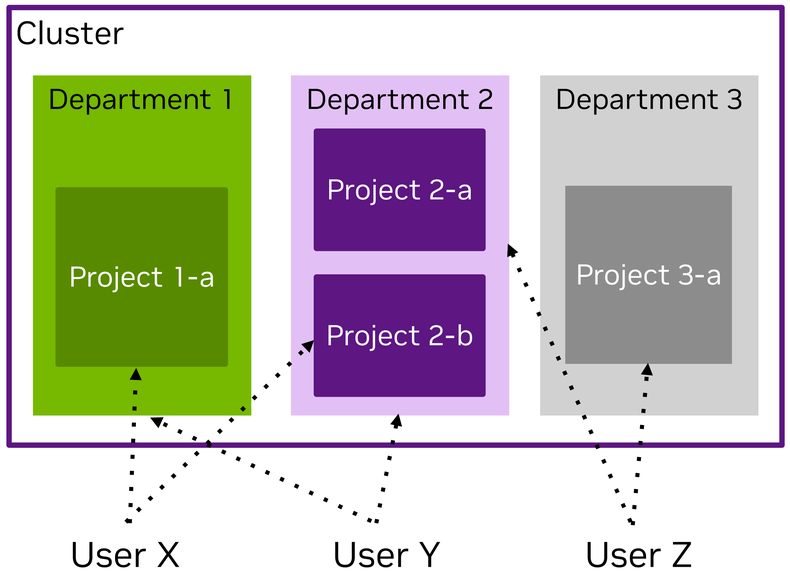
Managing Departments and Projects

NVIDIA Run:ai Scopes and Roles

Core Configurations

Resource Quotas

Onboard Users



Departments

- Contain Projects
- Assigned quota – recommend assigning greater than what is required by the aggregate Projects

Projects

- Assigned quota for specific initiatives
- Can be scoped to a single user, or to many
- Projects can have rules that control Workload behaviors (duration, node types, activity requirements)

Users

- Individuals with the ability to interact with Run:ai
- Administrators support cluster configuration and availability
- Users initiate and monitor the execution of Workloads
- A rich Role-Based Access Control system provides a wide variety of roles

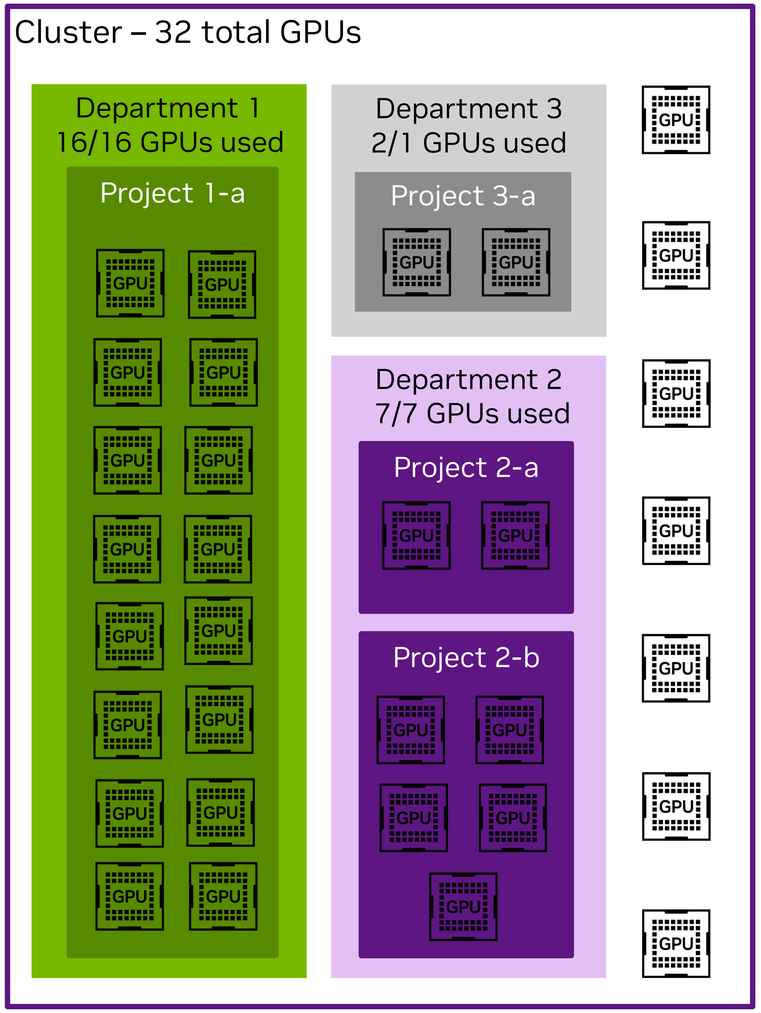
NVIDIA Run:ai Managing Quota

Flexibly Allocate GPUs between Users and Departments

Core Configurations

Resource Quotas

Onboard Users



Quota

- Maximum assigned resources for either a Department or Project
- Configurable for GPU devices, CPU cores, and CPU memory
- Oversubscription can be enabled at Department and/or Project, allowing unused resources to be leveraged when idle

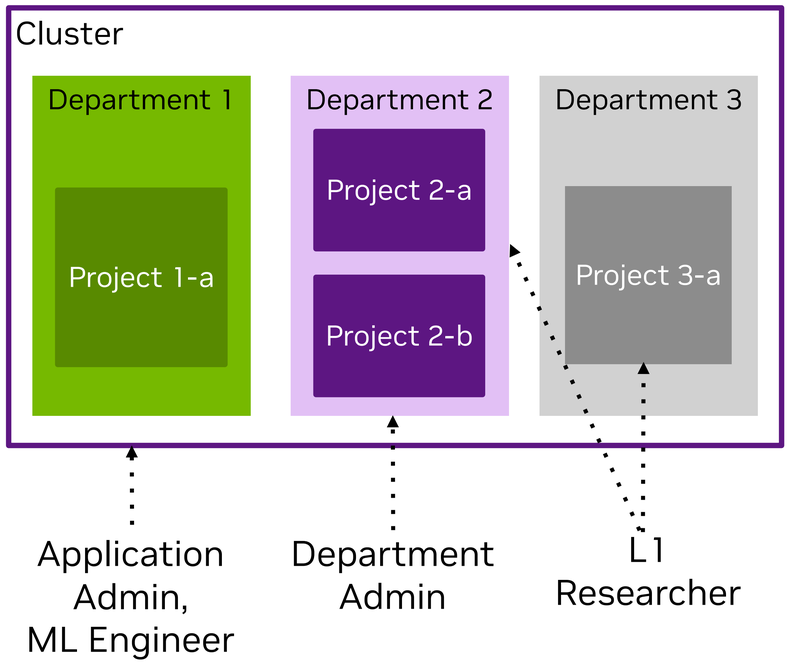
NVIDIA Run:ai Roles and Responsibilities

Customer admin and customer user personas

Core Configurations

Resource Quotas

Onboard Users



Admins are given Department Administrator and Application Administrator roles. These roles can:

- Create and edit projects
- Manage and assign access roles to users
- Create and edit components in the cluster, including compute resources, data sources and credentials
- Application admin can also create departments
- Customer roles cannot edit nodes or node pools on the cluster

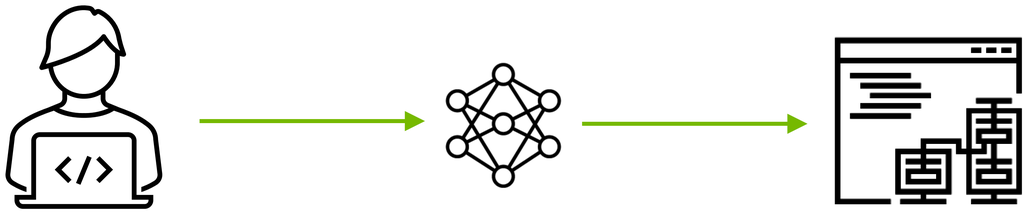
Customer User roles include:

- **L1 Researcher** – assigned to specific projects and can only submit workloads in those projects
- **Research Manager** – has view-only access to workloads and can create environments, resources, templates and data sources
- **ML Engineer** – cannot interact or create deployments on the cluster as inference is not available but can view departments, projects, node pools, nodes and dashboards within the cluster

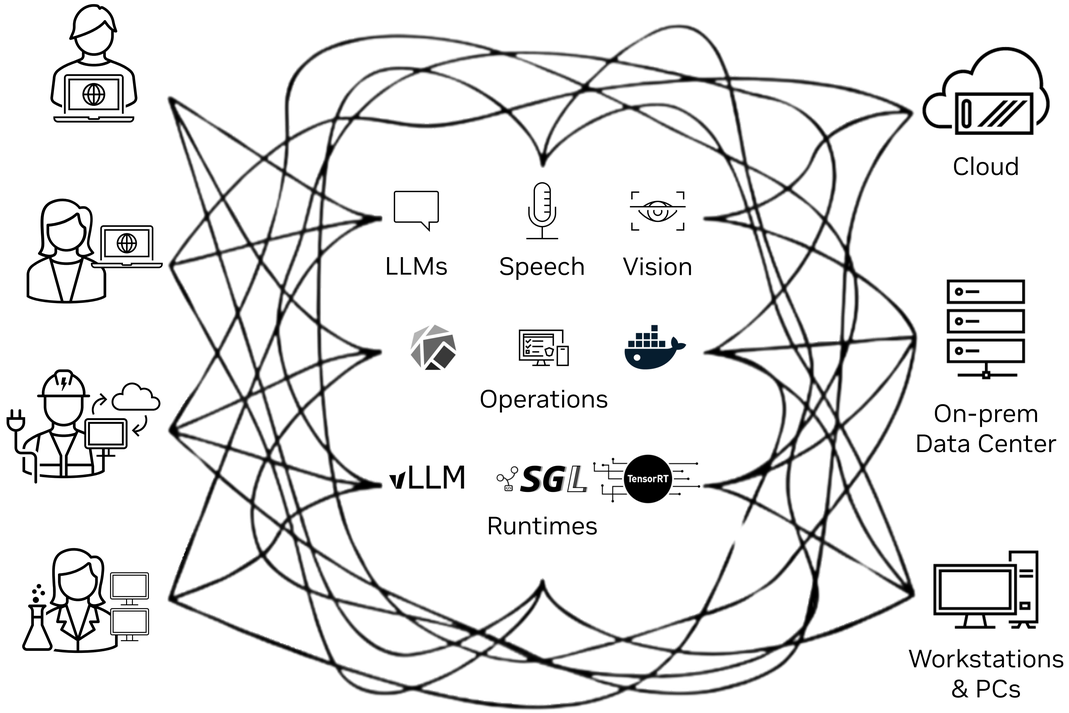
NIM - 雲原生 AI 微服務

DIY Works Until it Doesn’t

Prototyping



Operational Sprawl



Every new use case becomes a system of models with each team chooses different tools, frameworks, and runtimes, leading to massive complexity.

NVIDIA Building Blocks for AI

Blueprints

Research Assistant Agent

Customer Service Agent

Software Security Agent

Virtual Lab Agent

Video Analytics Agent

NVIDIA NeMo

Curator | Customizer | Evaluator | Guardrails | Retriever | Agent Toolkit

NVIDIA NIM & Nemotron

- Understanding & Reasoning
- Information Retrieval
- AI Safety
- Digital Humans
- Visual Content Generation
- Digital Biology
- Physical AI

AI Infrastructure

GPU | CPU | DPU | Networking

NVIDIA NIM

Scale models, not ops

Portable Run and scale cloud-native containers anywhere, maintaining control of data and apps

Easy Spin up production-ready endpoints in minutes with built-in observability, autoscaling and continuous updates

Enterprise Ready Gain confidence with NVIDIA-secured and verified software including monitoring and patching

Performant Leverage the best performing inference technology, preoptimized and continuously updated

vLLM NVIDIA TensorRT-LLM SGL

Microsoft Azure aws Google Cloud ORACLE DGX & DGX Cloud cisco DELL Technologies Hewlett Packard Enterprise Lenovo SUPERMICRO

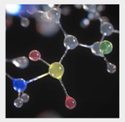
NVIDIA

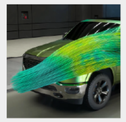
NVIDIA AI Enterprise

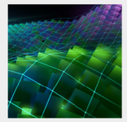
Cloud Native Software Platform for Production AI

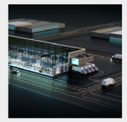

Deep Agents


Customer Service


Scientific Exploration


CAE Digital Twin


Data Analytics


Factory Digital Twin


Robotics Simulation

Application Development

Agentic AI
NVIDIA NeMo

Physical AI
NVIDIA Omniverse

Models
Cosmos
Nemotron
Frontier Models

Perception and Domain Skills | Data Processing | Foundational Frameworks

CUDA-X Libraries
cuDNN, cuOpt, cuML, cuDF, cuVS

Infrastructure Management

Model Serving
NVIDIA Dynamo-Triton

Workload and GPU Orchestration
NVIDIA Run:ai

Virtualization | Operators

Deploy Anywhere

Cloud
Data Center
Edge
Local Workstation

Features and Benefits

Lower infrastructure costs and reduce time to market while ensuring reliable, secure, and scalable operations

Build with Confidence

Security & Compliance

API Stability

Enterprise Support

Accelerate time to value

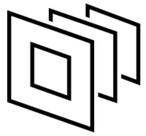
Ready-to-use, optimized microservices

Up to ***10x*** greater GPU availability for Developers

Optimize AI infrastructure utilization

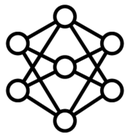
Up to ***5x*** GPU Utilization

Up to ***20X*** more workloads running



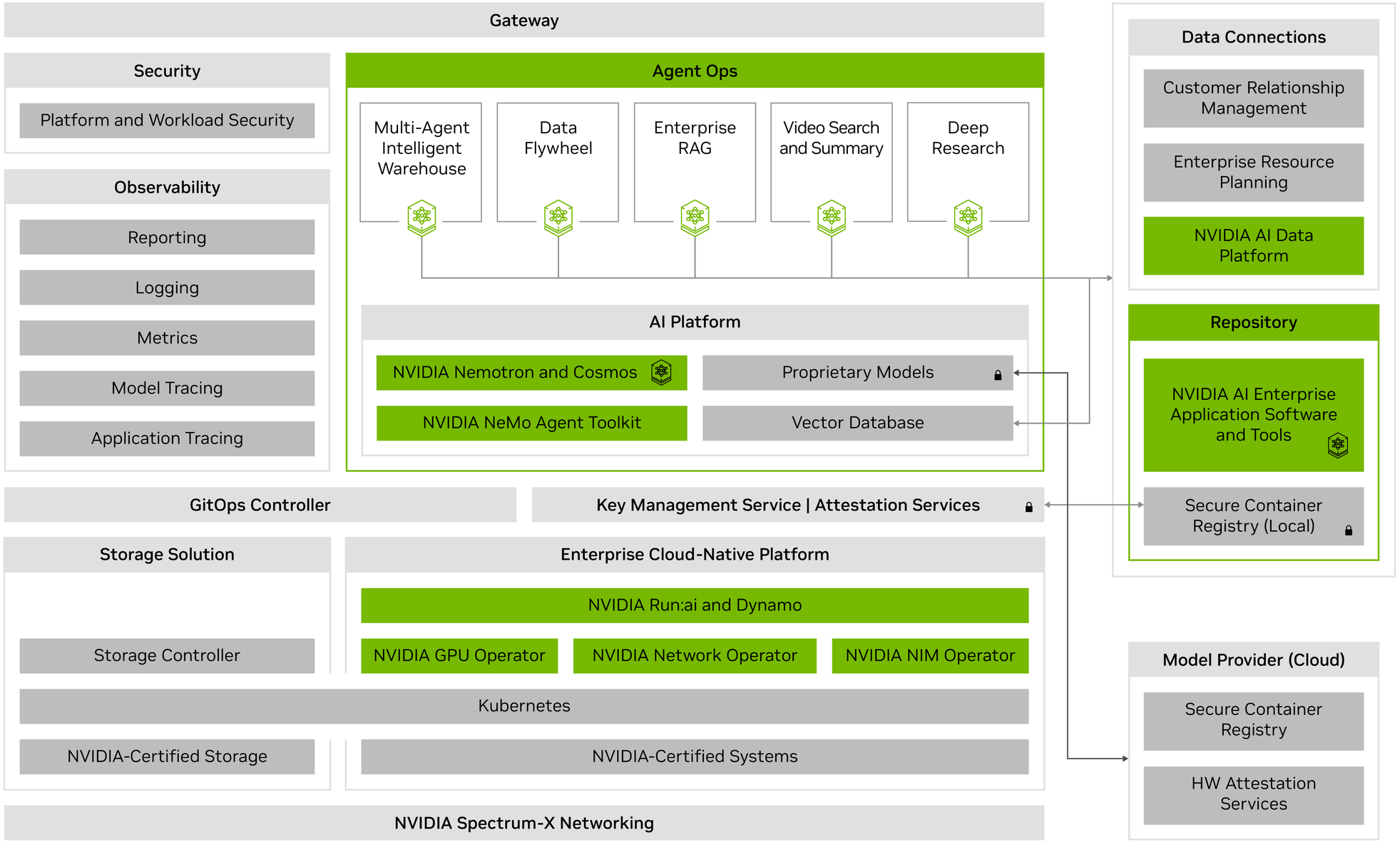
OpenUSD Libraries

Leading Open-Source AI Software



World-class Open Models

NVIDIA Enterprise AI Factory for Agents – Functional Architecture 2.0



NVIDIA Agent Toolkit

The Evolution of AI Agents

From LLMs and multi-turn prompts to long-running agents that modify themselves

ChatGPT

Nov 2022

Foundation LLMs,
Multi-turn Prompts

DeepSeek

Jan 2025

Reasoning Enables Reflection
For Better Outcomes

100x

OpenClaw

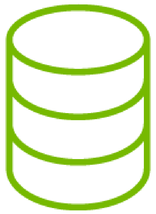
Jan 2026

Self Evolving Agents
Working on Your Behalf

1000x

Useful AI Has Arrived

The future workforce is biological and digital — working together, making each other extraordinary.



Informed

by the organization's knowledge and domain experience

Specialized

for proprietary workflows, open & fully controllable

Optimized

for frontier-level accuracy, lower cost at enterprise scale

Governed

so that they operate in a safer, more trustworthy manner

NVIDIA Agent Toolkit Pillars for Agentic AI

Open Models

Harness Accelerations

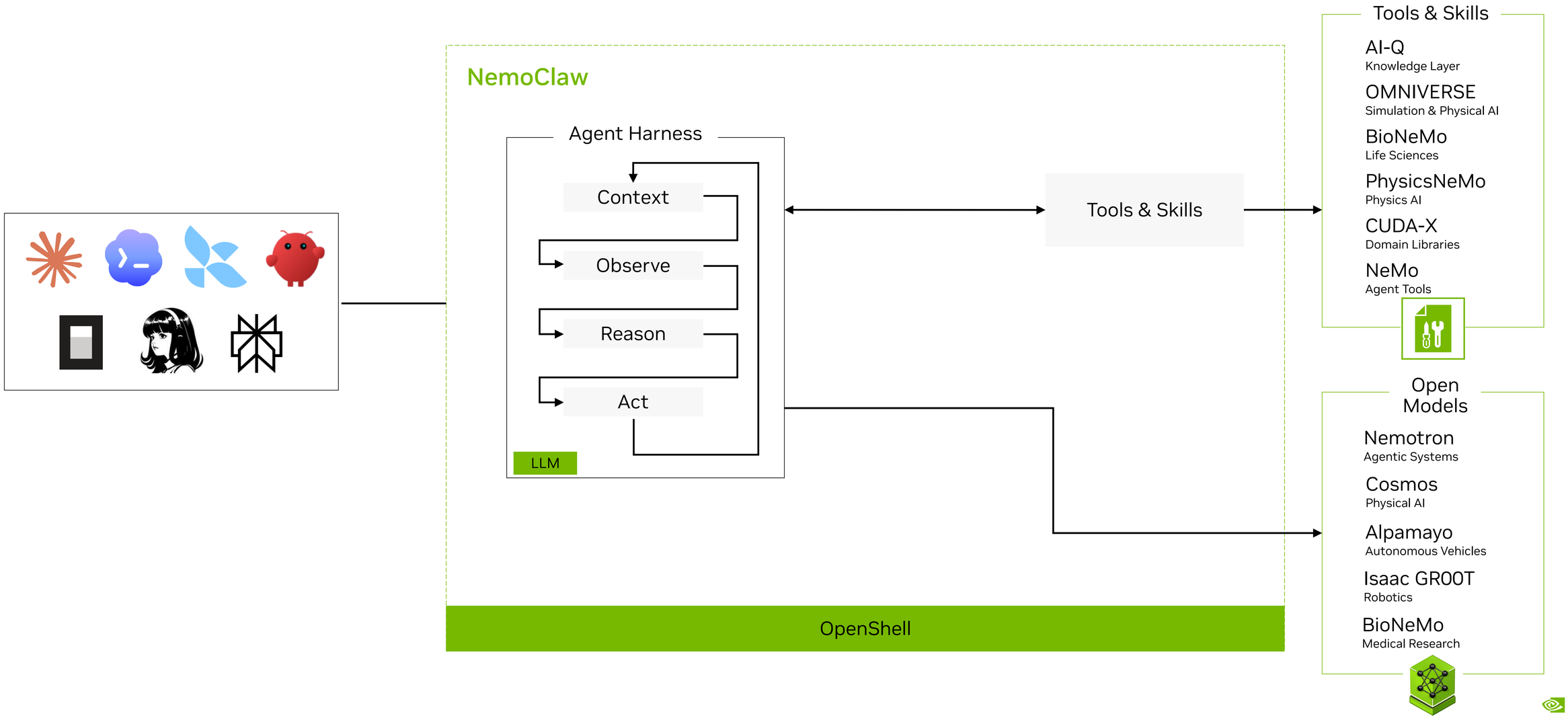
Tools & Skills

Runtimes & Orchestration

NemoClaw Blueprints

NVIDIA Agent Toolkit for Enterprise AI

Building Blocks for Specialized Agents



效益總結

BCM：叢集自動化維運基石

自動化極速部署

支援全校 GPU 伺服器的快速監控、作業系統配置與韌體升級，將數週的部署時間縮短至數小時。

BaseView 統一儀表板

提供單一可視化營運面板，即時掌握硬體健康度、使用趨勢與瓶頸，極小化 IT 維運負擔。

異質與雲端整合

靈活整合 Slurm 與 Kubernetes 容器化環境，並可隨需擴展至各大公有雲混合管理。

Run:AI：極大化校園算力價值

- **跨系所算力池化：** 打破硬體邊界，將全校 GPU 併入統一資源池進行公平調度。
- **Fractional GPU 切分：** 支援 GPU 記憶體動態切分，單張 GPU 可切分給數十位學生同時進行課程實作。
- **優先級與動態搶佔：** 大型科研專案可自動獲取高優先級算力，夜間自動調度閒置資源。

NIM：雲原生 AI 推論微服務

標準化容器部署

以開箱即用的 OCI 容器包裝主流大模型，大幅簡化地端推論部署流程。

極速推論效能

整合 TensorRT-LLM 引擎，提供極低延遲與超高吞吐量的推論服務。

開放 API 產業標準

提供標準 OpenAI 兼容 API，方便校方應用開發者無縫串接各項系統。

Agent Toolkit & NemoClaw：安全自主代理

- **NVIDIA Agent Toolkit：** 提供統一輕量庫與介面，快速將校務資料庫、科研工具與 Agent 進行對接串聯。
- **NemoClaw 沙盒安全堆疊：** 結合 OpenShell 安全執行環境，單一指令部署具備預設拒絕網路政策與權限隔離的防護網。
- **長時間運行助理 (Always-On Agents)：** 支援 Hermes Agent 與 Deep Agents 框架，讓 Agent 在安全環境中執行複雜任務。
- **在地端隱私保障：** 搭配 Nemotron 開源模型與 OpenShell，確保校園敏感資產與科研專利零外洩。

前瞻校園 AI 生態系建置藍圖

Optional subtitle goes here

Q1 基礎部署

導入 BCM，完成全校 GPU 伺服器叢集自動化維運環境

Q2 算力池化

上線 Run:ai，建立跨系所多租戶算力池與教學切分機制。

Q3 服務推論

藉由 NIM 部署標準化微服務，提供高吞吐推論 API

Q4 代理落地

運用 Agent Toolkit & NemoClaw 打造沙盒安全的自主校園 Agent。


NVIDIA