NVIDIA DGX最強免費管理軟體 BCM

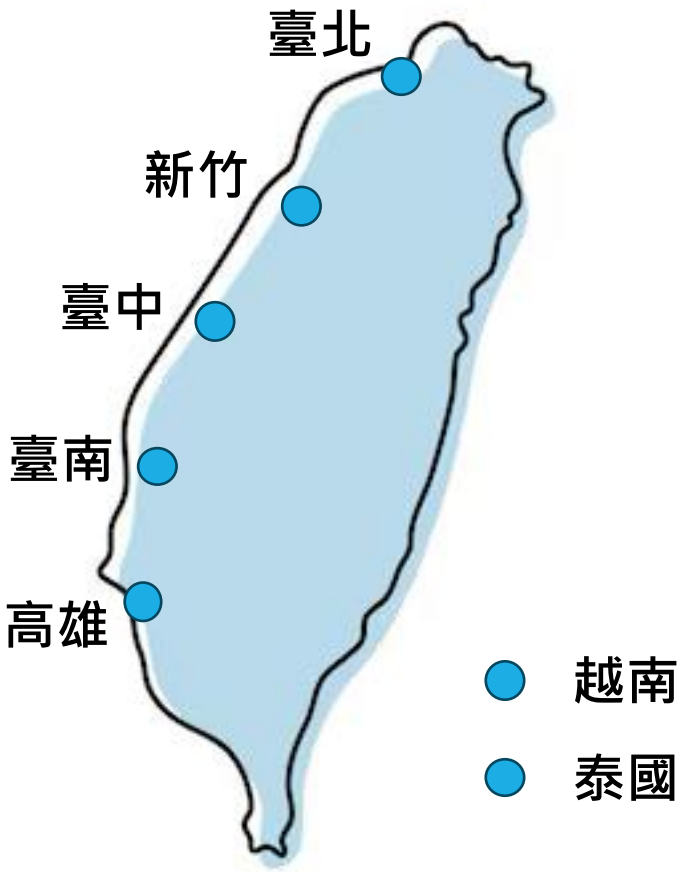
從基礎設施到算力賦能：
NVIDIA BCM 與 Run:ai 打造高效校園 AI 運算中心

January 2026

豐康科技 www.fongcon.com.tw

NVIDIA ELITE PARTNER

豐康科技-公司介紹



2012	豐康科技成立
2016	NVIDIA Quadro NPN
2017	NVIDIA TSLA & DGX NPN
2018	NVIDIA Elite NPN
2020	越南分公司成立
2021	資安團隊成立
2024	NVIDIA Enterprise Partner of the year
2025	泰國分公司成立

豐康科技-公司介紹

豐康科技
www.fongcon.com.tw

NVIDIA Enterprise partner of the year

硬、軟實力

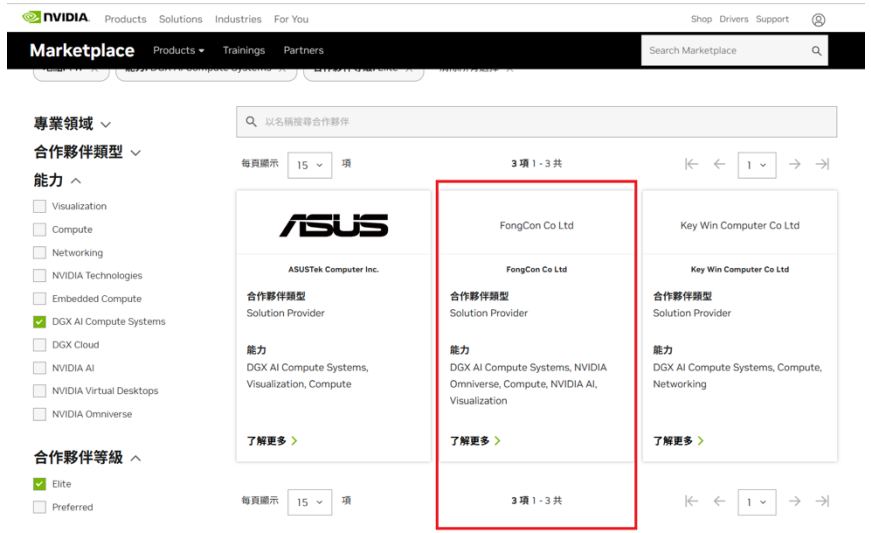
Nvidia 最佳夥伴

最佳服務

虛擬化

資安

各式硬體



信任

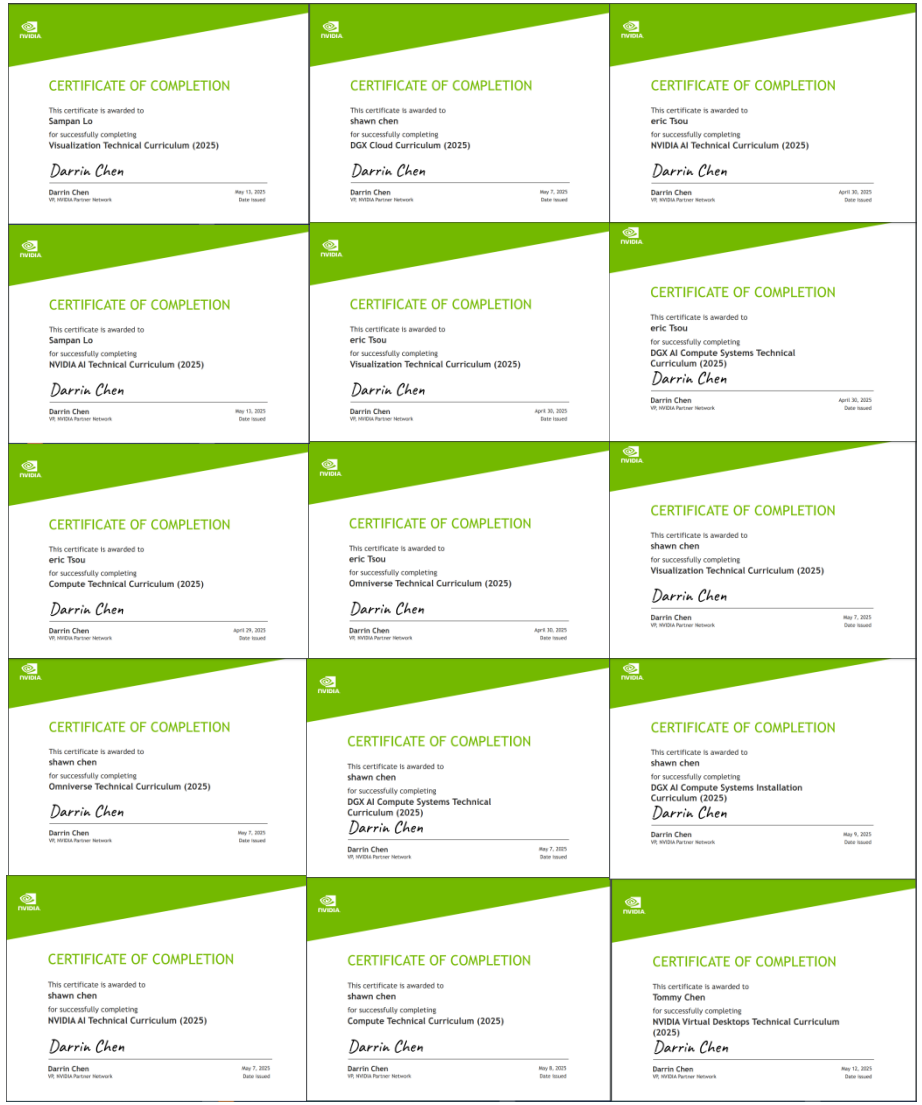
經驗

服務

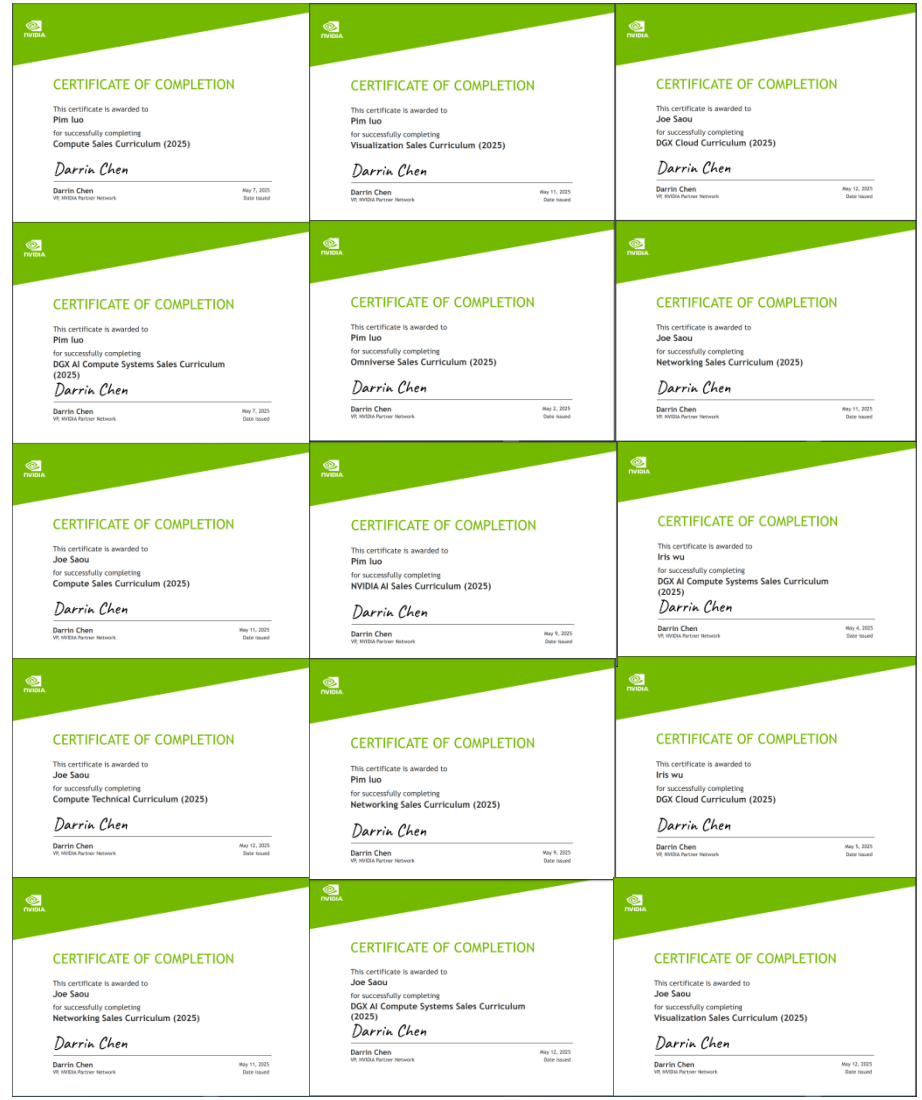
豐康科技-公司介紹

豐康科技
www.fongcon.com.tw

工程師證書



業務證書



豐康科技-案件分享

醫療產業

豐康科技 www.fongcon.com.tw

1

台北榮民總醫院
NVIDIA DGX H200

- 影像分析
- 基因分析
- 護理交班
- 藥物/病患GPT

2

台中榮民總醫院
NVIDIA DGX H200

- 影像分析
- 基因分析
- 護理交班
- 藥物/病患GPT

3

三軍總醫院
NVIDIA DGX H200

- 影像分析
- 基因分析
- 護理交班

4

中國醫藥大學
NVIDIA DGX B200

- 影像分析
- 基因分析
- **基因庫**
- 病患GPT

豐康科技-案件分享

政府

豐康科技 www.fongcon.com.tw

5

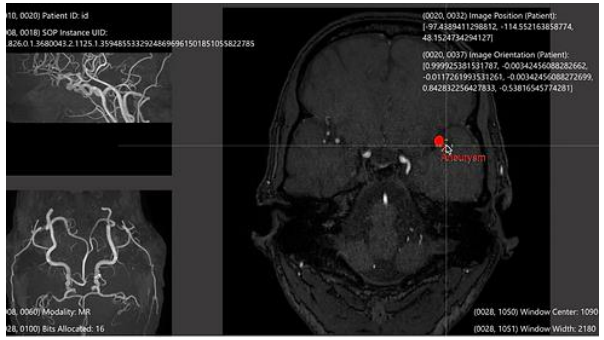
Linker Vision

Linker Vision – 高雄智慧城市

NVIDIA DGX H100

智慧交通

公安-煙霧偵測



醫療-腫瘤分析

自動駕駛-安全距離

6

中華電信

NVIDIA DGX-1

中華電信

- 算力出租
- 24hr 客服
- 氣象預測
- 視障服務
- AI 助手

豐康科技-服務項目

豐康科技
www.fongcon.com.tw

AI 伺服器

DGX B200 | B300

AI 軟體

Ubuntu

kubernetes

MLOps

SAS

APMIC

Linker Vision

Taiwan AI Labs
台灣人工智慧實驗室

基礎設施

電力設備

設施規劃

冷卻系統

AGENDA

豐康科技
www.fongcon.com.tw

1. 趨勢與挑戰: 校園 AI 建設的「困境」
2. Base Command Manager (BCM)：算力中心的「數位底座」
3. Run:ai：算力調度的「智慧大腦」
4. BCM + Run:ai 的 1+1>2 效應
5. 結語與未來展望

NVIDIA ELITE PARTNER

豐康科技
2026

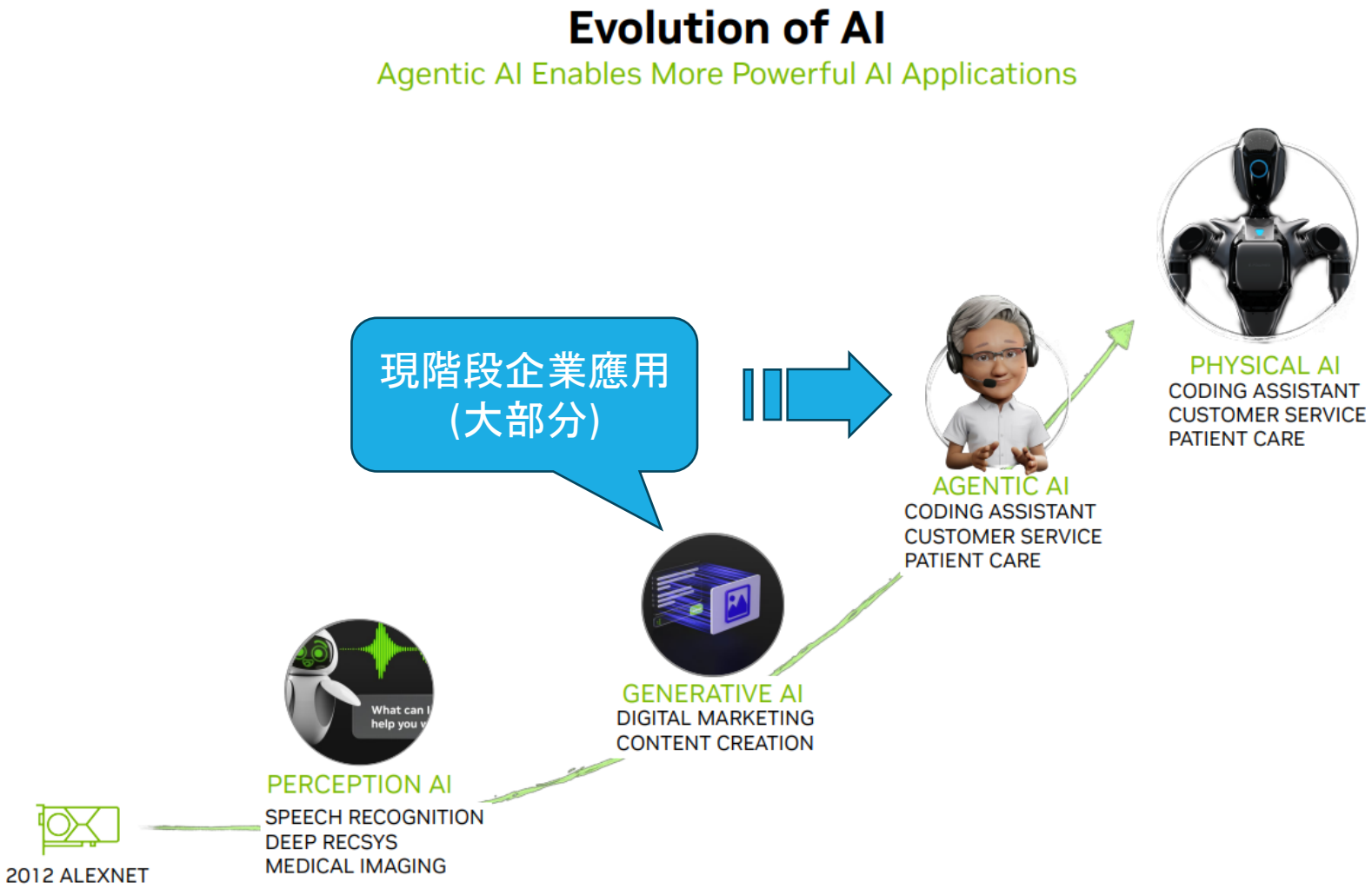
NVIDIA ELITE PARTNER

1

趨勢與挑戰

校園 AI 建設的「困境」

AI 應用發展階段

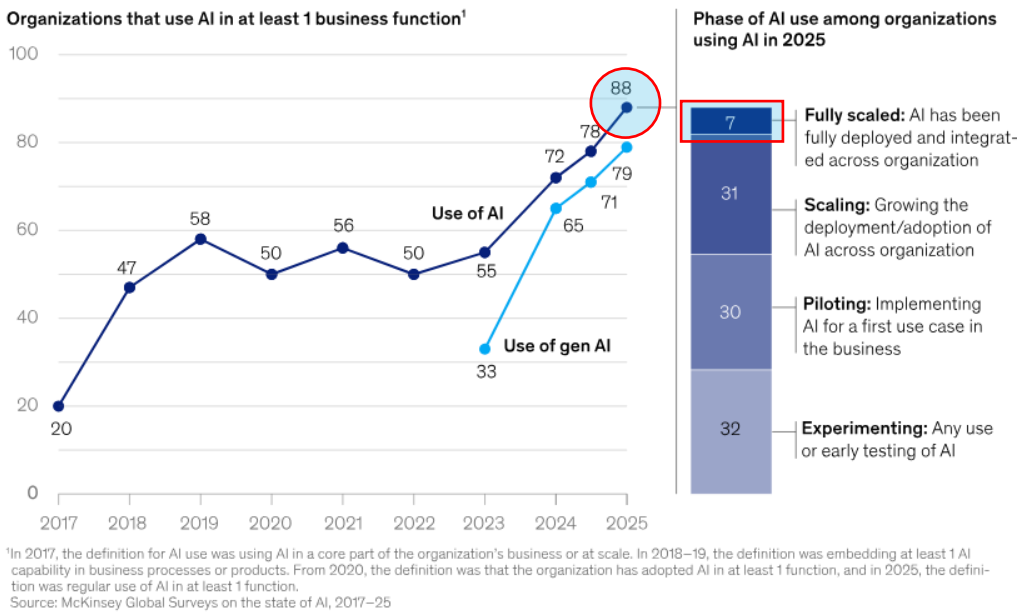


- Origin source: nvidia NPN training material

企業 AI 導入現況

Reported use of AI in at least one business function continues to increase.

Use of AI by respondents' organizations, % of respondents



[1]In 2017, the definition for AI use was using AI in a core part of the organization's business or at scale. In 2018–19, the definition was embedding at least 1 AI capability in business processes or products. From 2020, the definition was that the organization has adopted AI in at least 1 function, and in 2025, the definition was regular use of AI in at least 1 function.
Source: McKinsey Global Surveys on the state of AI, 2017–25

McKinsey & Company

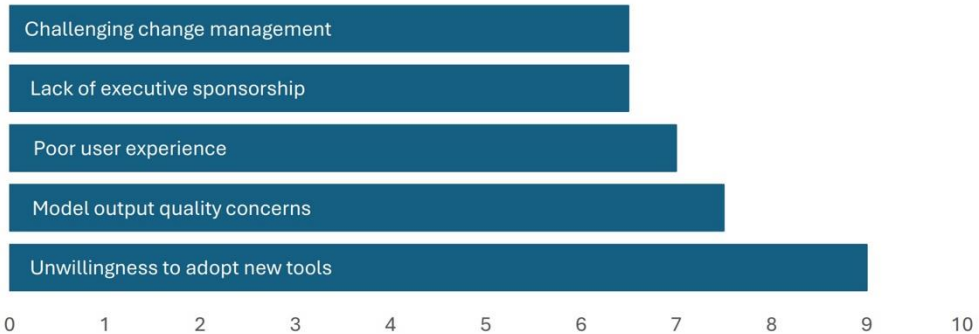
Ref https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai

Companies have invested billions into AI, 95% getting zero return

The Root Cause: The Learning Gap

The report's central explanation is that the core barrier is learning rather than infrastructure, regulation, or talent: "Most GenAI systems do not retain feedback, adapt to context, or improve over time."

Exhibit: Why GenAI pilots fail: top barriers to scaling AI in the enterprise
Users were asked to rate each issue on a scale of 1-10



Ref https://campustechnology.com/articles/2025/08/26/mit-report-most-organizations-see-no-business-return-on-gen-ai-investments.aspx

根據 GARTNR 統計：
儘管企業在 2024 年平均花費 190 萬美元在生成式 AI 計畫上，但不到 30% 的 AI 領導者報告其 CEO 對 AI 投資回報感到滿意。

Ref: https://www.gartner.com/en/articles/hype-cycle-for-artificial-intelligence

On-Prem AI 負載成長趨勢

Cloud repatriation	**83%**	are shifting workloads back on-prem[1]
Colocation adoption	**45%**	of GenAI workloads hosted in 3rd party data center[2]
Hybrid IT dominance	**62%**	are hybrid cloud, 32% planning to adopt within 12 months[3]
Control	**$$$**	Cloud cost overruns, compliance and sovereignty

[1] 2024 CIO Survey (Barclays)
[2,3] 2025 State of the Data Center (Foundry/Coresite)

- Origin source: nvidia NPN training material

2026 年校園 AI 建設的三大嚴峻挑戰

豐康科技
www.fongcon.com.tw

管理孤島

散

系所採購零散化，昂貴算力無法跨單位共享。

效率黑洞

錯

缺乏彈性分配機制，高階 GPU 淪為低效能文書開發機。

維運泥沼

難

驅動程式更新、網路拓樸除錯、容器環境衝突，耗盡 IT 人員精力。

最常被問到...

豐康科技
www.fongcon.com.tw

我知道AI要花錢,
但不知道刀口在哪...

為甚麼 GPU 總是用不滿?
明明排程還是很多...

GPU Server 買了,
然後呢...

我可以繼續用之前
的訓練流程嗎...

錢是砸了，
好像也沒預期的好...

看文件都需要 K8S,
我們不會惹...

核心思維：「**買到硬體只是開始，發揮算力價值才是關鍵。**」

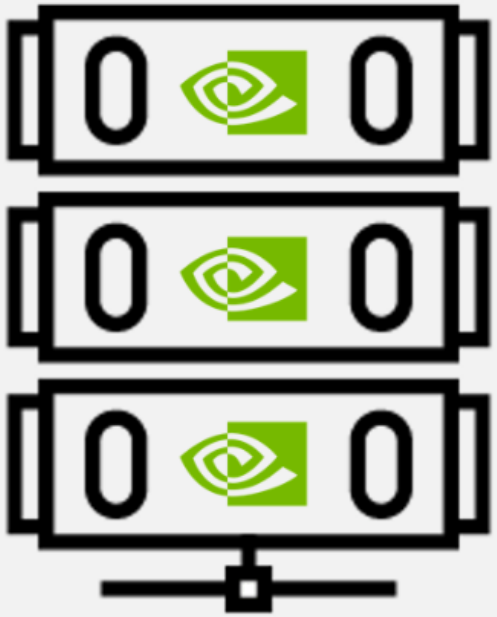
BASE COMMAND MANAGER (BCM)

算力中心的「數位底座」

豐康科技
2026

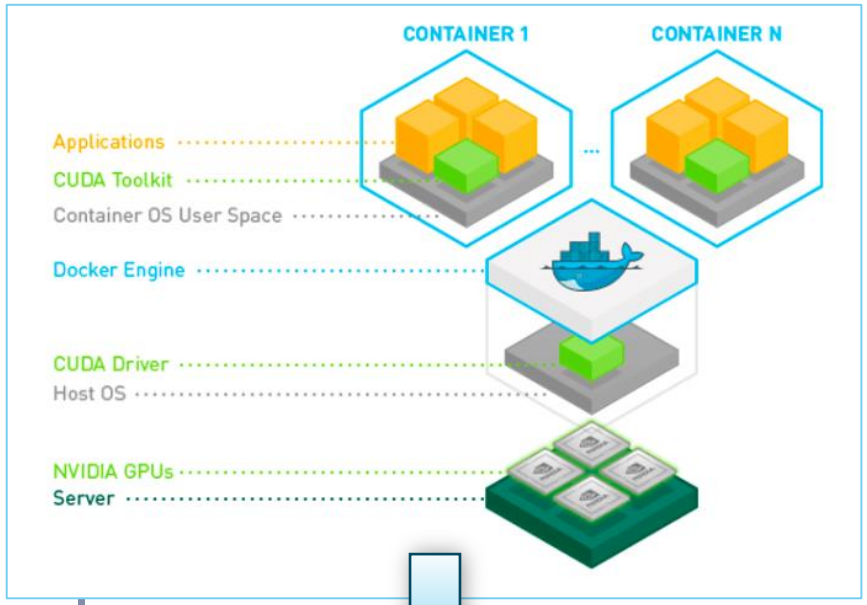
AI 導入關鍵要素

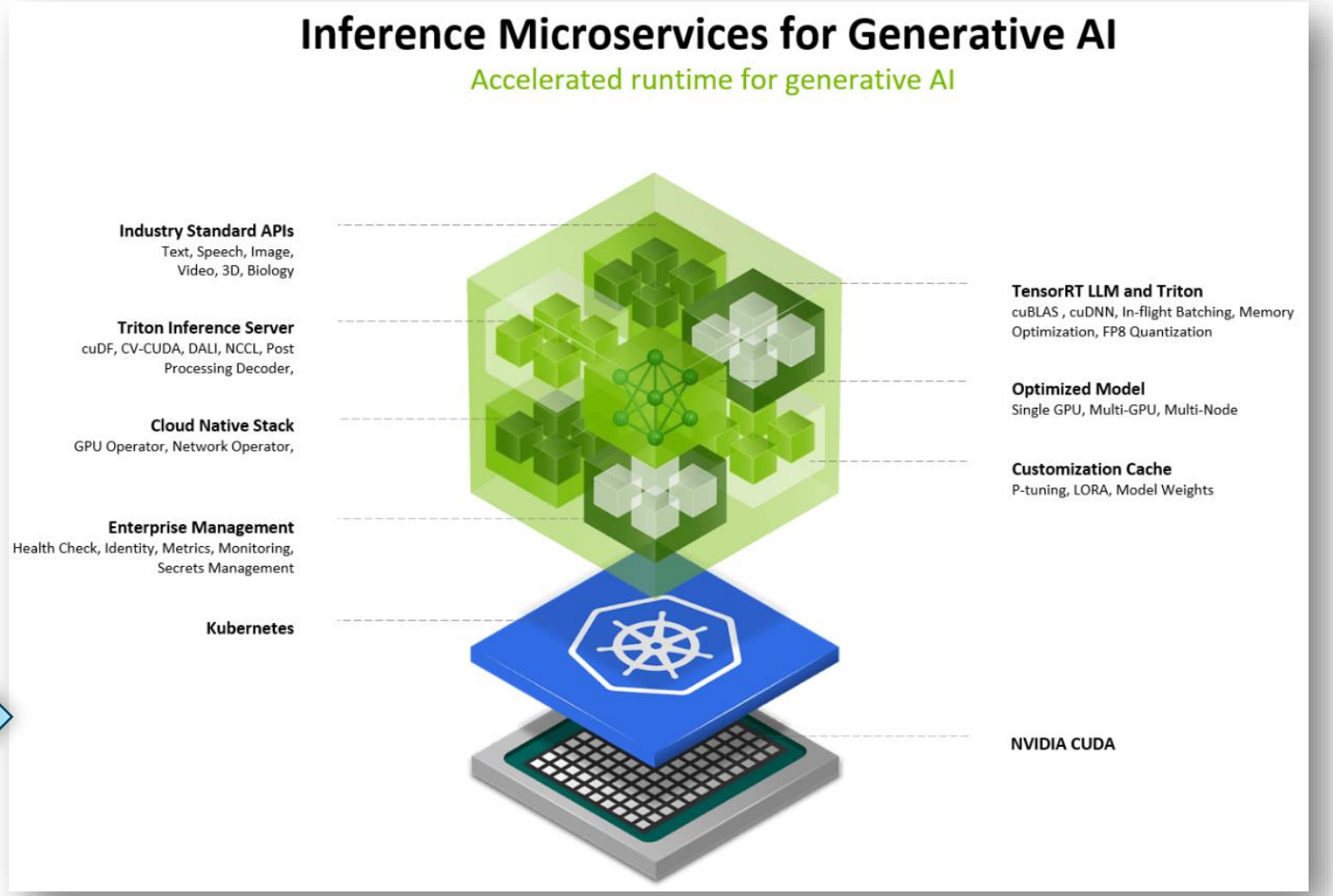
豐康科技
www.fongcon.com.tw

Accelerated Computing	Training and Inference Tools	AI Expertise
		

- Origin source: nvidia NPN training material

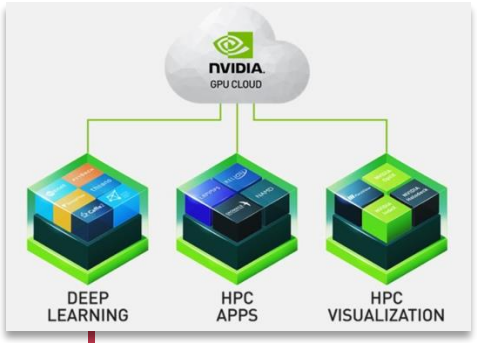
Nvidia Inference Microservices(NIM)

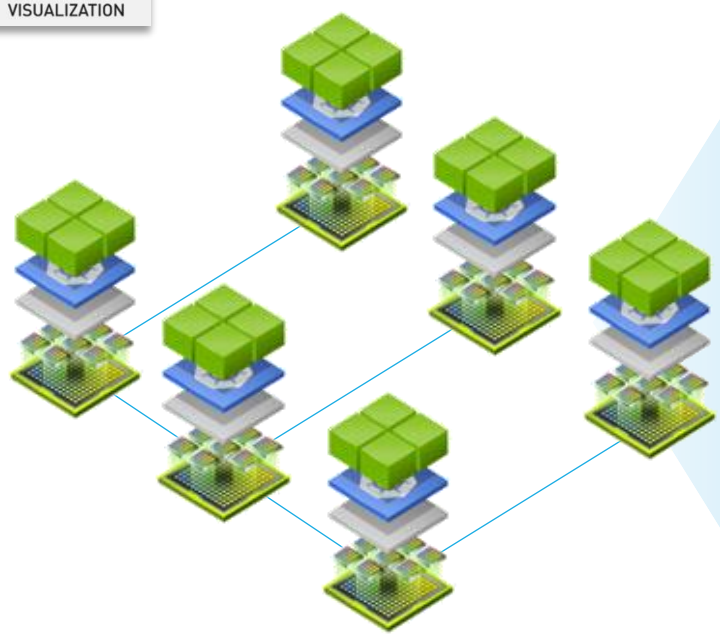


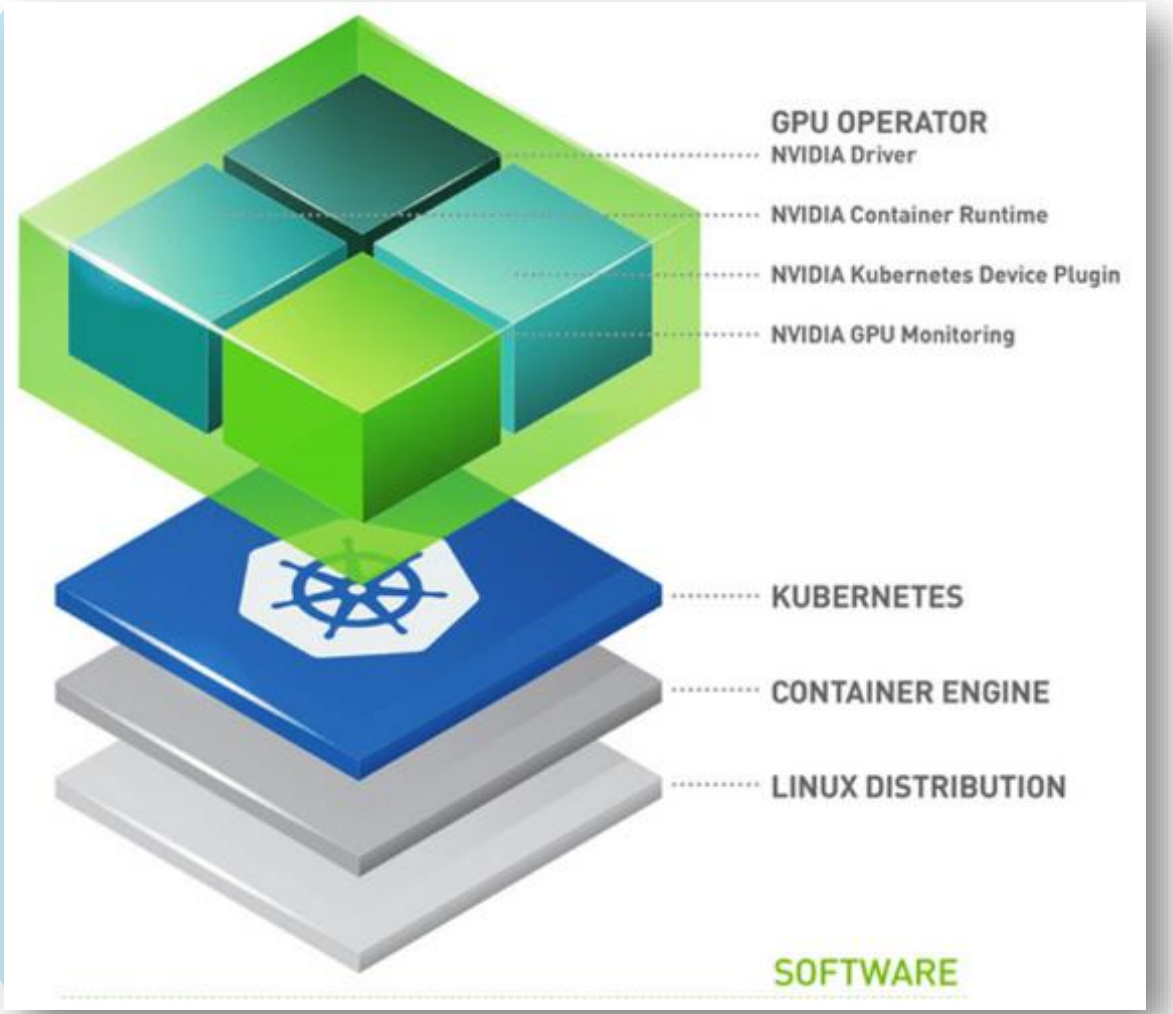


- Origin source: nvidia NPN training material

軟硬體分散架構

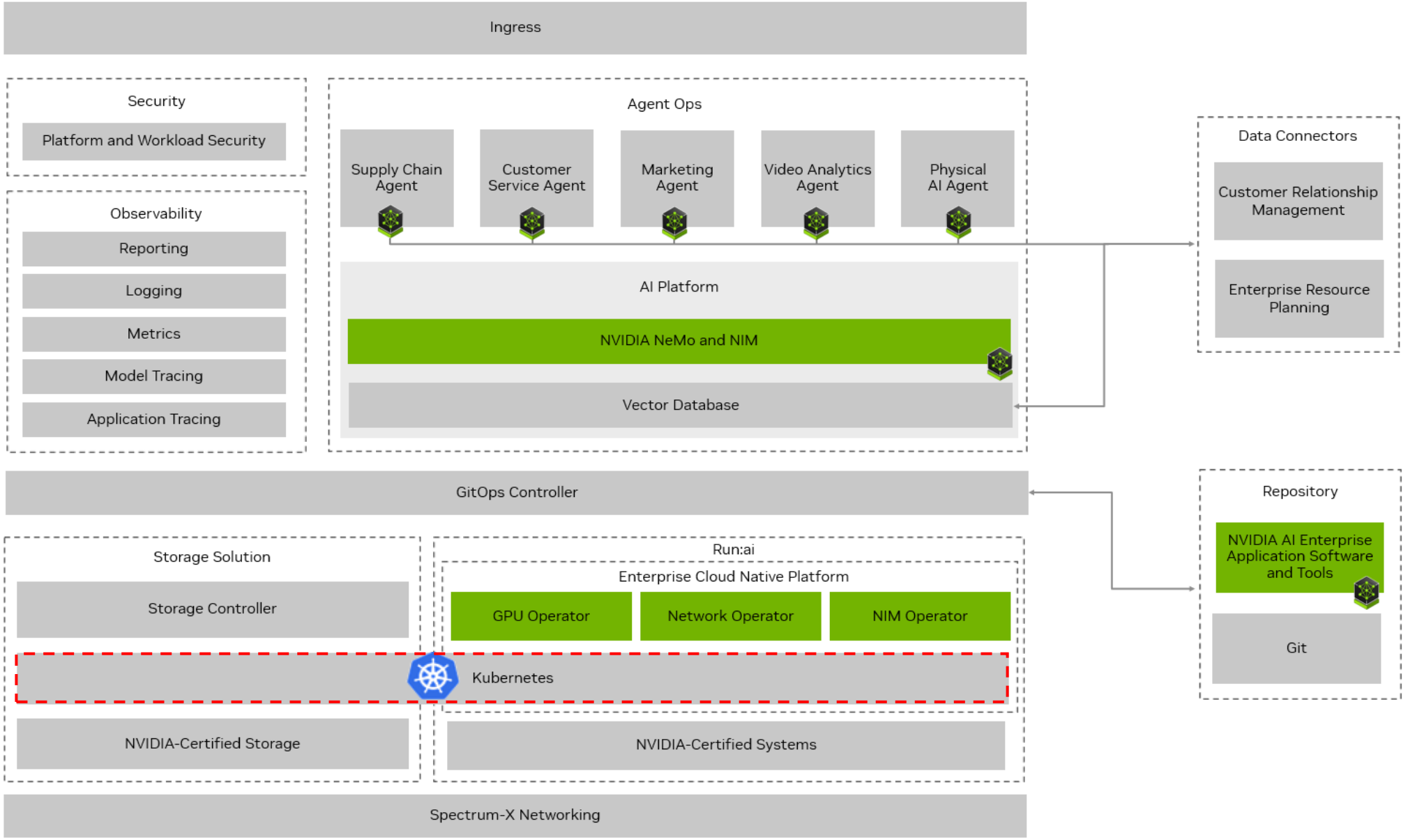






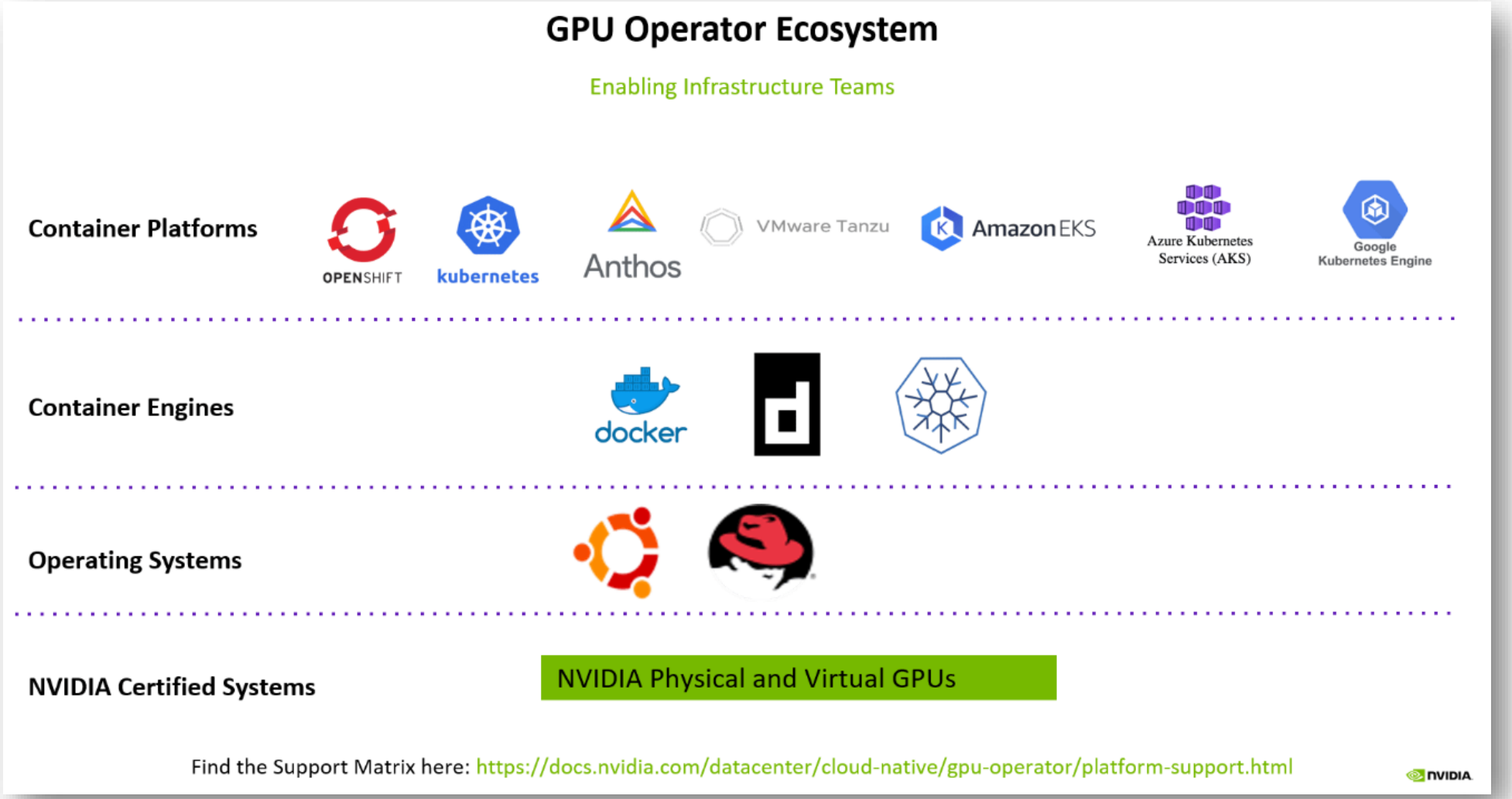
- Origin source: nvidia NPN training material

企業 AI 基礎架構



- Ref: https://docs.nvidia.com/ai-enterprise/planning-resource/ai-factory-white-paper/latest/agentic-ai-in-the-factory.html#enterprise-cloud-native-platform

GPU Operator 生態系統



- Origin source: nvidia NPN training material

Nvidia Base Command Manager (BCM)

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

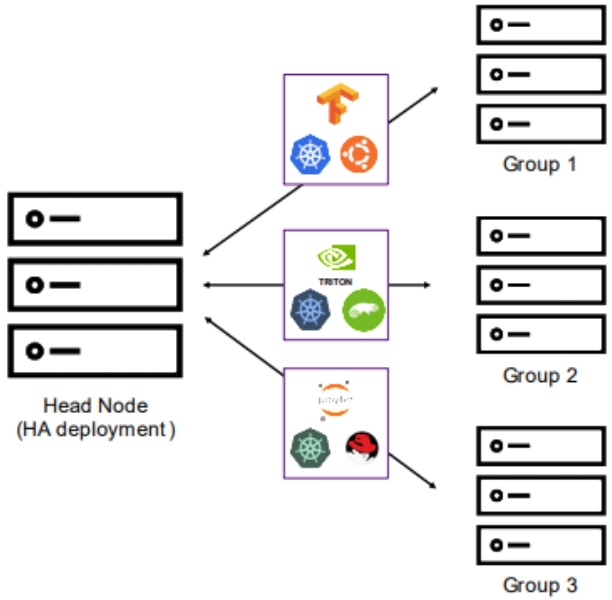
Edge

- Origin source: nvidia NPN training material

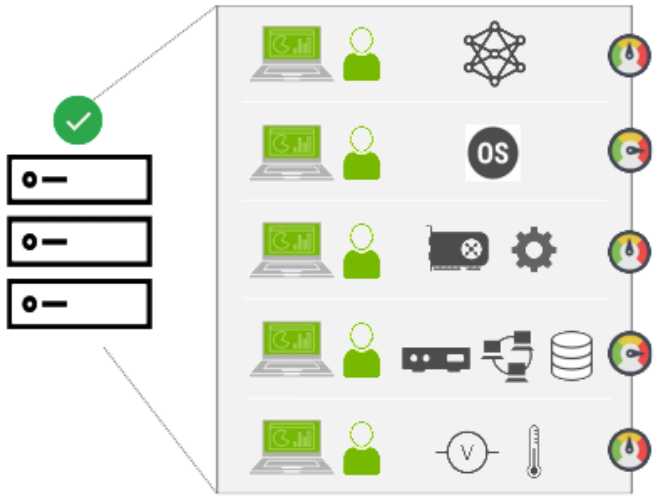
Infrastructure Management

豐康科技
www.fongcon.com.tw

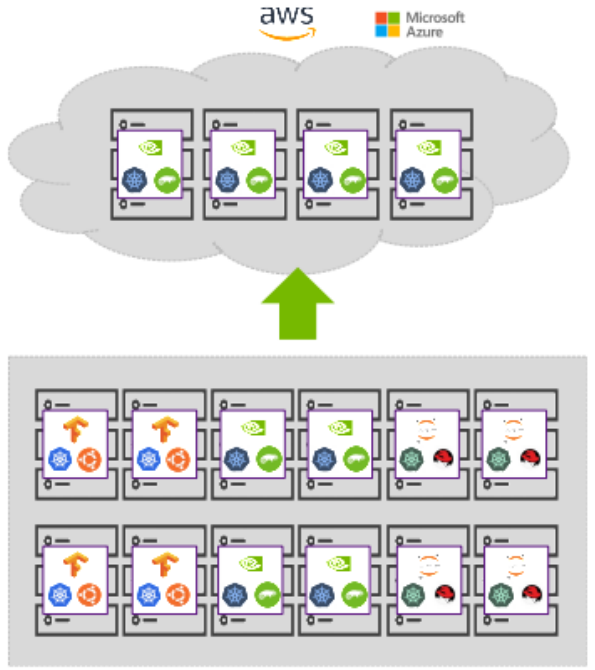
Image Management and Configuration



Utilization Metrics



Cloud Integration

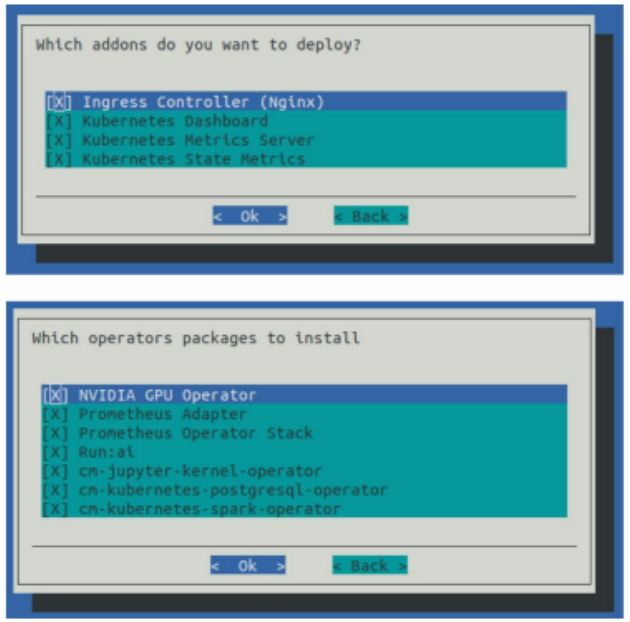


- Origin source: nvidia NPN training material

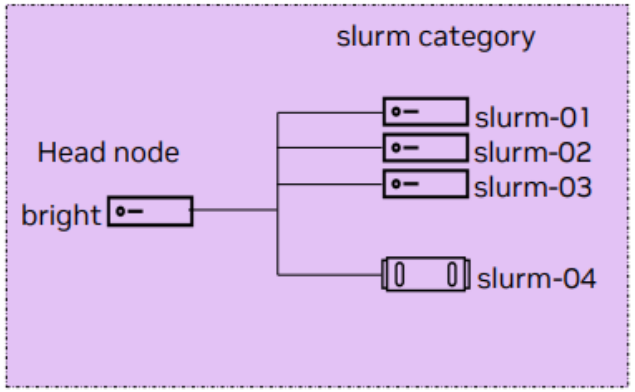
Workload Management

豐康科技
www.fongcon.com.tw

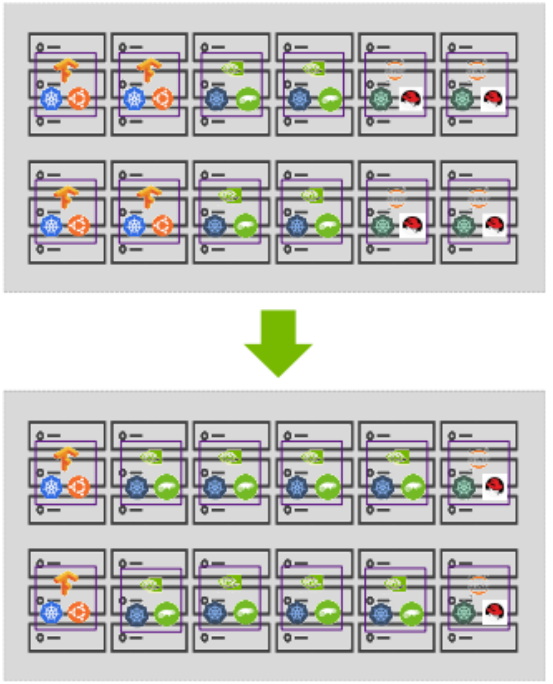
Kubernetes



Slurm



Auto Scaler

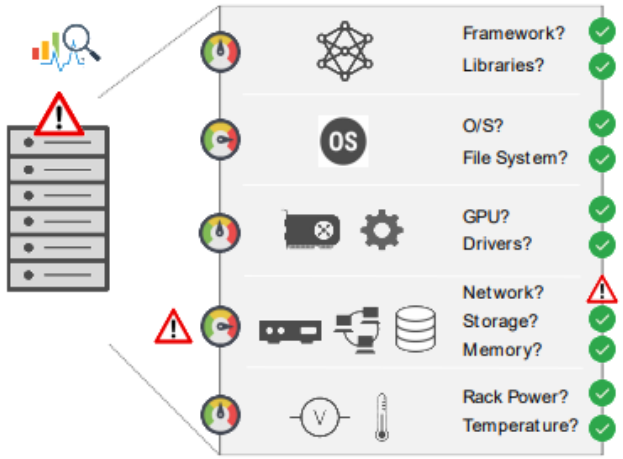


- Origin source: nvidia NPN training material

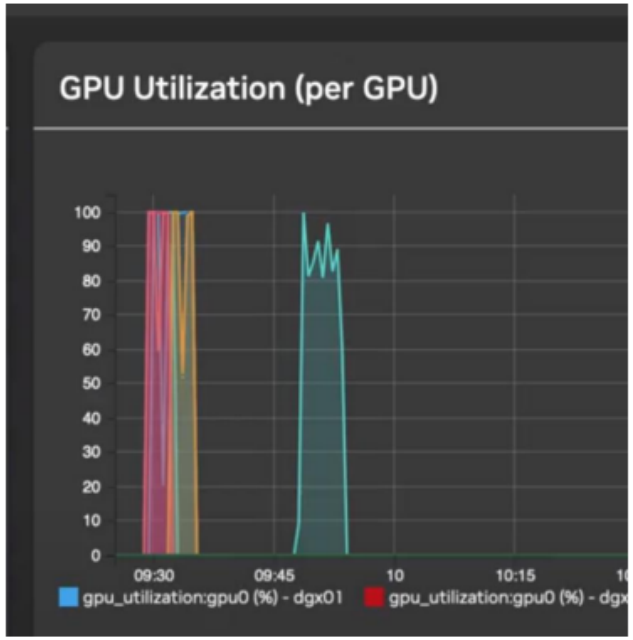
Resource Monitoring

豐康科技
www.fongcon.com.tw

Health Monitoring and Remediation



NVIDIA GPU Management and Monitoring



Reporting and Chargeback

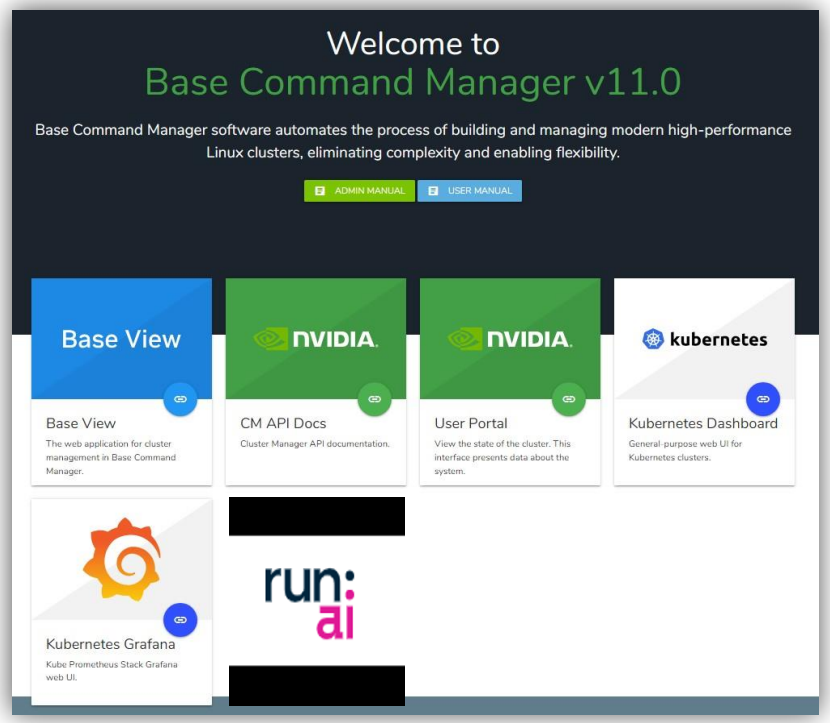
User	Jobs	Runtime (s)	CPU ($)
bob	1	1804	0.3608
charline	1	1804	0.7216
dennis	2	3606	1.0818
frank	1	1803	0.1803

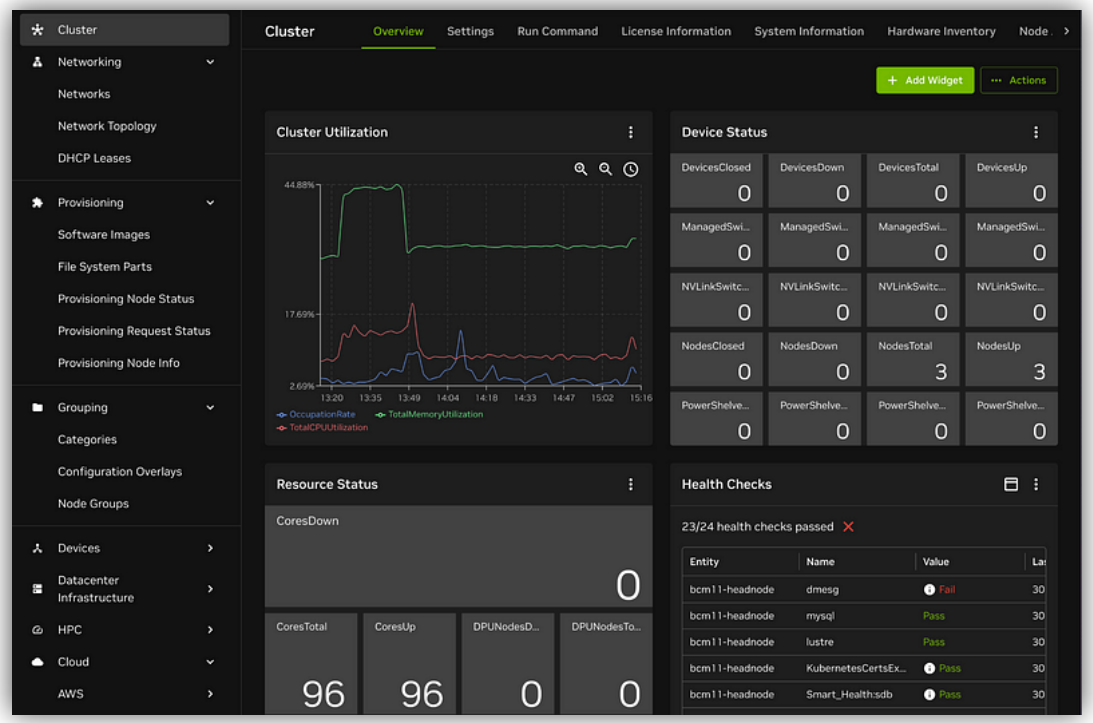
- Origin source: nvidia NPN training material

專為企業人工智慧基礎設施管理而構建

豐康科技 www.fongcon.com.tw

- **定位**：企業級叢集管理專家（前身為 Bright Cluster Manager）
- **管理情境**：多台 server 加入 BCM，不限廠牌，不限 VM, 實體機
- **深度監控**：整合各種 metrics (可以自定義配置介面)，即時偵測 GPU 功耗、溫度、NVLink 狀態，預先識別故障風險
- **GPU 整合**：各種 NVIDIA 的配件都可安裝 (e.g. GPU Driver)
- **Kubernetes 平台整合**：與節點帳號角色權限整合，可使用介面管理 Kubernetes





強大的整合管理介面

HPC workload manager

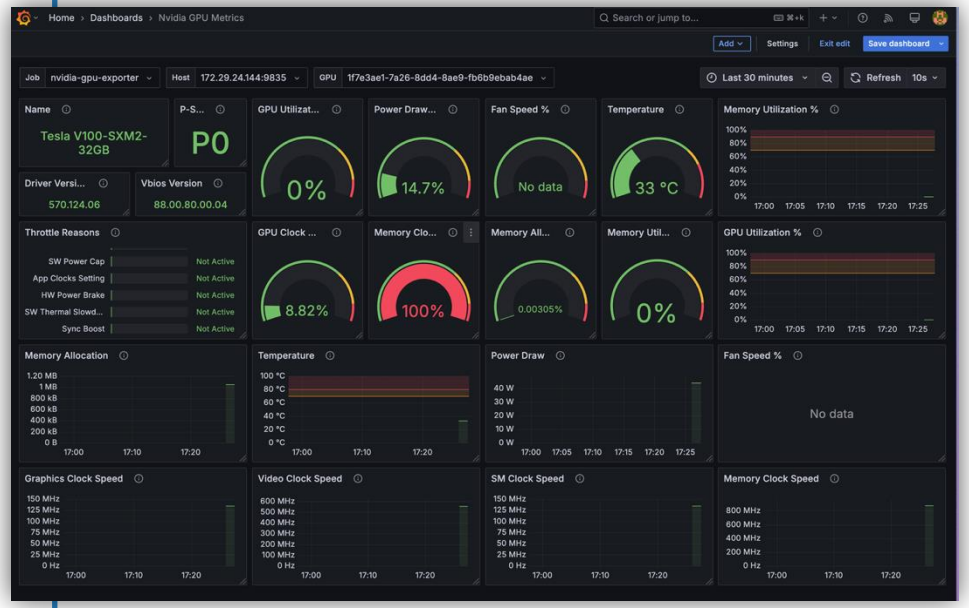
A workload management system is highly recommended to run compute jobs. Please choose the one that should be configured or choose 'None' to prevent configuration.

Please select workload manager:

PBS Works
PBS Pro

slurm workload manager
Slurm

IBM Spectrum LSF
IBM Spectrum LSF



Jupyter Wizard

1 Introduction — 2 Settings — 3 Node Selection — 4 VNC Installation — 5 Summary — 6 Deploy

Jupyter Setup Wizard

Upload Custom Configuration (Optional)

If you have a custom configuration for the wizard, you can upload it here. The wizard will use the custom configuration instead of the default one.

Upload Custom Config (YAML)

No file selected

Select File

Kubernetes Add User Wizard

✓ Select Cluster — ✓ Select User — ✓ Configure Security User Options — 4 Select Operators — 5 Summary — 6 Deploy

Available Operators

Please specify the operators to add to this user.

Spark-Operator

Cm-Kubernetes-Mpi-Operator

BCM 叢集架構

- **自動化部署**
  從 Bare-metal 裸機一鍵完成 OS、Driver、InfiniBand 與 K8s 安裝
- **統一管理配置**
  採用 PXE-boot，節點就算被改壞，重裝也很快
- **支持異質架構**
  （x86/ARM）與多代 GPU 共存管理
- **零停機更新**
  支持滾動式升級，確保校園 AI 服務不中斷

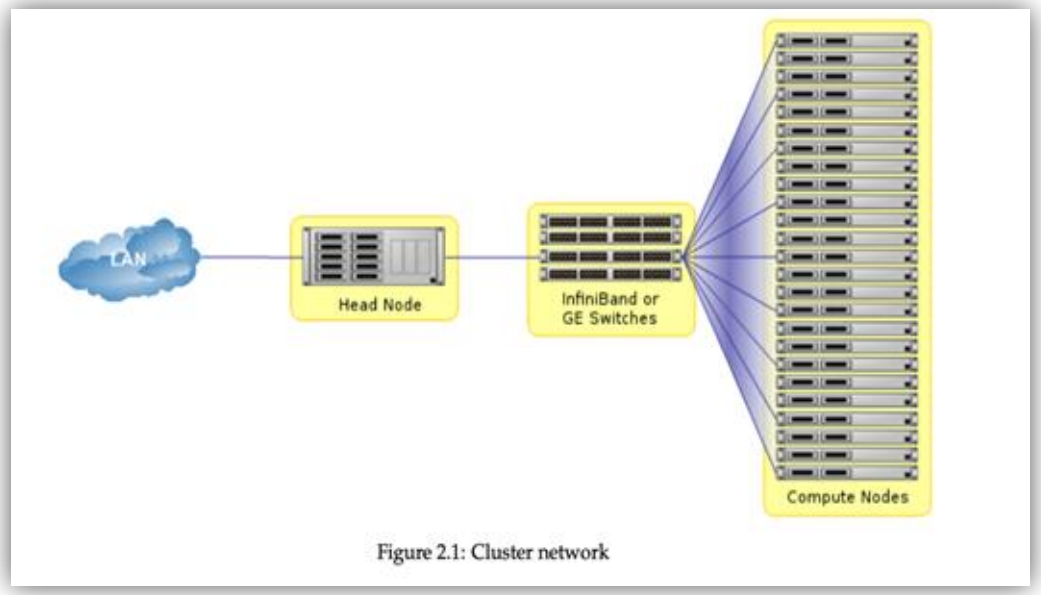
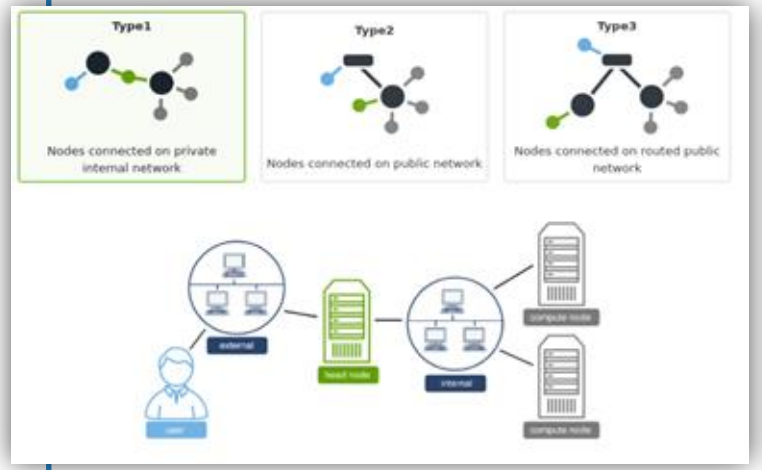
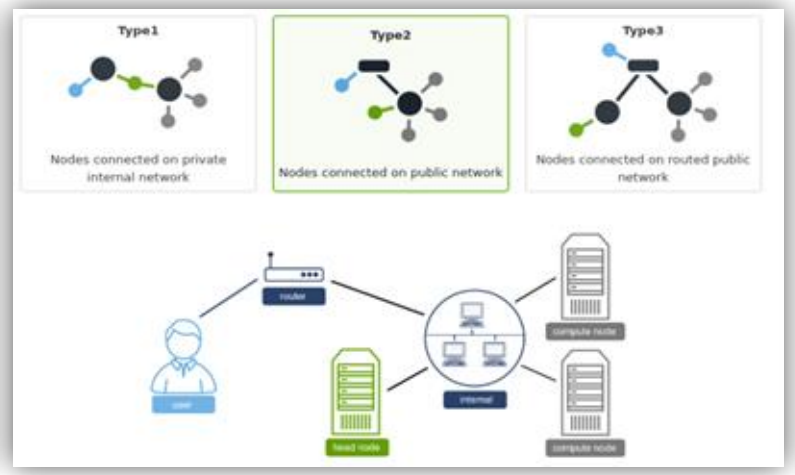
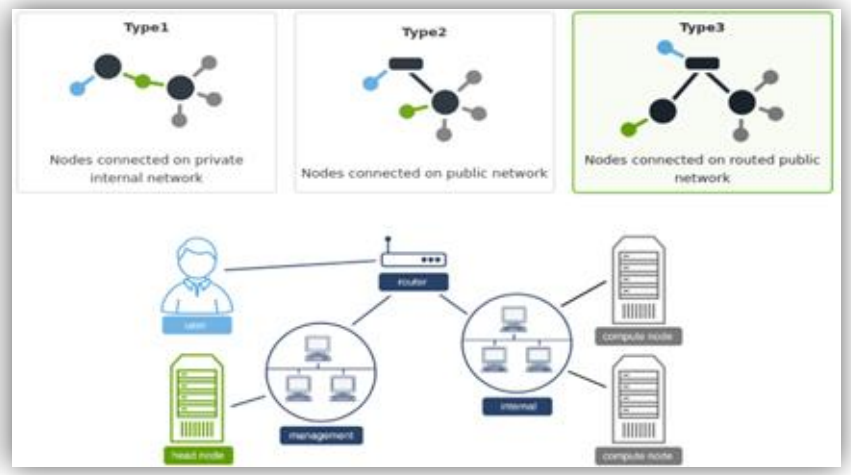


Figure 2.1: Cluster network





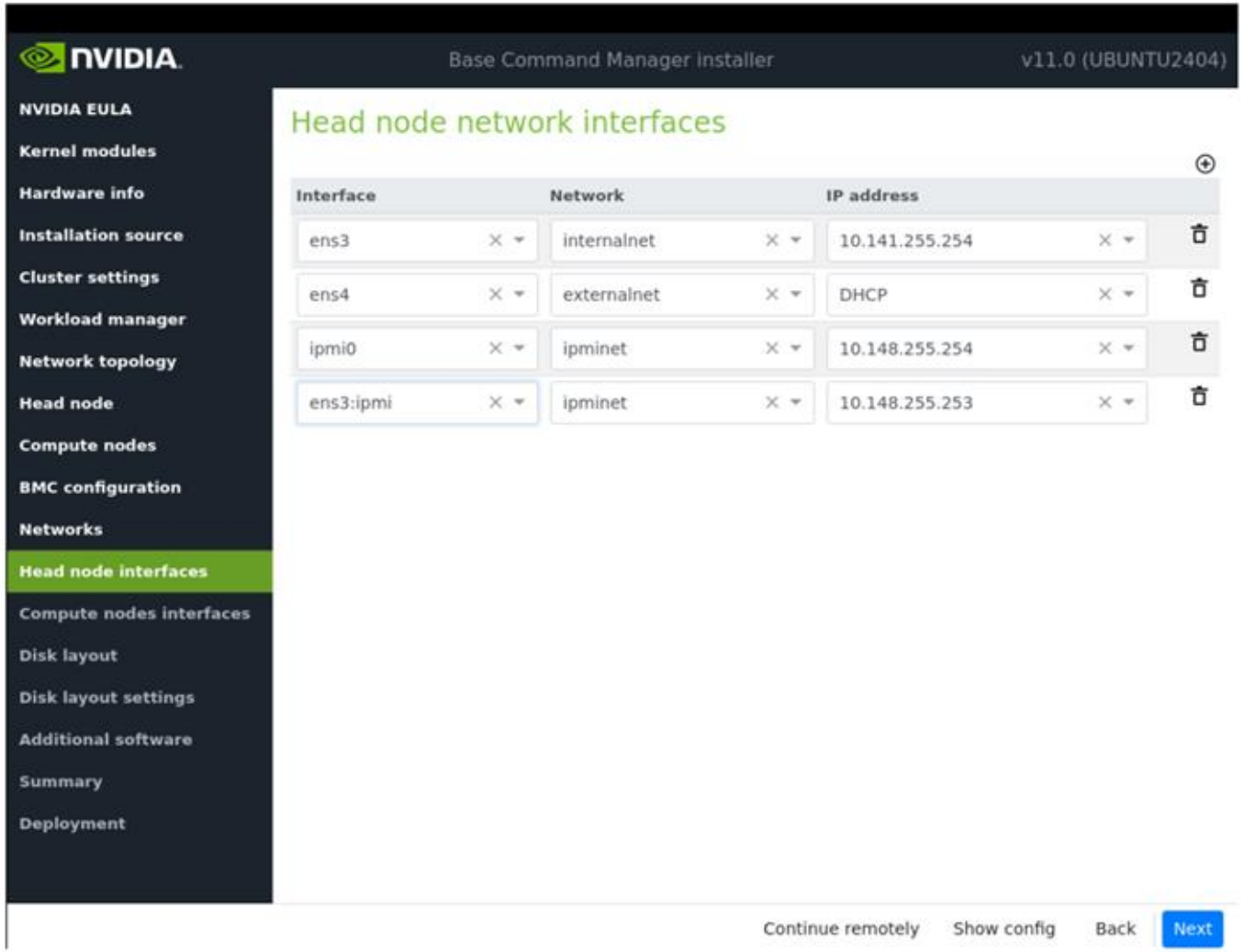


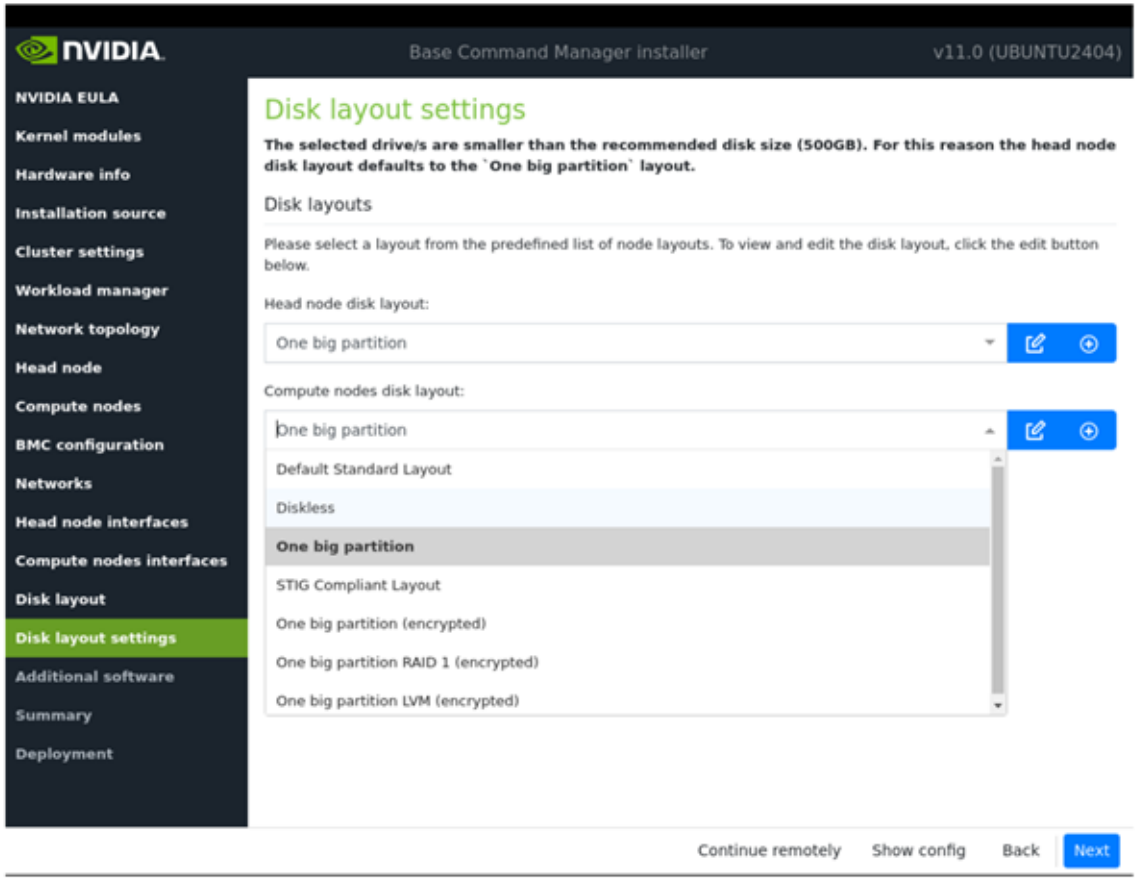
指令介面：cmsh (CLI)

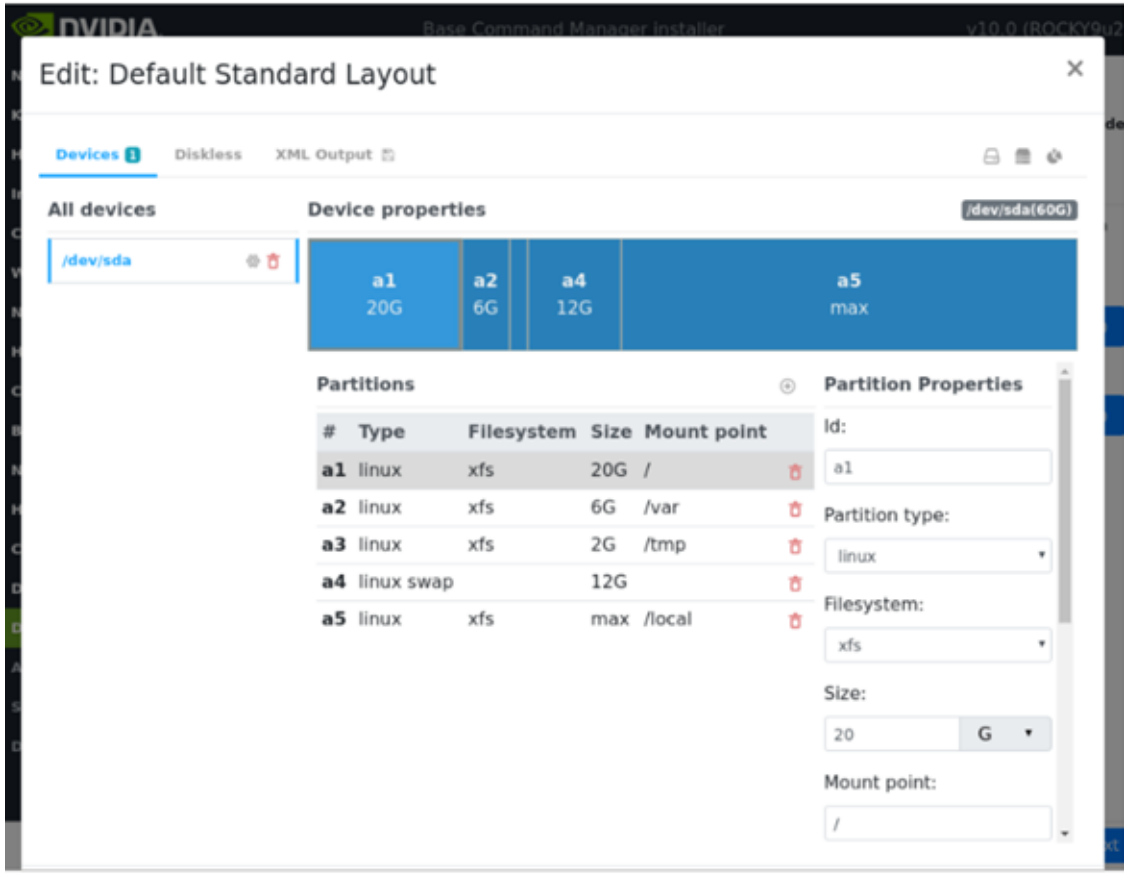
在 headnode 上面可以使用 `cmsh` 操作

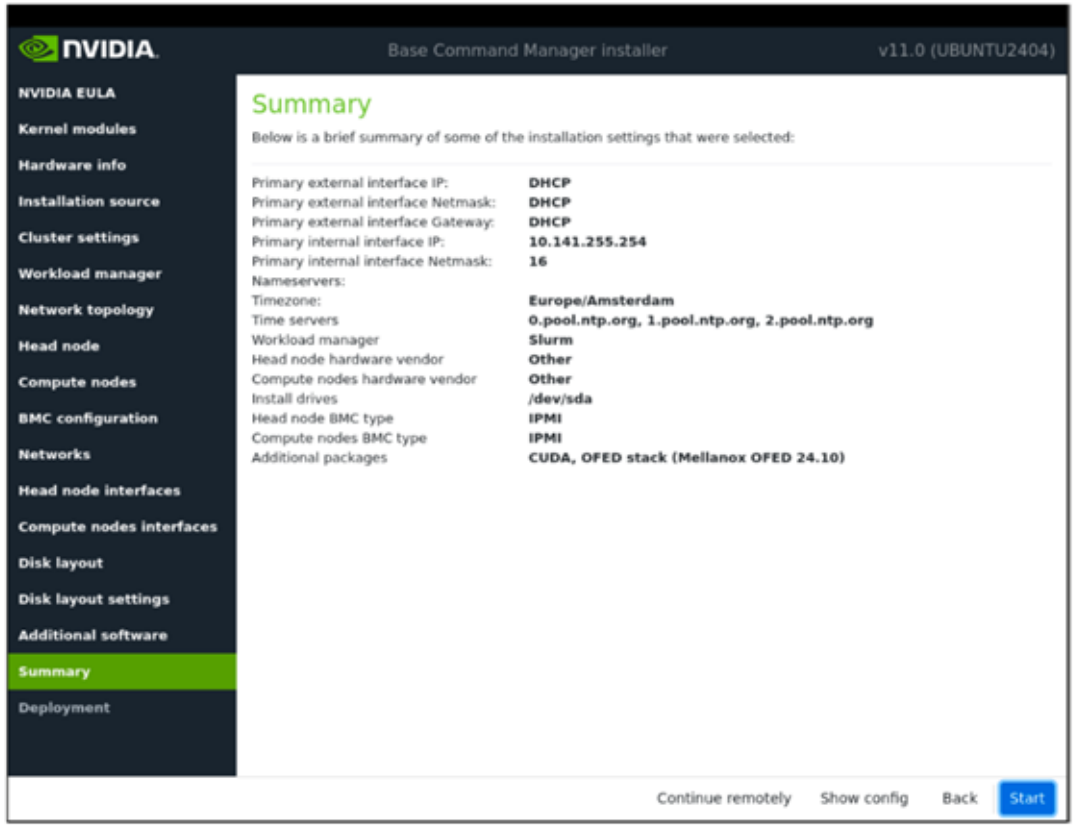
- **查看裝置是否 up**
  cmsh -c 'device status'
- **查看所有 network 資訊**
  cmsh -c 'device list -f hostname,ip,mac,network'
- **列出所有節點**
  cmsh -c "device list"
- **確認 healthcheck 狀態**
  cmsh -c "monitoring measurable; list healthcheck"

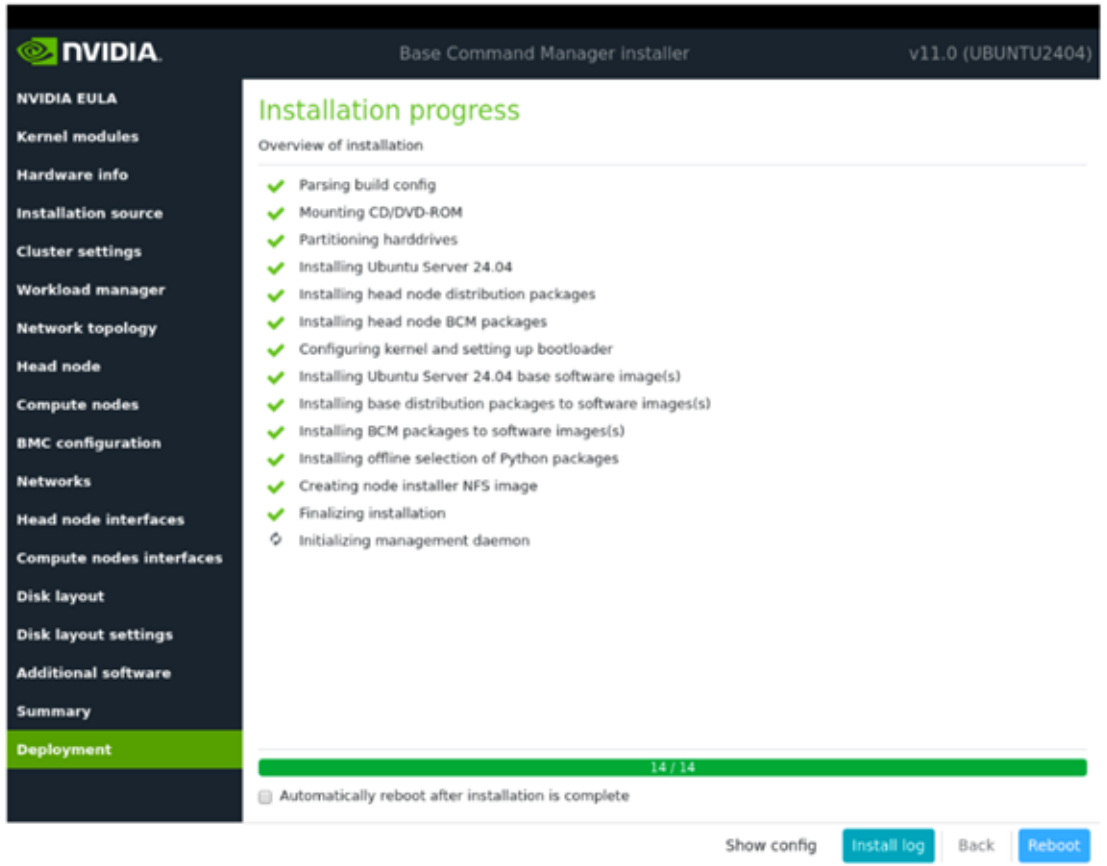
EXAMPLE: Head node 安裝




NVIDIA
Base Command Manager installer
v11.0 (UBUNTU2404)
NVIDIA EULA
Kernel modules
Hardware info
Installation source
Cluster settings
Workload manager
Network topology
Head node
Compute nodes
BMC configuration
Networks
Head node interfaces
Compute nodes interfaces
Disk layout
Disk layout settings
Additional software
Summary
Deployment
Disk layout settings
The selected drive/s are smaller than the recommended disk size (500GB). For this reason the head node disk layout defaults to the `One big partition` layout.
Disk layouts
Please select a layout from the predefined list of node layouts. To view and edit the disk layout, click the edit button below.
Head node disk layout:
One big partition
Compute nodes disk layout:
One big partition
Default Standard Layout
Diskless
One big partition
STIG Compliant Layout
One big partition (encrypted)
One big partition RAID 1 (encrypted)
One big partition LVM (encrypted)
Continue remotely
Show config
Back
Next


Edit: Default Standard Layout
Devices 1
Diskless
XML Output
All devices
/dev/sda
Device properties
/dev/sda(60G)
a1 20G
a2 6G
a4 12G
a5 max
Partitions
Type Filesystem Size Mount point
a1 linux xfs 20G /
a2 linux xfs 6G /var
a3 linux xfs 2G /tmp
a4 linux swap 12G
a5 linux xfs max /local
Partition Properties
Id:
a1
Partition type:
linux
Filesystem:
xfs
Size:
20
G
Mount point:
/

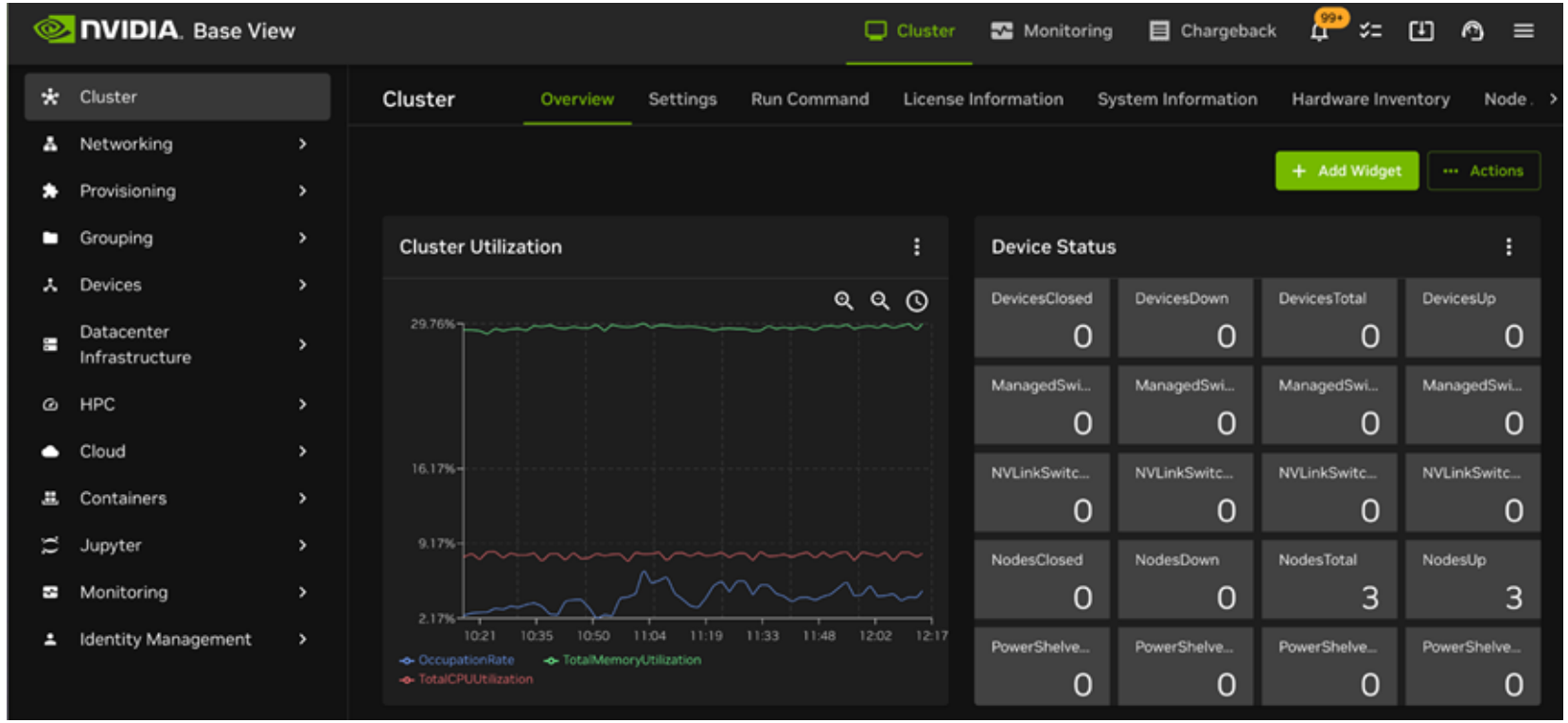

NVIDIA
Base Command Manager installer
v11.0 (UBUNTU2404)
NVIDIA EULA
Kernel modules
Hardware info
Installation source
Cluster settings
Workload manager
Network topology
Head node
Compute nodes
BMC configuration
Networks
Head node interfaces
Compute nodes interfaces
Disk layout
Disk layout settings
Additional software
Summary
Deployment
Summary
Below is a brief summary of some of the installation settings that were selected:
Primary external interface IP: DHCP
Primary external interface Netmask: DHCP
Primary external interface Gateway: DHCP
Primary internal interface IP: 10.141.255.254
Primary internal interface Netmask: 16
Nameservers:
Timezone: Europe/Amsterdam
Time servers 0.pool.ntp.org, 1.pool.ntp.org, 2.pool.ntp.org
Workload manager Slurm
Head node hardware vendor Other
Compute nodes hardware vendor Other
Install drives /dev/sda
Head node BMC type IPMI
Compute nodes BMC type IPMI
Additional packages CUDA, OFED stack (Mellanox OFED 24.10)
Continue remotely
Show config
Back
Start


NVIDIA
Base Command Manager installer
v11.0 (UBUNTU2404)
NVIDIA EULA
Kernel modules
Hardware info
Installation source
Cluster settings
Workload manager
Network topology
Head node
Compute nodes
BMC configuration
Networks
Head node interfaces
Compute nodes interfaces
Disk layout
Disk layout settings
Additional software
Summary
Deployment
Installation progress
Overview of installation
Parsing build config
Mounting CD/DVD-ROM
Partitioning harddrives
Installing Ubuntu Server 24.04
Installing head node distribution packages
Installing head node BCM packages
Configuring kernel and setting up bootloader
Installing Ubuntu Server 24.04 base software image(s)
Installing base distribution packages to software images(s)
Installing BCM packages to software images(s)
Installing offline selection of Python packages
Creating node installer NFS image
Finalizing installation
Initializing management daemon
14 / 14
Automatically reboot after installation is complete
Show config
Install log
Back
Reboot

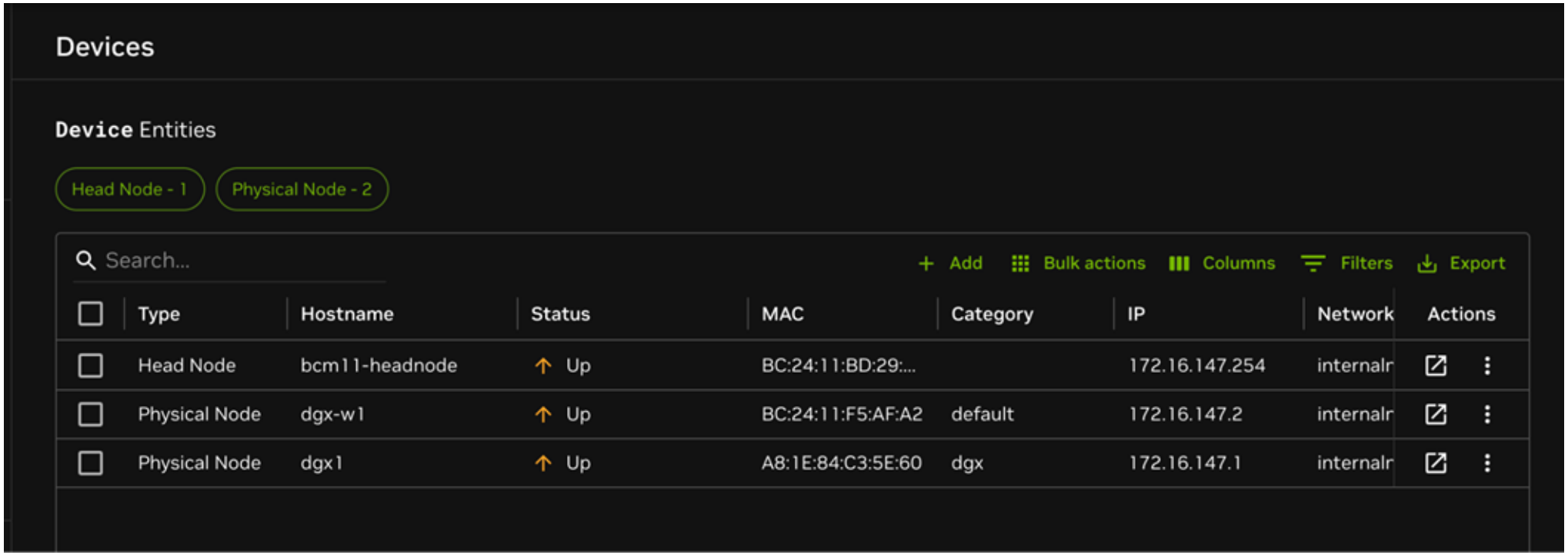
EXAMPLE: Base View (GUI)

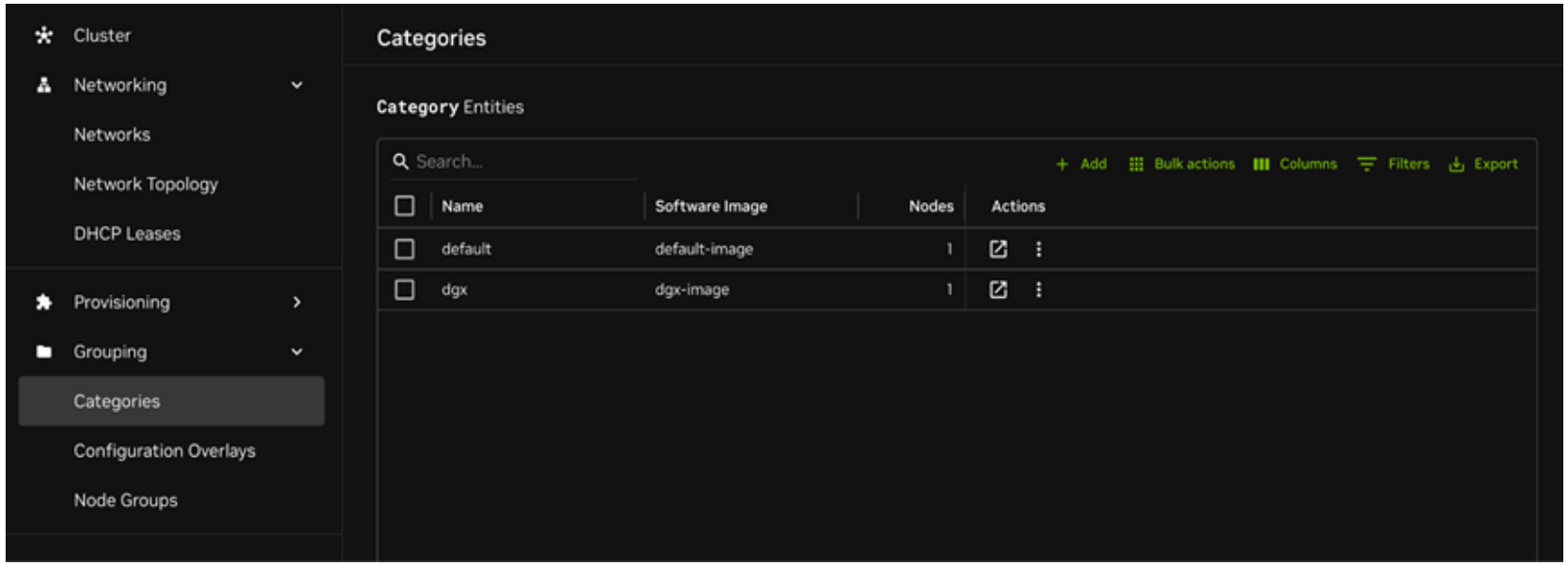
介面：

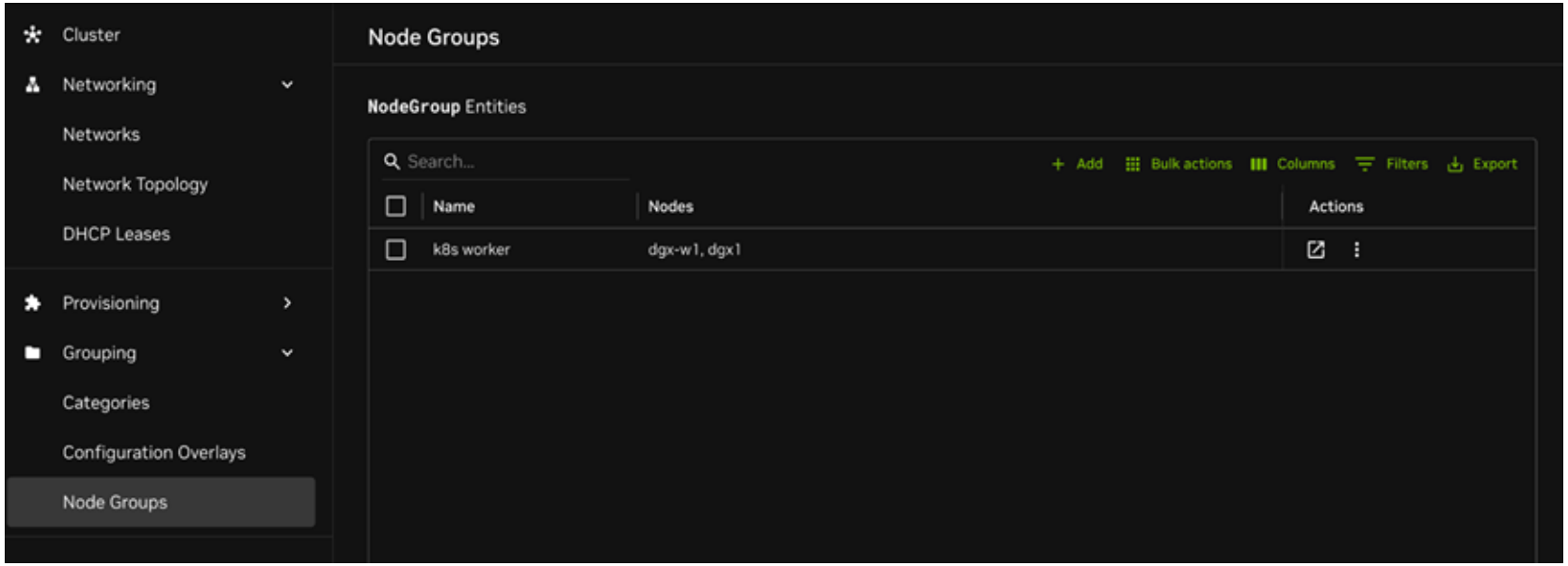
- Cluster：觀看整個 Cluster 的狀態，管理指令
- Networking：配置網路拓樸，檢查 DHCP lease 狀態
- Provisioning：設定開機用的 image
- Grouping：將機器依照性質分組
- Devices：查看機器狀態
- Datacenter Infrastructure：畫出機櫃圖，方便維運
- HPC：高速運算使用
- Cloud：與 public clouds 整合 (e.g. AWS, Azure)
- Containers：容器運算使用，如 Kubernetes or Docker
- Jupyter：啟用 Jupyter
- Monitoring：各節點的 Health check
- Identity Management：管理叢集帳號密碼

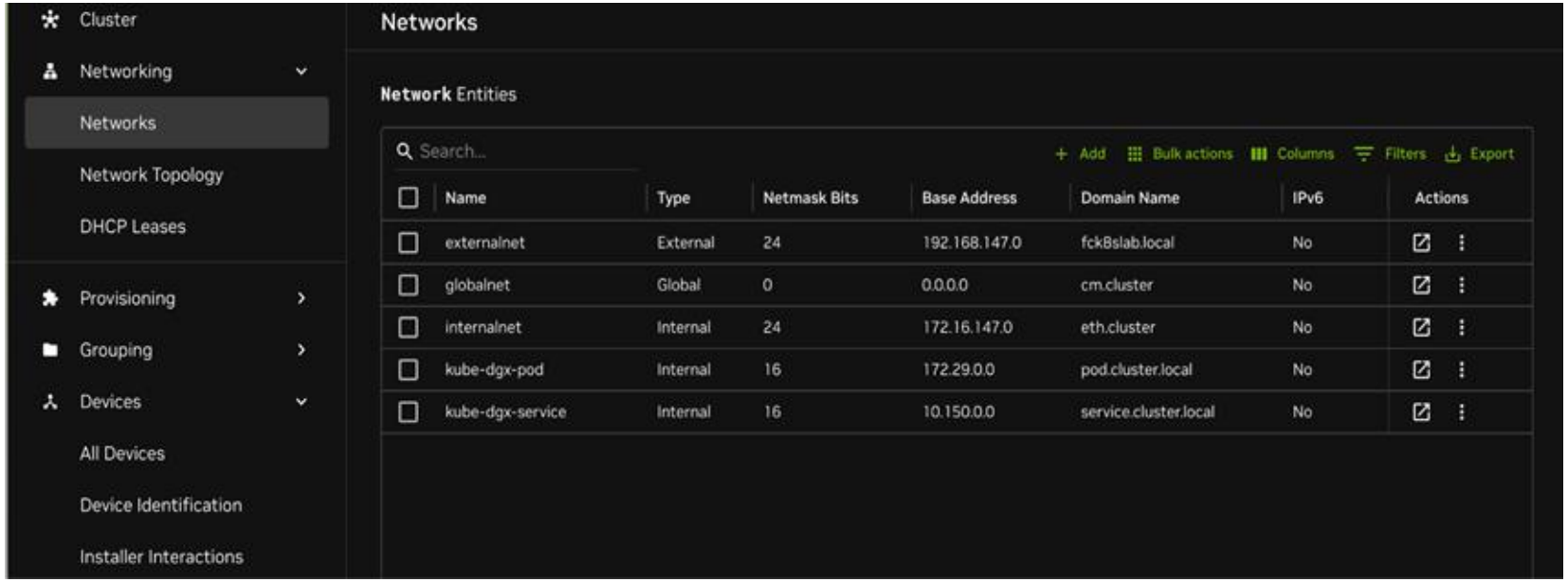


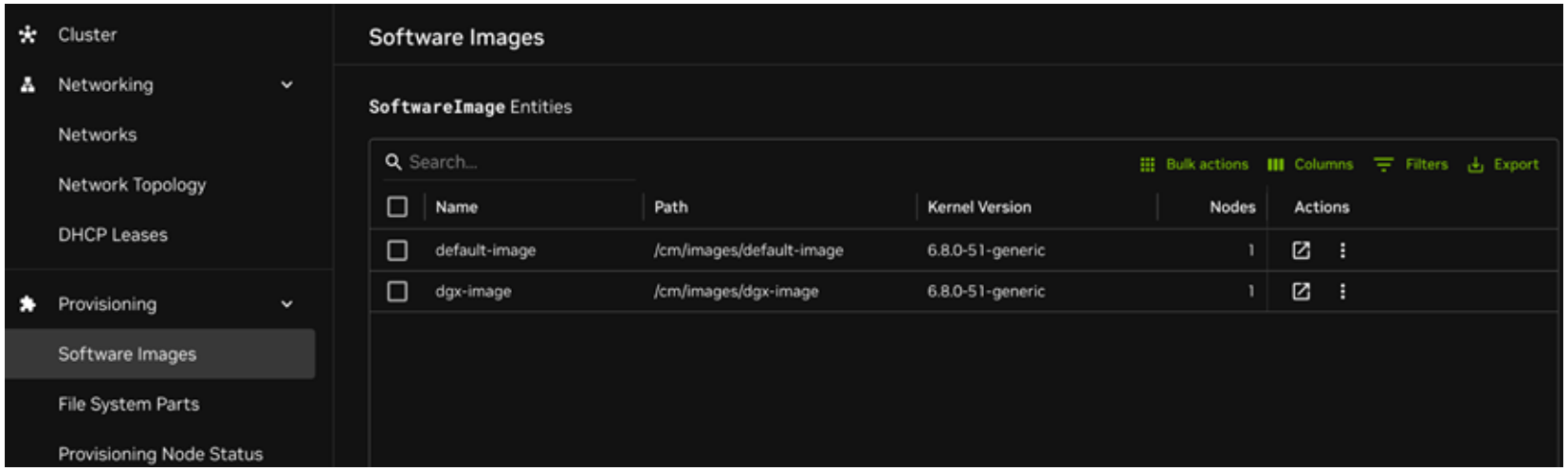
EXAMPLE: 節點管理

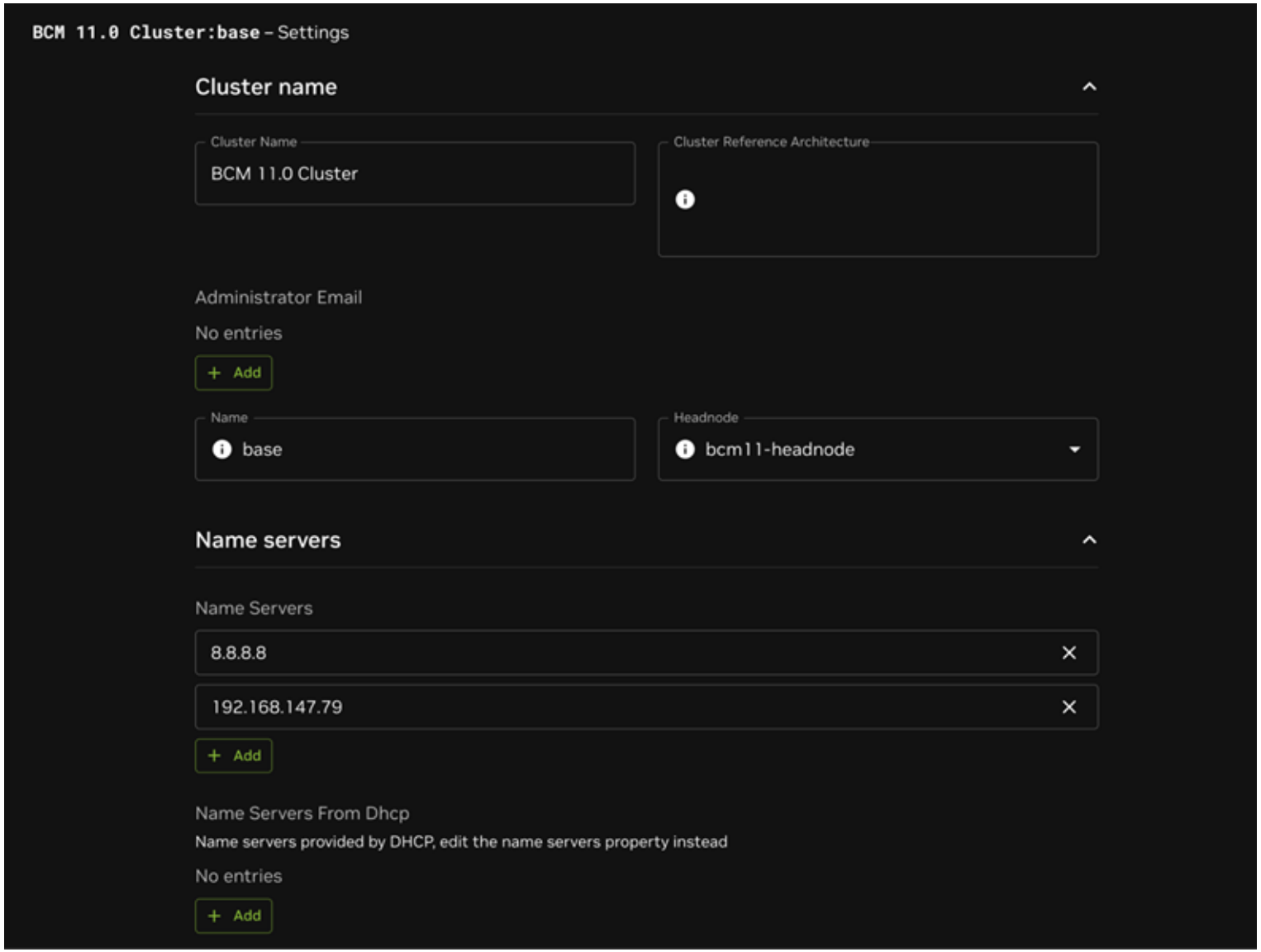



Cluster
Networking
Networks
Network Topology
DHCP Leases
Provisioning
Grouping
Categories
Configuration Overlays
Node Groups
Categories
Category Entities
Search...
Add
Bulk actions
Columns
Filters
Export
Name
Software Image
Nodes
Actions
default
default-image
1
dgx
dgx-image
1


Cluster
Networking
Networks
Network Topology
DHCP Leases
Provisioning
Grouping
Categories
Configuration Overlays
Node Groups
Node Groups
NodeGroup Entities
Search...
Add
Bulk actions
Columns
Filters
Export
Name
Nodes
Actions
k8s worker
dgx-w1, dgx1

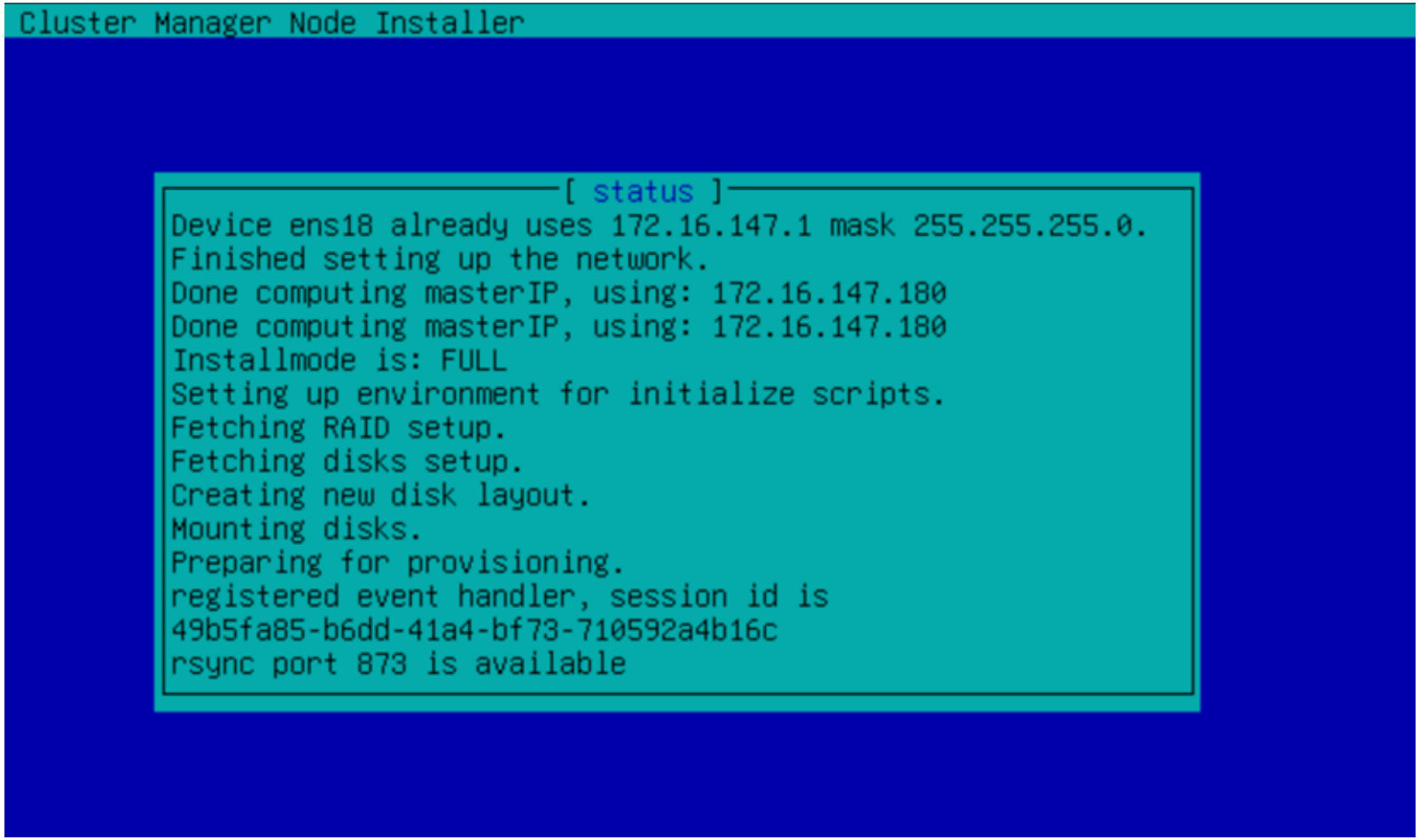

Cluster
Networking
Networks
Network Topology
DHCP Leases
Provisioning
Grouping
Devices
All Devices
Device Identification
Installer Interactions
Networks
Network Entities
Search...
Add
Bulk actions
Columns
Filters
Export
Name
Type
Netmask Bits
Base Address
Domain Name
IPv6
Actions
externalnet
External
24
192.168.147.0
fckBslab.local
No
globalnet
Global
0
0.0.0.0
cm.cluster
No
internalnet
Internal
24
172.16.147.0
eth.cluster
No
kube-dgx-pod
Internal
16
172.29.0.0
pod.cluster.local
No
kube-dgx-service
Internal
16
10.150.0.0
service.cluster.local
No

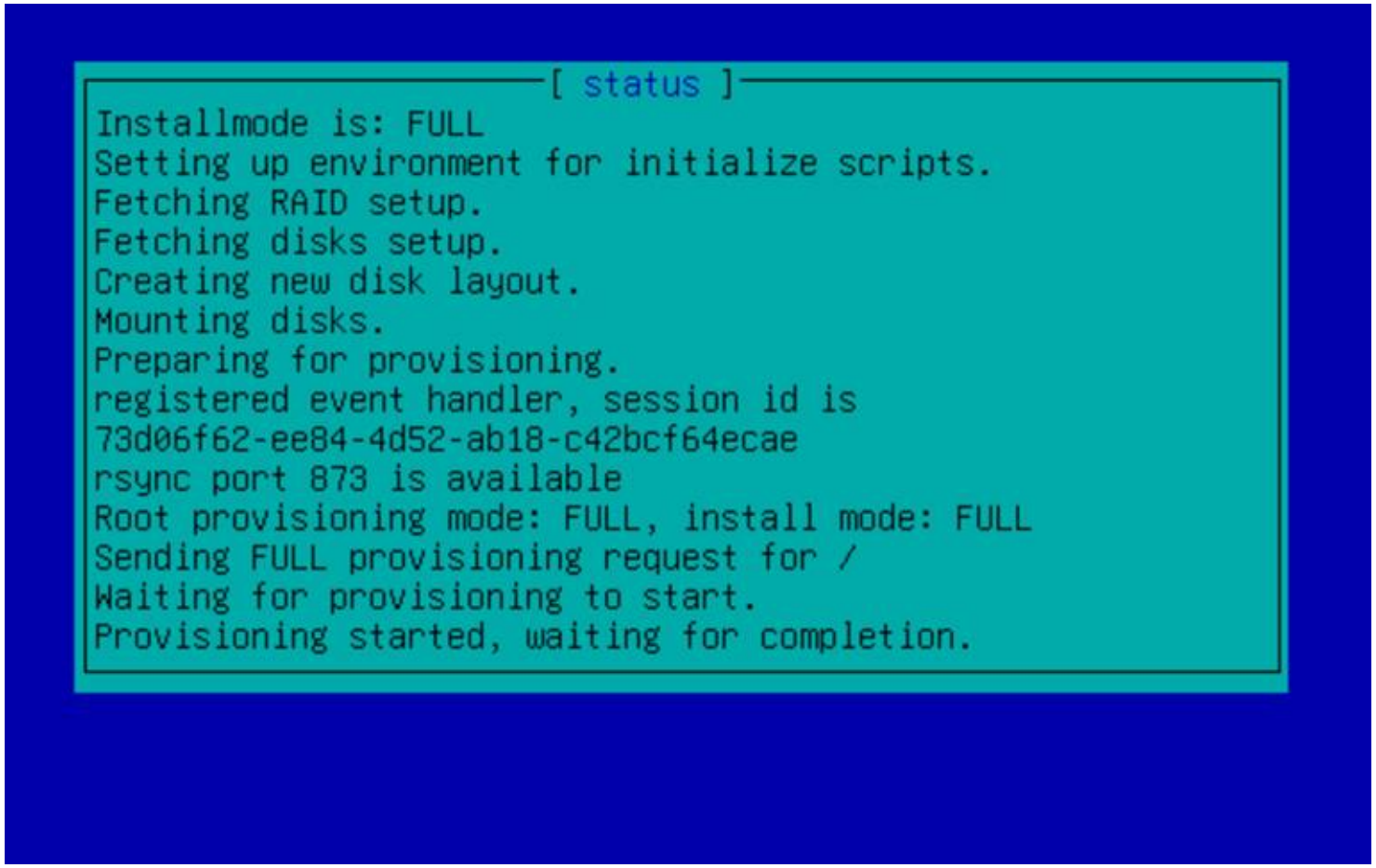

Cluster
Networking
Networks
Network Topology
DHCP Leases
Provisioning
Software Images
File System Parts
Provisioning Node Status
Software Images
SoftwareImage Entities
Search...
Bulk actions
Columns
Filters
Export
Name
Path
Kernel Version
Nodes
Actions
default-image
/cm/images/default-image
6.8.0-51-generic
1
dgx-image
/cm/images/dgx-image
6.8.0-51-generic
1


BCM 11.0 Cluster:base – Settings
Cluster name
Cluster Name
BCM 11.0 Cluster
Cluster Reference Architecture
Administrator Email
No entries
+ Add
Name
base
Headnode
bcm11-headnode
Name servers
Name Servers
8.8.8.8
192.168.147.79
+ Add
Name Servers From Dhcp
Name servers provided by DHCP, edit the name servers property instead
No entries
+ Add

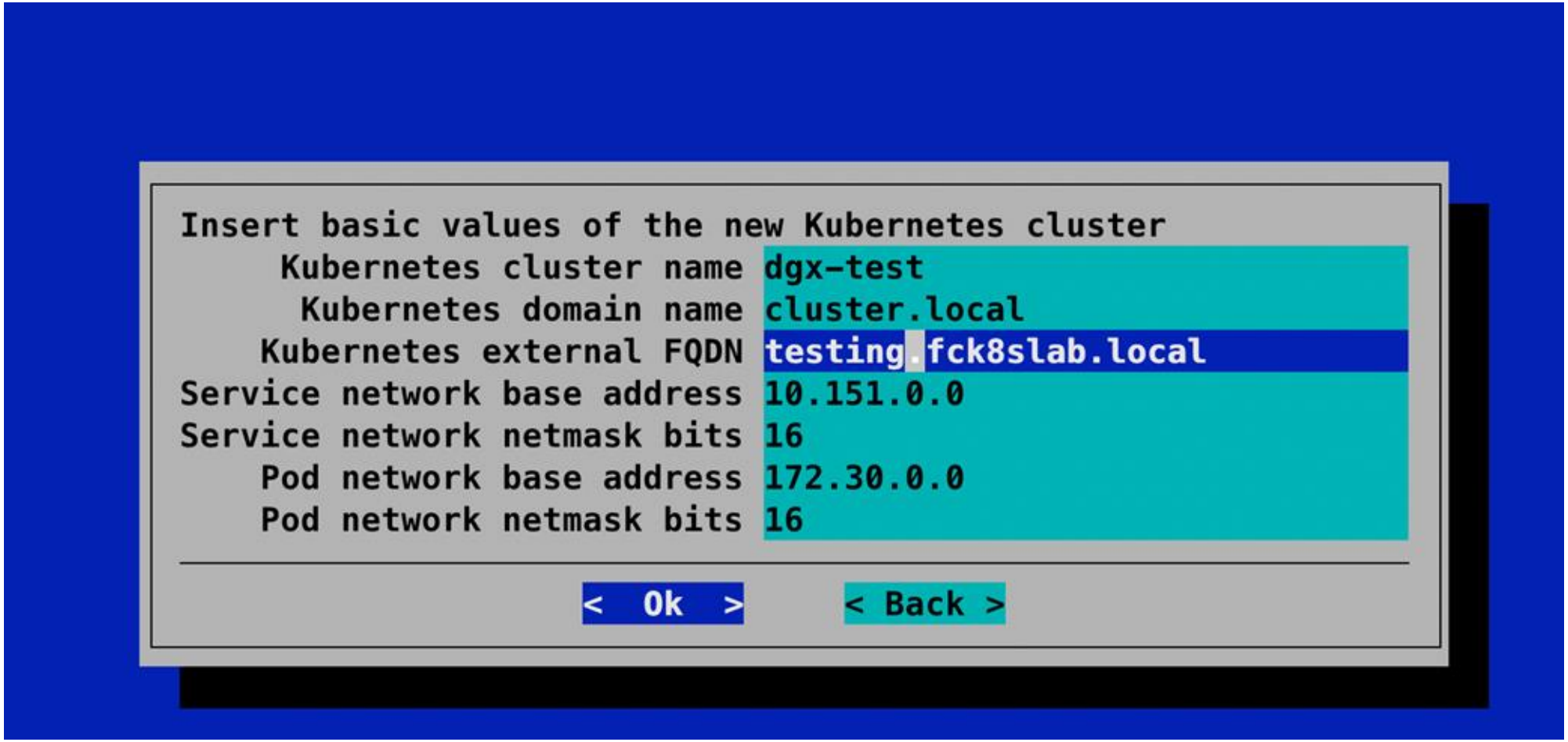
EXAMPLE: 節點 PXE-boot

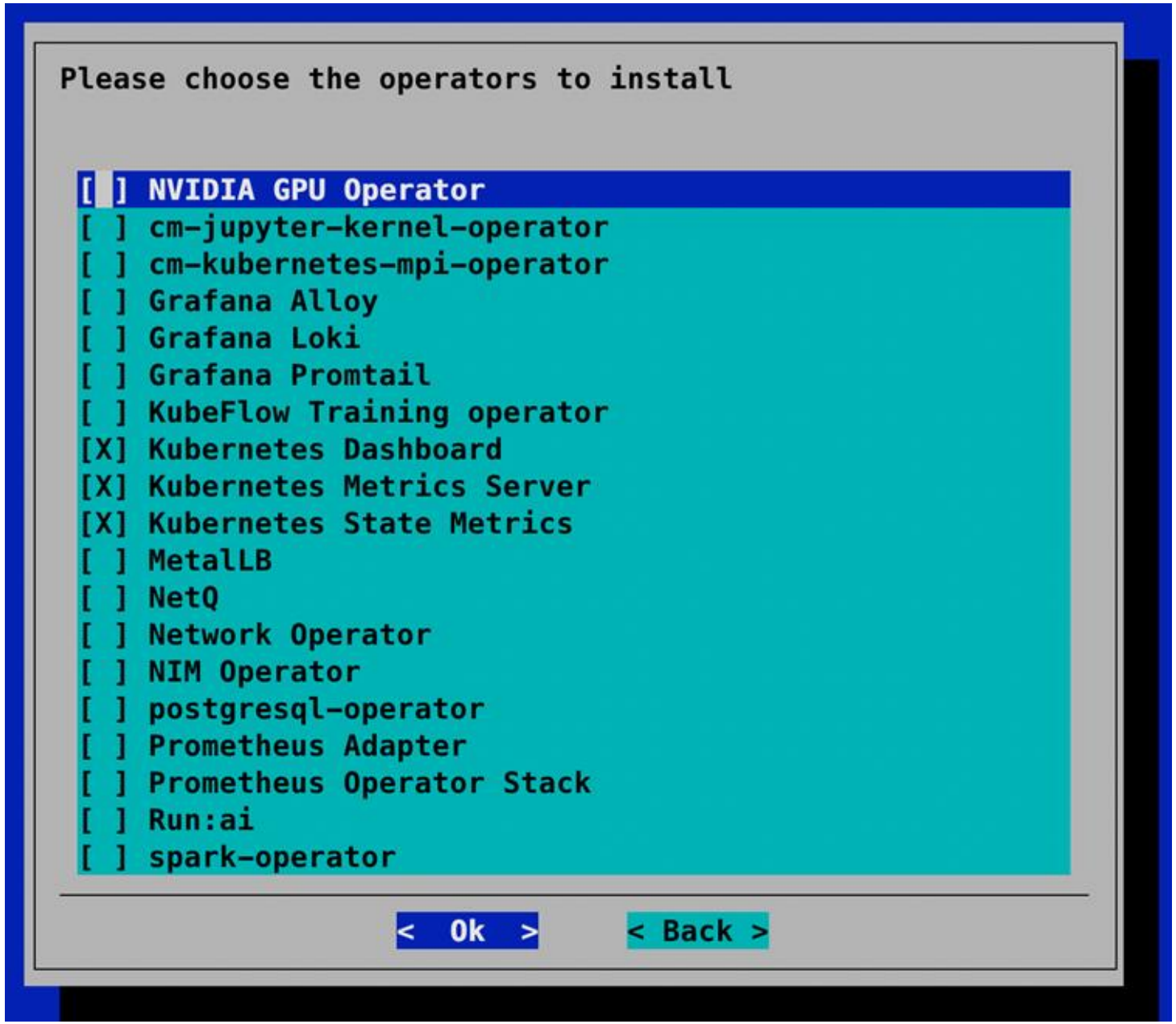
- compute node 只需要網路配置好，開機就會進行安裝


Cluster Manager Node Installer
[status]
Device ens18 already uses 172.16.147.1 mask 255.255.255.0.
Finished setting up the network.
Done computing masterIP, using: 172.16.147.180
Done computing masterIP, using: 172.16.147.180
Installmode is: FULL
Setting up environment for initialize scripts.
Fetching RAID setup.
Fetching disks setup.
Creating new disk layout.
Mounting disks.
Preparing for provisioning.
registered event handler, session id is
49b5fa85-b6dd-41a4-bf73-710592a4b16c
rsync port 873 is available


[status]
Installmode is: FULL
Setting up environment for initialize scripts.
Fetching RAID setup.
Fetching disks setup.
Creating new disk layout.
Mounting disks.
Preparing for provisioning.
registered event handler, session id is
73d06f62-ee84-4d52-ab18-c42bcf64ecae
rsync port 873 is available
Root provisioning mode: FULL, install mode: FULL
Sending FULL provisioning request for /
Waiting for provisioning to start.
Provisioning started, waiting for completion.

EXAMPLE: Kubernetes 安裝




Please choose the operators to install
[ ] NVIDIA GPU Operator
[ ] cm-jupyter-kernel-operator
[ ] cm-kubernetes-mpi-operator
[ ] Grafana Alloy
[ ] Grafana Loki
[ ] Grafana Promtail
[ ] KubeFlow Training operator
[X] Kubernetes Dashboard
[X] Kubernetes Metrics Server
[X] Kubernetes State Metrics
[ ] MetalLB
[ ] NetQ
[ ] Network Operator
[ ] NIM Operator
[ ] postgresql-operator
[ ] Prometheus Adapter
[ ] Prometheus Operator Stack
[ ] Run:ai
[ ] spark-operator
< Ok >
< Back >

```
To add users to the cluster use: refer to `cm-kubernetes-setup --help'

To use kubectl load the module file: kubernetes/default/1.32
Common URLs:
  - Kubernetes API server: https://bcm10-headnode.nvidia.com:10443
  - Kubernetes dashboard: https://dashboard.bcm10-headnode.nvidia.com:30443/
  - Kubernetes dashboard: https://0.0.0.0:30443/dashboard/
## Progress: 100

Took:     12:24 min.
Progress: 100/100
#################### Finished execution for 'Kubernetes Setup', status: completed

Kubernetes Setup finished!
```

RUN:AI

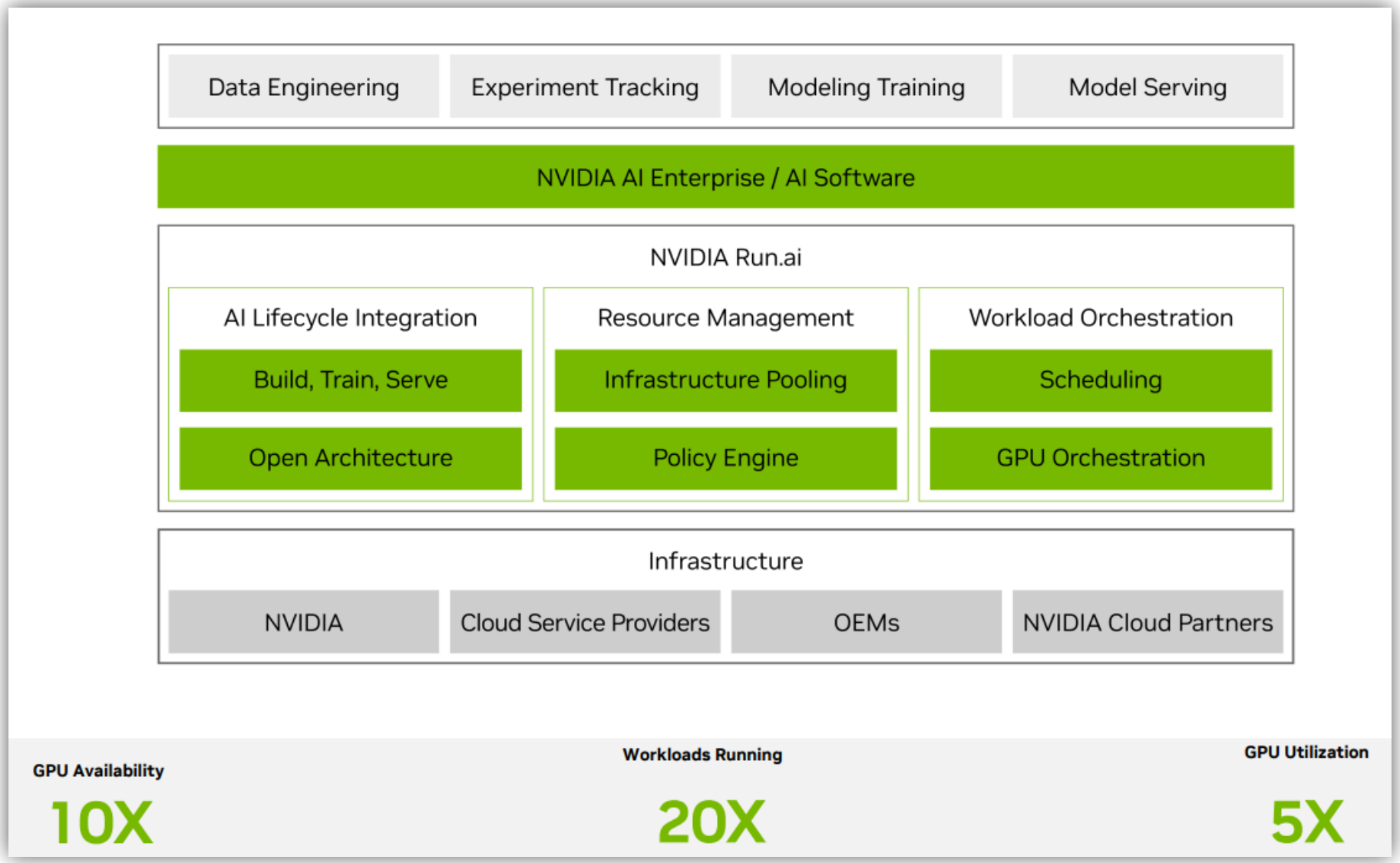
算力調度的「智慧大腦」

豐康科技
2026

NVIDIA ELITE PARTNER

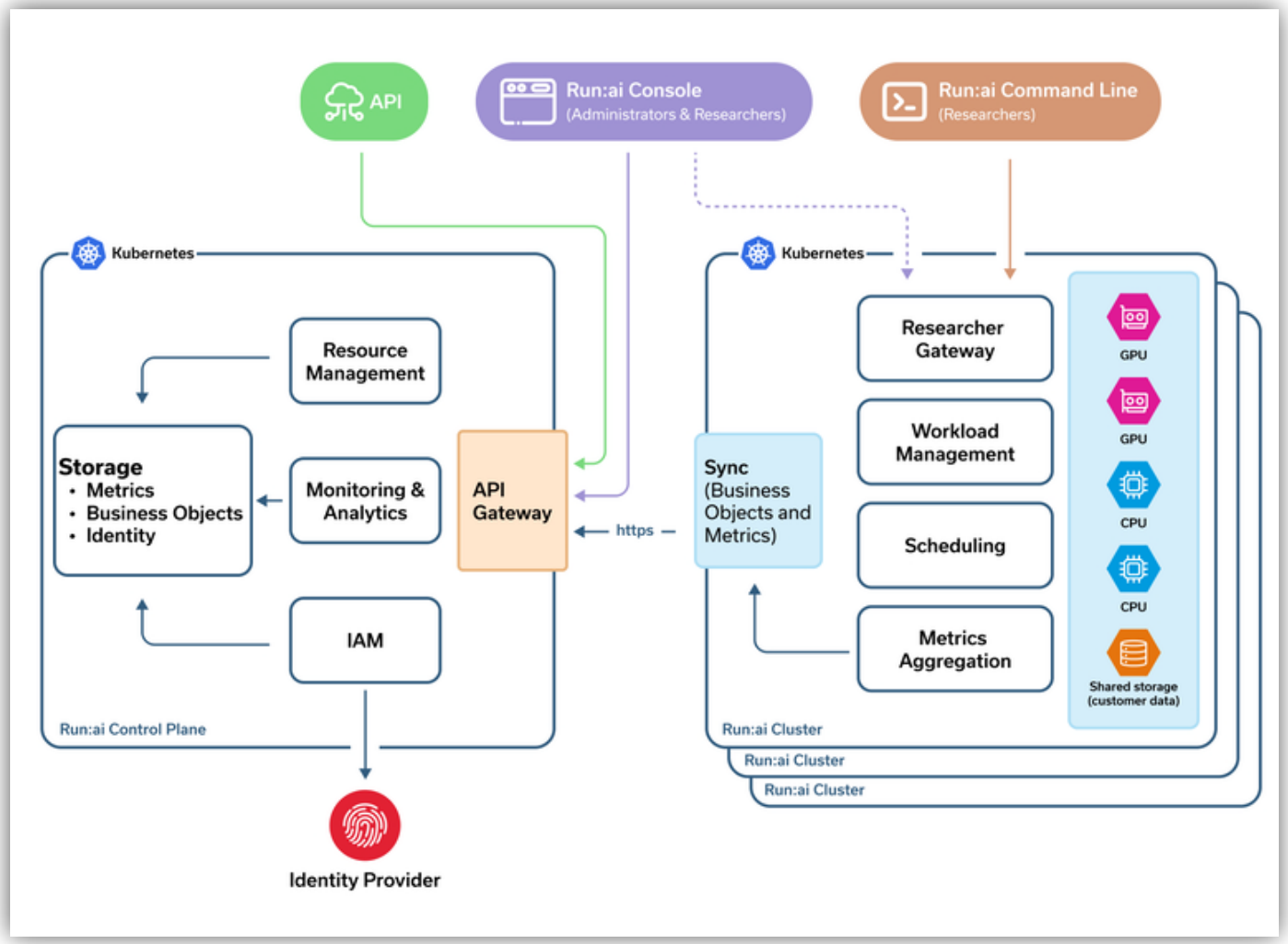
3

Nvidia Run:ai



AI工作負載與GPU編排

- **定位**：全球最強的 GPU 編排與資源管理平台
- **解決痛點**： 修正標準 Kubernetes 在 GPU 分配上的盲點
- **核心價值**： 將「硬體導向」的分配轉化為「任務導向」的自動化排程



最強大的 GPU 資源管理能力(沒有之一)

NVIDIA Run:ai 提供三種工作負載類型，對應到研究人員開發流程中的不同階段：

- Workspaces（工作區）：拿來做資料探索或模型實驗，適合互動式開發
- Training（訓練）：執行像是模型訓練或資料前處理這類需要大量資源的工作
- Inference（推論）：把訓練好的模型部署出來，提供即時推論服務用的

分類	Workspace ->	Training ->	Inference
用途	模型開發的第一階段；互動式、即時回饋	適用於長時間、無人看管的模型訓練	視為生產等級任務，排程優先度最高
IDE / 程式碼需求	可啟動自己選擇的 IDE（如 Jupyter、RStudio、VSCode）	需提供包含程式碼與相關資料集的容器映像檔	使用客製化容器映像檔或直接使用NIM
搶占政策	不可被搶占；確保開發任務不會被中斷	可被搶占；建議程式中加入 checkpoint	不可被搶占
配額政策	嚴格執行配額，除非明確設定，否則不得超出	有配額限制；預設可超出配額	嚴格依據分配的配額運作
超額行為	注意！若超出配額，可能會被中斷以讓出資源	若叢集有多餘資源，訓練可在超額狀況下執行	不支援超出配額執行
其他特點	設計用於持續且互動的開發工作	可搭配 Run:ai 的資料來源使用	支援自動擴展；確保低延遲與穩定可用性

Cluster

- 通常是一個Site的一座Kubernetes叢集
- 可為該Cluster底下的部門分配配額

Departments(可選階層)

- 包含多個專案（Projects）
  被上層 Cluster 分配配額（Quota）
  - 建議分配的配額應大於所有所屬專案所需配額的總和
- 可為自己部門底下的計畫分配配額

Projects

- 被上層 Department 分配配額
- 可為自己的 Project 內的特定任務或計畫分配配額
- 可以指定給單一使用者，也可以指定給多位使用者
- 專案可以設定規則來控制工作執行

Workloads

- 使用者在這個階層提交各種 Workload (Workspace, Training or Inference)

配額（Quota）

每個專案與部門都會針對每個節點資源池及資源類型設定一組保證的資源配額。

例如，專案 LLM-Train 在節點資源池 H100 的配額參數中，會指定該專案在該資源池中保證可使用:

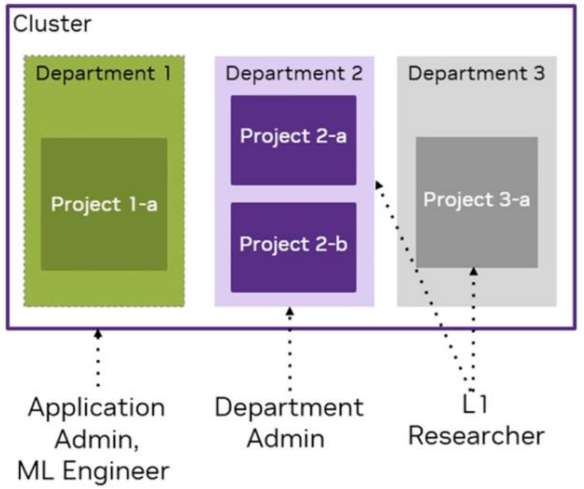
- GPU 數量
- CPU 核心數
- CPU 記憶體容量

超出配額（Over-quota）

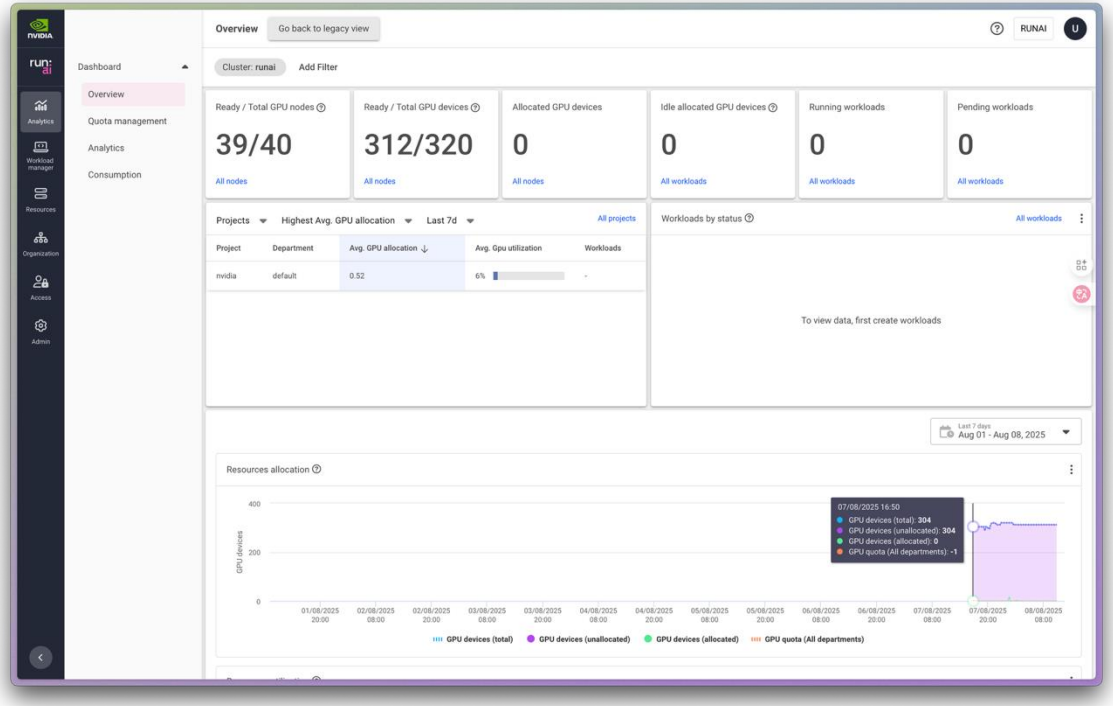
專案與部門可以使用任一節點資源池中未被使用的剩餘資源，超出其原本分配的配額。

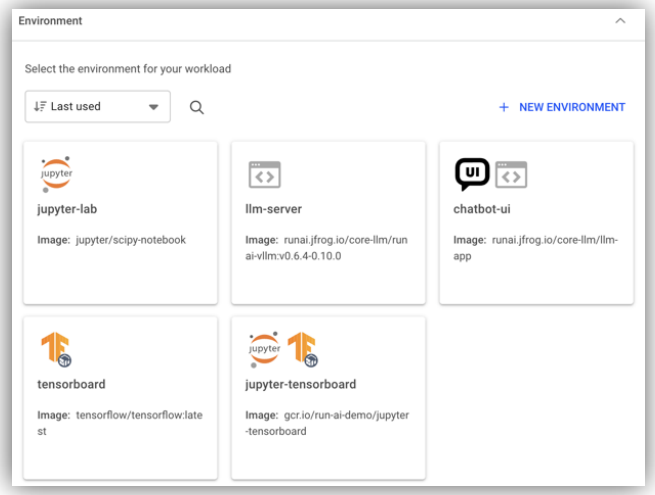
我們將這些資源稱為「超出配額資源」。

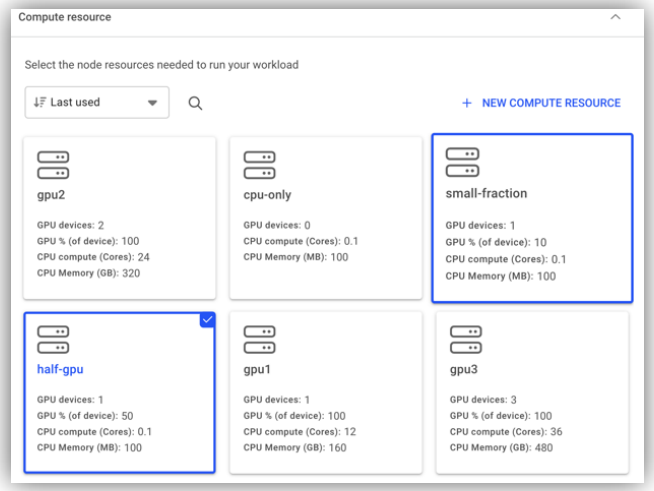
管理員可針對每個節點資源池，為每個專案與部門設定超出配額的參數。此設定 **預設是啟用** 的

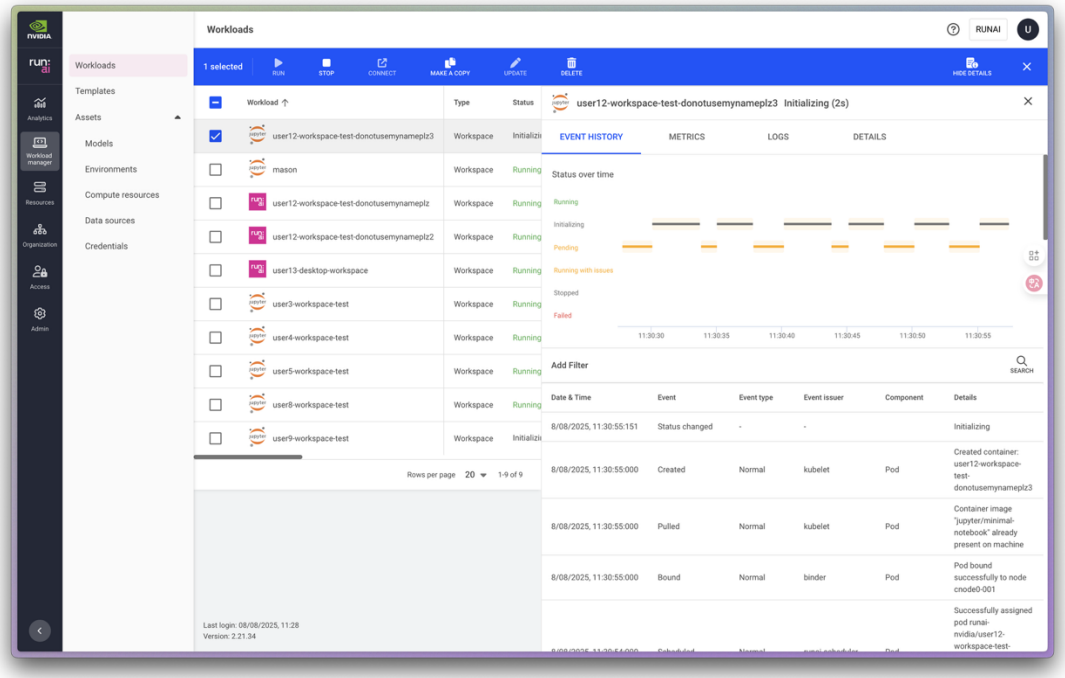


透過圖形介面操作


Overview
Go back to legacy view
RUNAI
Dashboard
Overview
Quota management
Analytics
Consumption
Cluster: runai
Add Filter
Ready / Total GPU nodes
39/40
Ready / Total GPU devices
312/320
Allocated GPU devices
0
Idle allocated GPU devices
0
Running workloads
0
Pending workloads
0
Projects
Highest Avg. GPU allocation
Last 7d
Workloads by status
To view data, first create workloads
Resources allocation
GPU devices (total)
GPU devices (unallocated)
GPU devices (allocated)
GPU quota (All departments)


Environment
Select the environment for your workload
Last used
+ NEW ENVIRONMENT
jupyter-lab
Image: jupyter/scipy-notebook
llm-server
Image: runai.jfrog.io/core-llm/runai-vllm:v0.6.4-0.10.0
chatbot-ui
Image: runai.jfrog.io/core-llm/llm-app
tensorboard
Image: tensorflow/tensorflow:latest
jupyter-tensorboard
Image: gcr.io/run-ai-demo/jupyter-tensorboard


Compute resource
Select the node resources needed to run your workload
Last used
+ NEW COMPUTE RESOURCE
gpu2
GPU devices: 2
GPU % (of device): 100
CPU compute (Cores): 24
CPU Memory (GB): 320
cpu-only
GPU devices: 0
CPU compute (Cores): 0.1
CPU Memory (MB): 100
small-fraction
GPU devices: 1
GPU % (of device): 10
CPU compute (Cores): 0.1
CPU Memory (MB): 100
half-gpu
GPU devices: 1
GPU % (of device): 50
CPU compute (Cores): 0.1
CPU Memory (MB): 100
gpu1
GPU devices: 1
GPU % (of device): 100
CPU compute (Cores): 12
CPU Memory (GB): 160
gpu3
GPU devices: 3
GPU % (of device): 100
CPU compute (Cores): 36
CPU Memory (GB): 480


Workloads
RUNAI
Templates
Assets
Models
Environments
Compute resources
Data sources
Credentials
1 selected
RUN
STOP
CONNECT
MAKE A COPY
UPDATE
DELETE
HIDE DETAILS
Workload
Type
Status
user12-workspace-test-donotusemynameplz3 Initializing (2s)
EVENT HISTORY
METRICS
LOGS
DETAILS
Status over time
Add Filter
SEARCH
Date & Time
Event
Event type
Event issuer
Component
Details

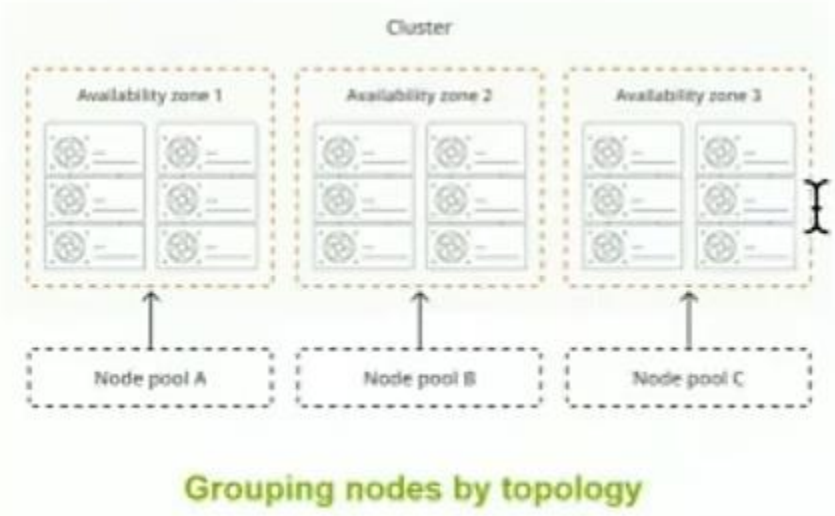
節點支援池

- Node Pools

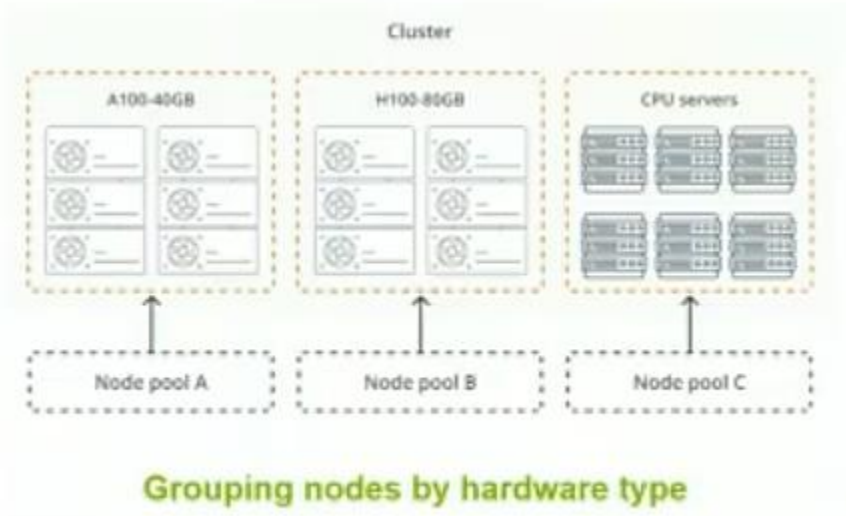
優點

節點池有助於有效管理異質資源。節點池是一種 Run:ai 結構，表示一組節點，這些節點使用預定義節點標籤（例如 NVidia GPU 類型）或管理員定義的節點標籤（任意鍵/值對）分組到資源桶中。

通常，分組節點具有共同的特徵或屬性，例如 GPU 類型或其他硬體功能（例如 Infiniband 連接），或代表鄰近群組（即透過本地超高速交換器互連的節點）。研究人員和機器學習工程師通常會使用節點池在特定資源類型上執行特定工作負載。



Grouping nodes by topology



Grouping nodes by hardware type

Project 內資源搶佔

豐康科技
www.fongcon.com.tw

被搶佔

STEP 1: 新增 workspace workload(under quota)
- Projects: proj1
- 任務名:proj1-8gpu-under-quota
- Environment: jupyter-lab
- Compute resource: gpu8
- CREATE WORKSPACE

STEP 2: 新增 training workload
- Projects: proj2
- 任務名:proj2-1gpu-under-quota
- Environment: jupyter-lab
- Compute resource: gpu1
- CREATE Training

STEP 3: 新增 workspace workload
- Projects: proj2
- 任務名:proj2-8gpu-under-quota
- Environment: jupyter-lab
- Compute resource: gpu8
- CREATE WORKSPACE

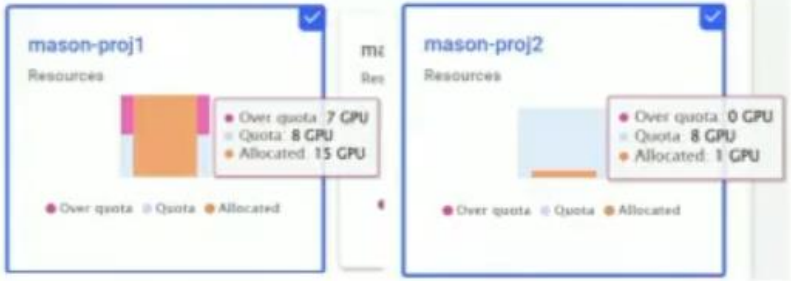
多 Project 資源搶佔

被搶佔

STEP 1: 新增 workspace workload(under quota)
- Projects: proj1
- 任務名:proj1-8gpu-under-quota
- Environment: jupyter-lab
- Compute resource: gpu8
- CREATE WORKSPACE

STEP 2: 新增 training workload(proj1 is LOW over quota priority)
- Projects: proj1
- 任務名:proj1-7gpu-over-quota
- Environment: jupyter-lab
- Compute resource: gpu7
- CREATE TRAINING

STEP 3: 新增 workspace workload
- Projects: proj2
- 任務名:proj2-1gpu-under-quota
- Environment: jupyter-lab
- Compute resource: gpu1
- CREATE WORKSPACE



STEP 4: 新增 workspace workload
- Projects: proj2
- 任務名:proj2-1gpu-under-quota2
- Environment: jupyter-lab
- Compute resource: gpu1
- CREATE WORKSPACE

BCM + RUN

AI 的 1+1>2 效應

豐康科技
2026

NVIDIA ELITE PARTNER

4

Role & Responsibility

豐康科技 www.fongcon.com.tw

Infrastructure Administrator

- **Centralized cluster management** - Manage all clusters from a single platform, ensuring consistency and control across environments.
- **Usage monitoring and capacity planning** - Gain real-time and historical insights into GPU consumption across clusters to optimize resource allocation and plan future capacity needs efficiently.
- **Policy enforcement** - Define and enforce security and usage policies to align GPU consumption with business and compliance requirements.
- **Enterprise-grade authentication** - Integrate with your organization's identity provider for streamlined authentication (Single Sign On) and role-based access control (RBAC).
- **Kubernetes-native application** - Install as a Kubernetes-native application, seamlessly extending Kubernetes for native cloud experience and operational standards (install, upgrade, configure).

Platform Administrator

- **AI Initiative structuring and management** - Map and set up AI initiatives according to your organization's structure, ensuring clear resource allocation.
- **Centralized GPU resource management** - Enable seamless sharing and pooling of GPUs across multiple users, reducing idle time and optimizing utilization.
- **User and access control** - Assign users (AI practitioners, ML engineers) to specific projects and departments to manage access and enforce security policies, utilizing role-based access control (RBAC) to ensure permissions align with user roles.
- **Workload scheduling** - Use scheduling to prioritize and allocate GPUs based on workload needs.
- **Monitoring and insights** - Track real-time and historical data on GPU usage to help track resource consumption and optimize costs.

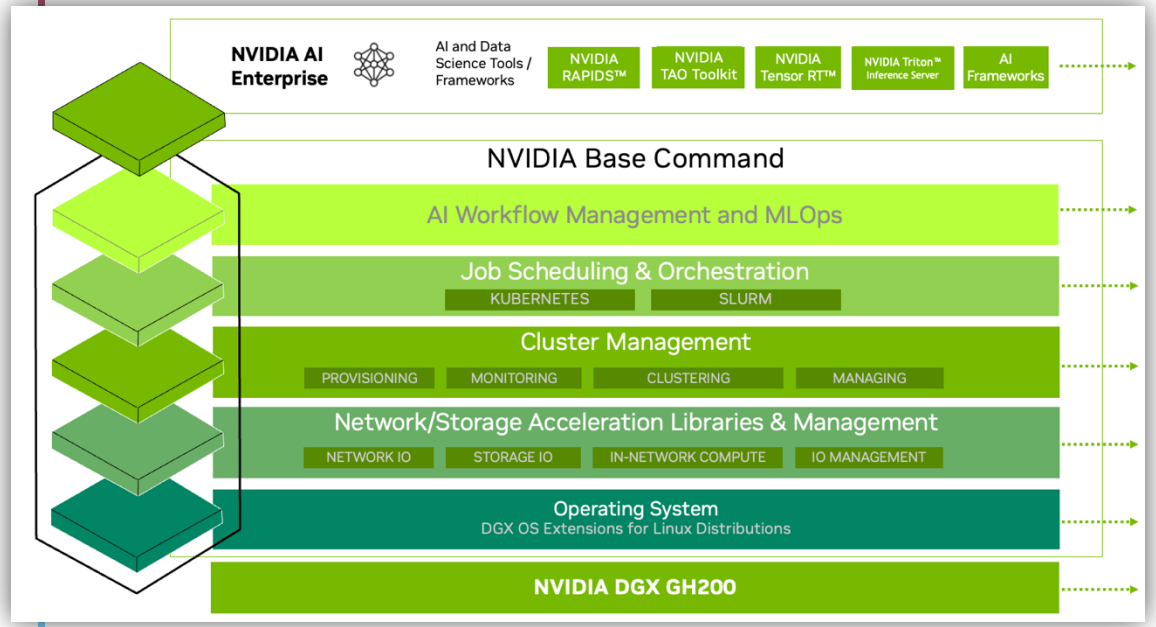
AI Practitioners

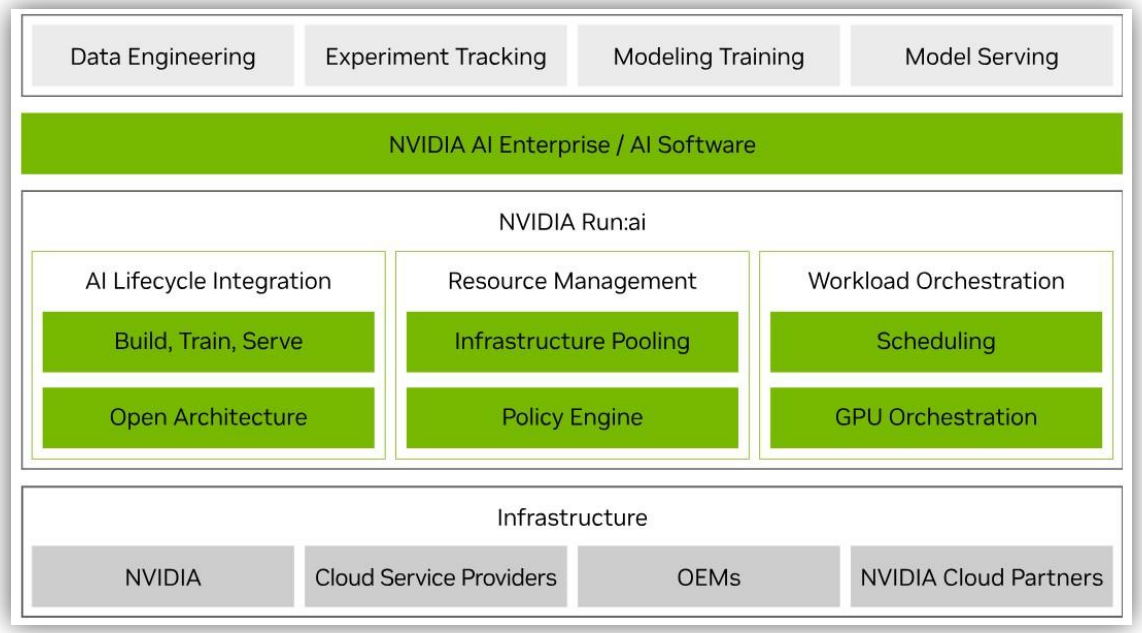
- **Optimized workload scheduling** - Ensure high-priority jobs get GPU resources. Workloads dynamically receive resources based on demand.
- **Fractional GPU usage** - Request and utilize only a fraction of a GPU's memory, ensuring efficient resource allocation and leaving room for other workloads.
- **AI initiatives lifecycle support** - Run your entire AI initiatives lifecycle – Jupyter Notebooks, training jobs, and inference workloads efficiently.
- **Interactive session** - Ensure an uninterrupted experience when working on Jupyter Notebooks without taking away GPUs.
- **Scalability for training and inference** - Support for distributed training across multiple GPUs and auto-scales inference workloads.
- **Integrations** - Integrate with popular ML frameworks - PyTorch, TensorFlow, XGBoost, Knative, Spark, Kubeflow Pipelines, Apache Airflow, Argo workloads, Ray and more.
- **Flexible workload submission** - Submit workloads using the NVIDIA Run:ai UI, API, CLI or run third-party workloads.

Ref: https://run-ai-docs.nvidia.com/self-hosted/getting-started/overview

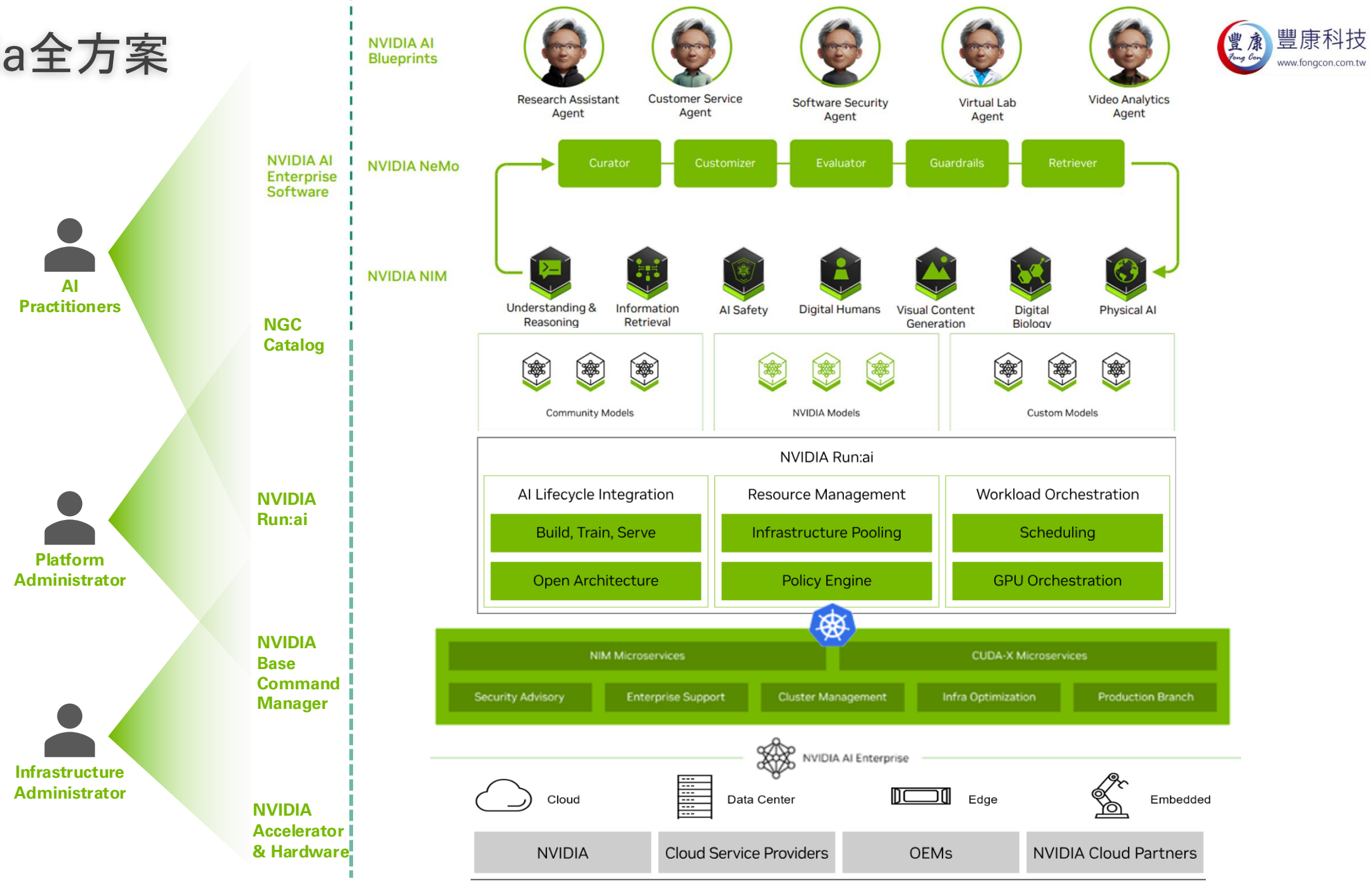
軟硬體協同：BCM 與 Run:ai 的 1+1 > 2

- **BCM 管「下層」**：確保伺服器活著、網路通著、硬體健康。
- **Run:ai 管「上層」**：確保每一分算力都被公平、高效地分配。
- **管理一致性**：減少 IT 部門在開源工具組合（如 K8s + SLURM）中的調教成本。





Nvidia全方案



結語

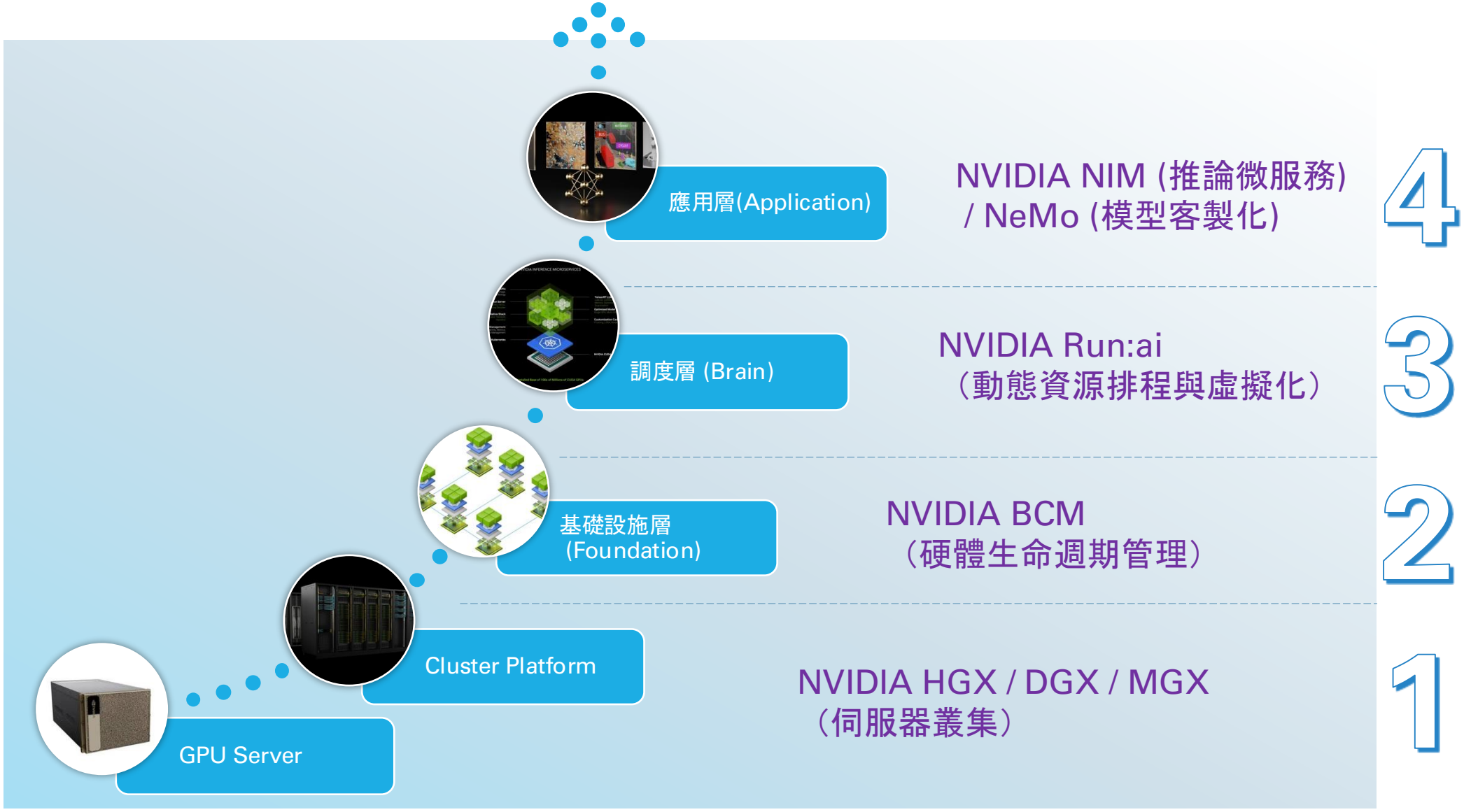
與未來展望

豐康科技
2026

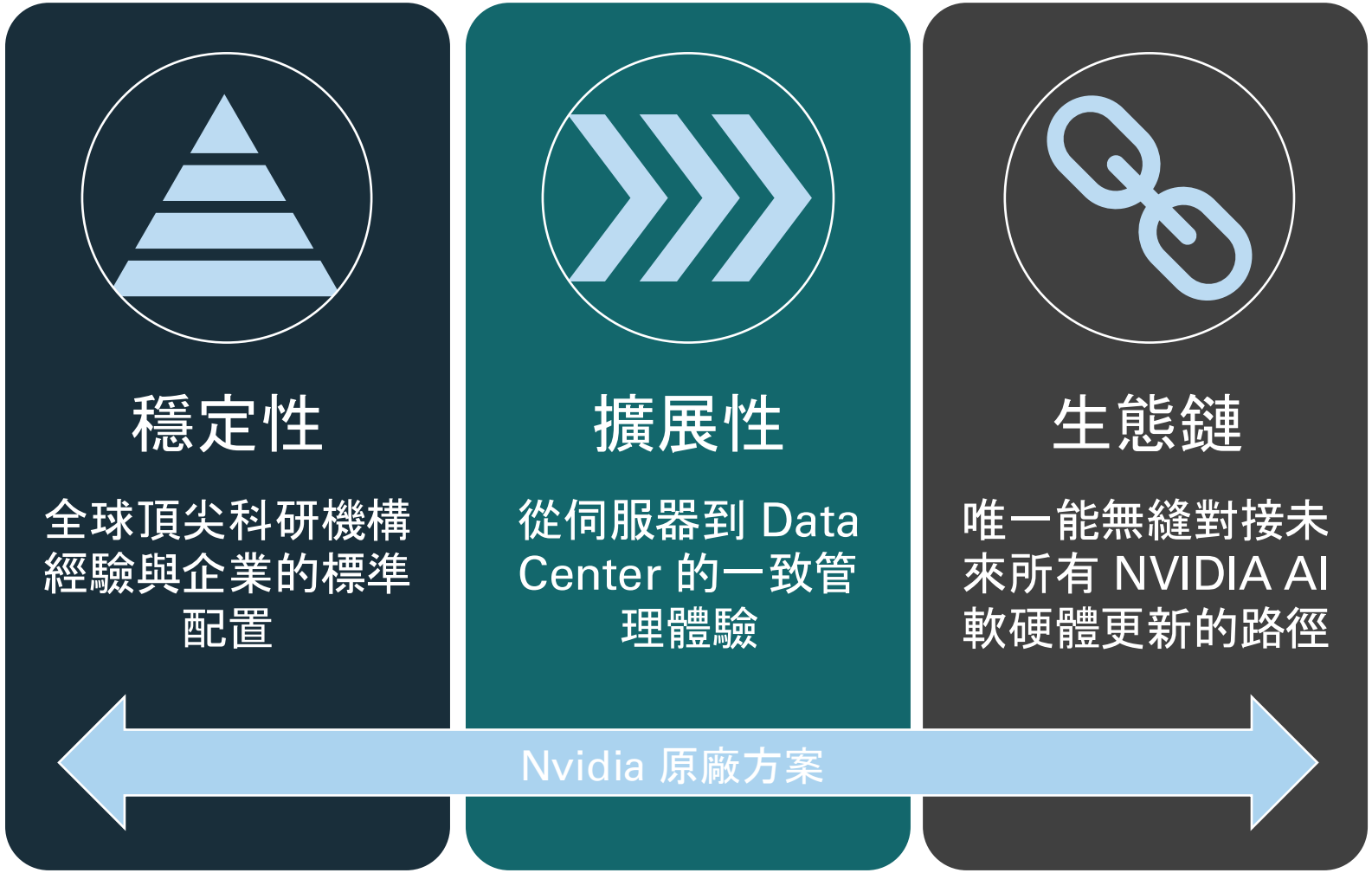
NVIDIA ELITE PARTNER

5

NVIDIA 全棧 AI 軟體解決方案架構



結論：選擇 NVIDIA 原廠方案的戰略意義



未來展望：邁向 Agentic AI 與 NIM 算力即服務 (MaaS)

豐康科技 www.fongcon.com.tw

2026 趨勢

學校不再只是租借 GPU, 而是提供「模型 API」

NIM 集成

在 BCM 與 Run:ai 穩定的環境上，快速部屬 NVIDIA NIM 微服務

讓師生專注於「創新」，而非「設定環境」

投資效益分析 (ROI)

豐康科技
www.fongcon.com.tw

硬體利用率	從傳統的 20-30% 提升至 80% 以上	80%⇧
維運人力	自動化部署減少 60% 重複性行政與設定工作	60%⇩
科研競爭力	自動化部署減少 60% 重複性行政與設定工作	∞ ⇨

結語

豐康科技
www.fongcon.com.tw

不是模型跑得最快，
而是平台能長期穩定營運。

豐康科技
www.fongcon.com.tw

Thank you !

NVIDIA ELITE PARTNER