Experiences and Lessons Learned in Operating GPU-Based HPC Systems

陳政宇
資深技術經理, 台智雲
Apr. 26th, 2024

TWNOG 5.0

Outline

- TWSC & TWCC Background
- AI-HPC and Cloud Hybrid Architecture Design
- Experiences and Lessons Learned
- GenAI Solution for Training and Inference

TWSC: OUR STORY

台智雲的前身，是科技部2017年起花了4年時間、耗資50億元打造、曾打進全球前20大超級電腦的「台灣杉二號」。

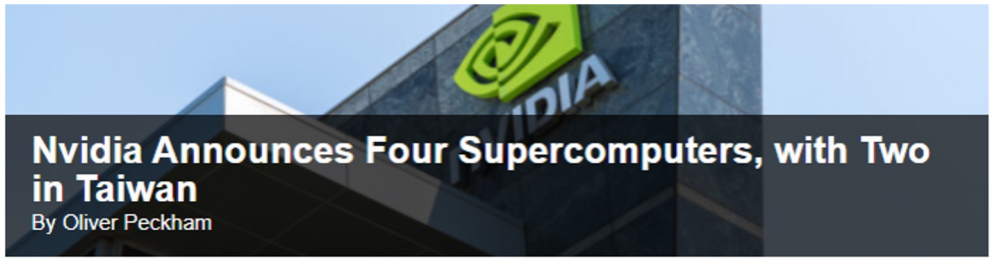
台灣第一家提供機敏資料落地、**國家主權管理**及**前瞻AI應用**發展的雲端高速運算及海量儲存之雲服務運營商。

透過國家級**TWCC 臺灣 AI 雲**平台資料中心，提供各種產業數位發展所需的AI智慧應用服務及雲架構解決方案。

AIHPC
in coming

Taiwania 4 HPC 2023	**#222** of Top500 2023/11
Taiwania 3 HPC 2020	**#181** of Top500 2021/6
Taiwania 2 AIHPC 2018	**#20** of Top500 2018/11

NARLabs 國家實驗研究院
國家高速網路與計算中心
National Center for High-performance Computing



Nvidia Announces Four Supercomputers, with Two in Taiwan

By Oliver Peckham

May 29, 2023

At the Computex event in Taipei this week, Nvidia announced four new systems equipped with its Grace- and Hopper-generation hardware, including two in Taiwan. Those two are Taiwania 4, powered by Nvidia's Grace CPU Superchip, and Taipei-1, based on Nvidia's H100 GPUs. The others: Helios and Israel-1.

Taiwan Computing Cloud for AI

TWCC TAIWAN COMPUTING CLOUD

TWS

HPC - Taiwania2

- 252 nodes / 2016 V100 GPUs
- 9 Nvidia DGX H100 (New in 2023)
- 10 PB Parallel file system
- EDR InfiniBand 100 Gbps
- 1.2 PUE (Warm Water Cooling)

Software Environment

- Slurm / Kubernetes
- Openstack
- Nvidia NGC Docker Images
- Ceph (Object & Block)
- Spectrum Scale (GPFS)

TOP 500 The List. Ranked 20th

The GREEN 500 Ranked 10th

HPC - compute node

- Intel Xeon Gold CPU x 2
- 768 GB memory
- 240 GB SSD + 4TB NVMe
- Nvidia Tesla V100 w/32GB x 8
- EDR InfiniBand 100 Gbps x 4
- Dual Port 10Gb Ethernet

MPI / AI Framework

- OpenMPI / Intel oneAPI
- Tensorflow / PyTorch
- Nvidia NGC images
- .....and more

TAIWANIA 2
9
petaflops
Rmax

NCHC NARLabs 財團法人國家實驗研究院 國家高速網路與計算中心 National Center for High-performance Computing

台灣杉二號 TAIWANIA 2

Current Dev. for Day0

Cluster Planning

Day 0 Op

Day 1 Op

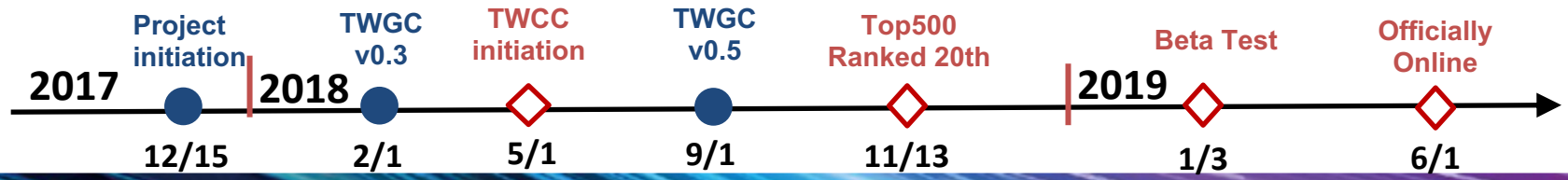
Day 2 Op

Cluster Planning

- Taiwan GPU Cloud

TWGC

10 Nvidia DGX V100 16GB



Cluster Planning

- A simple AI pipeline

Cluster Planning

TWCC TAIWAN COMPUTING CLOUD

- Taiwan Computing Cloud

Big Data (Cloud, Hadoop, Spark)

- Using CPU to run computing jobs
- High I/O Throughput(Read and Write)
- 3Vs(Volume, Velocity, Variety)

Cloud

- On-demand cloud service
- Reliable infrastructure
- Cyber security

AI Training(AIHPC/GPU Container)

- GPU-Accelerated Computing
- Write once read many and small files (e.g. images)
- Large-scale training jobs
- Interactive interfaces

Current Dev. for Day1

The birth of TWCC

Day 0 Op

Day 1 Op

Day 2 Op

TWCC Software Stack

Virtual Compute Service

VCS (Virtual Compute Service)

雲端佈建，多重選擇

動動手指、點點滑鼠，即刻間擁有您專屬的工作環境。提供 Linux Ubuntu、CentOS...等多種作業系統供您建立，CPU 數量、記憶體容量、硬體資源任您選擇！

放心使用，安全無疑慮

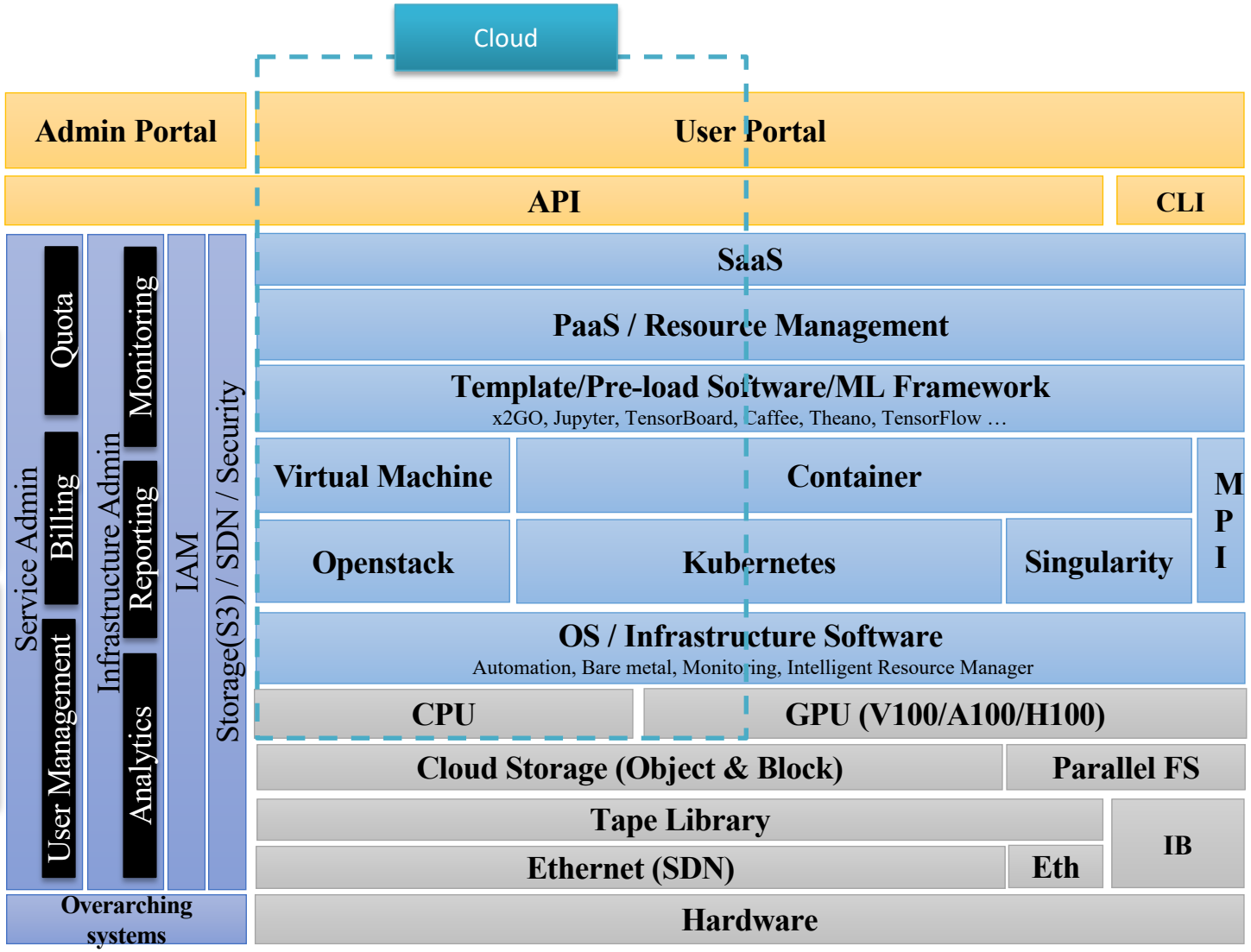
透過安全殼層（SSH）並搭配鑰匙對的方式連線虛擬運算個體、對敏感性資料傳輸加密並築牆抵禦惡意程式，層層防護讓您安心使用。

輕鬆備份，迅速還原

映像檔備份系統提供您在雲端保存虛擬運算個體的設定、重要資料，能快速從災難中復原、重建，損失程度大幅降低！

自動平衡負載，服務不中斷

擁有負載平衡機制，可設定系統資源使用條件，自動偵測並彈性擴展縮減處理環境、調配資源，無需時時擔心計算工作因系統超載而被迫中斷。



TWCC Software Stack

Interactive Container

Scheduled Container

CCS (Container Compute Service)

輕鬆使用服務

透過 TWCC 入口網站、API、CLI (Command Line Interface)，皆可建立容器運算服務。有別於傳統使用超級電腦資源時僅能藉由 CLI 操作運算，TWCC 供您自由選擇熟悉的介面，輕鬆堆疊應用。

快速部署工作環境

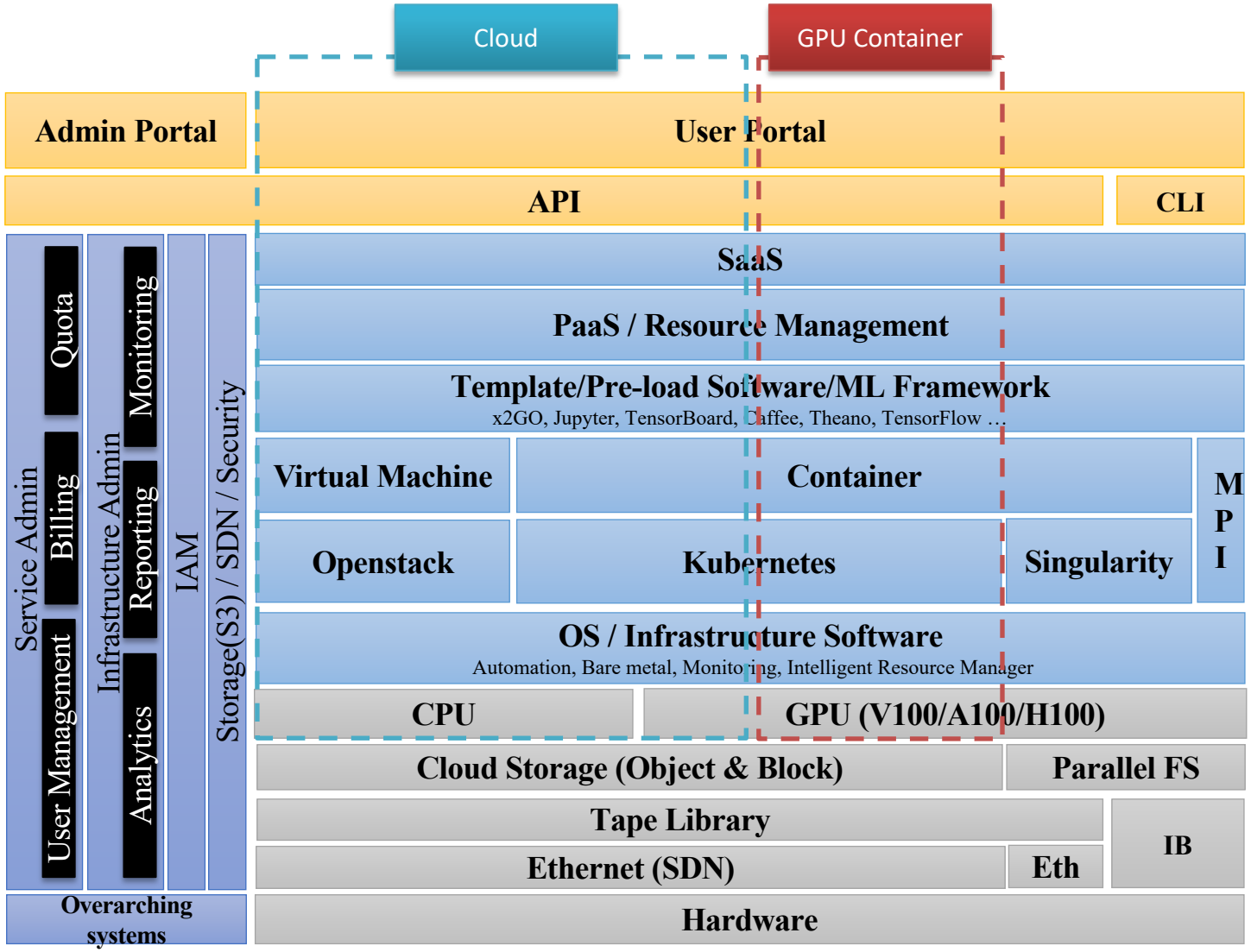
採用 Kubernetes 架構，並導入 Nvidia 優化 AI 軟體堆疊，串接與世界接軌的虛擬化技術，短時間內便能快速部署開發工作環境，相較於傳統的方式可節省 3 倍的時間，並能彈性地轉換平台。

多樣化的 AI 框架

提供 Nvidia 優化之 TensorFlow、Caffe、CUDA、Torch、PyTorch、TensorRT 、TensorRT Server、CNTK、MXNet、Theano、DIGITS、RAPIDS 等等的 AI 框架，無需費心安裝，並能滿足不同模型訓練與推論的需求。

多項底層優勢

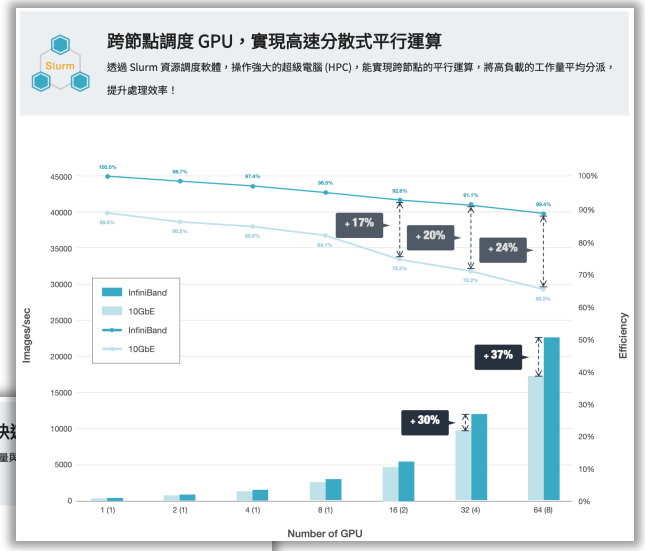
- 租戶隔離的網路架構，安全使用。
- Kubernetes 經常性的滾動升級容器，功能不斷優化。
- Kubernetes 與 HPC 超融合架構，GPU 容器開發程式，並能大量佈署至 HPC 進行跨節點運算。

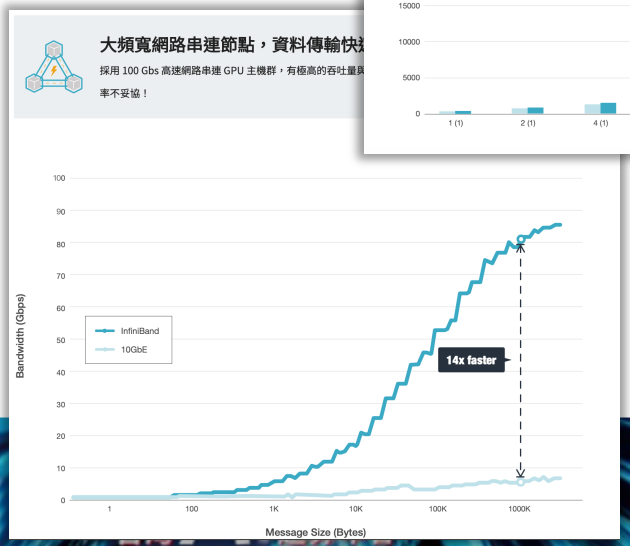


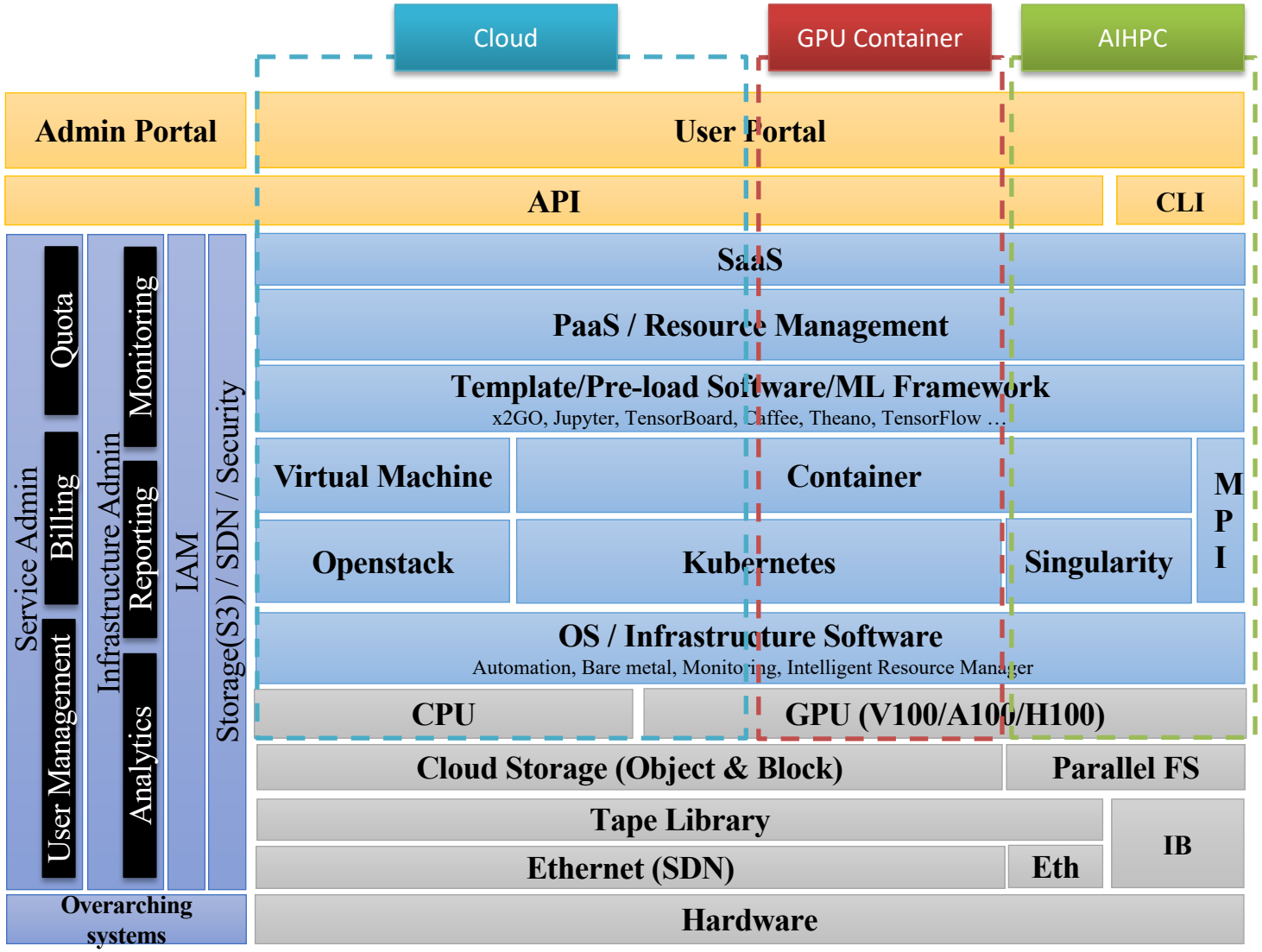
TWCC Software Stack

HPC Job

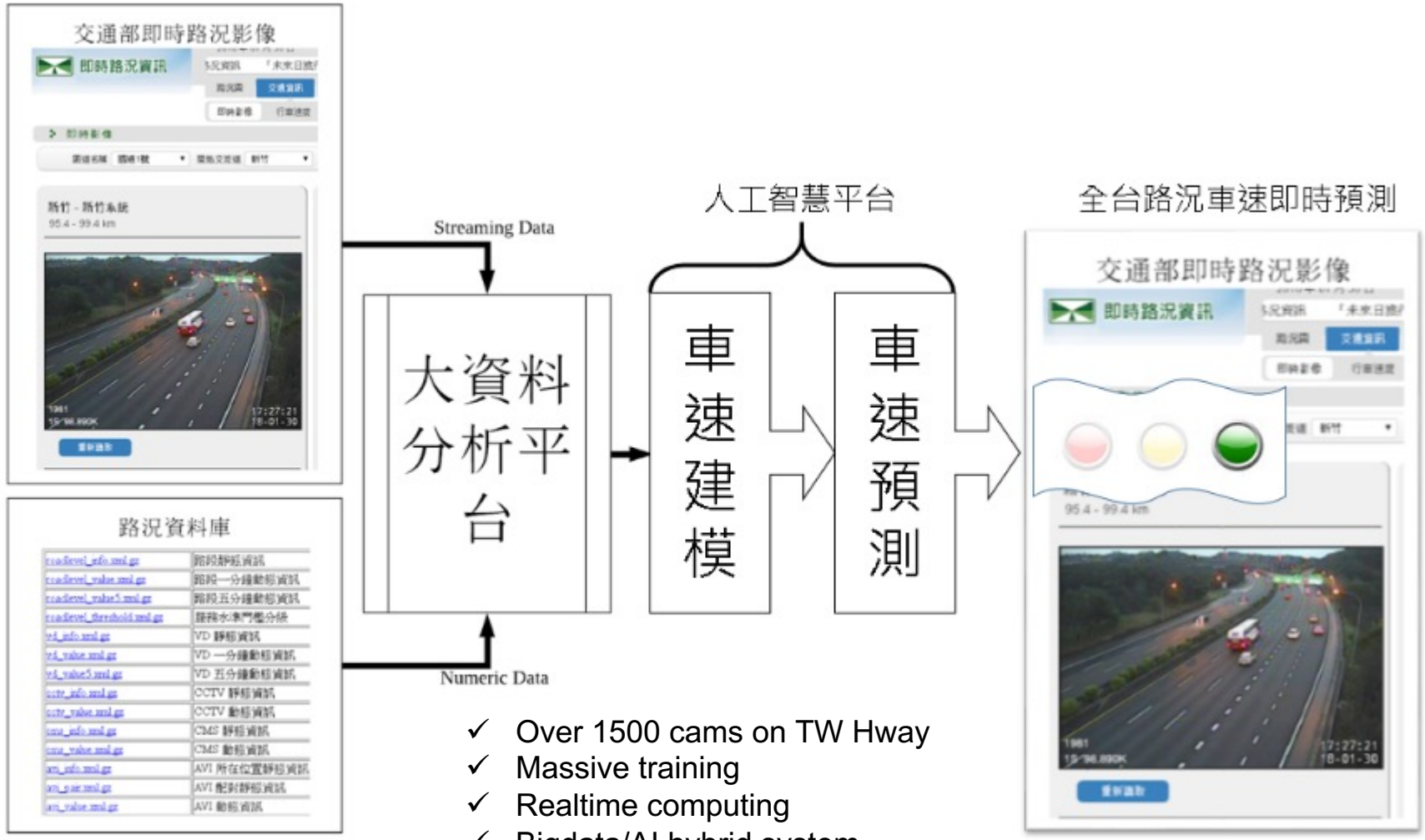
HPC (High-Performance Computing)







Example Case: Highway Speed realtime prediction



- ✓ Over 1500 cams on TW Hway
- ✓ Massive training
- ✓ Realtime computing
- ✓ Bigdata/AI hybrid system

Lessons learned in Hybrid-Architecture

- Performance considerations
  - HPC
    - NUMA
    - IB Topology
  - Kubernetes
    - NUMA
    - SRIOV
  - Openstack
    - NUMA
    - Passthrough A100/H100 to Openstack VM
- GPU Resource Management
- Security

HPC – NUMA

- GPU-CPU Affinity

```
[root@v100 slurm]# cat gres.conf
Name=gpu File=/dev/nvidia0 Cores=[0-17]
Name=gpu File=/dev/nvidia1 Cores=[0-17]
Name=gpu File=/dev/nvidia2 Cores=[0-17]
Name=gpu File=/dev/nvidia3 Cores=[0-17]
Name=gpu File=/dev/nvidia4 Cores=[18-35]
Name=gpu File=/dev/nvidia5 Cores=[18-35]
Name=gpu File=/dev/nvidia6 Cores=[18-35]
Name=gpu File=/dev/nvidia7 Cores=[18-35]
```

```
[root@v100 slurm]# nvidia-smi topo -m
        CPU Affinity    NUMA Affinity
GPU0 ...  0-17          0
GPU1 ...  0-17          0
GPU2 ...  0-17          0
GPU3 ...  0-17          0
GPU4 ...  18-35         1
GPU5 ...  18-35         1
GPU6 ...  18-35         1
GPU7 ...  18-35         1
```

- CPU Isolation
  - Reserved for OS, Parallel File System, Monitoring etc.

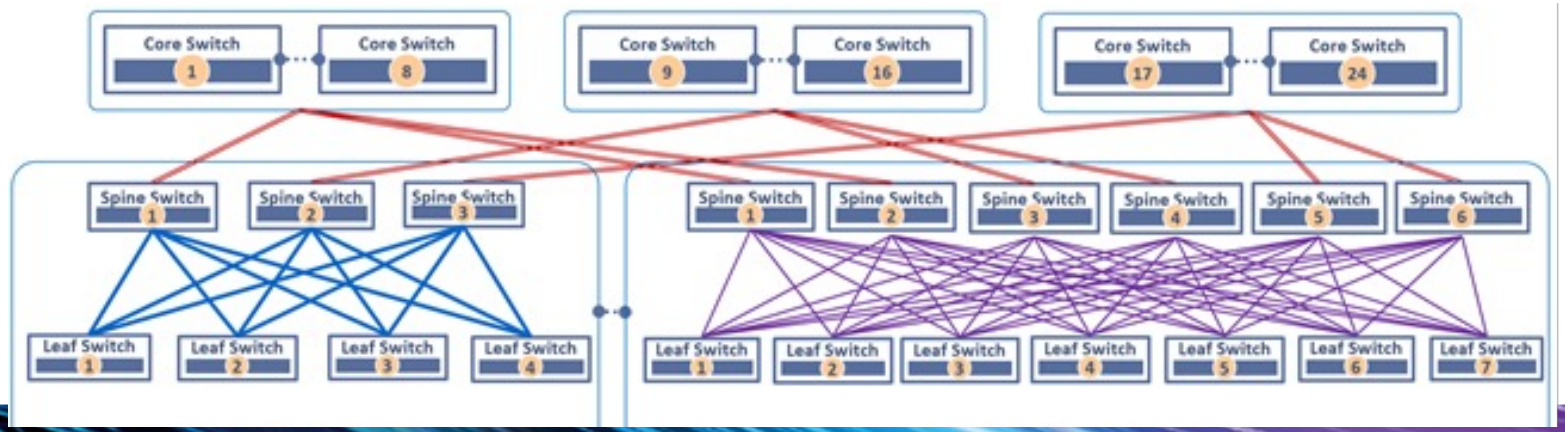
```
[root@v100 slurm]# cat slurm.conf
…
NodeName=v100 Gres=gpu:8 Sockets=2 CoresPerSocket=18 ThreadsPerCore=1 CpuSpecList=17,35 …
…
```

HPC – Infiniband Topology

- Defining the leaf switches and nodes are important.

```
[root@v100 slurm]# cat topology.conf
…
SwitchName=IBISL12   Switches=IBISL12[1-3]
SwitchName=IBISL121  Switches=IBLF120[1-2],IBLF1207
SwitchName=IBISL122  Switches=IBLF120[3-4],IBLF1207
SwitchName=IBISL123  Switches=IBLF120[5-6],IBLF1207
SwitchName=IBLF1201  Nodes=gn12[01-05]
SwitchName=IBLF1202  Nodes=gn12[05-09]
…
```

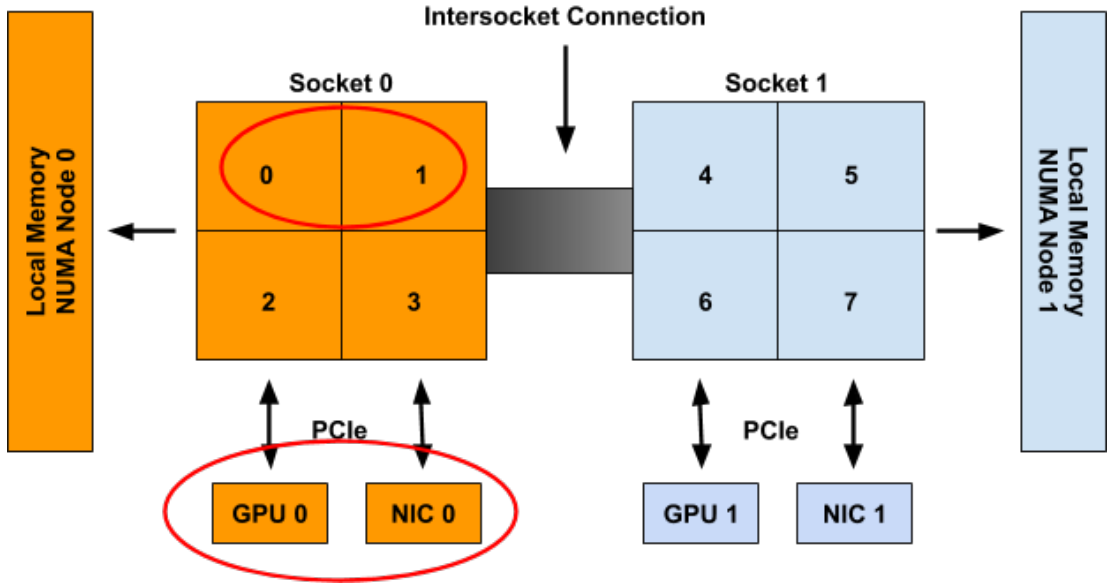
```
[root@v100 slurm]# cat slurm.conf
…
TopologyPlugin=topology/tree
TopologyParam=TopoOptional
…
[root@v100 slurm]# $ srun "env"
…
5: SLURM_TOPOLOGY_ADDR=IBISL12.IBISL121.IBLF1201.gn1201
5: SLURM_TOPOLOGY_ADDR_PATTERN=switch.switch.switch.node
```



Kubernetes – NUMA

- GPU-CPU Affinity

```
[root@v100 kubernetes]# cat config.yaml
…
featureGates:
  CPUManager: true
cpuManagerPolicy: static
cpuManagerReconcilePeriod: 5s
topologyManagerPolicy: best-effort
…
kubeReserved:
  cpu: 500m
…
```



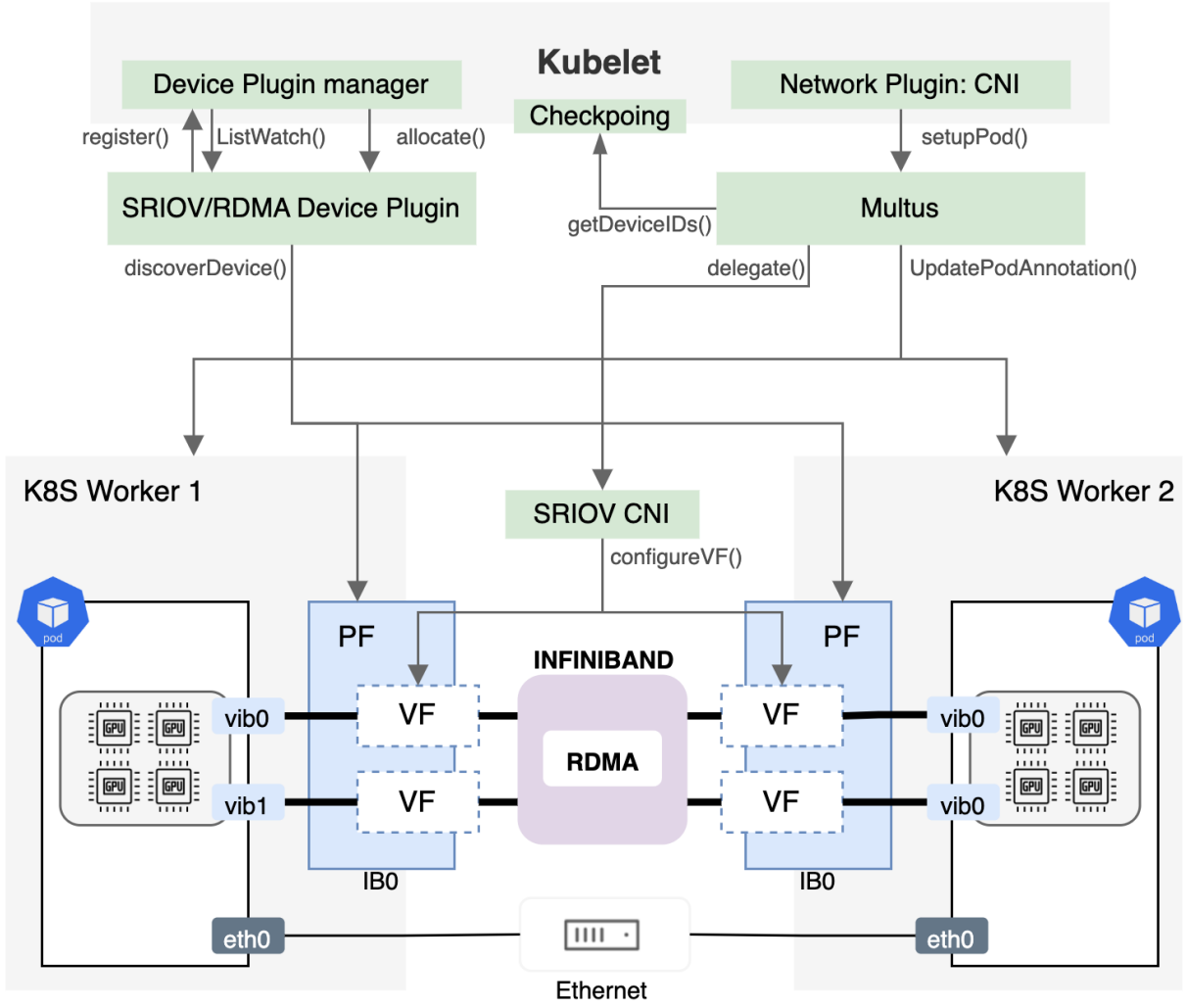
Ref: https://kubernetes.io/blog/2020/04/01/kubernetes-1-18-feature-topoloy-manager-beta/

K8S Over IB (SR-IOV)

```
kind: NetworkAttachmentDefinition
metadata:
  annotations:
    k8s.v1.cni.cncf.io/resourceName: twcc.ib/mlnx_sriov_ib
spec:
  config: '{
      "type": "ib-sriov",
      "ipam": {
        "type": "whereabouts",
        "rdmaIsolation": true,
```

```
kind: ConfigMap
"resourceList": [{
        "resourceName"; "mlnx_sriov_ib",
         "selectors": {
              "pfNames": ["ib0", "ib1", "ib2", "ib3"],
              "LinkTypes": ["infiniband"],
              "isRdma": true,
              "devices": ["1018"]
```

```
template:
    metadata:
      annotations:
        k8s.v1.cni.cncf.io/networks: sriov-conf-0@vib0
 resources:
          requests:
            twcc.ib/mlnx_sriov_ib: 1
```



Openstack – NUMA

- /etc/nova/nova.conf

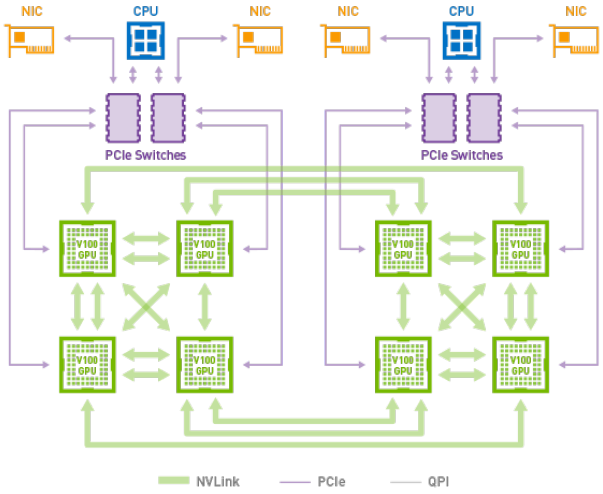
```
vcpu_pin_set=0-16,17-34
enabled_filters=<...>,NUMATopologyFilter
```

- GPU-CPU Affinity

```
openstack flavor create --disk 100 --vcpus 14 --ram 186368 \
--property aggregate_instance_extra_specs:pinned='true' \
--property hw:cpu_policy='dedicated' \
--property pci_passthrough:alias='V100:2' \
--property hw:numa_nodes=2 \
<flavour-name>
```

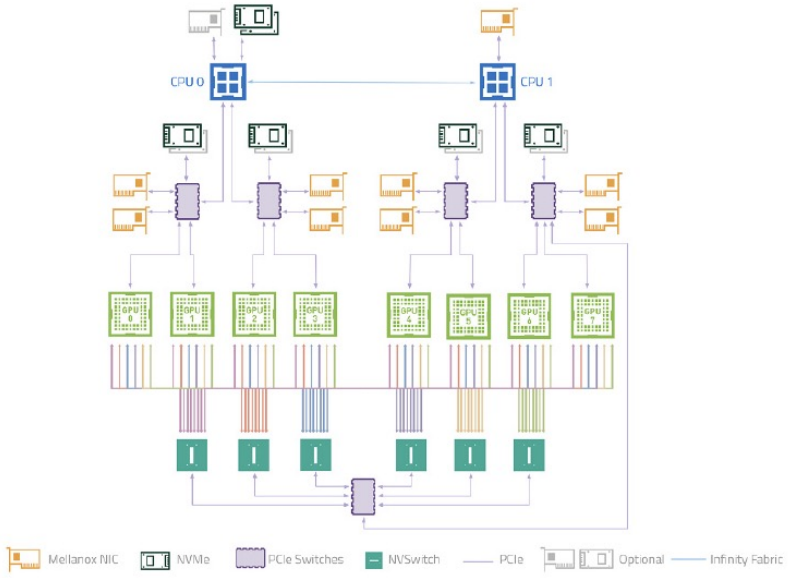
NVIDIA GPU Topology

V100 GPU



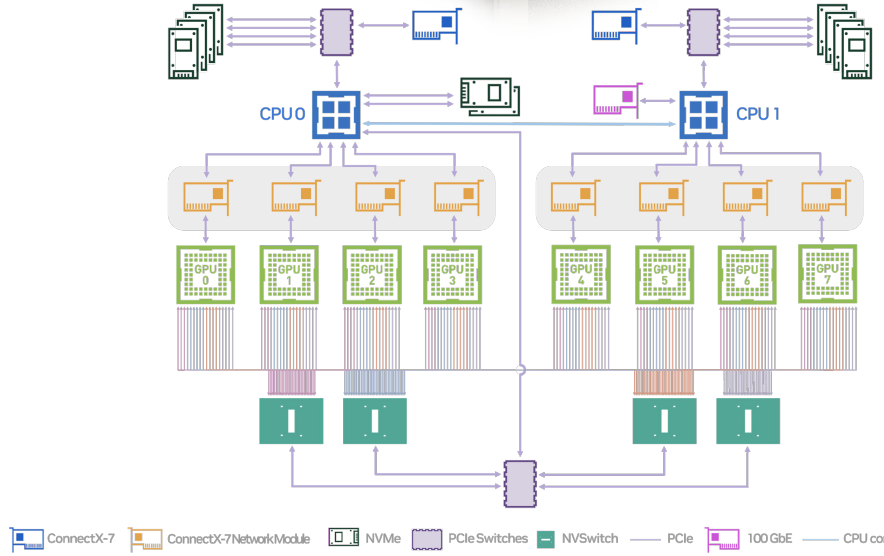
GPU Interconnect up to 300GB/s

A100 GPU



GPU Interconnect up to 600GB/s

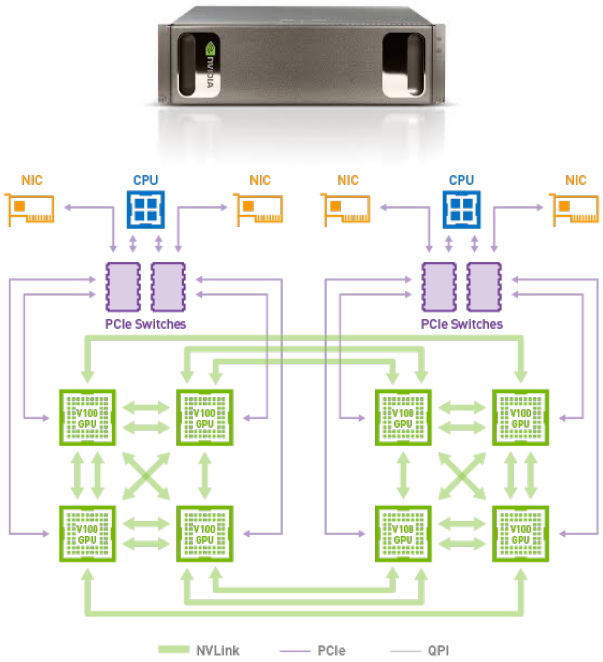
H100 GPU



GPU Interconnect up to 900GB/s

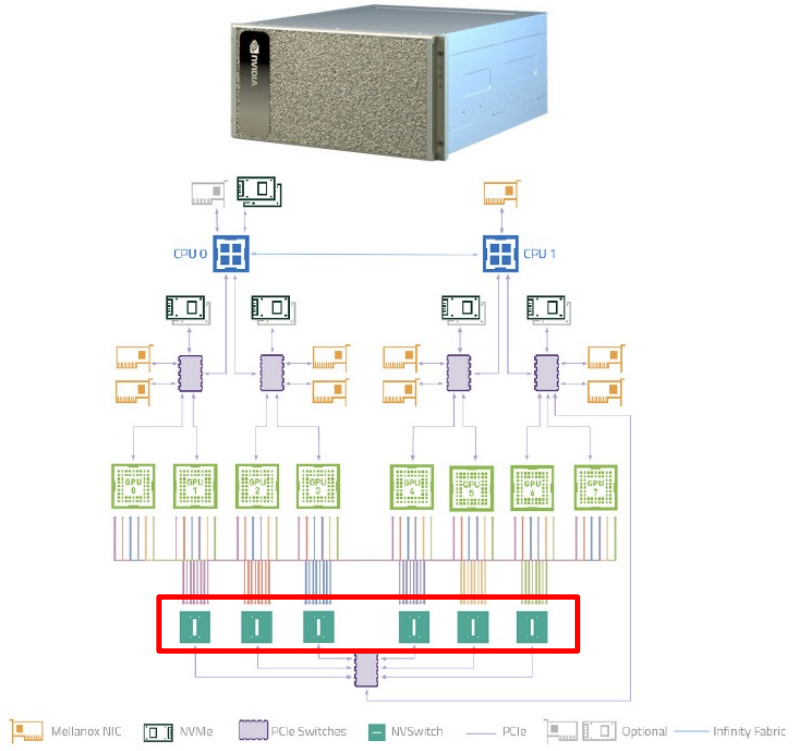
NVIDIA GPU Topology

V100 GPU



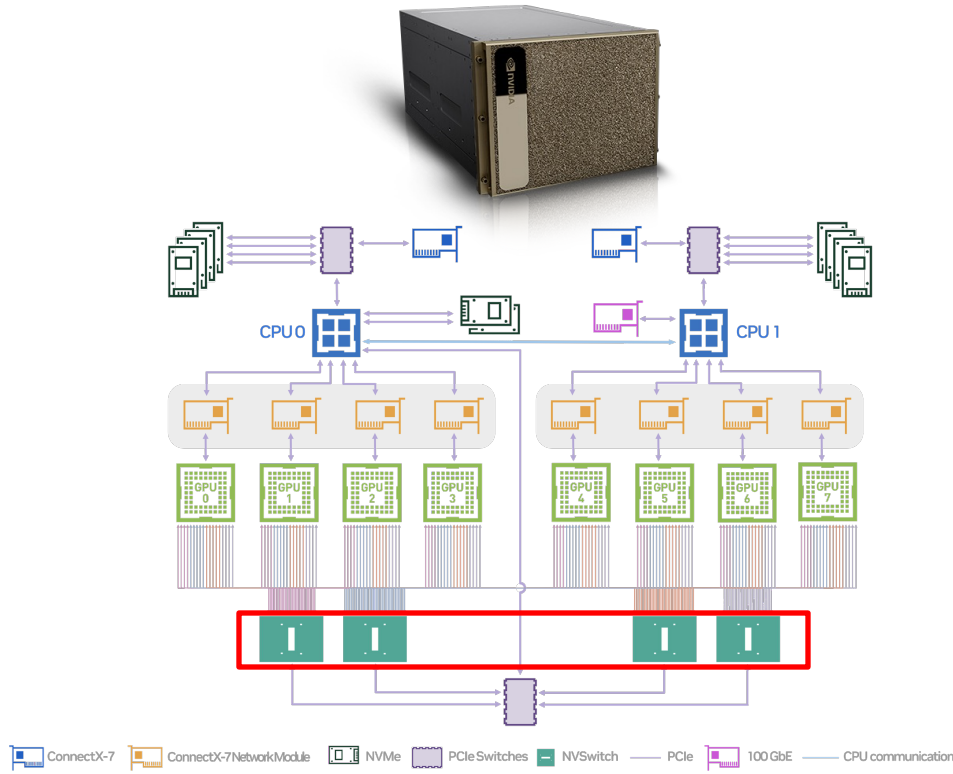
GPU Interconnect up to 300GB/s

A100 GPU



GPU Interconnect up to 600GB/s

H100 GPU



GPU Interconnect up to 900GB/s

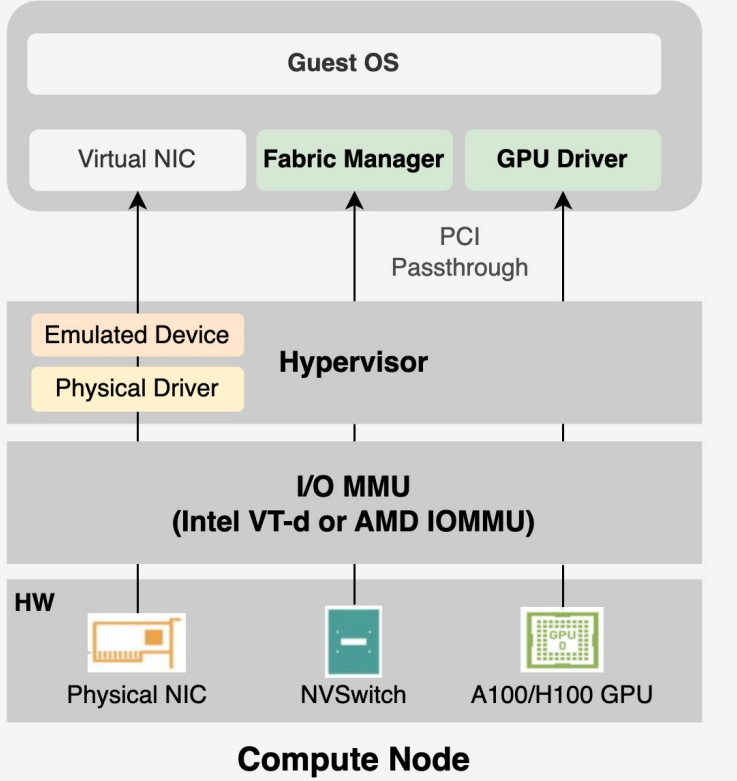
Passthrough A100/H100 to Openstack VM

- A100 NVSwitch Devices

```
c4:00.0 Bridge: NVIDIA Corporation Device 1af1 (rev a1)
c5:00.0 Bridge: NVIDIA Corporation Device 1af1 (rev a1)
c6:00.0 Bridge: NVIDIA Corporation Device 1af1 (rev a1)
c7:00.0 Bridge: NVIDIA Corporation Device 1af1 (rev a1)
c8:00.0 Bridge: NVIDIA Corporation Device 1af1 (rev a1)
c9:00.0 Bridge: NVIDIA Corporation Device 1af1 (rev a1)
```

- H100 NVSwitch Devices

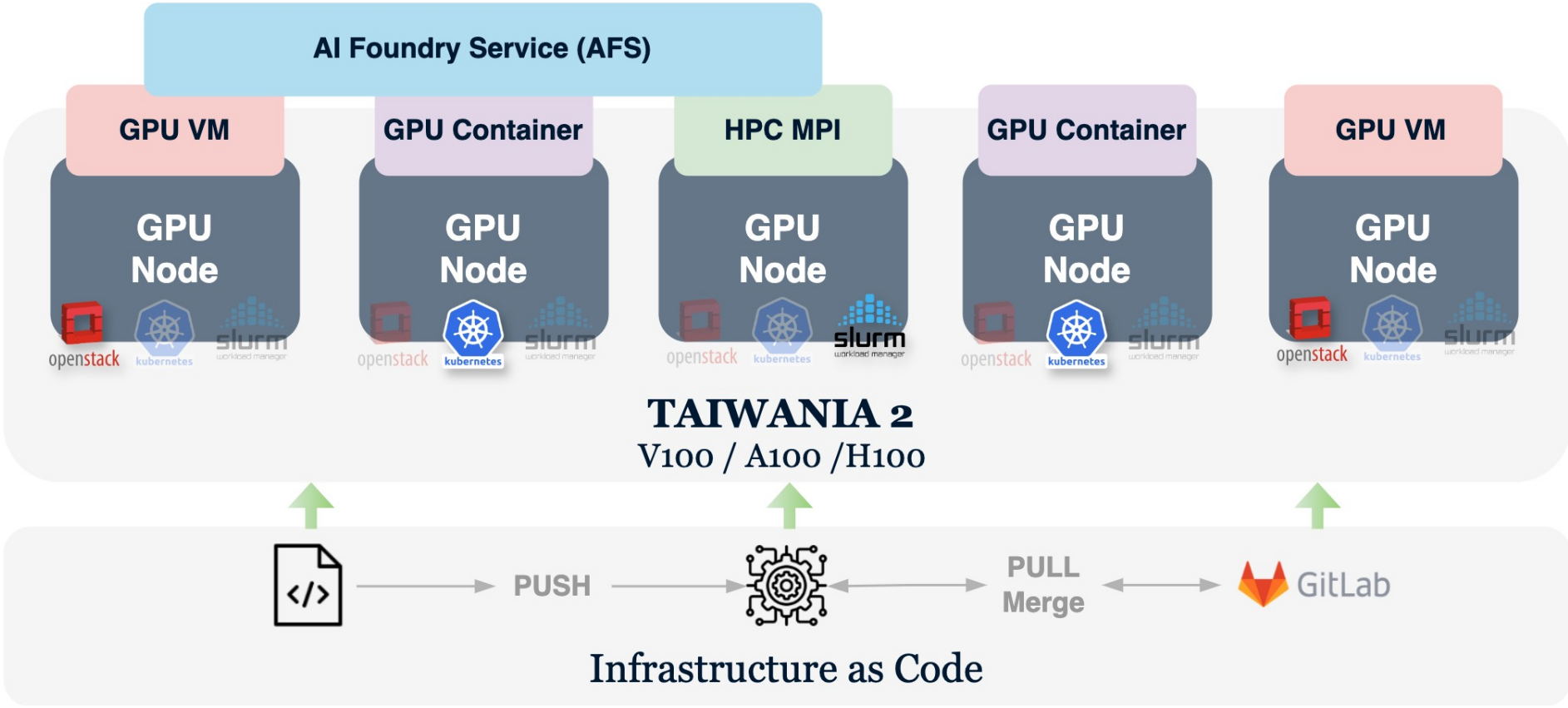
```
07:00.0 Bridge: NVIDIA Corporation Device 22a3 (rev a1)
08:00.0 Bridge: NVIDIA Corporation Device 22a3 (rev a1)
09:00.0 Bridge: NVIDIA Corporation Device 22a3 (rev a1)
0a:00.0 Bridge: NVIDIA Corporation Device 22a3 (rev a1)
```



- /etc/nova/nova.conf

```
[pci]
alias = { "vendor_id":"10de", "product_id":"20b0", "device_type":"type-PF", "name":"A100" }
alias = { "vendor_id":"10de", "product_id":"1af1", "device_type":"type-PCI", "name":"NVSWITCH" }
[pci]
alias = { "vendor_id":"10de", "product_id":"2330", "device_type":"type-PF", "name":"H100" }
alias = { "vendor_id":"10de", "product_id":"22a3", "device_type":"type-PCI", "name":"NVSWITCH" }
```

Fully Utilized GPU Resource



Rolling Upgrade and Patented Tech. (US 11,513,858 B2)

TWCC ISO And Security Certifications

全球

ISO 27001:2013
資訊安全管理系統

BS10012:2017
個人資訊管理系統

ISO 27017:2015
雲端服務資訊安全

ISO 27018:2014
個人隱私資料保護

ISO 9001
品質管理系統

ISO 50001
能源管理系統

DCOS-4
數據中心營運標準

歐美

美國健康保險流通與責任法案
HIPAA

歐盟一般資料保護規範
GDPR

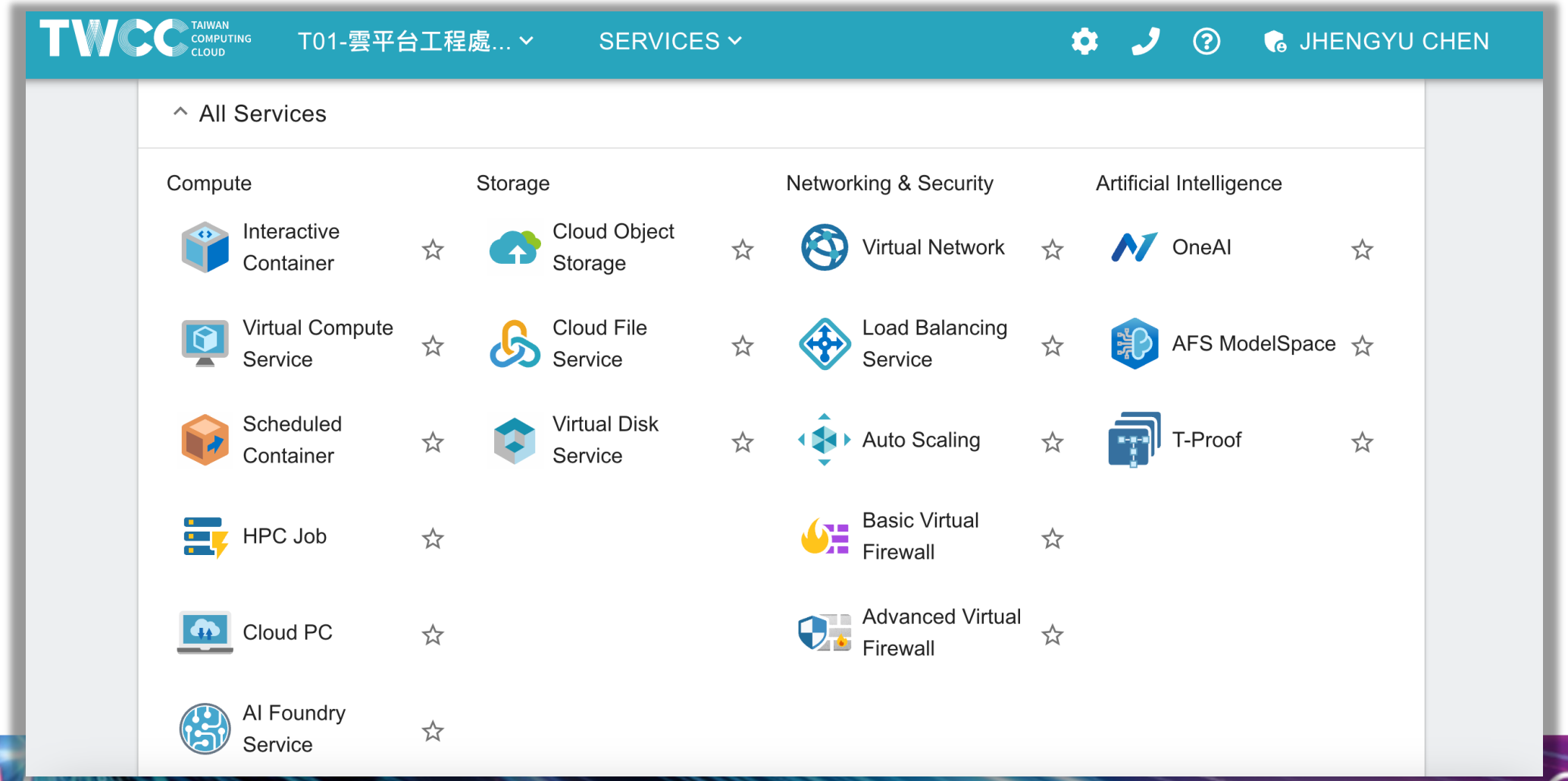
中華民國

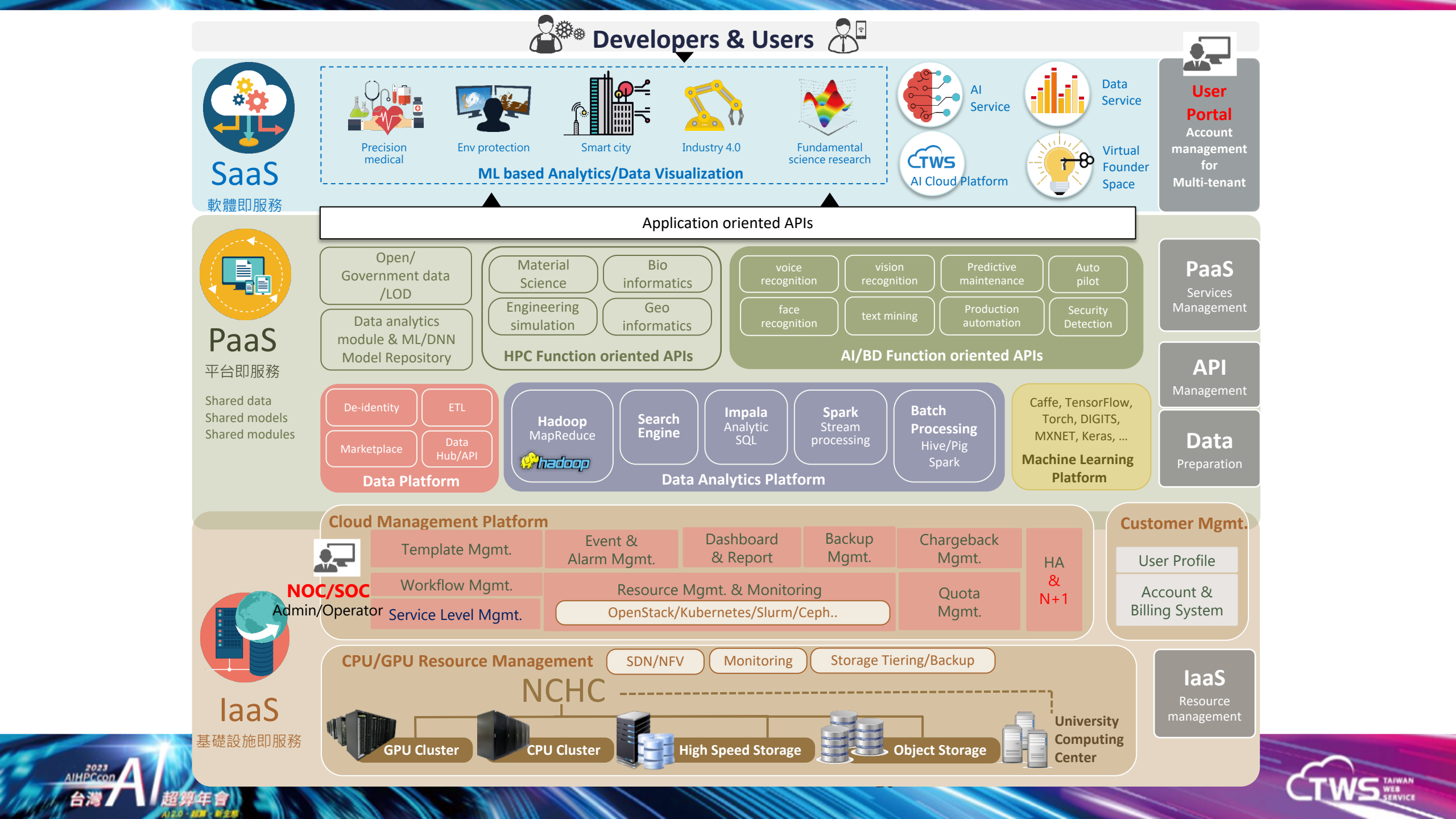
個人資料保護法
PDPA

資通安全管理法
CSMA

ISO 9001 Plus Award 2019
SGS Taiwan
QMS Transformation and Innovation Management Benchmark
變革創新管理品質典範獎

TWCC Services




Developers & Users
SaaS
軟體即服務
Precision medical
Env protection
Smart city
Industry 4.0
Fundamental science research
ML based Analytics/Data Visualization
AI Service
Data Service
AI Cloud Platform
Virtual Founder Space
User Portal
Account management for Multi-tenant
Application oriented APIs
PaaS
平台即服務
Shared data
Shared models
Shared modules
Open/ Government data /LOD
Data analytics module & ML/DNN Model Repository
Material Science
Bio informatics
Engineering simulation
Geo informatics
HPC Function oriented APIs
voice recognition
vision recognition
Predictive maintenance
Auto pilot
face recognition
text mining
Production automation
Security Detection
AI/BD Function oriented APIs
PaaS
Services Management
API
Management
Data
Preparation
De-identity
ETL
Marketplace
Data Hub/API
Data Platform
Hadoop MapReduce
Search Engine
Impala Analytic SQL
Spark Stream processing
Batch Processing Hive/Pig Spark
Data Analytics Platform
Caffe, TensorFlow, Torch, DIGITS, MXNET, Keras, ...
Machine Learning Platform
Cloud Management Platform
Template Mgmt.
Event & Alarm Mgmt.
Dashboard & Report
Backup Mgmt.
Chargeback Mgmt.
HA & N+1
Workflow Mgmt.
Service Level Mgmt.
Resource Mgmt. & Monitoring
OpenStack/Kubernetes/Slurm/Ceph..
Quota Mgmt.
Customer Mgmt.
User Profile
Account & Billing System
NOC/SOC
Admin/Operator
IaaS
基礎設施即服務
CPU/GPU Resource Management
SDN/NFV
Monitoring
Storage Tiering/Backup
NCHC
GPU Cluster
CPU Cluster
High Speed Storage
Object Storage
University Computing Center
IaaS
Resource management

Current Dev. for Day2

Generative AI

Day 0 Op

Day 1 Op

Day 2 Op

AFS

GenAI Solution for Training and inference

TWSC AFS: AI Foundry Service

IndustrialGPT Solutions

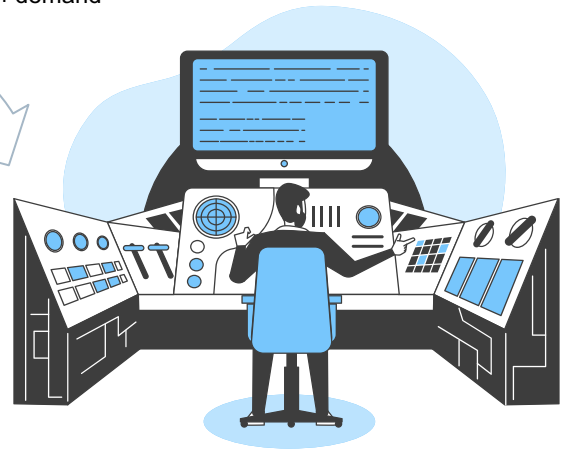
Step 1.
Upload training data

Data Collection

Platform

Provided by
TWSC

Step 2.
Deploy on Premise* / On-demand



Self Mgmt. Deploy

NOTE: Premise* deployment can be easily operated in on-premise **AFS Appliance**

01 Full Control

02 No Code

03 Formosa LLM

04 Wide Adoption

Standing on Formosa LLM
Pursue Excellence

企業專屬大語言模型優化 ▸▸▸▸▸▸▸▸▸▸▸▸▸▸▸▸▸▸▸▸▸▸▸▸▸▸▸▸▸▸ 部署上線

模型訓練資料
格式整理
(jsonl'input':'target')

Data Collection

選擇 FFM 大語言模型
進行企業專屬模型優化

Fine-tuning

模型驗證

Evaluation
Playground

模型部署與推論

Deploy

AFS 解決方案

AFS Platform 雲端訓練
企業大模型優化方案

AFS Cloud 雲端推論
企業大模型部署方案

AFS Appliance 地端推論
AFS ModelSpace 雲端推論

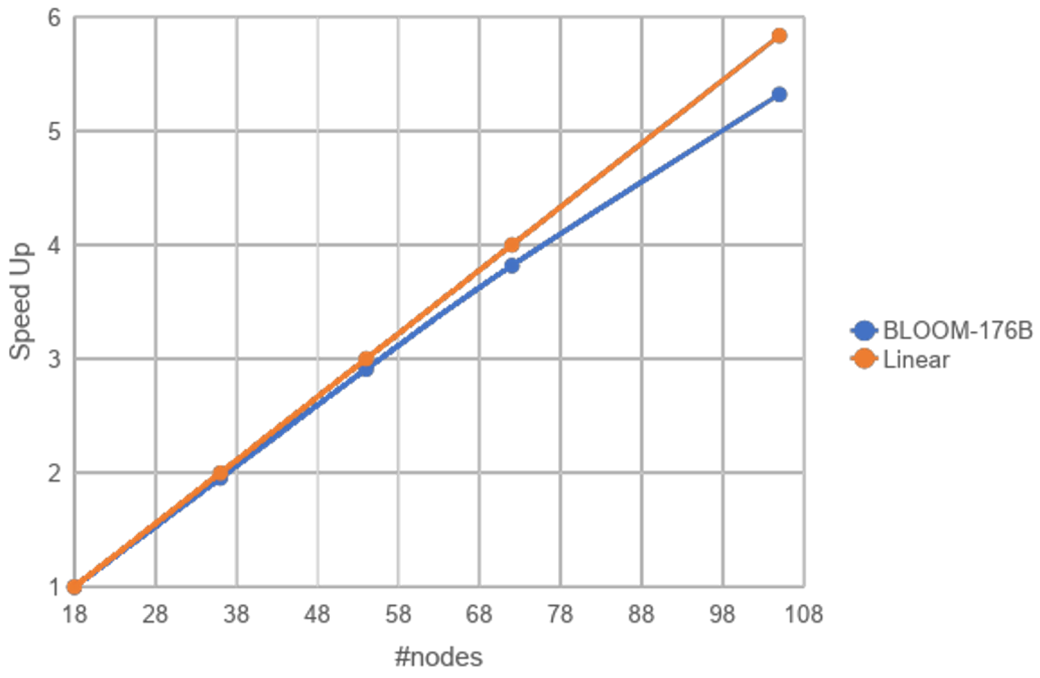
AFS ModelSpace

Utilize CCS with InfiniBand ability
Proprietary LLM fast deployment

Implementing GPT-3 level LLM in TAIWANIA 2

Due to the large model with 176 billion parameters, it is not possible to train directly on any single GPU. It requires precise model segmentation and efficient distributed training.

- Training can achieve linear acceleration.
- Training and inference of the 176 billion parameter model can be run on TWCC.



Model	Nodes	TP	PP	MBS	Training time/step (seconds)	Samples per second	TFLOPs
1.3B	1	1	1	8	26.64	19.23	58.53
1.3B	2	1	1	8	14.92	34.35	52.28
1.3B	4	1	1	8	6.73	76.05	57.89
7.1B	1	2	1	2	250.10	4.09	56.16
7.1B	4	2	1	4	65.28	15.69	53.80
7.1B	8	2	1	8	38.03	26.93	46.17
176B	18	8	18	1	777.23	2.64	53.15
176B	36	8	18	1	396.40	5.16	52.11
176B	54	8	18	1	266.30	7.68	51.66
176B	72	8	18	1	203.26	10.08	50.81
176B	105	8	35	2	145.70	14.04	48.56

Thanks!

Do you have any questions?

sales@twsc.io
https://www.twsc.io


TWS
TAIWAN
WEB
SERVICE