Weighing AI delivery between Openstack & Kubernets

Feimin Yuan (John)
F5
26 Aug 2025

Agenda

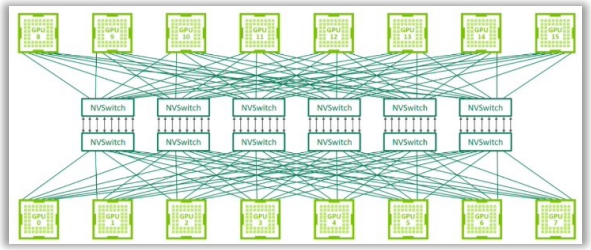
Where we are at running GPU infra

Opensource infra becomes the norm

Network evolve with accelerated stack (DPU)

Where we are at running GPU infra

GPU infra scales are vastly different



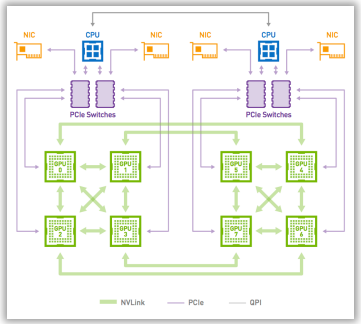
Datacenter scale

HPC or K8s on bare-metal

Low-latency E-W full mesh network

Close-source technology

Astronomically high cost



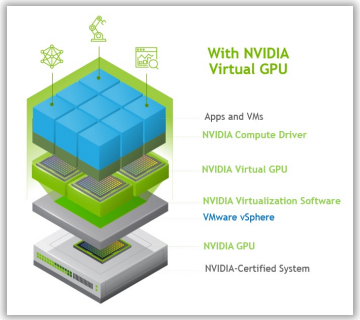
Cluster scale

Bare-metal, K8s on bare-metal

Multiple host high-speed network

Mixed close-source and open-source technology

High in investment & operation



Host scale

VMs, K8s on VMs

Often within single host network

Prefer Open-source technology

Moderate in investment & Operation

BECAUSE

GPU infra are tilting more to inferencing

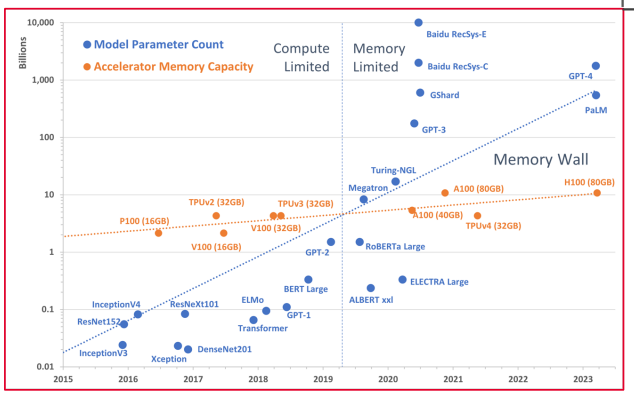
Foundation Large Models

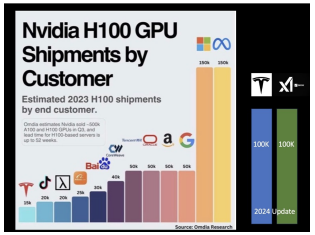
Meta Llama 2 70B

OpenAI ChatGPT 4.0

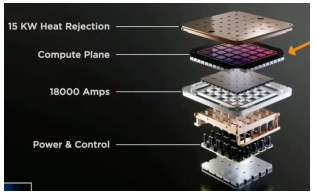
Mainstream LLMs see in its training size will slow down, although the current hunger for GPU cluster scale is still massive.

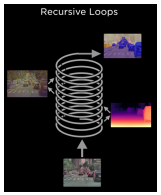
Focus is on approximating human abilities and becoming next generation of internet platforms





Industry specific models

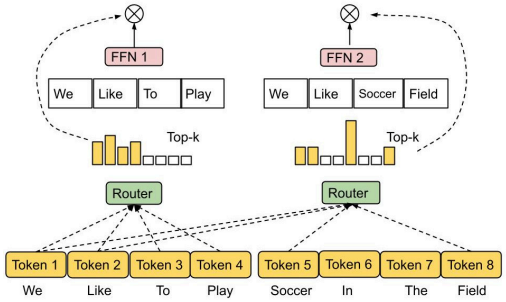




Industry models focus on delivering and continuous delivering services and functions in a close feedback loop with both models and datasets.

e.g. Tesla FSD trained on Dojo architecture requires continuous new data retraining, model evolution.

Open source models lowered access barrier for more enterprise to develop customized human facing models

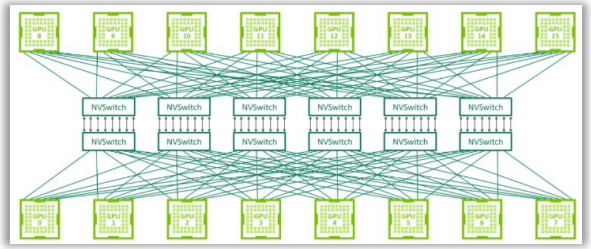


MISTRAL AI_ 8X22

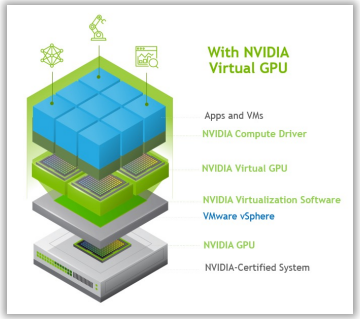
Open Weight GPT-OSS Models

More open source models

Change from cluster scale to host scale

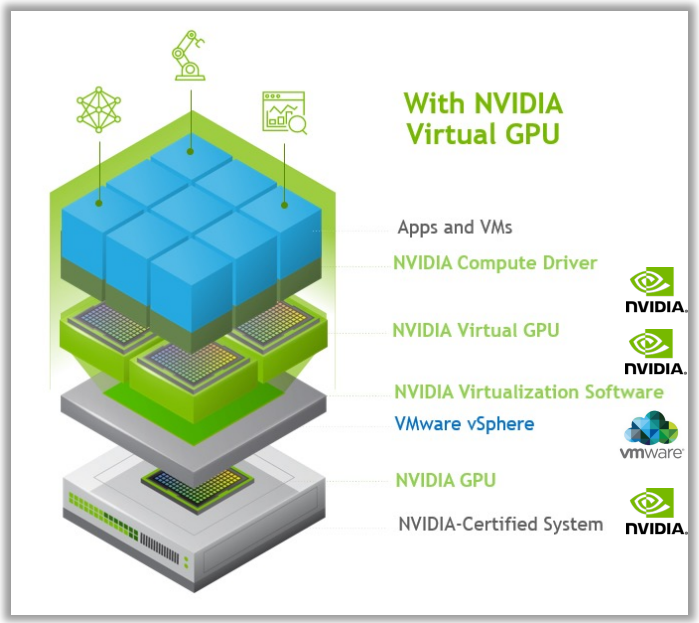


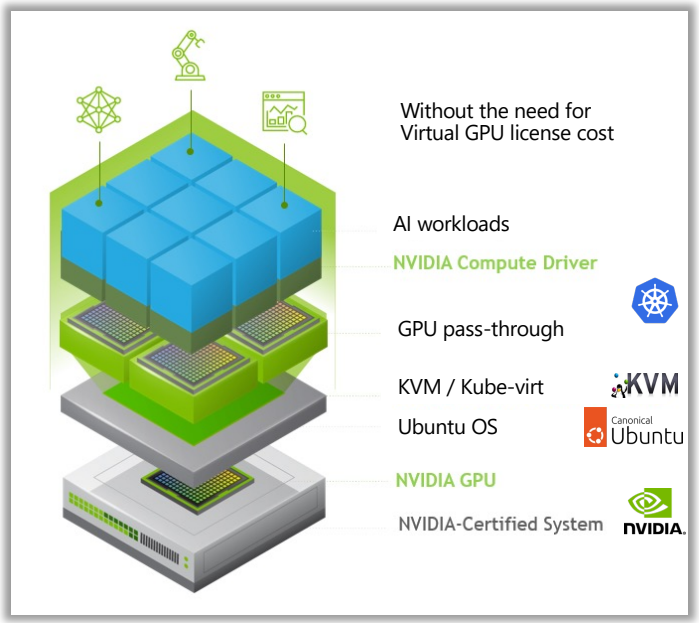
AI LLM provider	Max GPUs/Instance	GPUs Type
OpenAI GPT	10k GPUs	Nvidia H100



AI cloud provider	Max GPUs/Instance	GPUs Type
Provider 1	8 GPUs	Nvidia H100, A100
Provider 2	8 GPUs	Nvidia H100, A100
Provider 3	16 GPUs	Nvidia A100 (A2)
Provider 4	8 GPUs	Nvidia H100, A100
Provider 5	4 GPUs	Nvidia Tesla GPUs

Reduction of scale brings more opp for open infrastructure





Opensource infra becomes the norm

Host scale brings more rational investment in AI infra

OBJECTIVE 1

Drive down cost

Leverage open-source platforms and tools

Maximize the use of server resource/components

OBJECTIVE 2

Reduce Complexity

Simplify GPU infra software stack

Adopt proven technologies

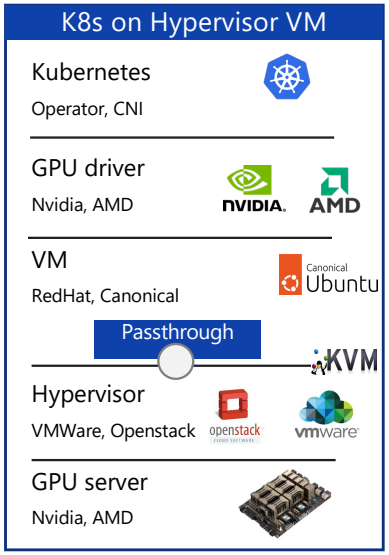
OBJECTIVE 3

Deliver Security

Extend isolation and control into Host tenant Network(not just DC network)

Remove unnecessary hops(CPU, host memory)

Choices of infra are not made but tried

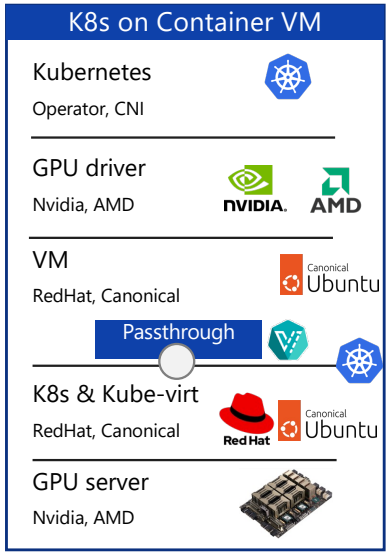


1

K8s on Hypervisor VM

Openstack manages hardware resource and GPU pass through

K8s manages AI network and workloads

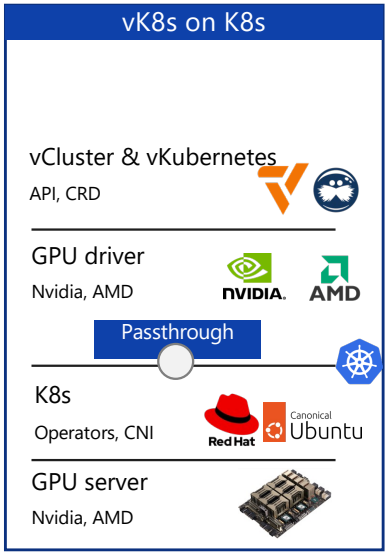


2

K8s on Container VM

K8s manages hardware resource and GPU with kube-virt

Tenant K8s inside VM manages AI network and workloads



3

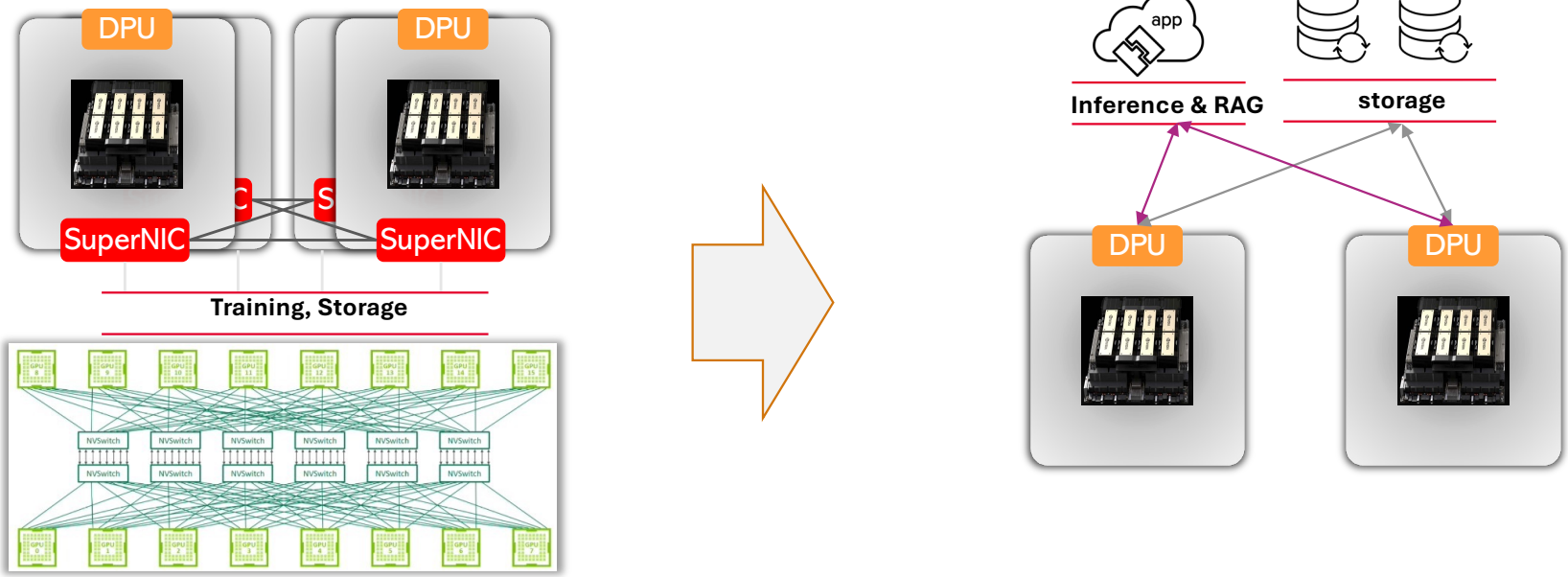
vK8s on K8s

K8s manages hardware resource, GPU operator and network

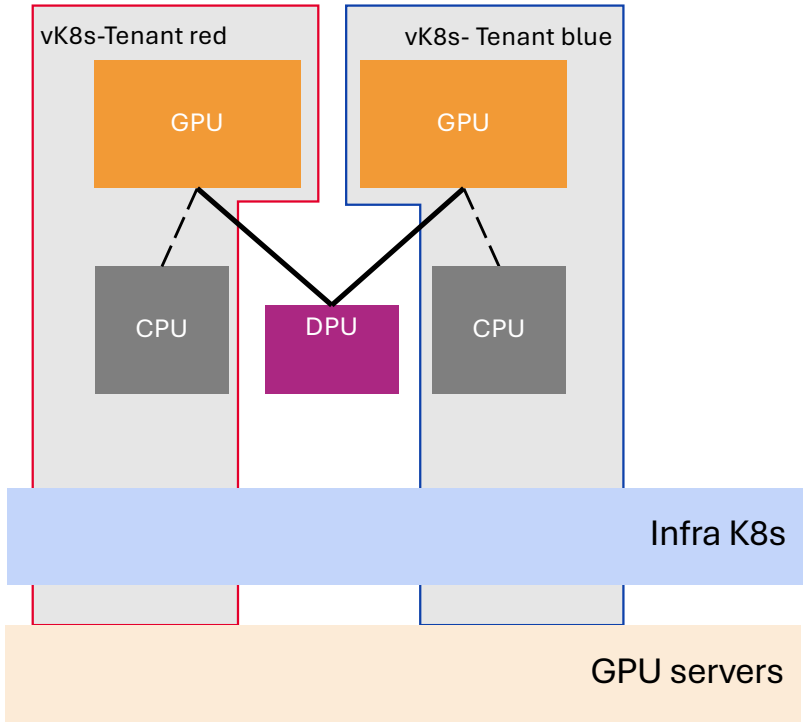
vK8s through customized API endpoint manages AI workloads

Network evolve with accelerated stack (DPU)

Network stack evolves from cluster scale to host scale



Tenancy is easy with opensource tools



- Tenancy with k8s in VMs are easy to achieve.
- But network delivery and security is not.
- DPU as the new pathway for accelerated AI network becomes the key point of control

Putting the infra stack to test

Server-1

- OS: Ubuntu 22.04

  NVIDIA DOCA

  DPU Firmware

  F5 BNK GA Software
- Kubernetes: 1.30
- Calico CNI: 3.29.1
- Nginx: 1.29.0

Server-2

- OS: Ubuntu 22.04
- Kubernetes: 1.30
- Calico CNI: 3.29.1
- Nginx: 1.29.0

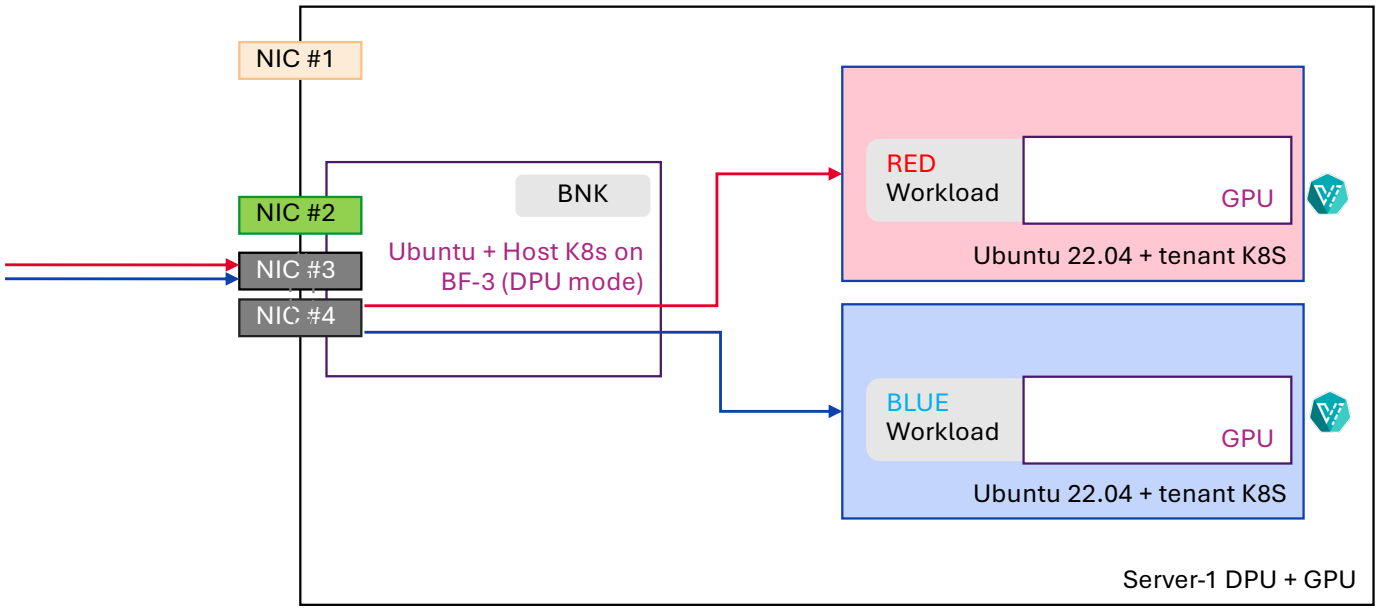
Server-3

- OS: Ubuntu 22.04
- Kubernetes: 1.30
- Cilium CNI: 1.17.6
- Nginx: 1.29.0

Server Specification List

Server Type	Specifications
Lenovo SR675	CPU : AMD EPYC 9334 32C x2 Memory : 2TB Disk : 960GBx2 & 7.68TBx2 GPU : NVIDIA L40S DPU : NVIDIA BlueField-3 B3220
Lenovo SR675	CPU : AMD EPYC 9334 32C x2 Memory : 2TB Disk : 960GBx2 & 7.68TBx2 Network : - ConnectX-7 Infiniband GPU : NVIDIA L40S
Lenovo SR675	CPU : AMD EPYC 9334 32C x2 Memory : 2TB Disk : 960GBx2 & 7.68TBx2 Network : - ConnectX-7 Infiniband GPU : NVIDIA L40S

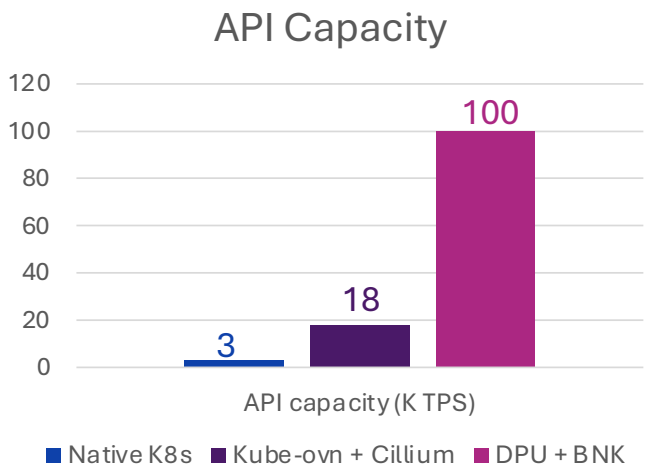
Deliver AI workloads fast and securely

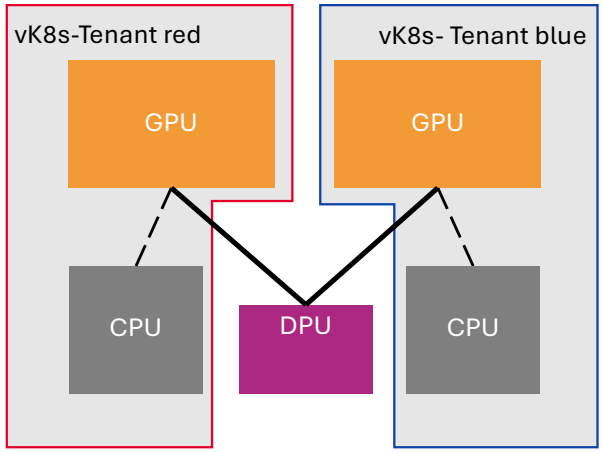


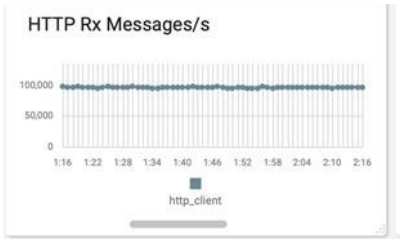
Tenant Network 1

Tenant Network 2

AI API delivery metrics - Capacity

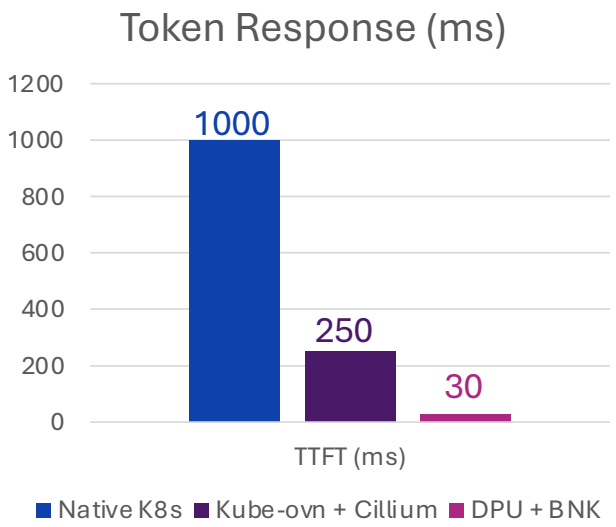


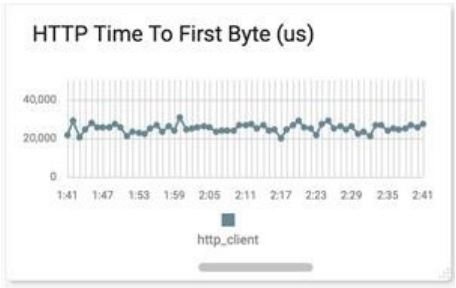




- An increase of **30+** times helps delivery of AI API at large scale on a single function/platform.
- Maintaining same performance whether it is a large tenant or multiple small scale teanants.

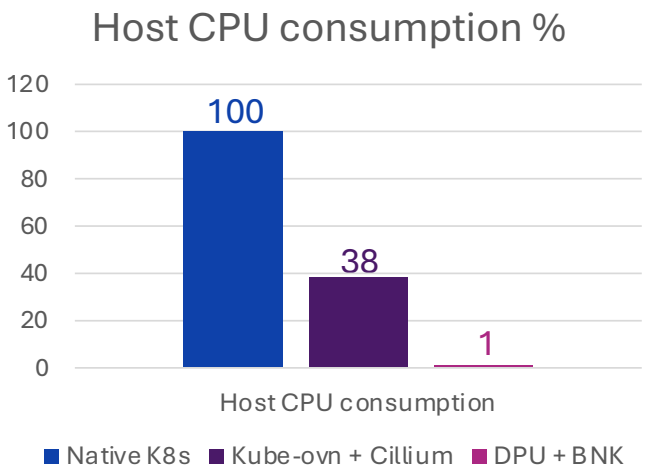
AI API delivery metrics – Token Response

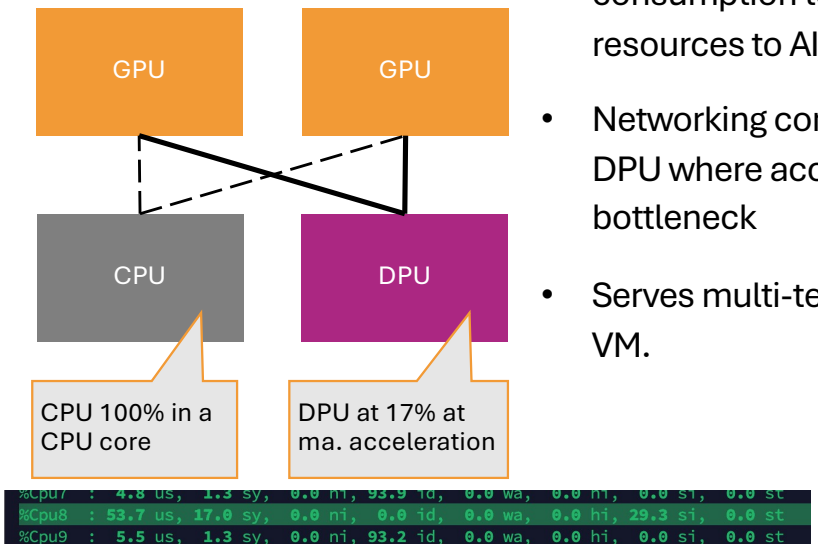




- A reduction of **3000%** helps delivery of Token response faster to user.
- Allows AI systems to quickly return inference and continue to next batch of data.
- Impacts overall LLM user experience significantly.

AI API delivery metrics – Host CPU consumption





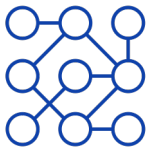
- A reduction of **99%** in Host CPU consumption to allocate server resources to AI workloads.
- Networking completely offloaded to DPU where acceleration removing bottleneck
- Serves multi-tenancy with GPU, in VM.

F5 BIG-IP Next for Kubernetes

Deployed on NVIDIA BlueField-3 DPUs

F5 and NVIDIA deliver high-performance networking and security for AI factories, simplifying operations through multi-tenancy and optimized GPU utilization.

Solving the unique challenges AI workloads bring to data centers



Traffic Management

Intelligent Load-balancing

Ingress / Egress

High-Performance

Scale and Flexibility

Seamless scale-out

Multi-tenancy

Security

Firewall with DDoS mitigation

Intrusion Prevention

Encryption and Certificates

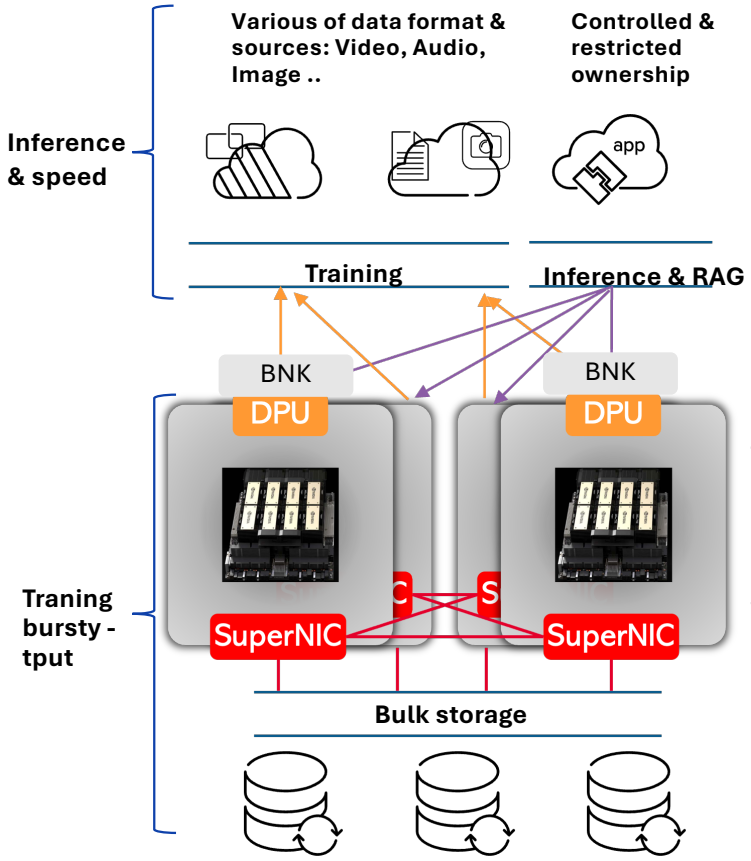
Observability

Traffic analytics and capture

Unified traffic management and security visibility

A practice book for getting AI infra ready for inferencing

Data & model owners want SECURITY, EFFICIENCY & RESILIENCY from AI cloud.



Security

Data security requirement is paramount for tenants.

- Enterprise
- Industrial
- Government departments

Secure E2E.

- Secure isolated store required.
- Restricted access.

Efficiency

Different data format & granularity.

- Image
- Text
- Audio
- Video
- Raw sensor stream

Bulk transfer.

- Fast & simultaneous transfer into server storage(cache) and GPU memory.

Resiliency

Reliable inteligent traffic loading agent

- Restoring network context of training inference jobs instantaneously within and across AI servers.
- Parallel data loading to host storage(cache) during training running process.

N-S

Shared compute

Shared network

Shared storage

E-W

Owning timeshare of GPU centric resources

shared flat network (IB, RoCE)

speak file and not object

Multi-tenant

Enterprise & industry & government

Foundation model & Platform