Affordable ML Platform

ML Platform on affordable hardwares

Zhiya

Agenda

- What’s an affordable ML Platform?
- Who needs this ML Platform?
- Key components (Which part is necessary and why)
- Key technical points
  - Scalable Container Environment
  - GPU Sharing

What's an affordable ML Platform?

- ML Platform
  - Manage the pipeline of experiment, development, deployment
- Affordable ML Platform
  - With single or few GPUs
  - All about sharing

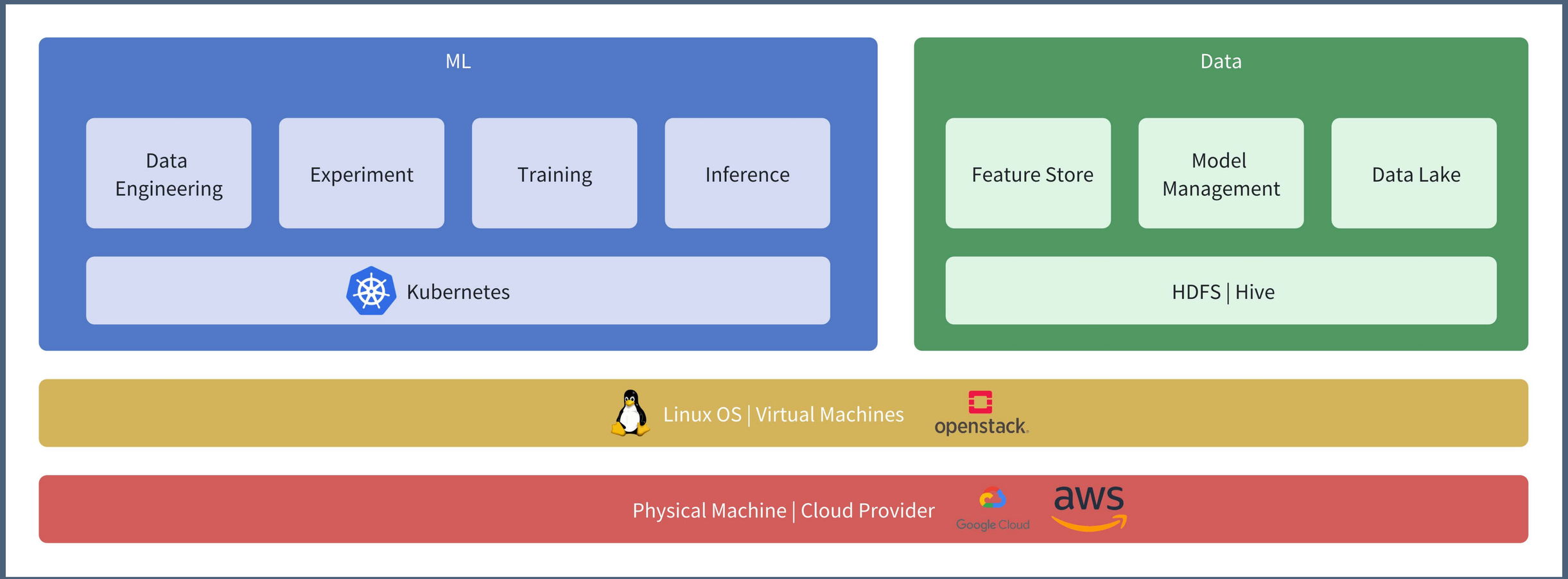
Who needs this ML Platform?

- GPUs are expensive
- GPUs are idle out of working hours
- GPUs are idle during working hours
- GPUs are too powerful

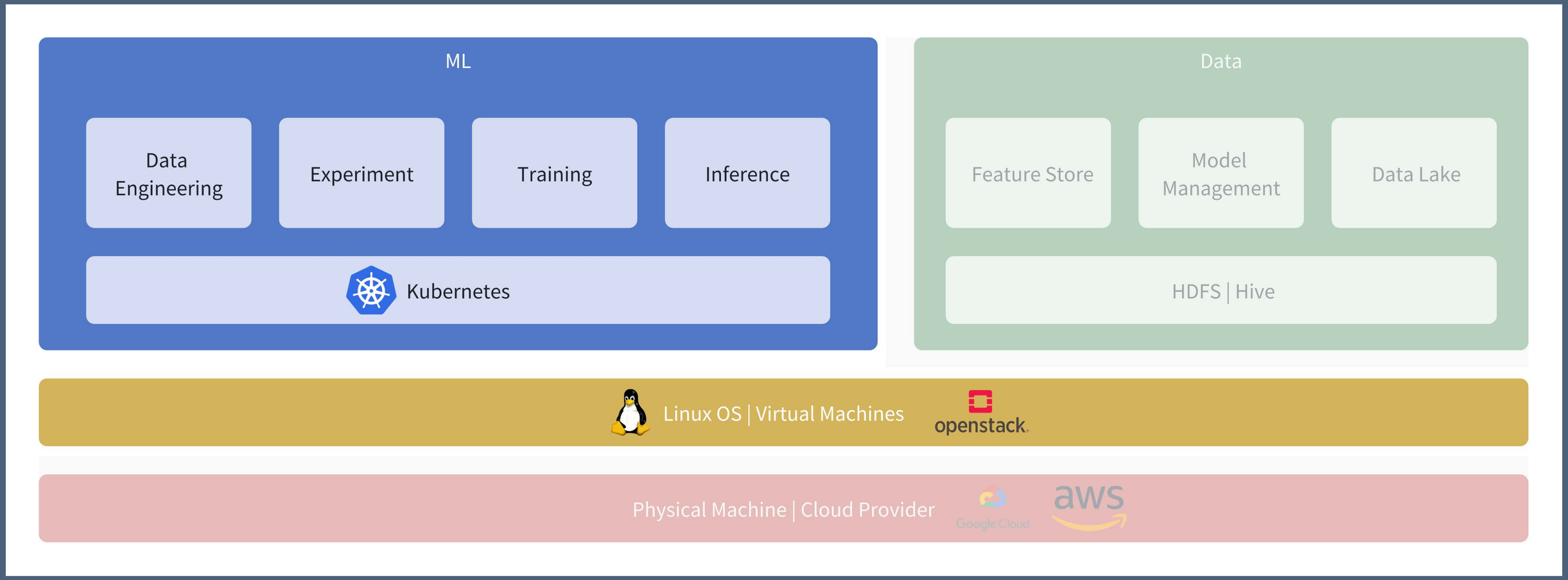
Who needs this ML Platform?

- Startups and Small Businesses
- Educational Institutions
- Non-Profit Organizations
- Freelancers and Consultants

Key Components



Key Components



Key Tech Points

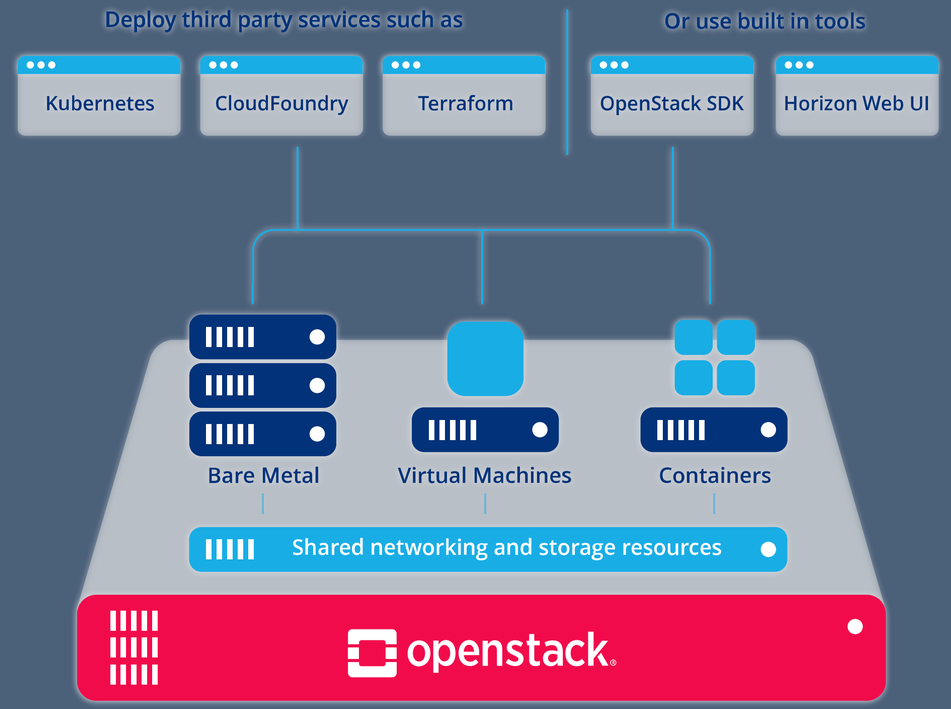
Scalable Container Environment

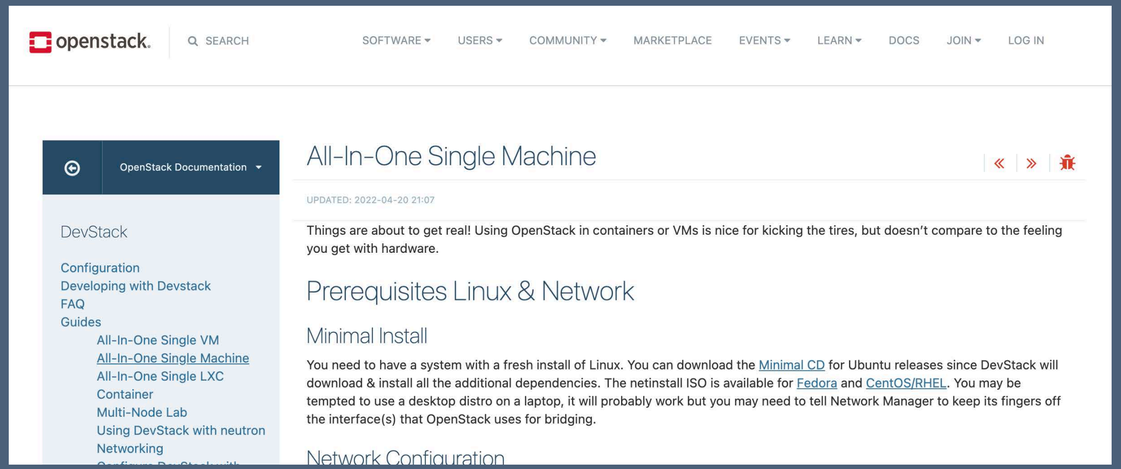
- Affordable business scenes
  - Educational Institutions: Single machine with few GPU card
  - Startups and Small Businesses: Few PCs with GPU
  - Freelancers and Consultants: PC with 1 GPU
- Conflicts
  - Kubernetes minimal nodes = 3

Key Tech Points

Scalable Container Environment

- Introduce Openstack
  - Affordability: Single machine compatibility
  - Compatibility: Mix of VM/Container/Bare Metal
  - Scalability: From single machine to multi-node cluster
  - Flexibility: Easy to remove

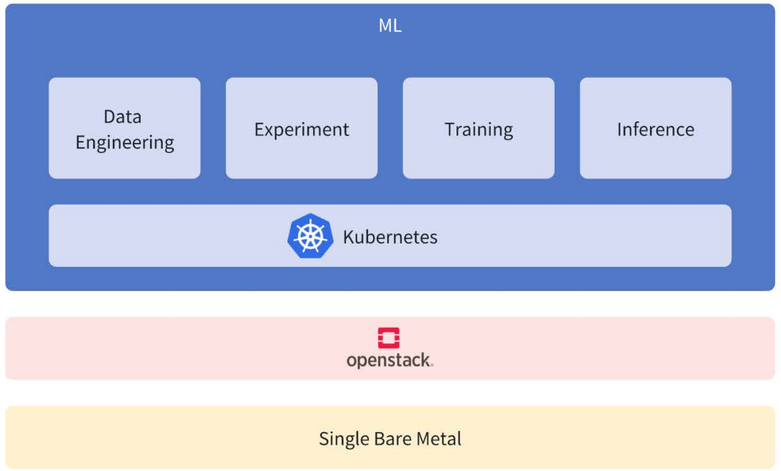




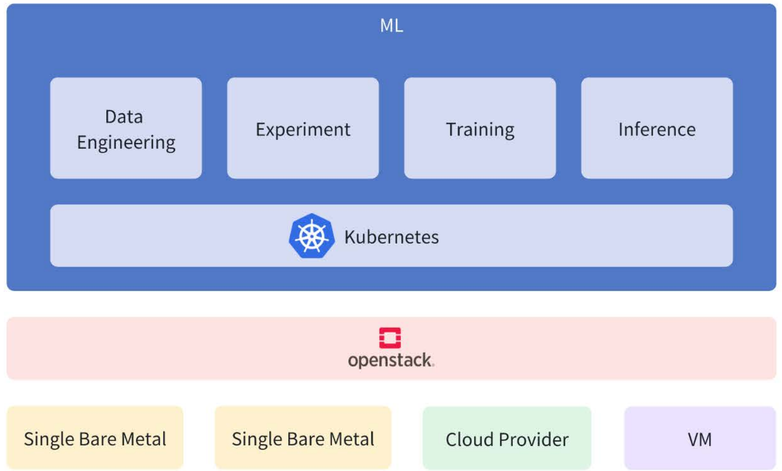
Key Tech Points

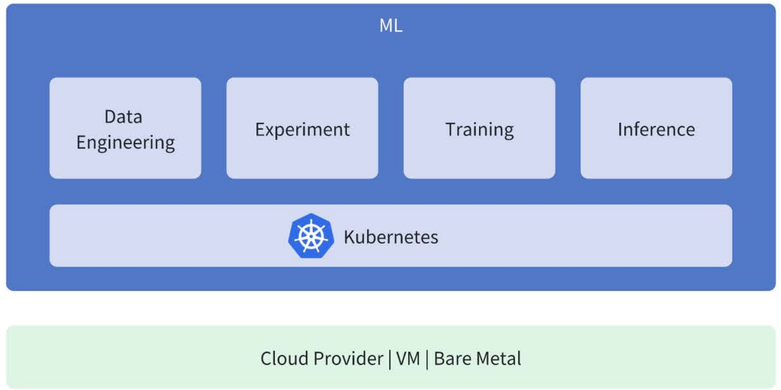
Scalable Container Environment

Present



Future





Key Tech Points

GPU Sharing

- Official Solution
  - Multi-instance GPU: Inference
  - GPU time-sharing: Traning
  - Nvidia MPS: Experiment
- Tencent TKE GaiaGPU

	Multi-instance GPU	GPU time-sharing	NVIDIA MPS
General	Parallel GPU sharing among containers	Rapid context switching.	Parallel GPU sharing among containers
Isolation	A single GPU is divided in up to seven slices and each container on the same physical GPU has dedicated compute, memory, and bandwidth. Therefore, a container in a partition has a predictable throughput and latency even when other containers saturate other partitions.	Each container accesses the full capacity of the underlying physical GPU by doing context switching between processes running on a GPU.	NVIDIA MPS has limited resource isolation, but gains more flexibility in other dimensions, for example GPU types and max shared units, which simplify resource allocation.
Suitable for these workloads	Recommended for workloads running in parallel and that need certain resiliency and QoS.	GPU time-sharing is optimal for scenarios where full isolation and continuous GPU access might not be necessary, for example, when multiple users test or prototype workloads without idling costly GPUs.	Recommended for batch processing for small jobs because MPS maximizes the throughput and concurrent use of a GPU. MPS allows batch jobs to efficiently process in parallel for small to medium sized workloads..

Key Tech Points

GPU Sharing

- Multi-instance GPU
  - Partitioned into up to seven separate GPU Instances
  - MIG allows multiple vGPUs to run in parallel on a single GPU
  - High performance professional GPU only

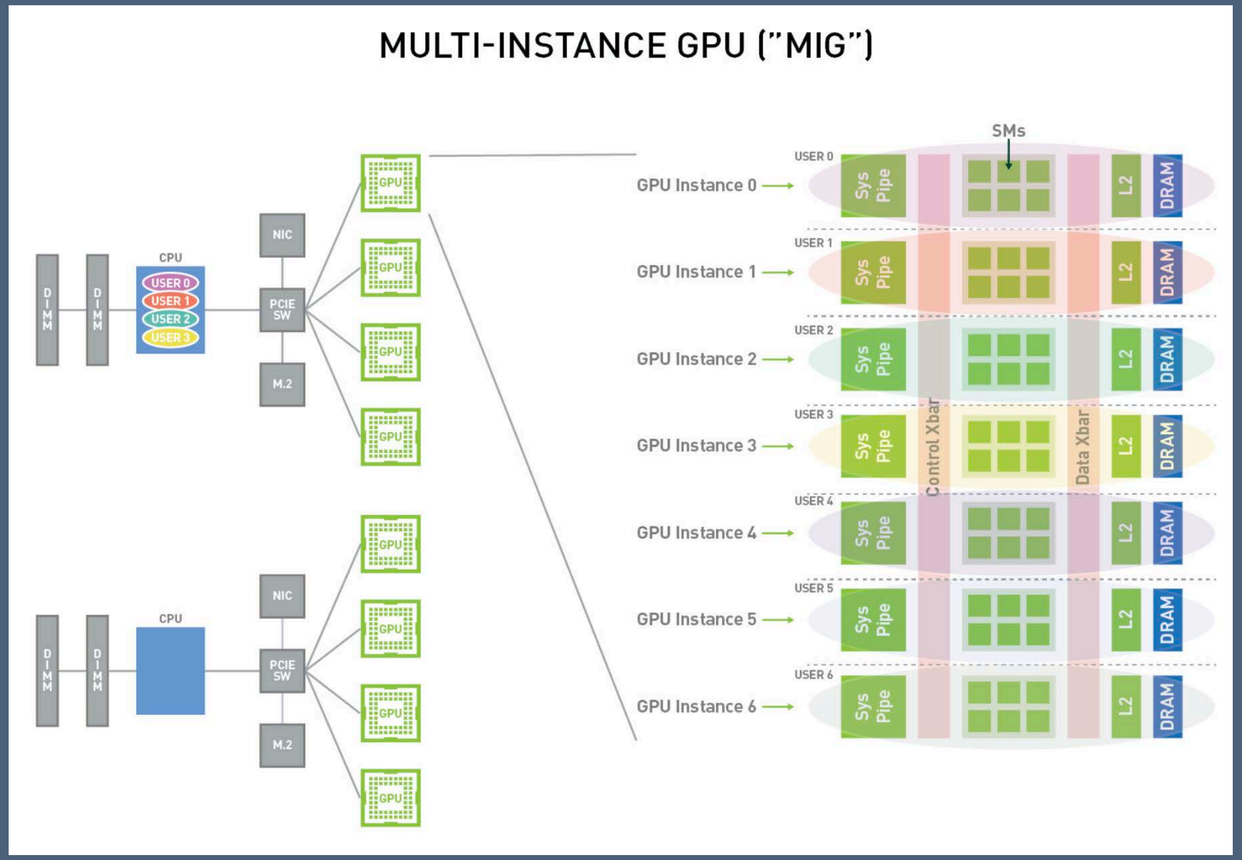


Table 1. Supported GPU Products

Product	Architecture	Microarchitecture	Compute Capability	Memory Size	Max Number of Instances
H100-SXM5	Hopper	GH100	9.0	80GB	7
H100-PCIE	Hopper	GH100	9.0	80GB	7
H100-SXM5	Hopper	GH100	9.0	94GB	7
H100-PCIE	Hopper	GH100	9.0	94GB	7
H100 on GH200	Hopper	GH100	9.0	96GB	7
A100-SXM4	NVIDIA Ampere	GA100	8.0	40GB	7
A100-SXM4	NVIDIA Ampere	GA100	8.0	80GB	7
A100-PCIE	NVIDIA Ampere	GA100	8.0	40GB	7
A100-PCIE	NVIDIA Ampere	GA100	8.0	80GB	7
A30	NVIDIA Ampere	GA100	8.0	24GB	4

Key Tech Points

GPU Sharing

- GPU time-sharing: Traning
  - No mem and fault isolation
  - Professional GPUs only

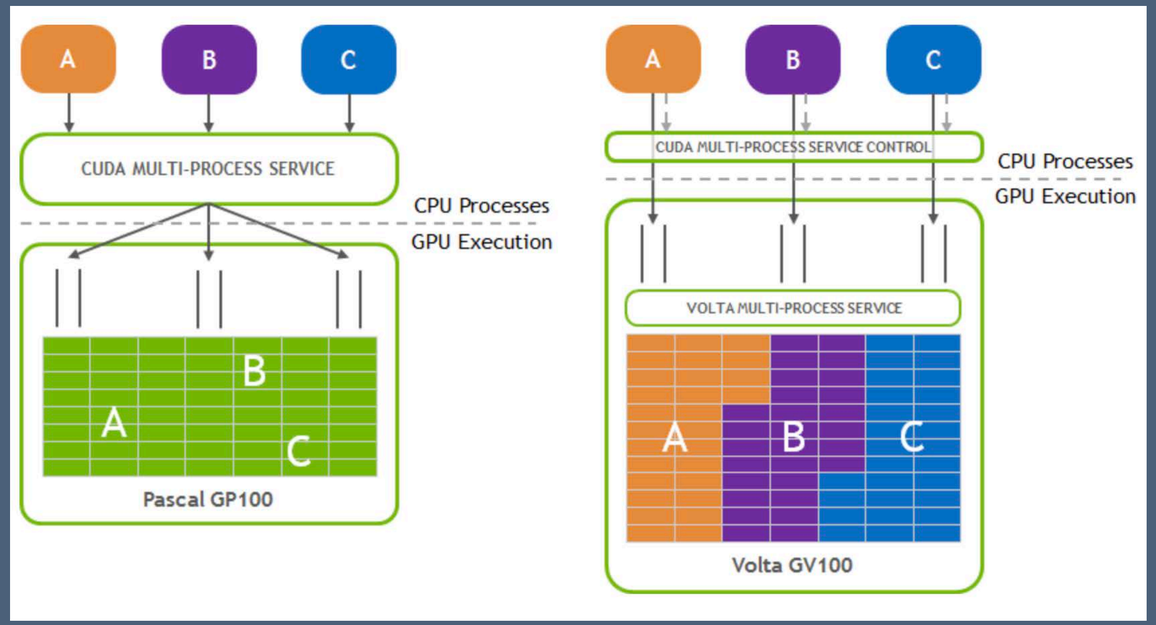
GPU[2, 3]	vGPU 17	vGPU 16[4]	vGPU 15	vGPU 14	vGPU 13	vGPU 12	vGPU 11	vGPU 10	vGPU 9	vGPU 8	vGPU 7	vGPU 6	vGPU 5	GRID 4	GRID 3	GRID 2
NVIDIA A800 PCIe 80GB	-	-	✔	-	-	-	-	-	-	-	-	-	-	-	-	-
NVIDIA A800 PCIe 80GB liquid cooled	-	-	✔	-	-	-	-	-	-	-	-	-	-	-	-	-
NVIDIA A800 HGX 80GB	-	-	✔	-	-	-	-	-	-	-	-	-	-	-	-	-
NVIDIA A100 HGX 80GB	-	-	✔	✔	✔	✔	-	-	-	-	-	-	-	-	-	-
NVIDIA A100 PCIe 80GB	-	-	✔	✔	✔	-	-	-	-	-	-	-	-	-	-	-
NVIDIA A100 PCIe 80GB liquid cooled	-	-	✔	-	-	-	-	-	-	-	-	-	-	-	-	-
NVIDIA A100X	-	-	✔	✔	-	-	-	-	-	-	-	-	-	-	-	-
NVIDIA A100 HGX 40GB	-	-	✔	✔	✔	✔	✔[10]	-	-	-	-	-	-	-	-	-
NVIDIA A100 PCIe 40GB	-	-	✔	✔	✔	✔	✔[10]	-	-	-	-	-	-	-	-	-
NVIDIA A40	✔	✔	✔	✔	✔	✔	-	-	-	-	-	-	-	-	-	-
NVIDIA A30	-	-	✔	✔	✔	-	-	-	-	-	-	-	-	-	-	-
NVIDIA A30X	-	-	✔	✔	-	-	-	-	-	-	-	-	-	-	-	-
NVIDIA A16	✔	✔	✔	✔	✔	-	-	-	-	-	-	-	-	-	-	-
NVIDIA A10	✔	✔	✔	✔	✔	✔[9]	-	-	-	-	-	-	-	-	-	-
NVIDIA A2	✔	✔	✔	✔	-	-	-	-	-	-	-	-	-	-	-	-
NVIDIA H800 PCIe 80GB[6]	-	-	✔	-	-	-	-	-	-	-	-	-	-	-	-	-
NVIDIA H100 PCIe 80GB	-	-	✔	-	-	-	-	-	-	-	-	-	-	-	-	-
NVIDIA L40S[4]	✔	✔	-	-	-	-	-	-	-	-	-	-	-	-	-	-
NVIDIA L40[7]	✔	✔	✔	-	-	-	-	-	-	-	-	-	-	-	-	-
NVIDIA L20[8]	✔	✔	-	-	-	-	-	-	-	-	-	-	-	-	-	-
NVIDIA L4[6]	✔	✔	✔	-	-	-	-	-	-	-	-	-	-	-	-	-
NVIDIA L2[8]	✔	✔	-	-	-	-	-	-	-	-	-	-	-	-	-	-
NVIDIA RTX 5000 Ada[4]	✔	✔	-	-	-	-	-	-	-	-	-	-	-	-	-	-
NVIDIA RTX 5880 Ada	✔	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
NVIDIA RTX 6000 Ada[7]	✔	✔	✔	-	-	-	-	-	-	-	-	-	-	-	-	-
NVIDIA RTX A6000	✔	✔	✔	✔	✔	✔	-	-	-	-	-	-	-	-	-	-
NVIDIA RTX A5500	✔	✔	✔	✔	-	-	-	-	-	-	-	-	-	-	-	-
NVIDIA RTX A5000	✔	✔	✔	✔	✔	✔[9]	-	-	-	-	-	-	-	-	-	-
Quadro RTX 8000	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	-	-	-	-	-	-
Quadro RTX 8000 passive	✔	✔	✔	✔	✔	✔	✔	✔	-	-	-	-	-	-	-	-
Quadro RTX 6000	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	-	-	-	-	-	-
Quadro RTX 6000 passive	✔	✔	✔	✔	✔	✔	✔	✔	-	-	-	-	-	-	-	-
Tesla V100	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	-	-	-	-
Tesla T4	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔[11]	-	-	-	-	-
Tesla P100	-	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	-	-	-
Tesla P40	-	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	-	-	-
Tesla P6	-	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	-	-	-
Tesla P4	-	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	-	-	-
Tesla M60	-	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
Tesla M10	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	-	-
Tesla M6	-	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
GRID K2	-	-	-	-	-	-	-	-	-	-	-	-	-	✔	✔	✔
GRID K1	-	-	-	-	-	-	-	-	-	-	-	-	-	✔	✔	✔

✔ GPU is supported
- GPU is not supported

Key Tech Points

GPU Sharing

- Nvidia MPS
  - Supported by all current GPUs
  - High performance
  - All context in the same GPU context



Key Tech Points

GPU Sharing

- Tencent TKE GaiaGPU
  - Supported by all current GPUs
  - Complete isolation
  - CUDA hijacking
  - Opensource

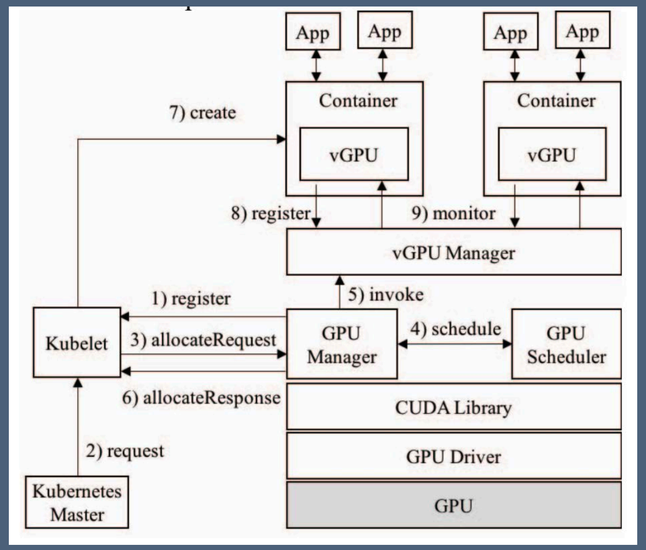


TABLE I. THE INTERCEPTED CUDA DRIVER APIS

	CUDA Driver API	Description
Memory-related APIs	cuMemAlloc	Allocates device memory.
	cuMemAllocManaged	Allocates memory that will be automatically managed by the Unified Memory system.
	cuMemAllocPitch	Allocates pitched device memory.
	cuArrayCreate	Creates a 1D or 2D CUDA array.
	cuArray3DCreate	Creates a 3D CUDA array.
	cuMipmappedArrayCreate	Creates a CUDA mipmapped array.
Computing resources-related APIs	cuLaunch	Launches a CUDA function.
	cuLaunchKernel	Launches a CUDA function.
	cuLaunchCooperativeKernel	Launches a CUDA function where threads blocks can cooperate and synchronize as they execute.
	cuLaunchGrid	Launches a CUDA function.
Device info-related APIs	cuDeviceTotalMem	Returns the total amount of memory on the device.
	cuMemGetInfo	Gets free and total memory.

J. Gu, S. Song, Y. Li and H. Luo, "GaiaGPU: Sharing GPUs in Container Clouds," 2018 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Ubiquitous Computing & Communications, Big Data & Cloud Computing, Social Computing & Networking, Sustainable Computing & Communications (ISPA/IUCC/BDCloud/SocialCom/SustainCom), Melbourne, VIC, Australia, 2018, pp. 469-476, doi: 10.1109/BDCloud.2018.00077.

Summary

ML Platform for everyone

An affordable ML Platform running on 3 PCs with 3 RTX1060

Thanks