Accelerating AI Business Outcomes

Through Strategic AI Data Center Infrastructure Deployment

INFINITIX CEO Wenyu Chen

議題大綱

- About Infinitix
- The market of AI Computing
- The Challenges AI Computing Cloud Operation
- To Plan an AI Computing Cloud
- Q & A

About Infinitix

INFINITIES → INFINITIX

market share of AI platform in Taiwan

builder of national AI cloud

Leading Customers

Semi-conductor

UMC
REALTEK
Himax
fitipower
SiS

Manufacturing

中鋼公司
台塑企業 FORMOSA PLASTICS GROUP
DELTA

Finance Service

永豐金控 SinoPac Holdings
iPASS 一卡通

Academic

中央研究院 ACADEMIA SINICA
NATIONAL CHENG KUNG UNIVERSITY 1931
TAIPEI TECH
I-SHOU UNIVERSITY
TAMKANG UNIVERSITY
TZU CHI UNIVERSITY
CHINA MEDICAL UNIVERSITY

AI Data Center

AICC
Artificial Intelligence Computing Center
moda 數位發展部 Ministry of Digital Affairs

Government

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
交通部中央氣象局 CENTRAL WEATHER BUREAU
金門縣政府

Healthcare

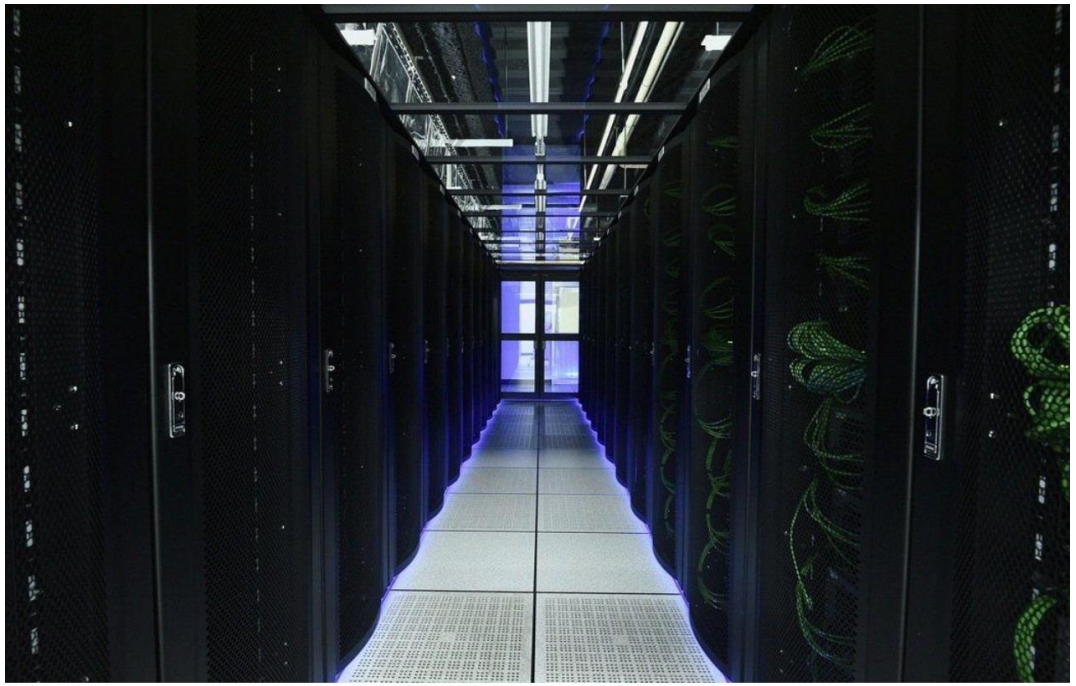
Transportation

Taoyuan Airport 桃園機場
Metro Taipei
metro Taoyuan

Energy

台灣電力公司

Taiwan’s Ministry of Digital Affairs Built the AI Computing Cloud with INFINITIX for the ecosystem of startups

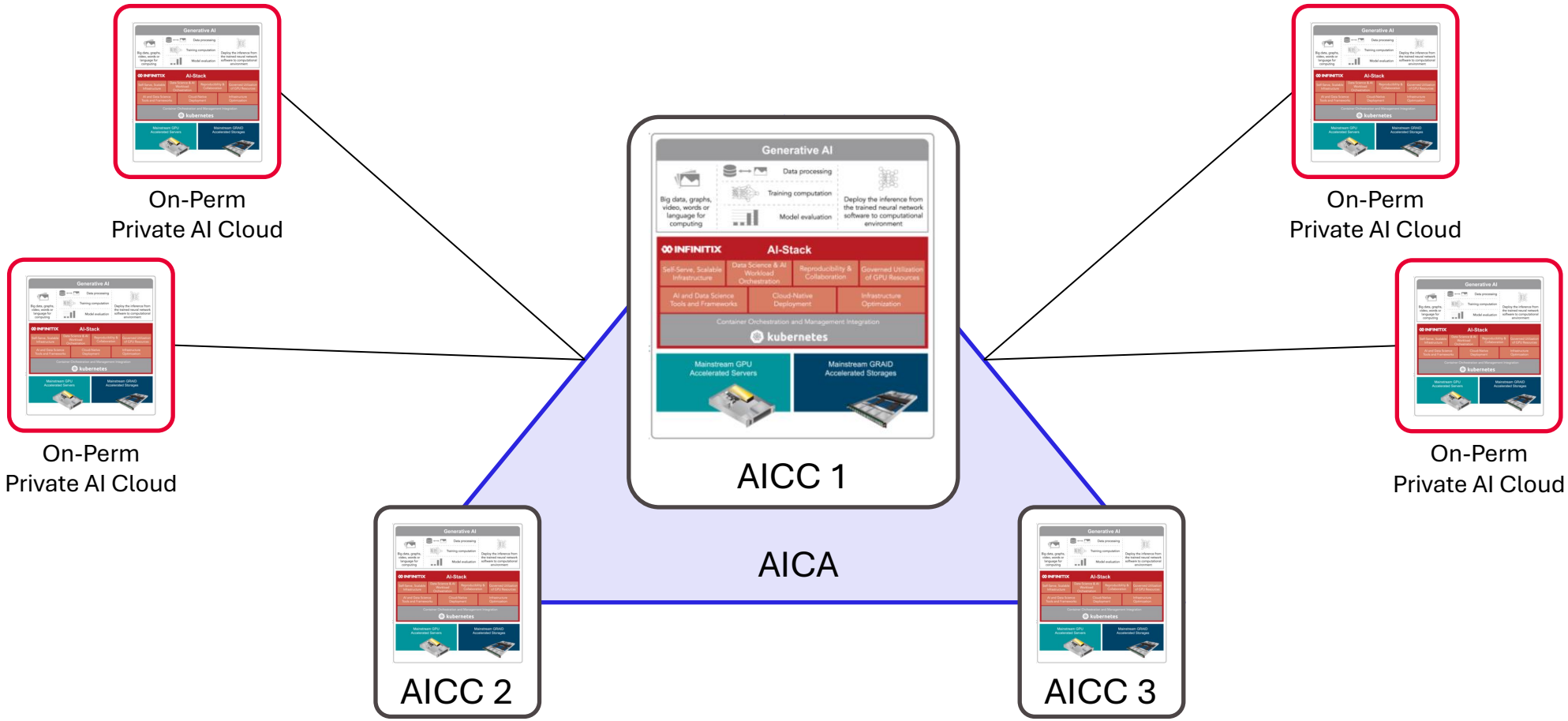


Supported for H100 and MI300X GPU

https://money.udn.com/money/story/5640/8114582

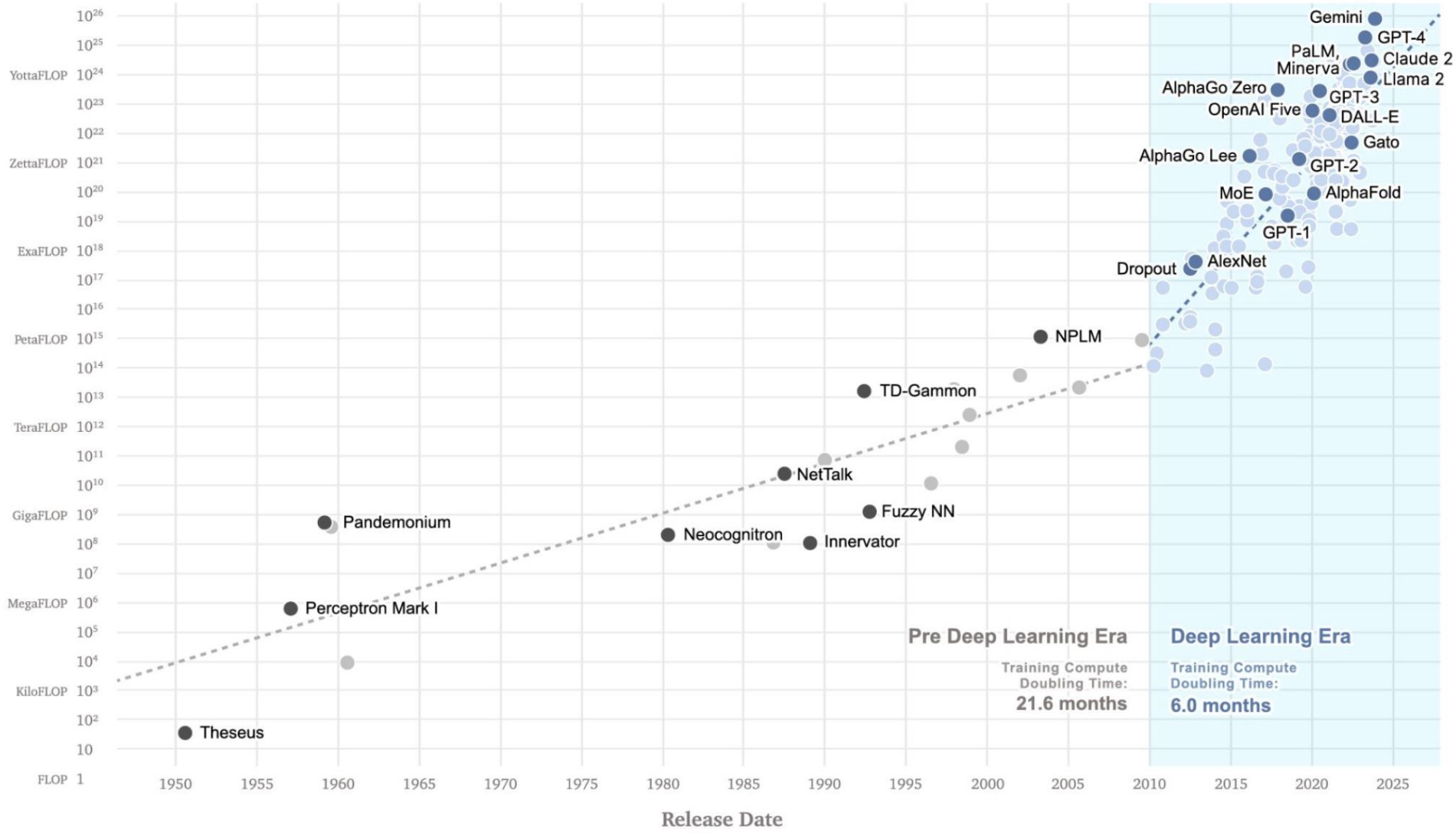
Infinitix Assist to Build Asia's Most Advanced H200 AI Computing Cloud in Taiwan

Initiate the cooperation alliance with AI Computing Cloud

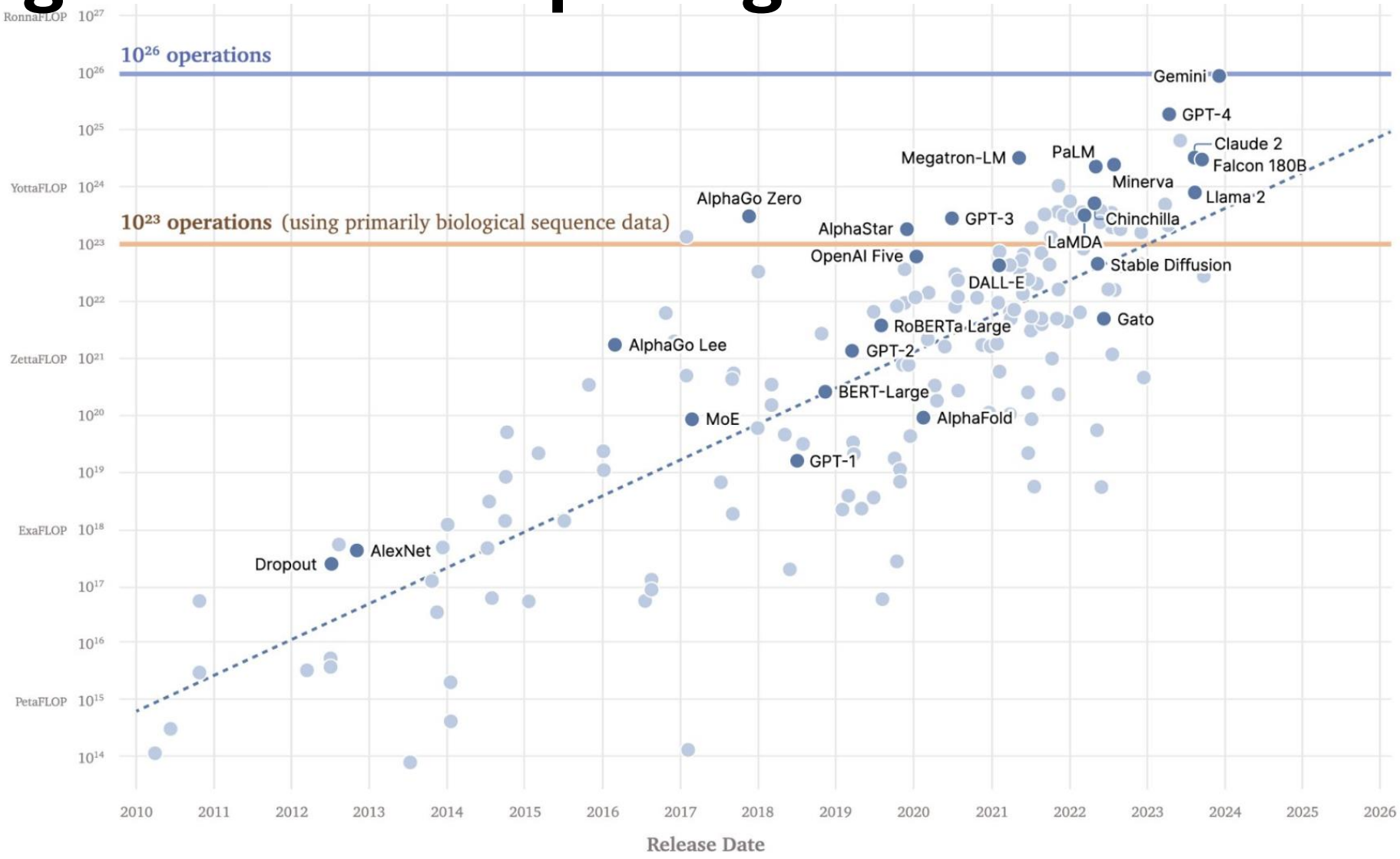


The market of AI Computing

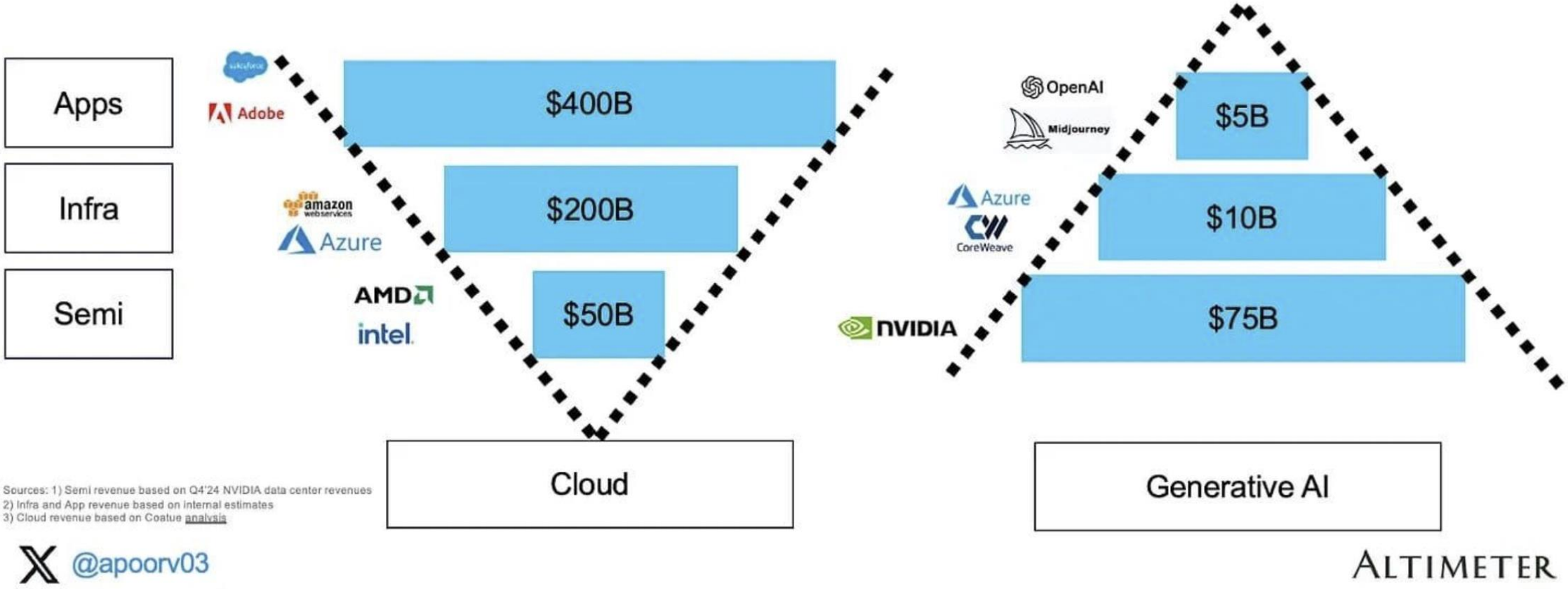
Rapid growth in computing demand



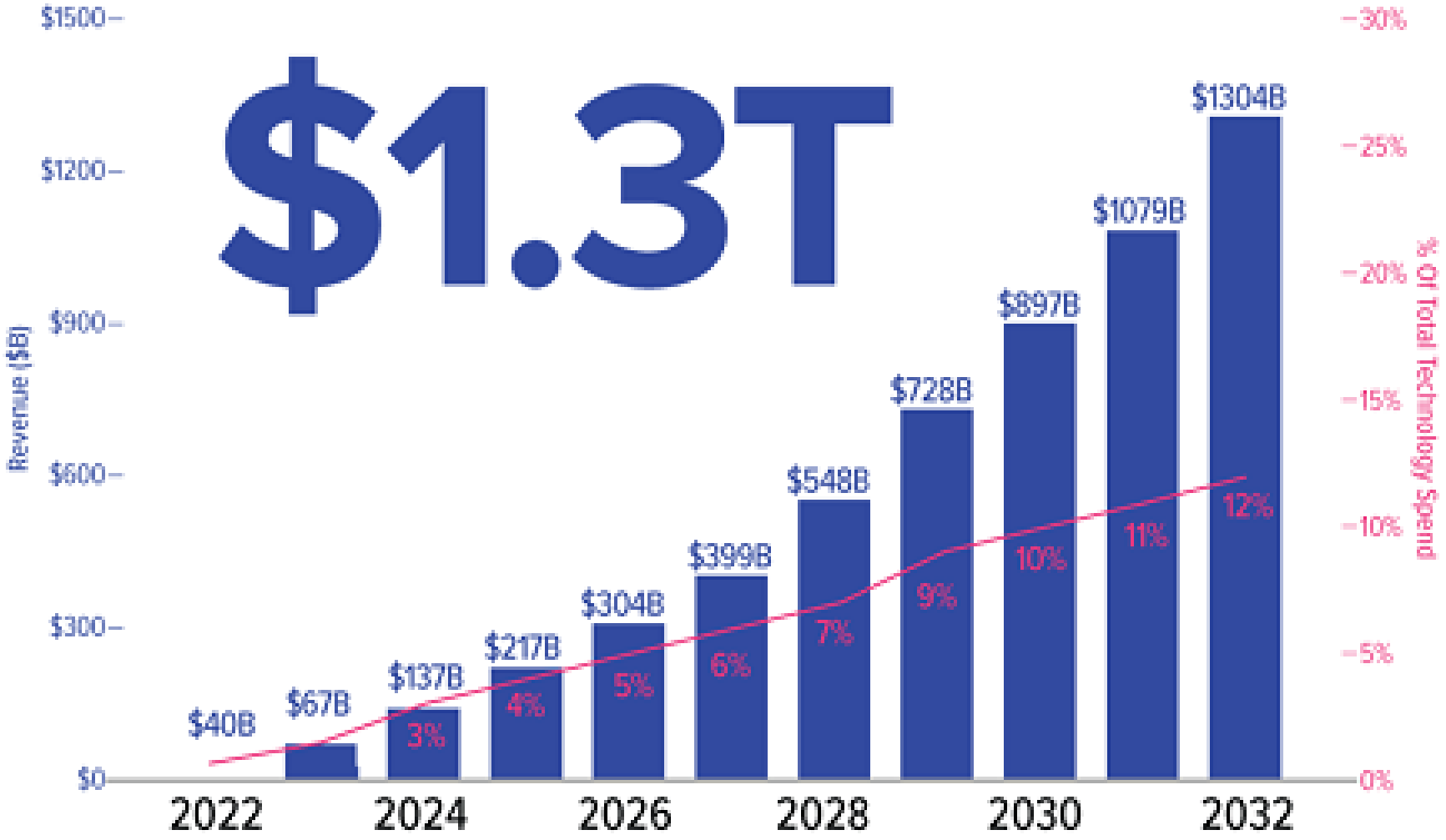
Rapid growth in computing demand



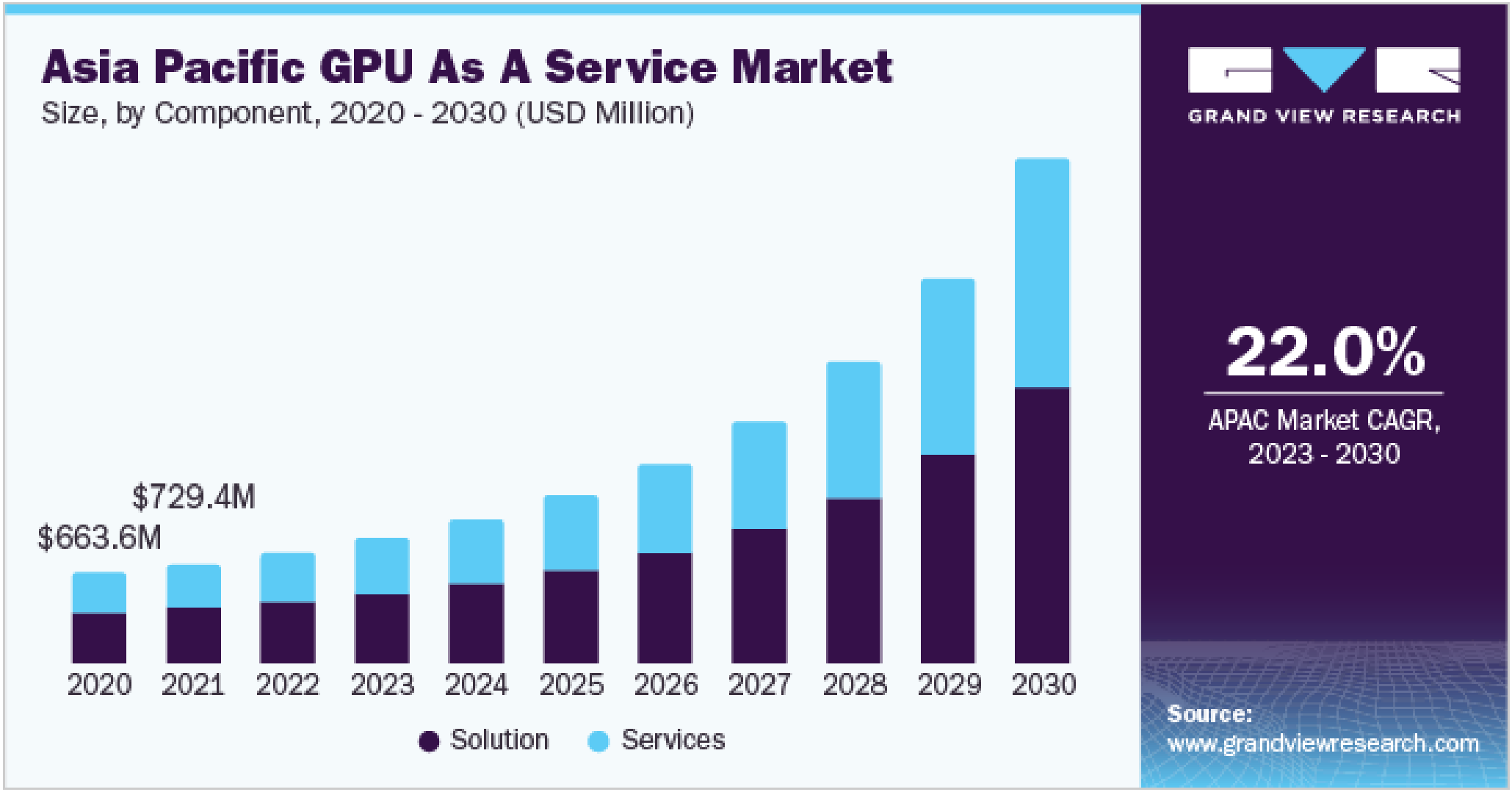
AI Market Overview



Generative AI Market Overview

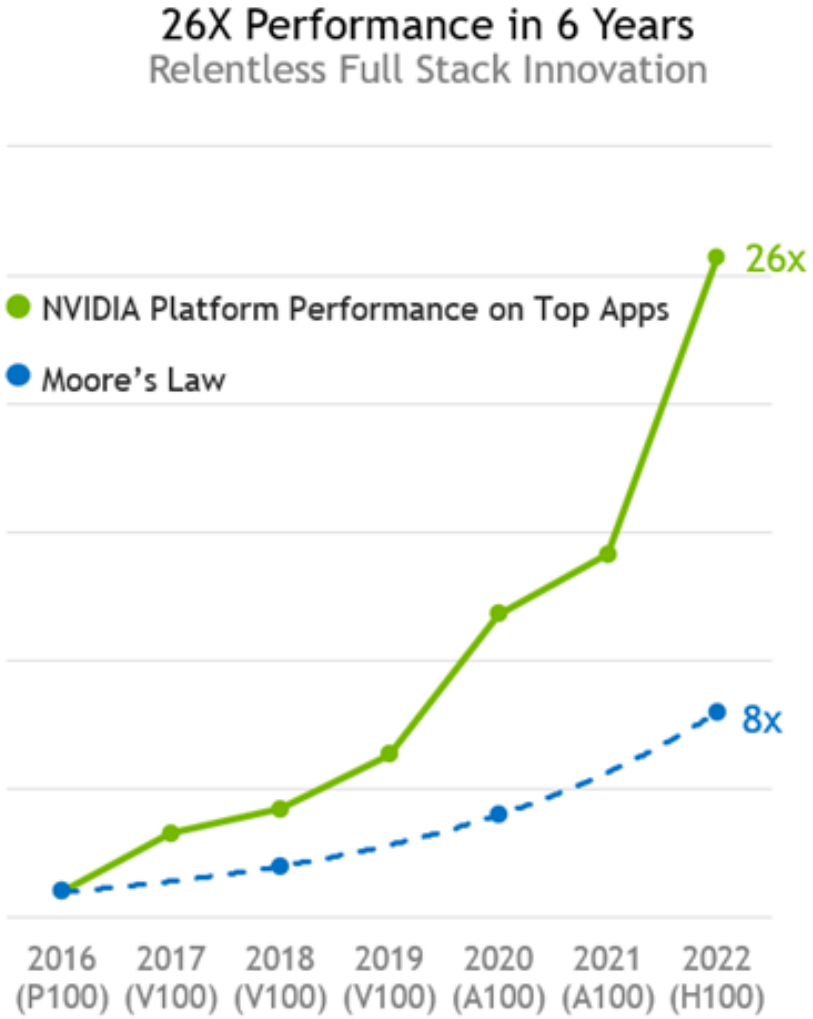


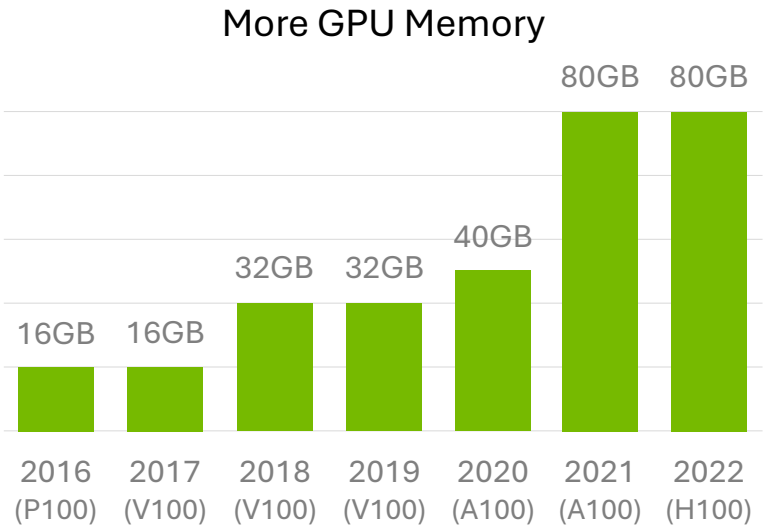
Asia-Pacific GPU-as-a-service Market Share

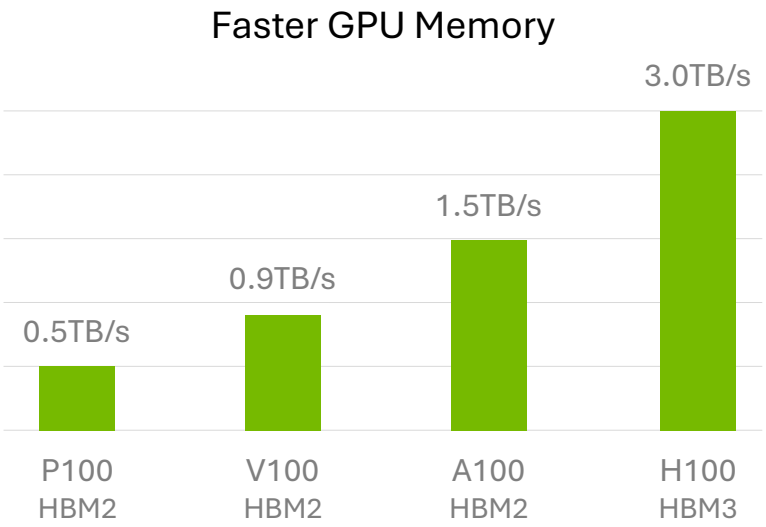


The Challenges of AI Computing Cloud Operation

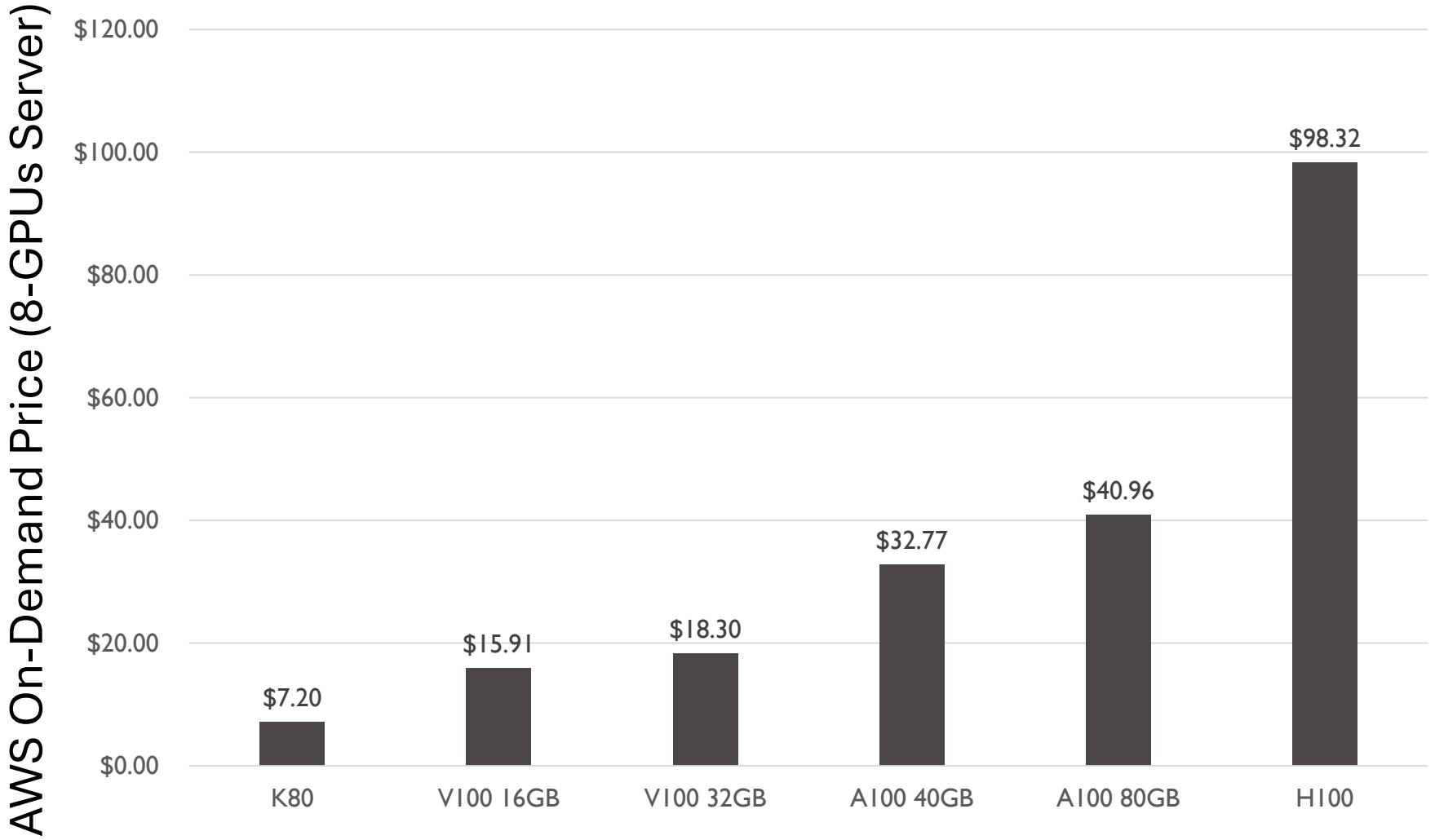
AI Infra. has become extremely more performant





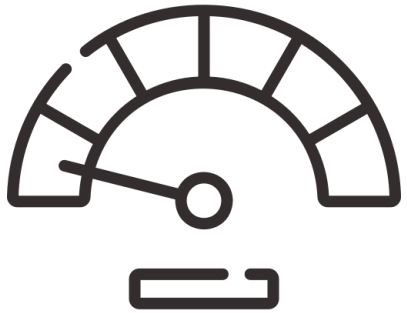


10x increase in GPU costs



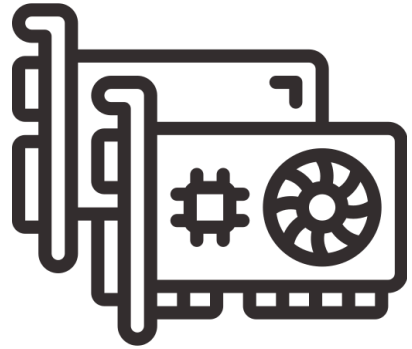
Challenges with AI Infrastructure Management

Lack of controls and prioritization



Low utilization, high cost

Difficult to visibility and better decision making



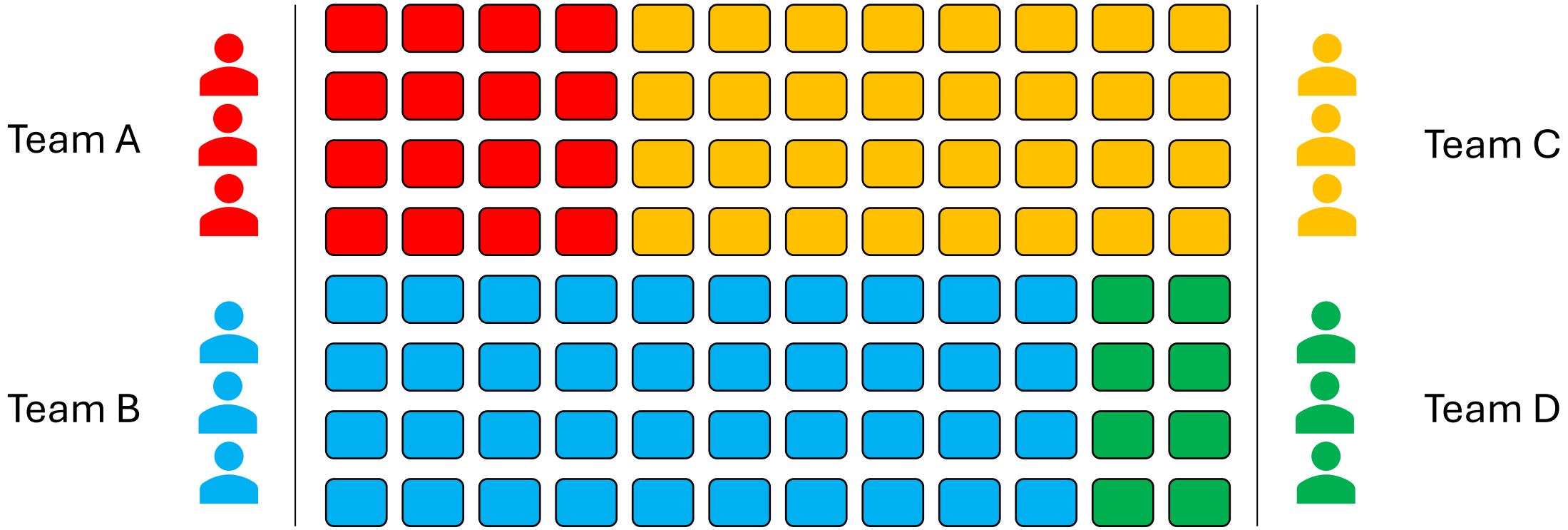
Users are still in need for more GPUs

GPU Pooling from siloed to collaborative efforts

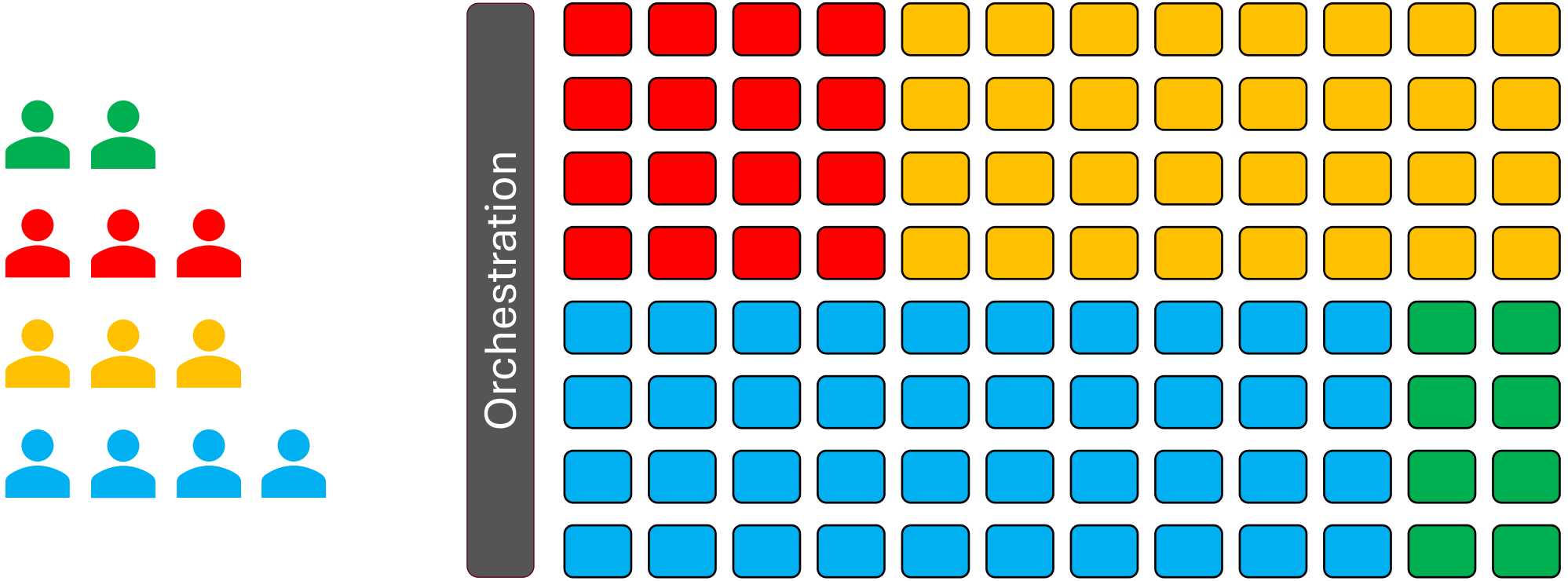
Siloed Infrastructure → Shared Clusters

On-Demand Compute → Reserved Clusters

GPU Pooling from siloed to collaborative efforts



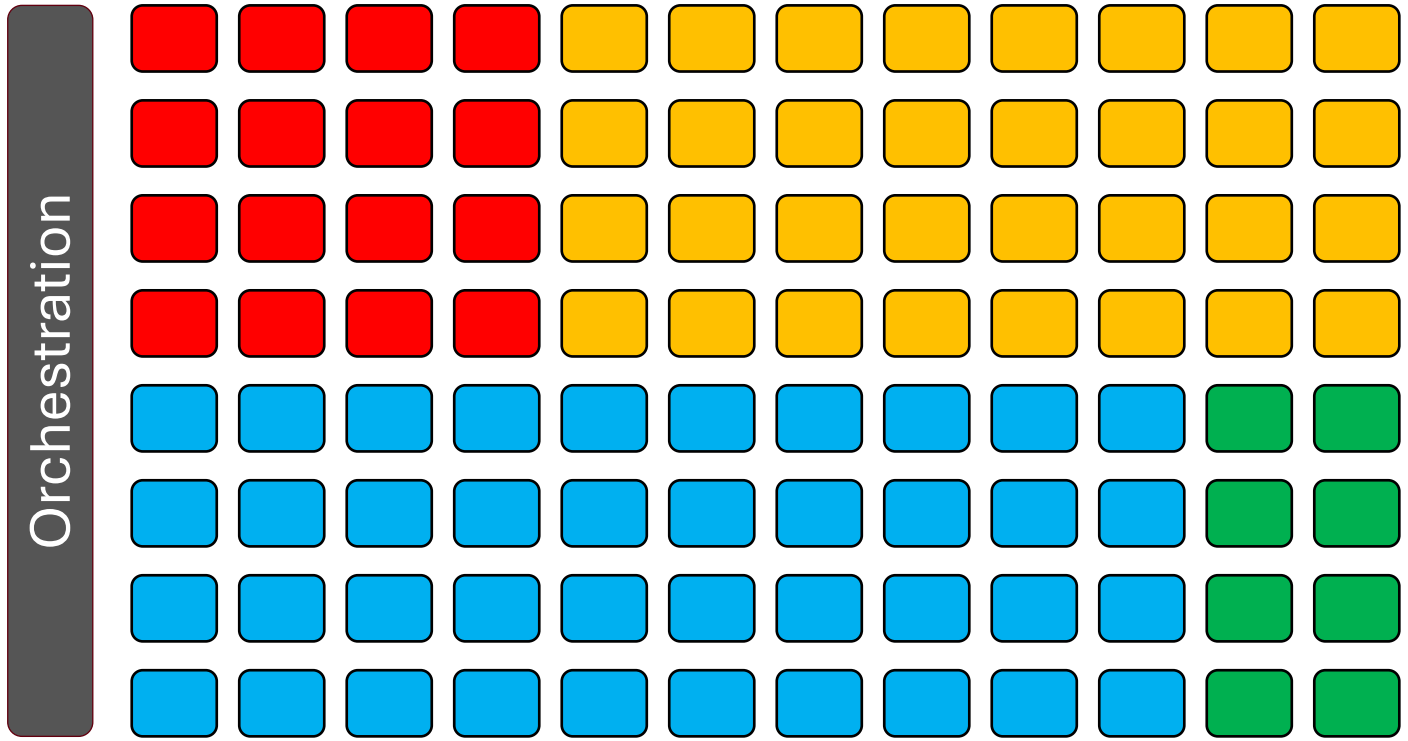
GPU Pooling + Orchestration



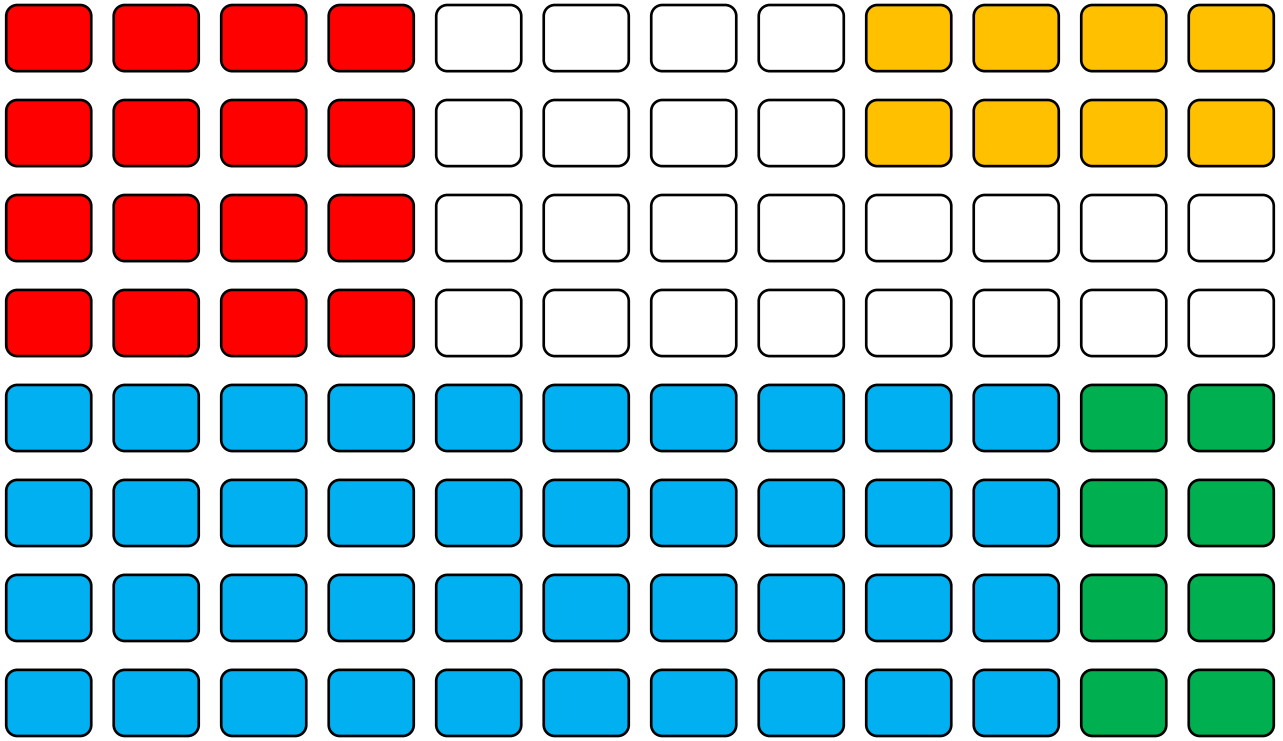
GPU Pooling + Orchestration

Orchestration capabilities

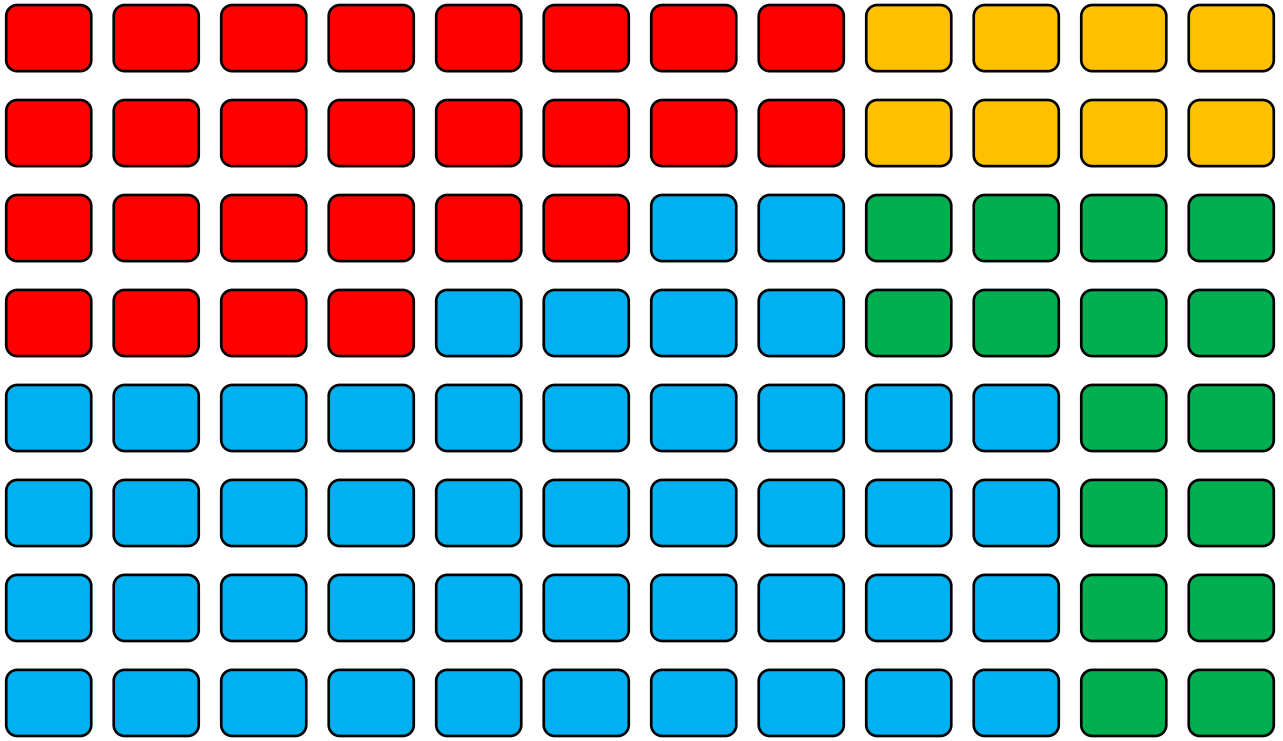
- Controls on resource allocations
- Workload prioritization
- Job queueing
- Workload monitoring and execution
- Accounting and reporting



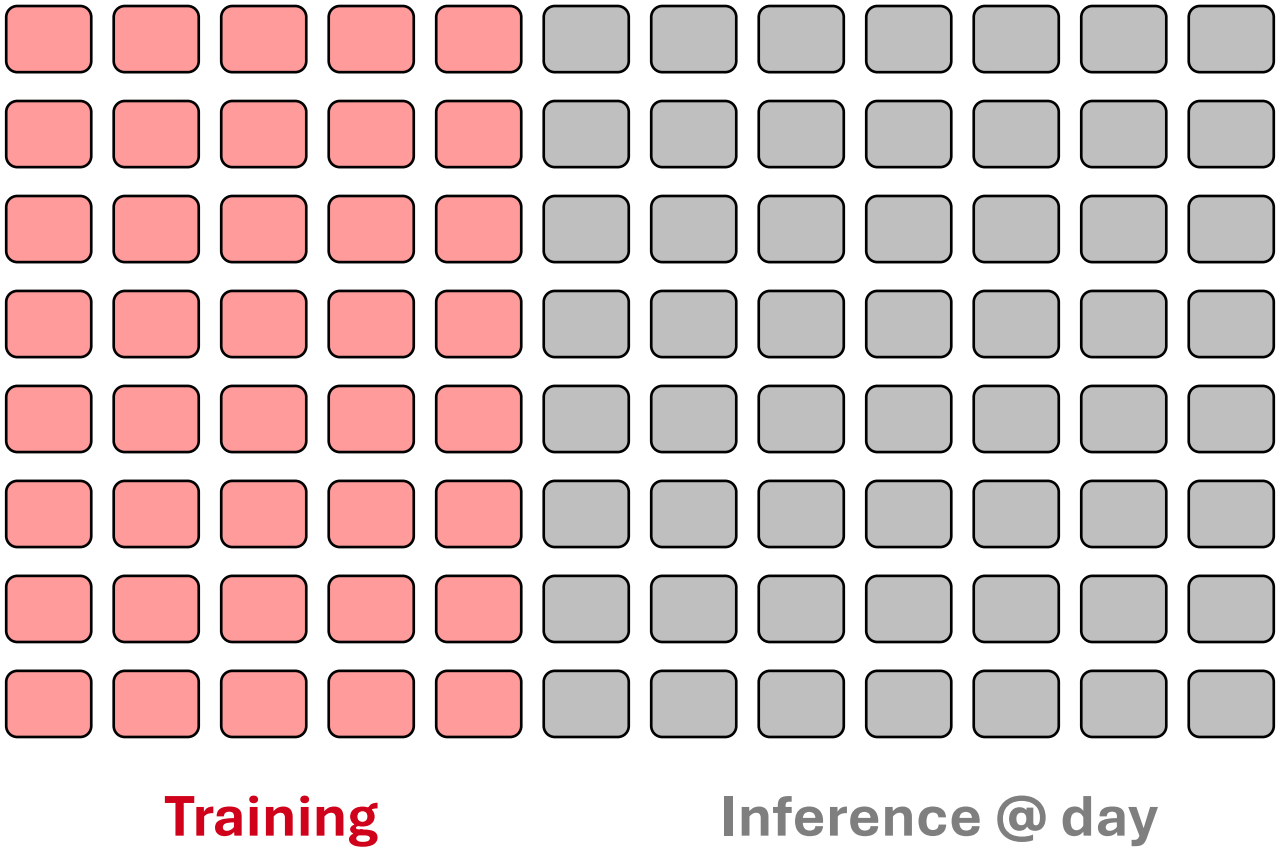
Repurposing resources between different teams



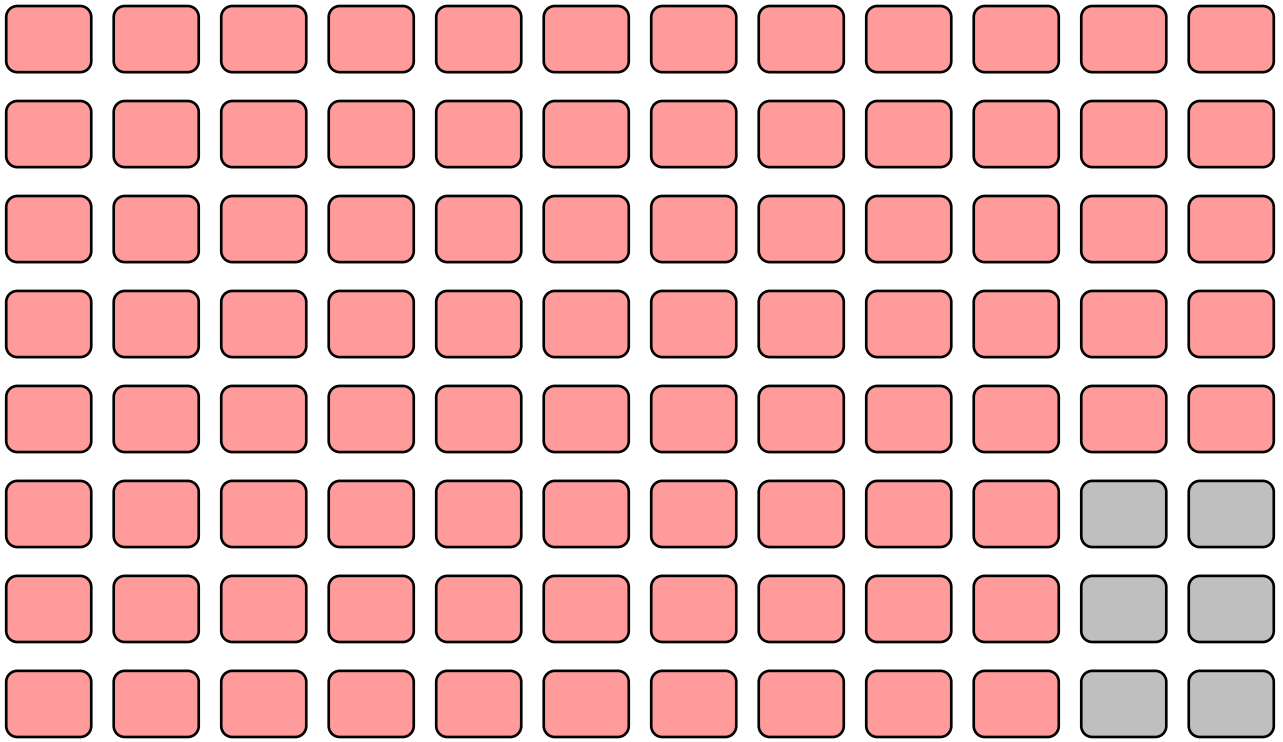
Repurposing resources between different teams



Repurposing resources between different workloads



Repurposing resources between different workloads



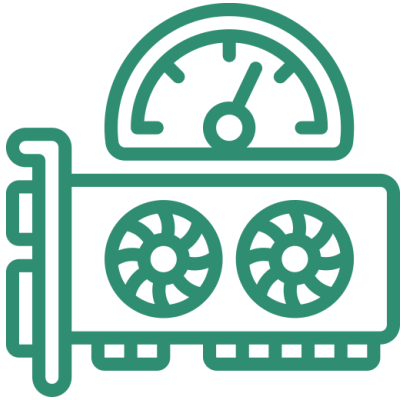
Training

Inference @ night

The Benefits

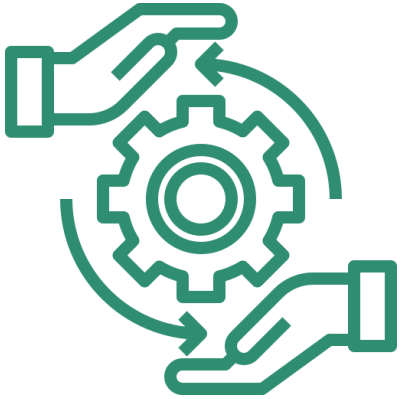
Higher Efficiency

Through Sharing and repurposing resources



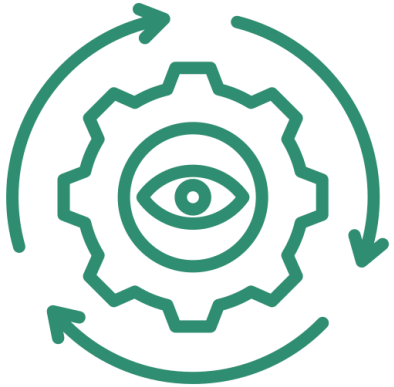
More GPU Accessibility

Users become more productive with easier access to more GPUs



Controls & Governance

Ability to align resources with business goals



Centralized Visibility

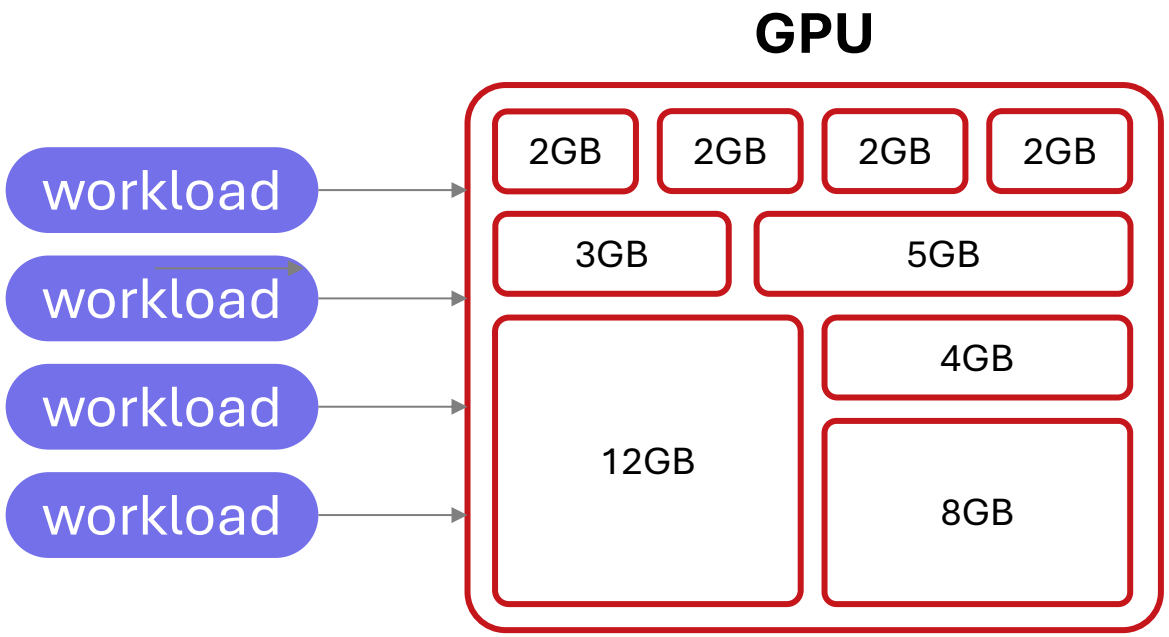
Better planning and decision making

Not all workloads need whole powerful GPUs

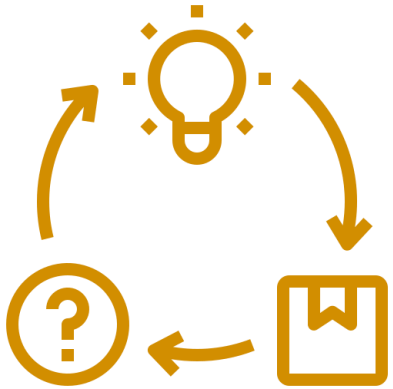
AI-Stack isolates the GPU elasticity into many GPU slices to divide workload requirements.

Multiple workloads share a single GPU

- Notebooks
- Inference workloads
- GPU slicing type:
  - AI-Stack GPU Slicing (software isolation)
  - Hardware isolation

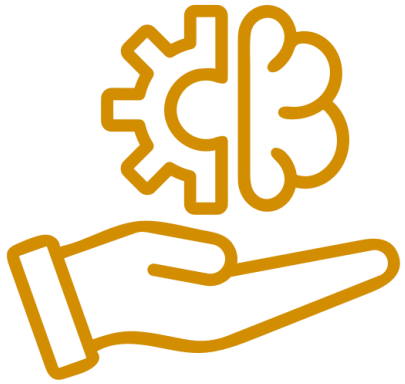


Support for the entire AI lifecycle



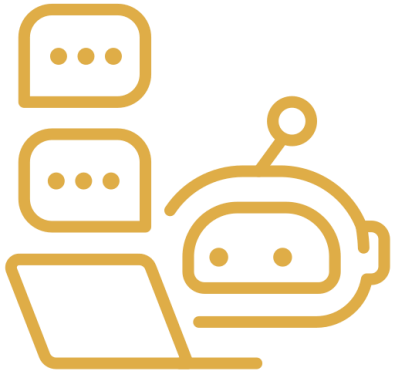
Model Development

Dev & debug in IDE tools like Jupyter notebooks, VSCode, PyCharm etc



Fine-tuning & Training

Run long model tuning or training workloads as batch jobs



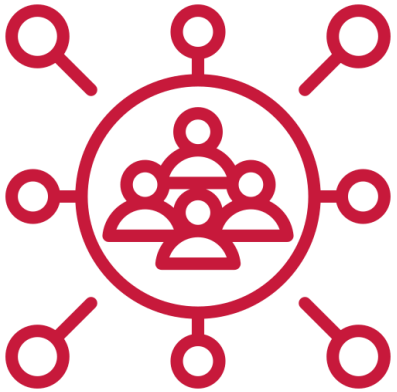
Prompt Engineering

Experiment with language and GenAI models through prompt engineering

Serving in Production

Deploy models in production to serve business applications

The Keys for Operation AI Infrastructure Platform



Resource Pooling

Centralize GPUs into a single cluster to simplify management and increase efficiency

Workload Scheduler

Repurpose resources and prioritize workloads according to business goals

GPU Slicing

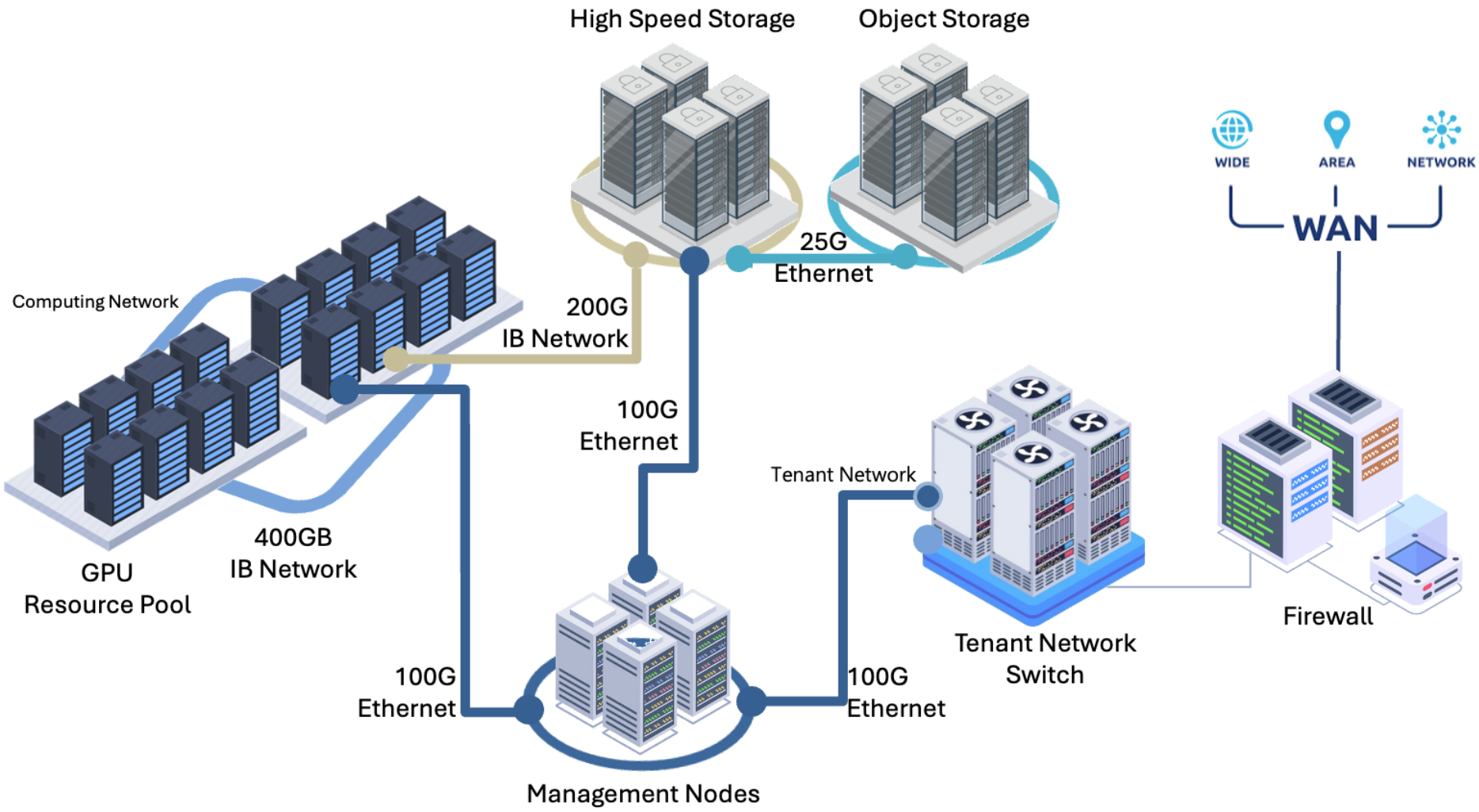
Run more notebooks or inference servers on the same infrastructure

Tooling and Integration

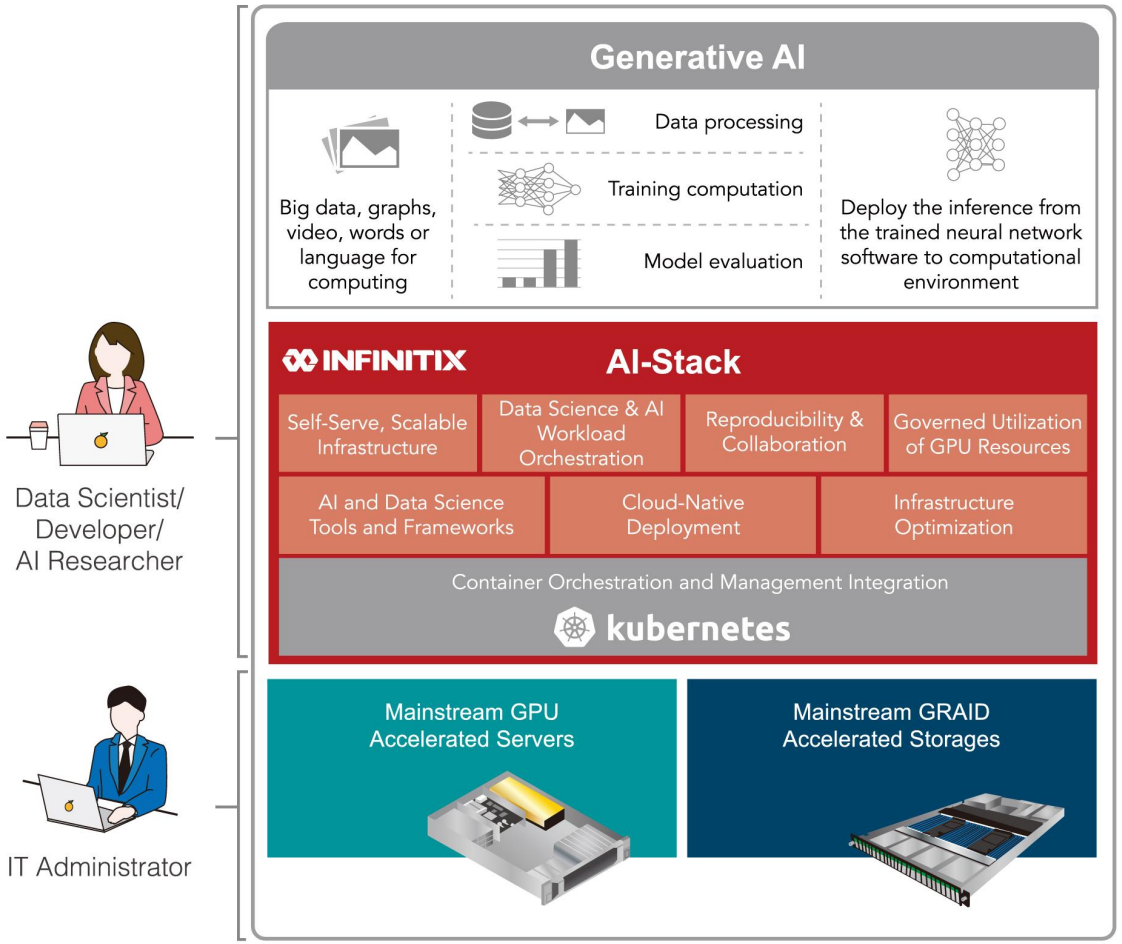
Support the entire AI lifecycle and maintain openness and flexibility to support new tooling

To Plan an AI Computing Cloud

AI Computing Cloud Architecture Blueprint



The Platform for AI computing cloud



- MLOps Platform for AI Development Accelerates Software Development
- Provides a complete MLOps development process from prototype to production
- Easily scales with built-in computing power
- Provides a virtualized pool of computing resources
- Maximizes system performance with high availability and reliability

The best AI-Ready Enterprise Platform

AI-Stack pairs GPU and the Enterprise MLOps benefits of workload orchestration, self-serve infrastructure, GPU optimization, and collaboration with the cost-effective scale from containerization on mainstream accelerated servers and storages.

> For Data Scientists & AI Researchers

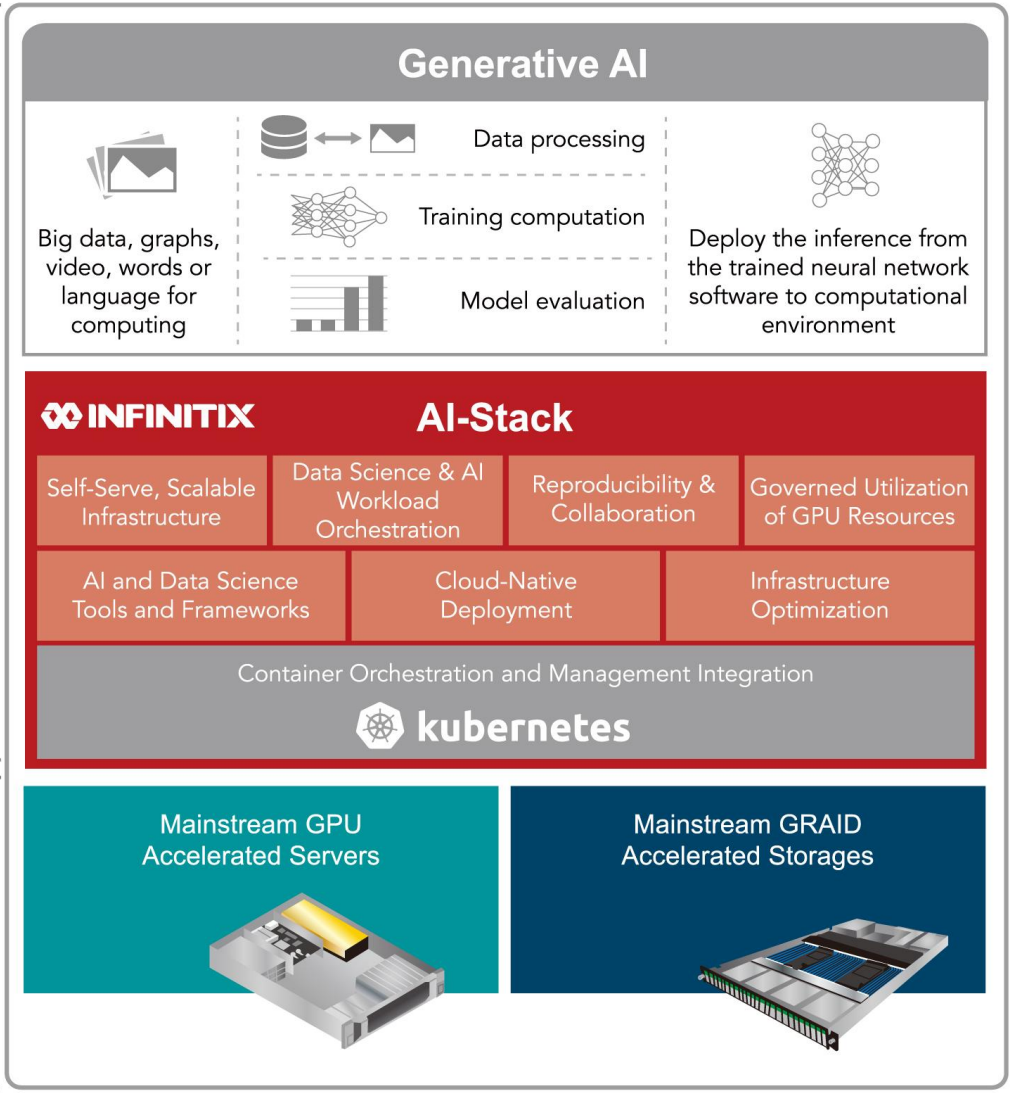
Focus on research instead of dev ops.
Launch AI-Stack on-demand with a container configured with the latest data science tools, frameworks, and GPUs.

Data Scientist/
Developer/
AI Researcher

> For IT

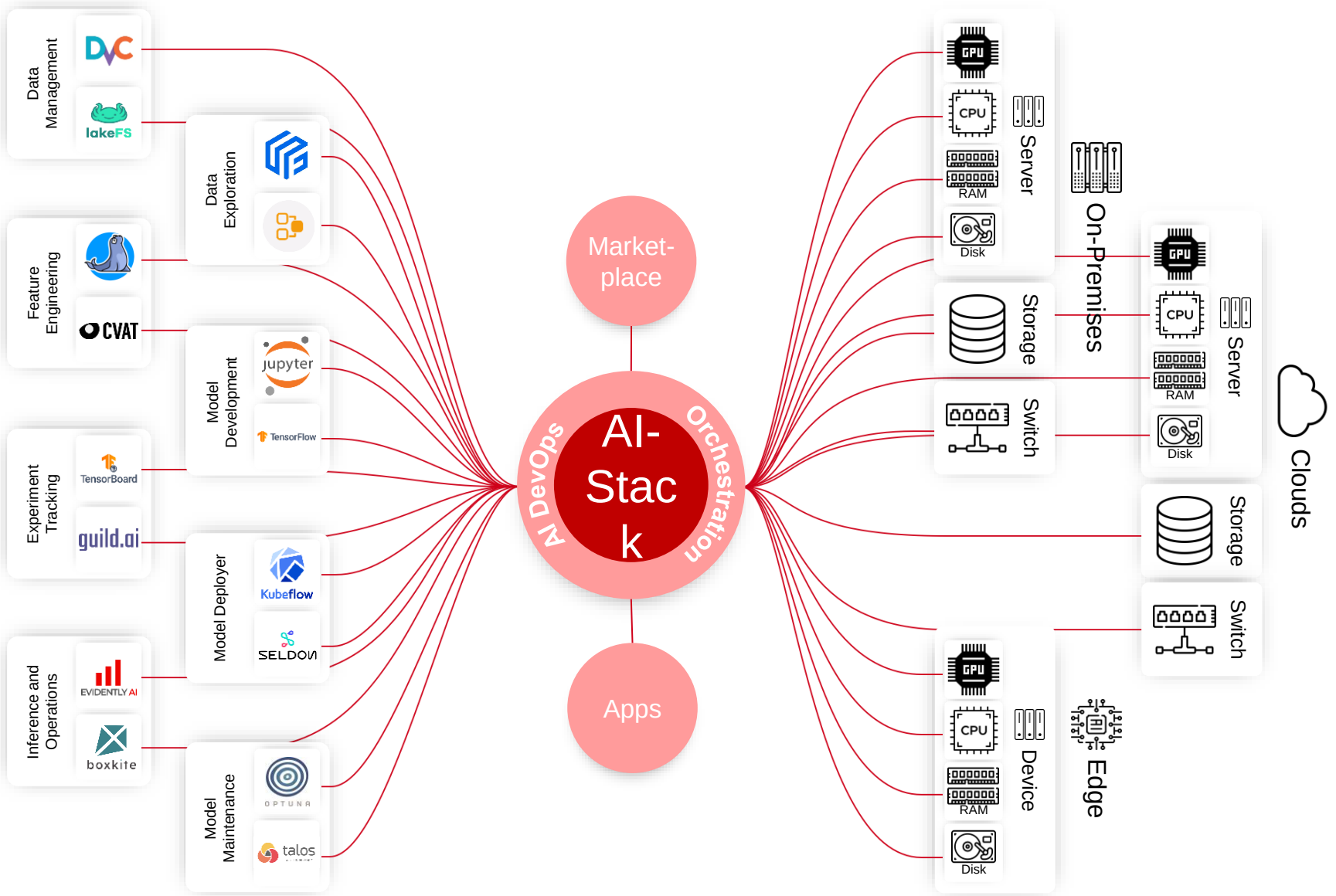
Get the confidence of enterprise-grade security, manageability, and support.
AI-Stack is validated to run on Kubernetes and deployed on industry-leading GPU systems.

IT Administrator



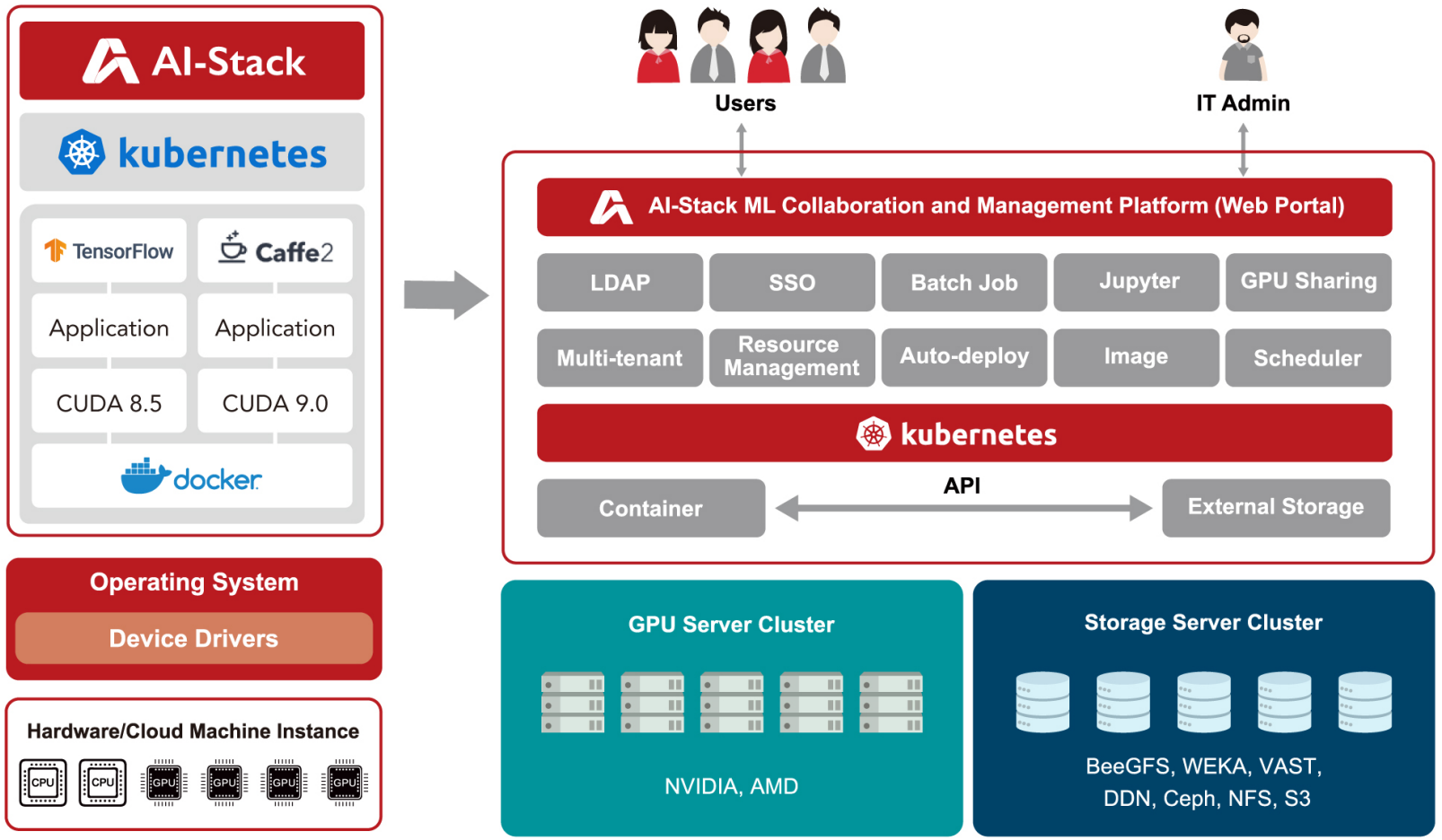
AI-Stack covers all AI DevOps needs in a scalable way.

AI-Stack provides a foundation for AI, whether on-premises, clouds, or edge, allowing organizations to have their AI resources on a single, unified platform that supports AI at all stages of development, from building and training models to running inference in production.



Empower your AI Teams.

We've built a software layer that abstracts AI hardware away from data scientists and ML engineers, letting Ops and IT simplifies the delivery of computing resources for any AI workload and any AI project.



Help the team manage GPU resources more efficiently

Seamlessly connect with commonly used AI development tools

Quota Limit
2500m CPU
Unlimited RAM
4096MiB GPU

Resource isolation, authority management and quota limit

Automated execution, pre-set single or batch inference

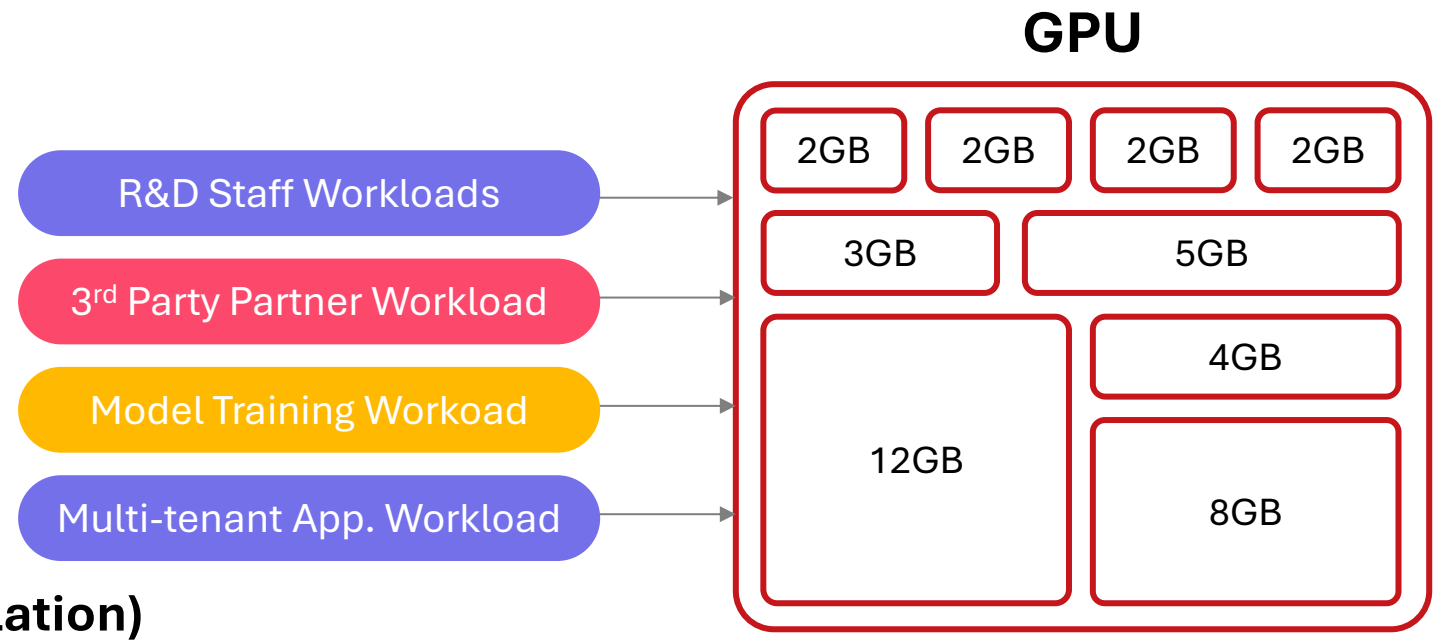
GPU Slice Technology Maximizes GPU Usage

Single GPU supports multi-workload at the same time

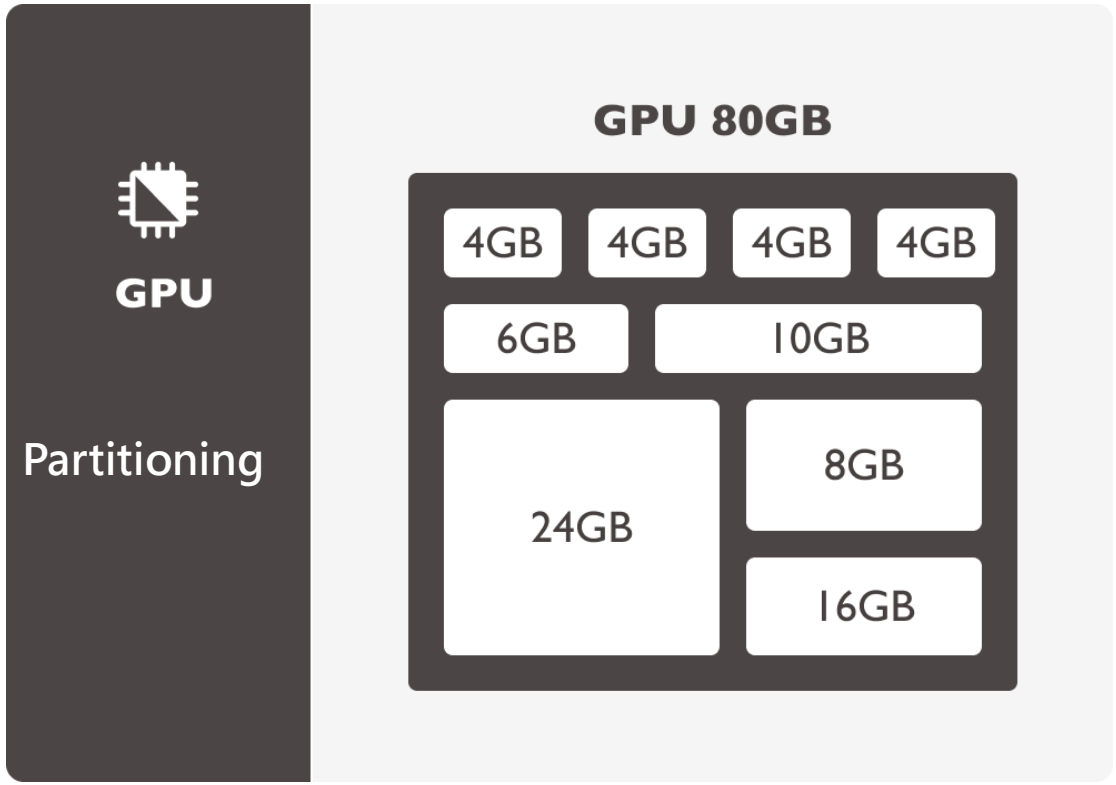
- **Inference workloads**
- **Model training load**
- **Multi-tenant application load**
- **R&D staff development load**

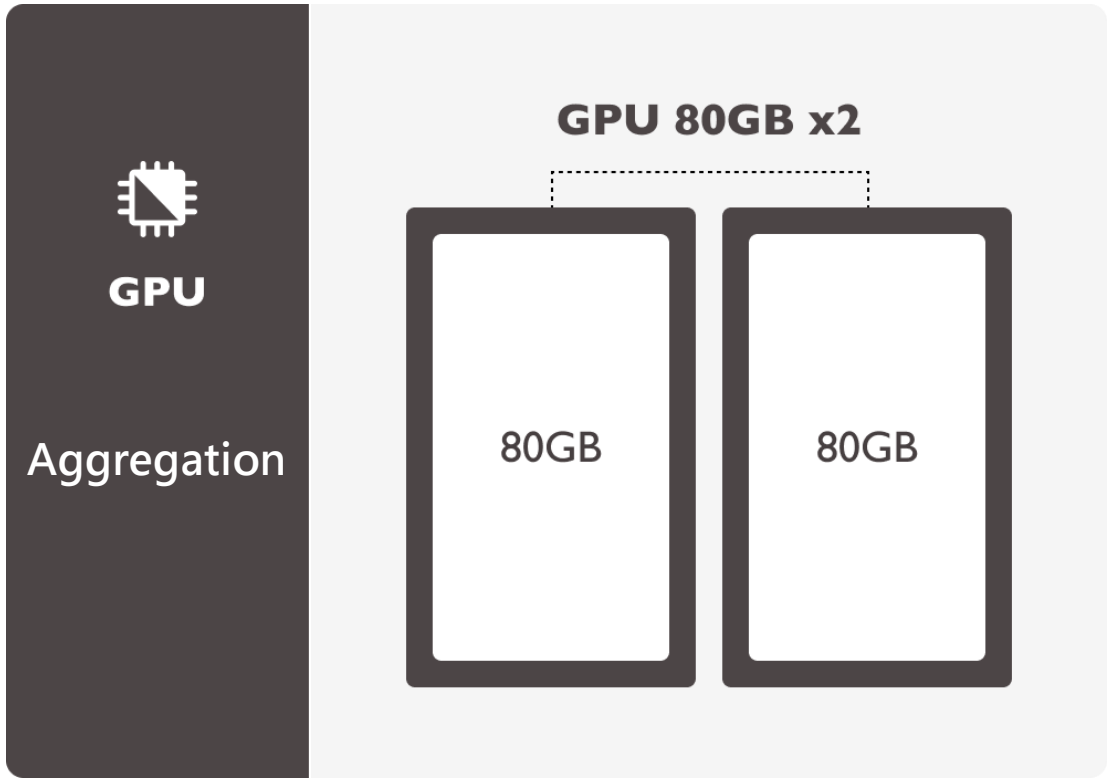
Supported GPU slicing technology

- **AI-Stack GPU slicing (software isolation)**
- **Hardware isolation**



AI-Stack: GPU Partitioning and Aggregation Technology for Maximum Efficiency





Utilization rate ↑

One GPU → Multi GPU Fractions

Computing performance ↑

Multi GPUs → One GPU Aggregation

AI-Stack enhances GPU efficiency when helping enterprises implement AI

90% ↑

GPU Utilization

10x ↑

Workload execution

1min ↓

Development Environment Setup

10x ↑

Enhanced ROI

GPU Utilization

GPU Partitioning Increases Utilization 30% → 90%

Workload execution

Multiple users and tasks increase efficiency by 10x

DevEnv Setup

Reduces setup time from 2 weeks → 1 min

Enhanced ROI

Boosts return on investment by 10x

與我們聯絡

INFINITIX

AI-Stack

FB

LinkedIn

40

Thank you.