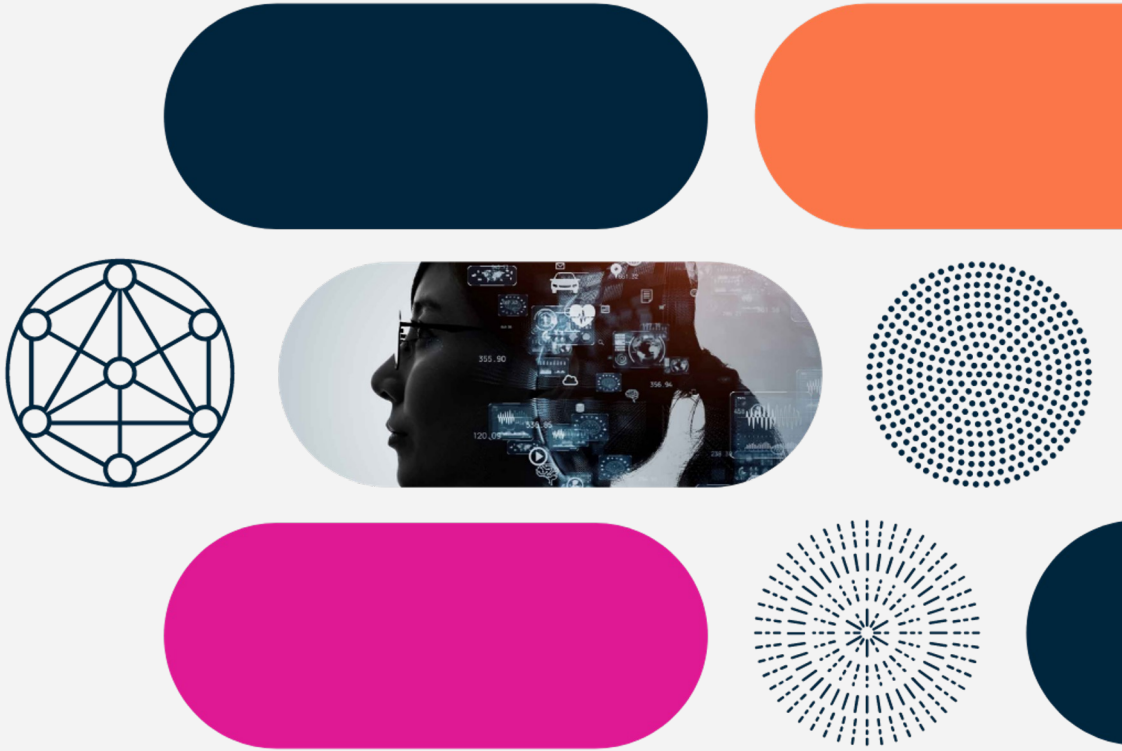
run:ai

Accelerating AI Workflows on AI Data Center Infrastructure



Ronen Dar
Guy Salton
March, 2024

Agenda

01
Intro

02
Challenges and best practices in operating AI Infrastructure

03
Demo

Ronen Dar

Co-Founder & CTO, Run:ai

- Lives near Tel Aviv, in Israel
- Since 2018, Co-Founder & CTO at Run:ai
- PhD & Postdoc in Information Theory, background in Chip Startups

Guy Salton

Director of Solutions Engineering

- Lives in Tel Aviv, Israel
- Kubernetes enthusiast
- CS and Devops background in B2B startups

AI Infrastructure Orchestration

Pooling and sharing AI compute resources across clouds and on-premise to deliver unmatched ML productivity and infrastructure availability

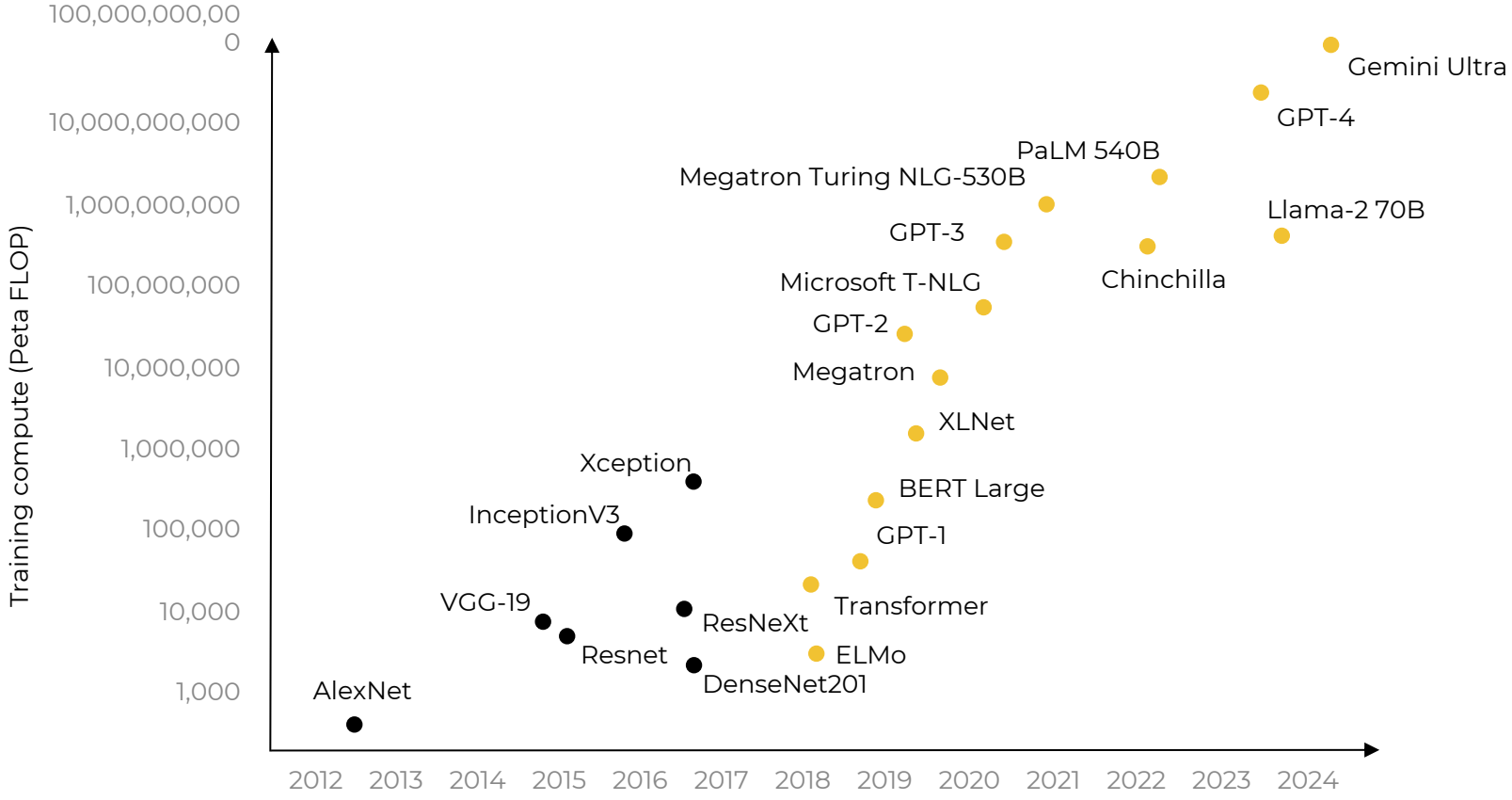
Companies Accelerating AI

Adobe
BNY MELLON
ZEBRA
SONY
CRÉDIT AGRICOLE
USAA
zenseact
Massachusetts Institute of Technology
UNIVERSITY OF OXFORD
xiaomi
mobileye An Intel Company
NOKIA

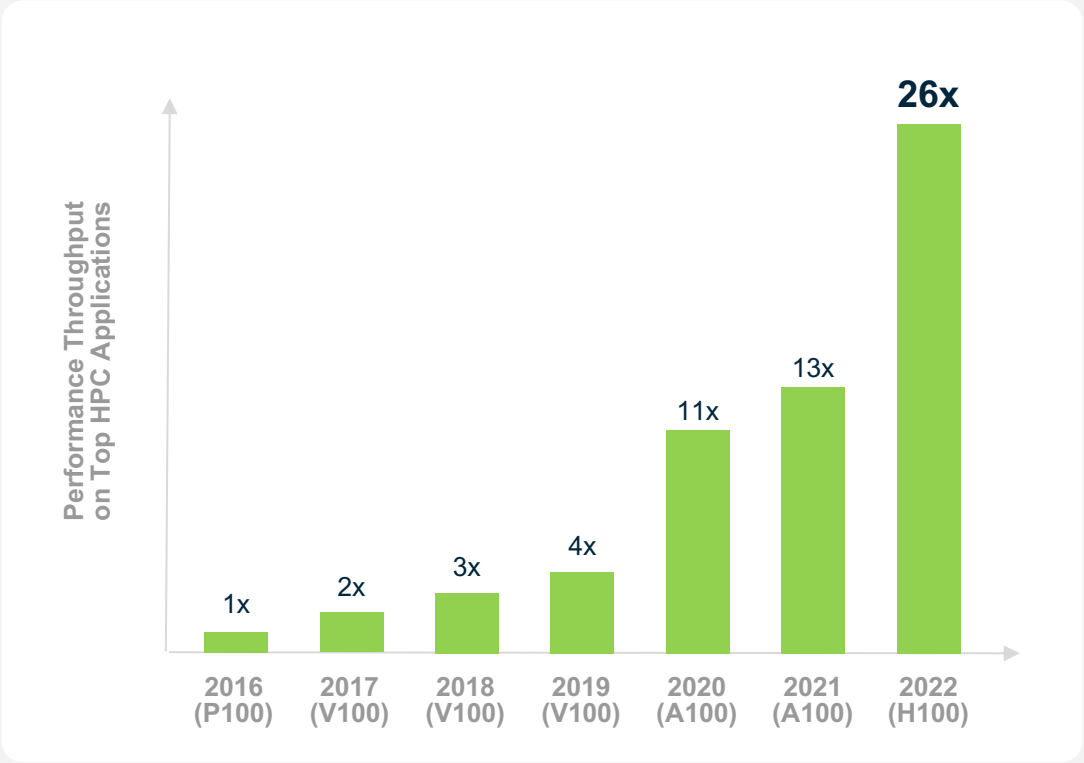
run:ai

Certified for NVIDIA SuperPODs

Explosive growth in compute requirements for AI



AI Infrastructure has become extremely more performant

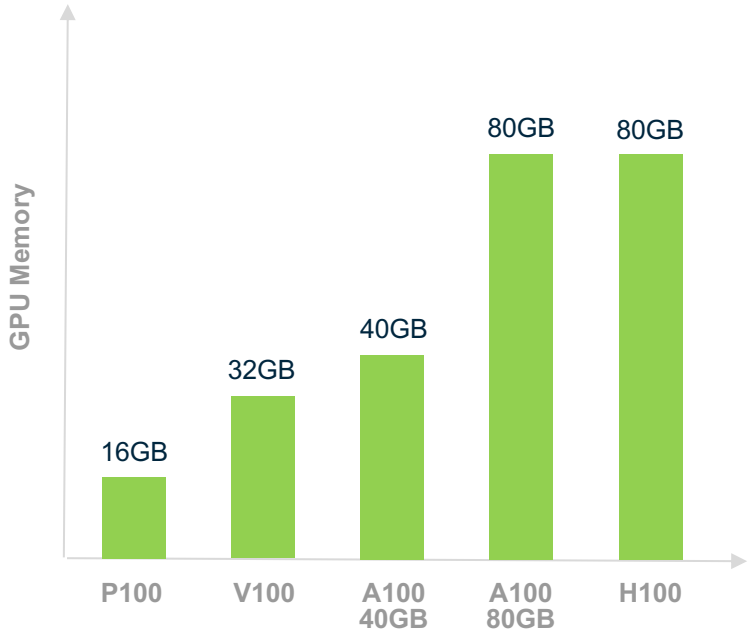


Higher Throughput

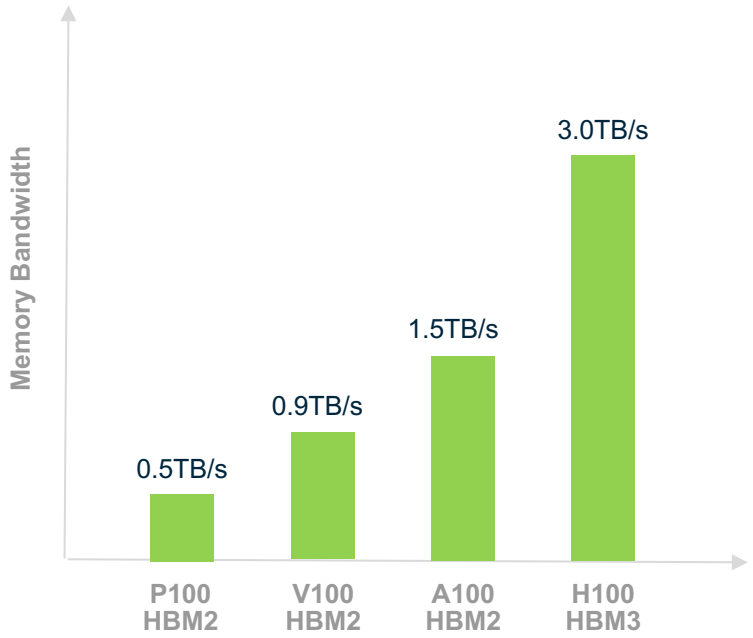
* https://developer.nvidia.com/blog/fueling-high-performance-computing-with-full-stack-innovation/

Larger GPUs with higher memory bandwidth

More GPU Memory



Faster GPU Memory



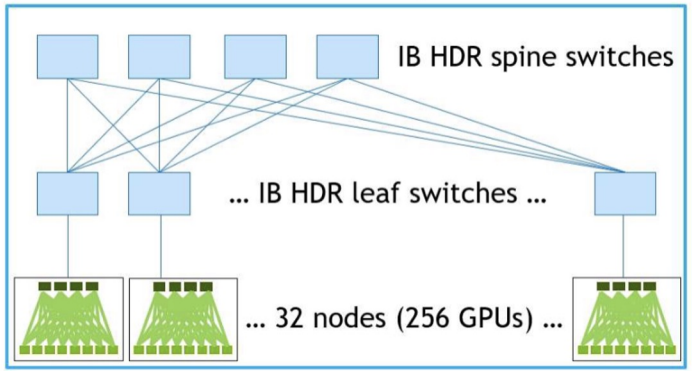
* https://resources.nvidia.com/en-us-tensor-core/gtc22-whitepaper-hopper

Increased performance for SuperPODs

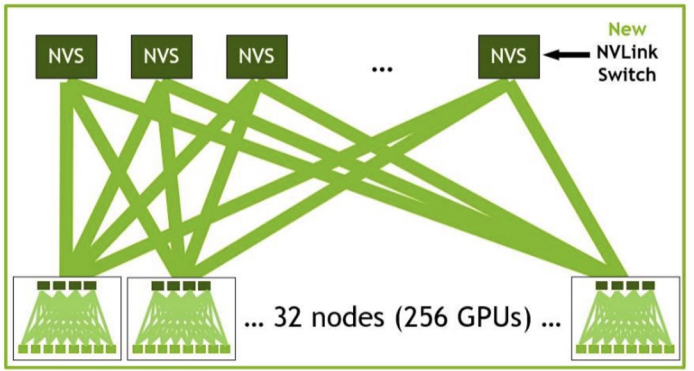
	A100 SuperPod			H100 SuperPod			Speedup	
	Dense PFLOP/s	Bisection [GB/s]	Reduce [GB/s]	Dense PFLOP/s	Bisection [GB/s]	Reduce [GB/s]	Bisection	Reduce
1 DGX / 8 GPUs	2.5	2,400	150	16	3,600	450	**1.5x**	**3x**
32 DGXs / 256 GPUs	80	6,400	100	512	57,600	450	**9x**	**4.5x**

- **9x in all-to-all exchanges**
- **4.5x in all reduce throughput**

32 A100 DGX SuperPOD

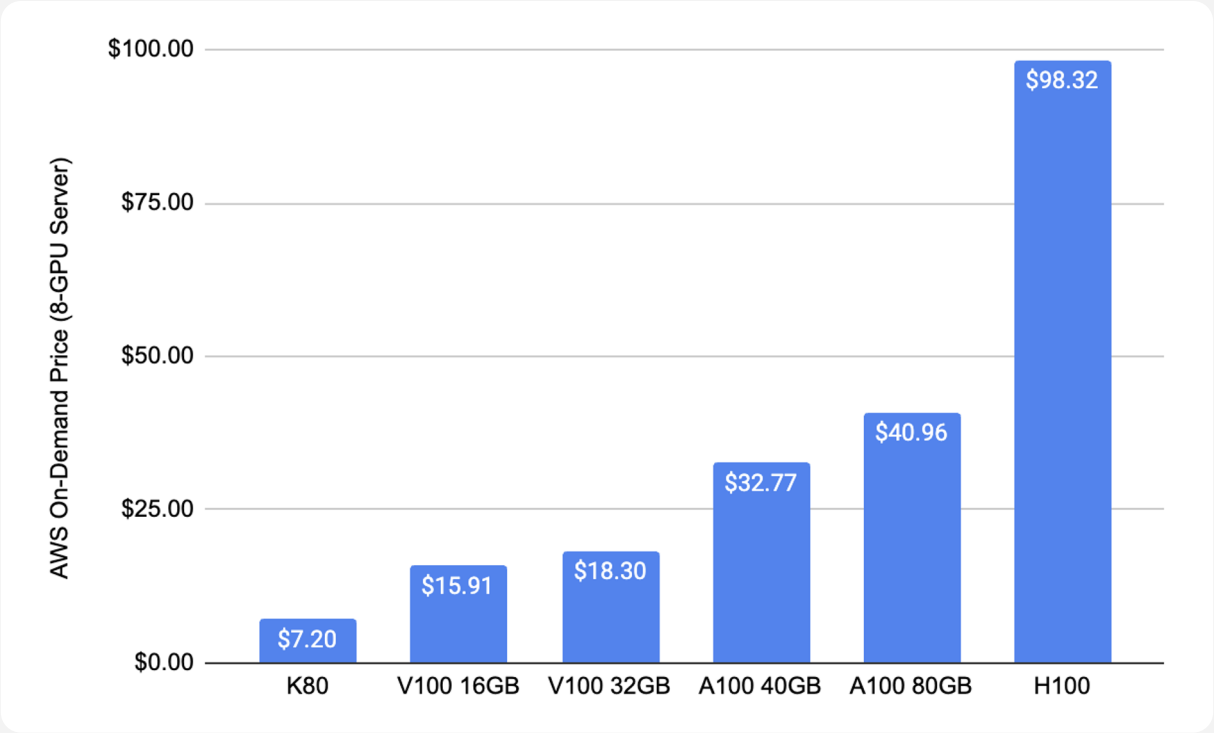


32 H100 DGX SuperPOD



* Data taken from https://resources.nvidia.com/en-us-tensor-core/gtc22-whitepaper-hopper

10x increase in GPU costs



Challenges with AI Infrastructure Management

Lack of controls and prioritization

Low utilization, high Cost

Visibility and better decision making

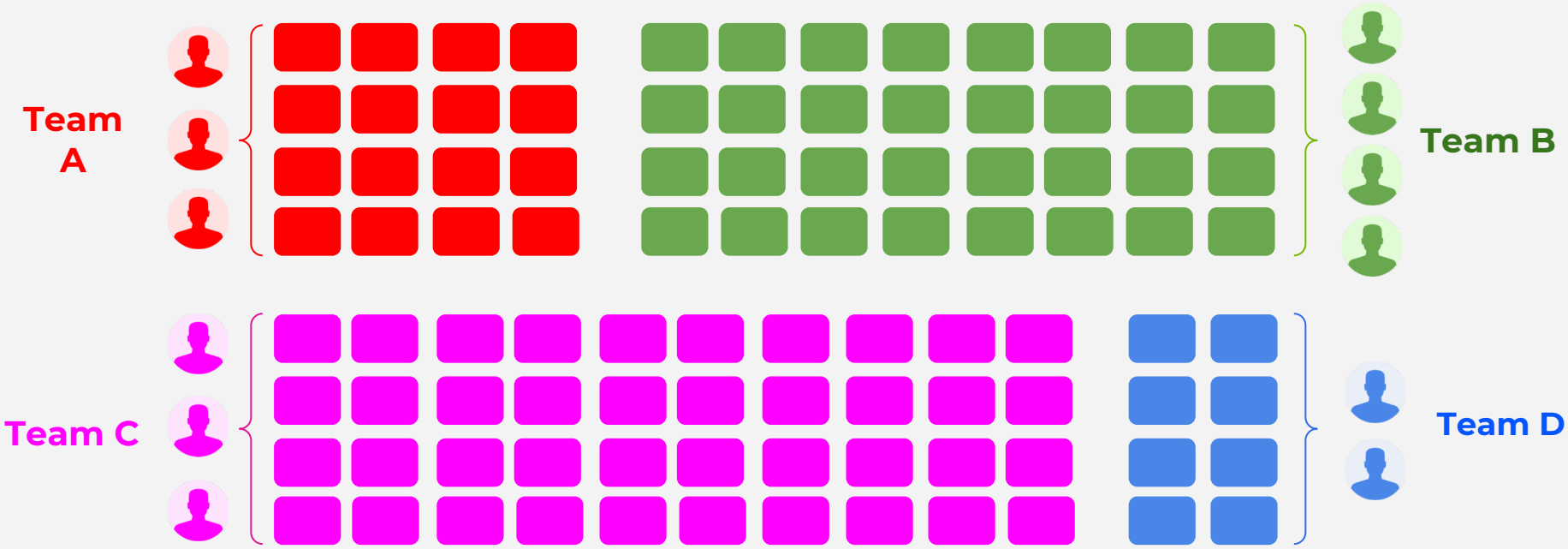
Users are still in need for more GPUs

GPU pooling – from siloed AI to collaborative efforts

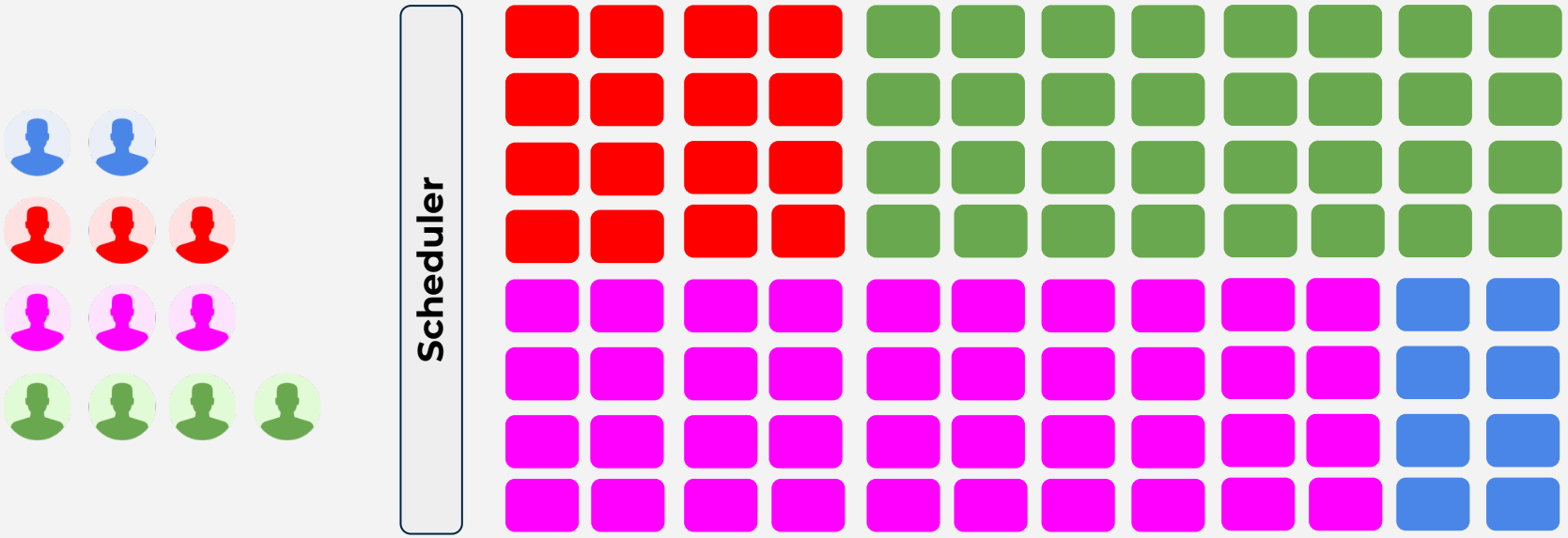
Siloed Infrastructure → **Shared Clusters**

On-Demand Compute → **Reserved Clusters**

GPU pooling



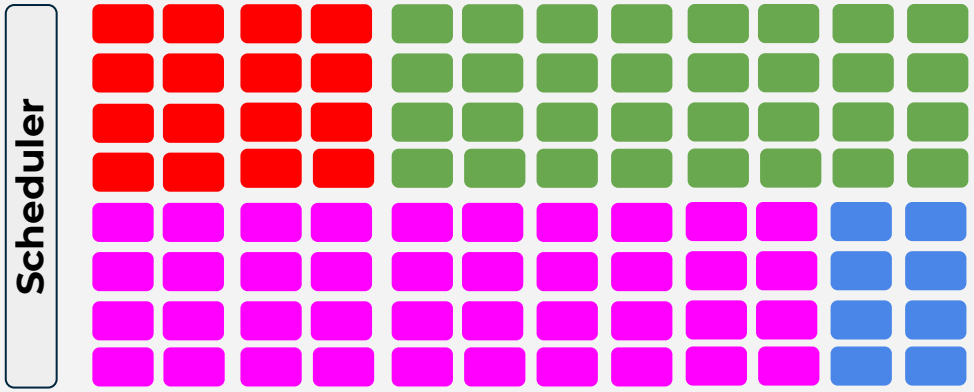
GPU pooling + Orchestration



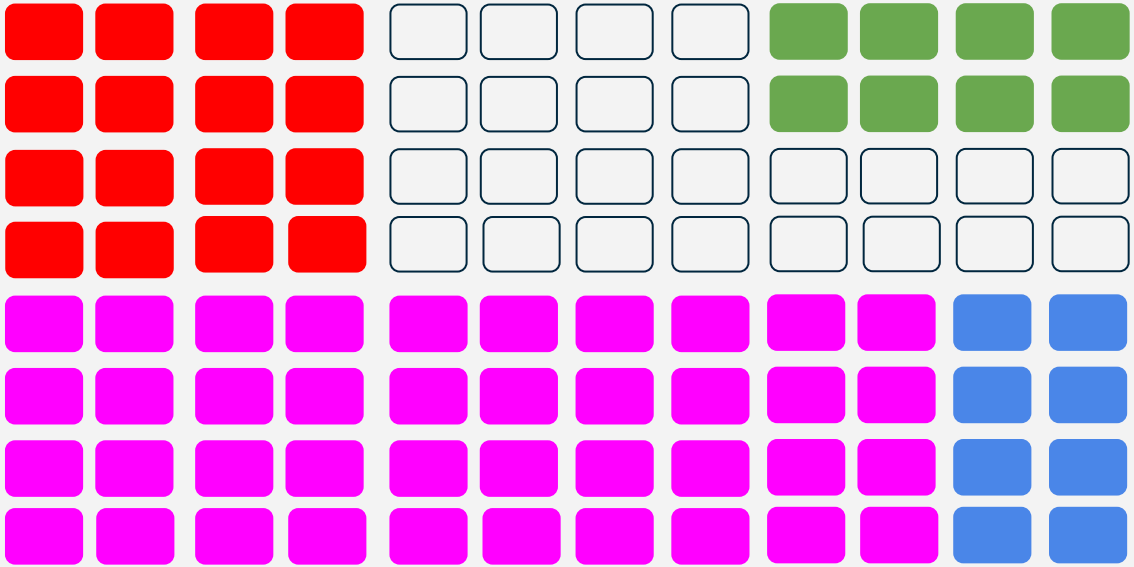
GPU pooling + Orchestration

Orchestration capabilities

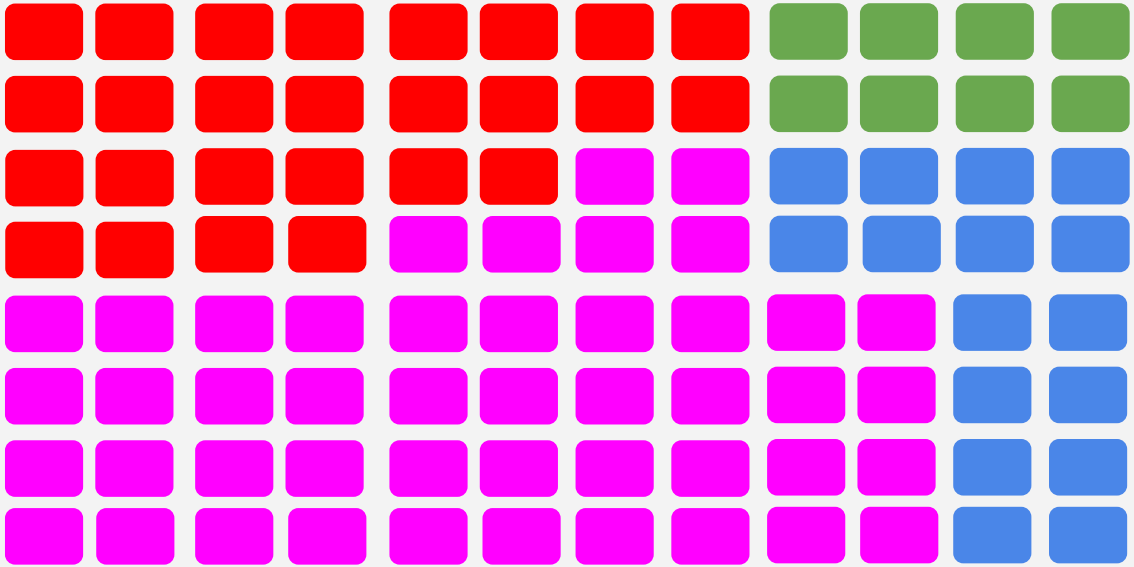
- Controls on resource allocations
- Workload prioritization
- Job queueing
- Workload monitoring and execution
- Accounting and reporting



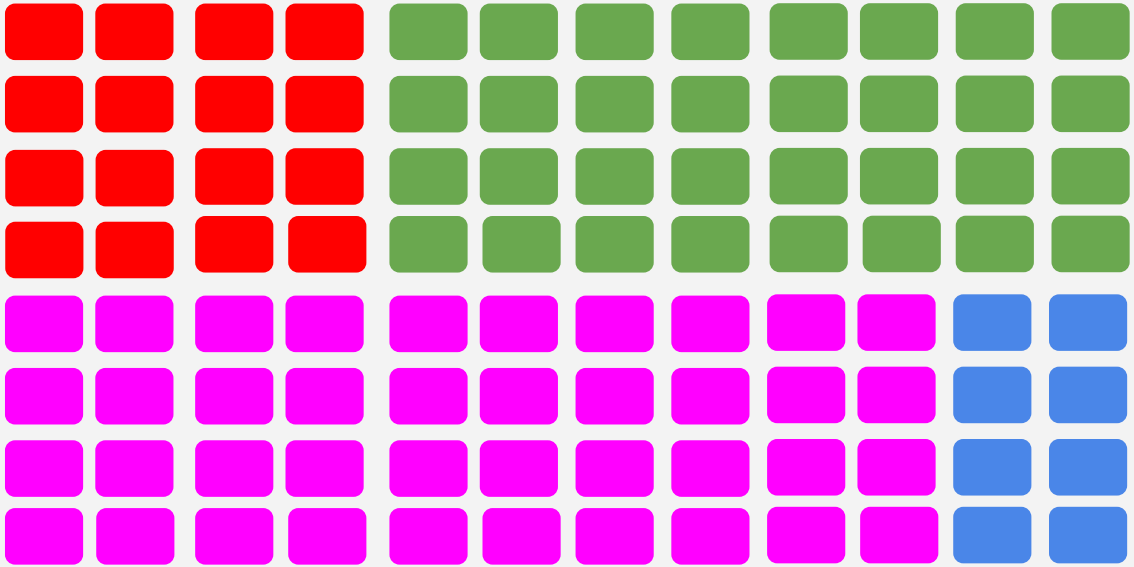
Repurposing resources between different **teams**



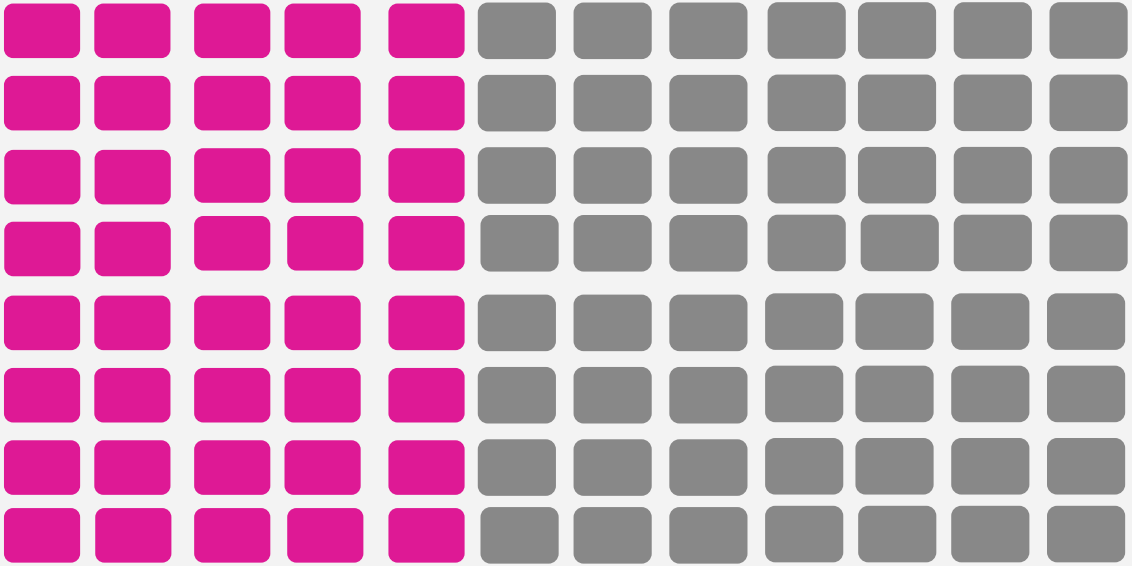
Repurposing resources between different **teams**



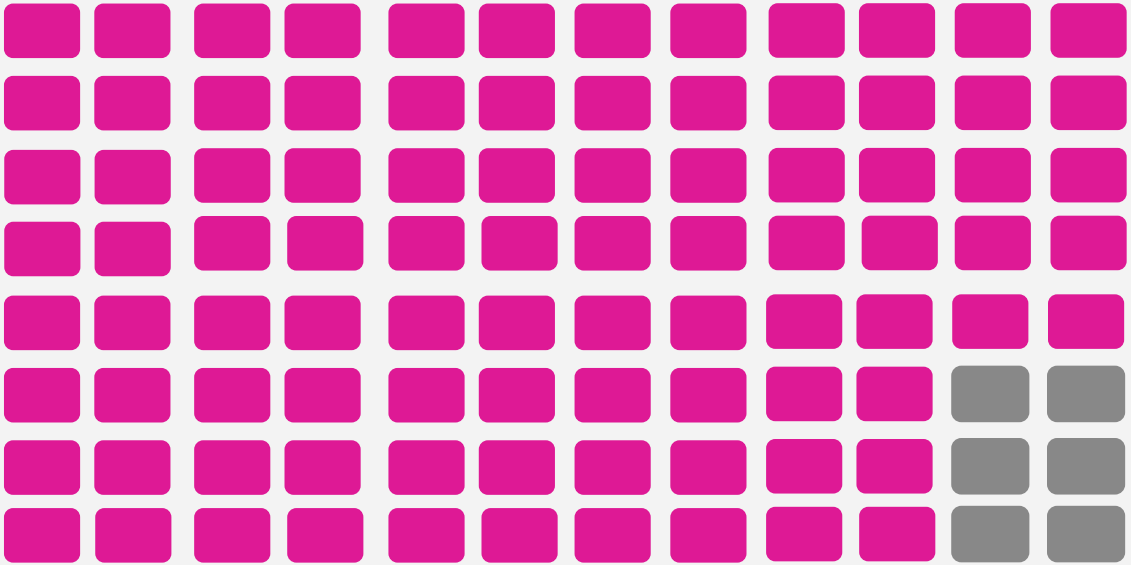
Repurposing resources between different **teams**



Repurposing resources between different **workloads**



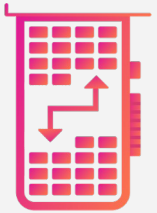
Repurposing resources between different **workloads**



Benefits

Higher Efficiency

Through sharing and repurposing resources

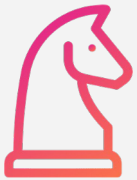


More GPU Accessibility

Users become more productive with easier access to more GPUs

Controls & Governance

Ability to align resources with business goals



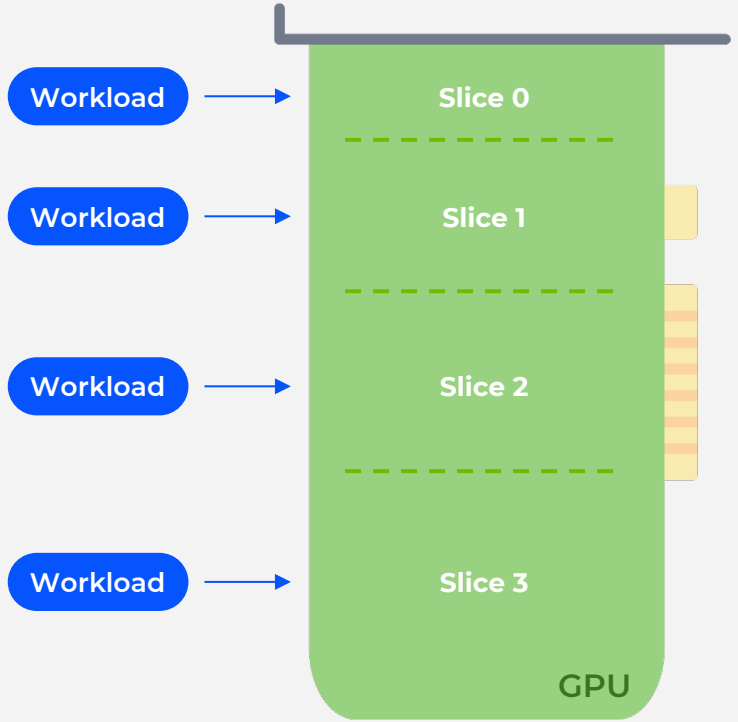
Centralized Visibility

Better planning and decision making

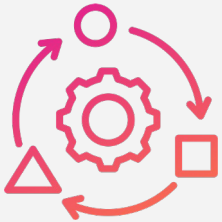
Not all workloads need whole powerful GPUs

Multiple workloads share a single GPU

- Notebooks
- Inference workloads
- GPU slicing methods:
  - Fractional GPUs (software isolation)
  - MIG (hardware isolation)

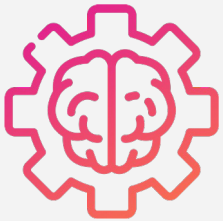


Support for the entire AI lifecycle



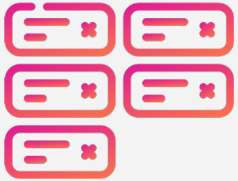
Model Development

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



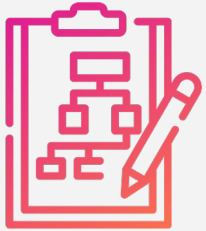
Fine-tuning & Training

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

Tooling

Training Framework

PyTorch

PyTorch Lightning

HOROVOD

DeepSpeed

NVIDIA NEMO

RAY

IDE Tools

jupyter

Visual Studio Code

PC

Experiment Tracking

W&B

comet

TensorBoard

mlflow

Pipeline Orchestration

Apache Airflow

Kubeflow

Serving Tools

vLLM

NVIDIA TRITON INFERENCE SERVER

TensorRT

Text Generation Inference

Open Source LLMs

Llama-2-7b-chat-hf

Llama-2-70b-hf

Falcon-40b-instruct

Mixtral-8x7B-Instruct-v0.1

The Keys for Operating AI Infrastructure Platforms

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

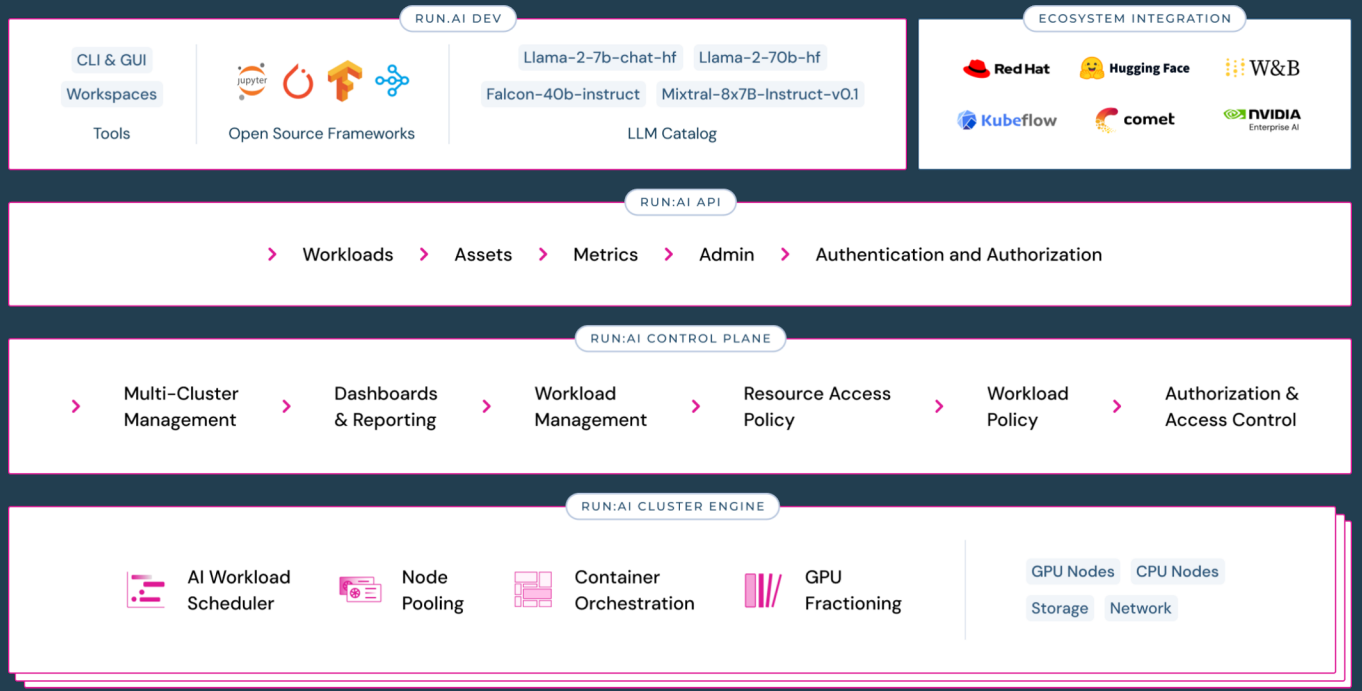
Fractional GPUs

Run more notebooks or inference servers on the same infrastructure

Tooling & Integrations

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

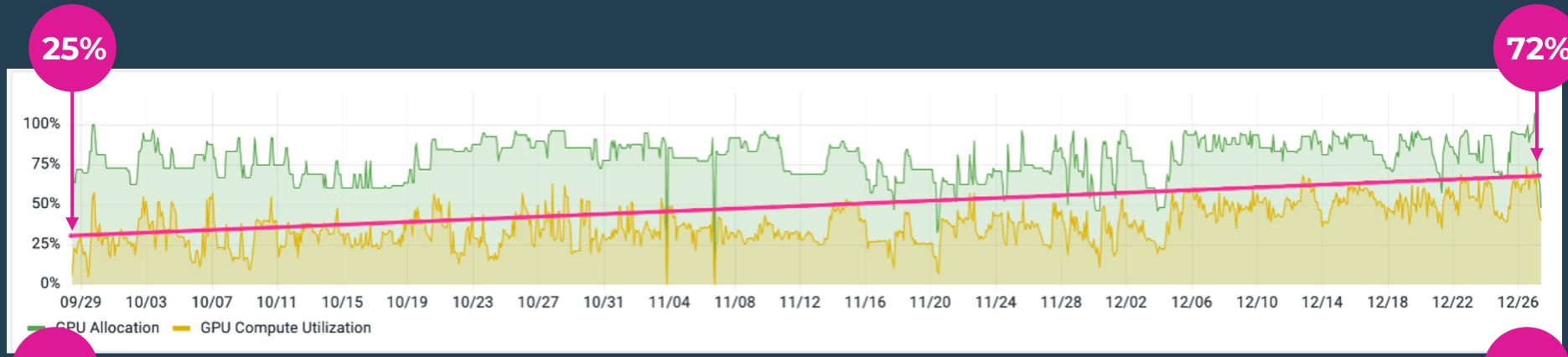
Introducing the Run:ai Platform

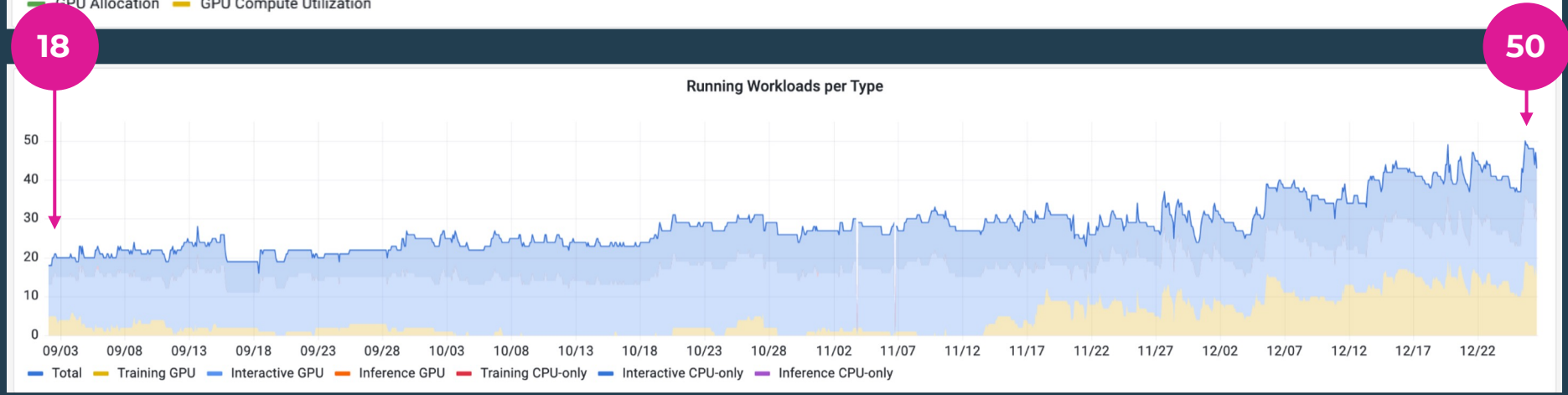


AI Infrastructure

Demo

Higher utilization, more workloads





run:ai

Visit us at
Booth 1408

Monday 10am
Accelerating AI Workflows on AI Data Center Infrastructure
Omri G. & Ersin Y. from Adobe

Tuesday 3pm
Throughput Performance Benchmarking: Pre-Training Foundational Large Language Models on Kubernetes
Ronen D. & Raz R.

Wednesday 2pm
Accelerating AI Workflows on AI Data Center Infrastructure
Ronen D. & Guy S.

On-Demand
Expert Perspectives on the Evolution of AI Infrastructure
Panel

On-Demand
Considerations for Choosing LLM Serving Technologies
Ekin K.