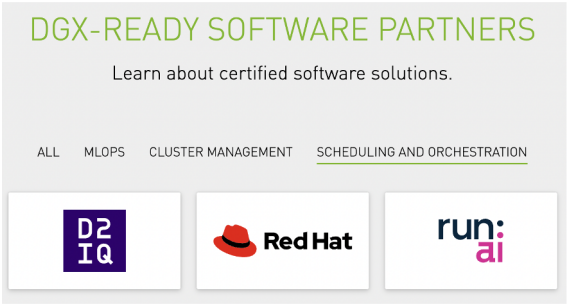
run:ai

Centralized Compute Management for the AI Era

Agenda

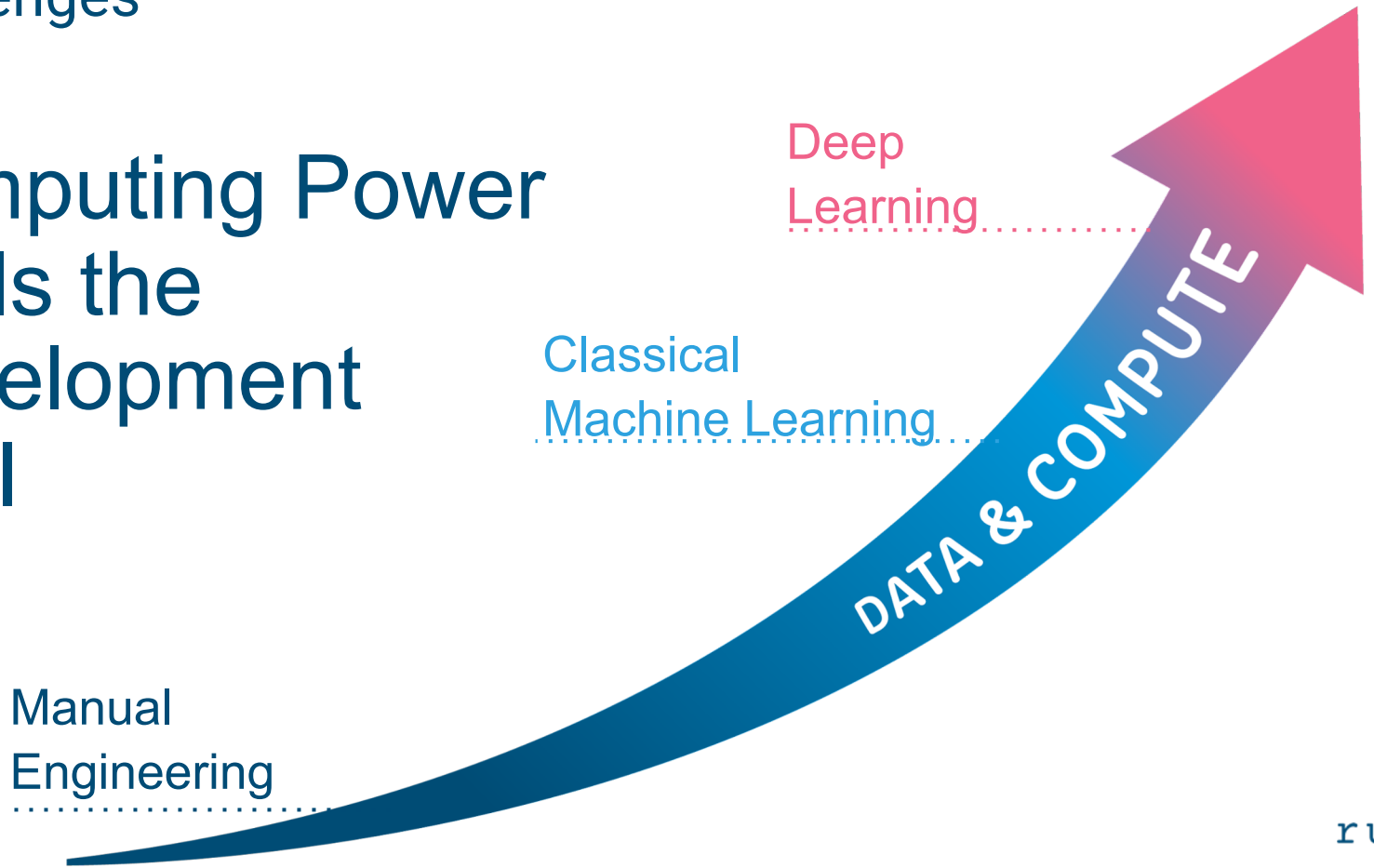
- Who are we?
- A glimpse into our customers
- Today’s challenges in the AI era
- RUN:AI’s unique proposition
- RUN:AI for its partners

Run:AI at-a-Glance



- HQ is in Israel with offices in NY and Boston
- Approx 100 employees mostly R&D, growing rapidly…
- Our own IP
- Experts in K8's & cloud native technologies (building GPU clusters for more than 3 years)
- **DGX certified software vendor**
- **Premier Inception partner with nVIDIA**
- Active customers in ALL leading AI verticals (Healthcare, Automotive, Defence, Oil & Gas, Research, Retail and FSI)

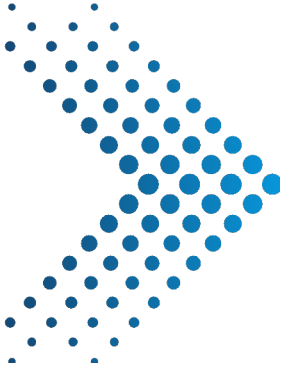
Computing Power Fuels the Development of AI



Challenges

From Traditional Software Engineering to Data Science

Interactive Development

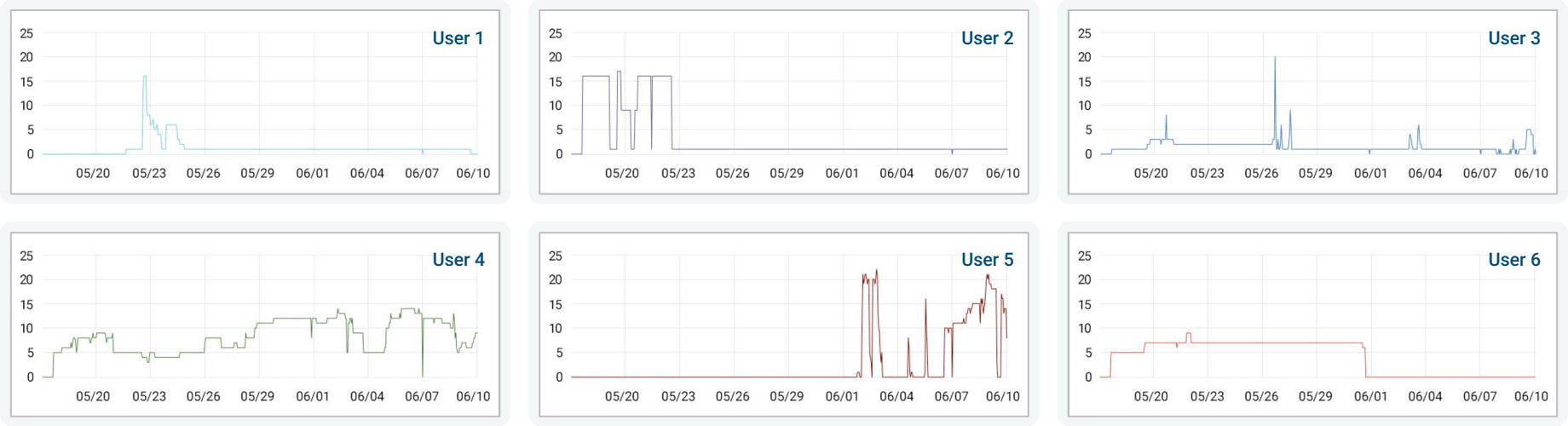


Experimentation

Challenges

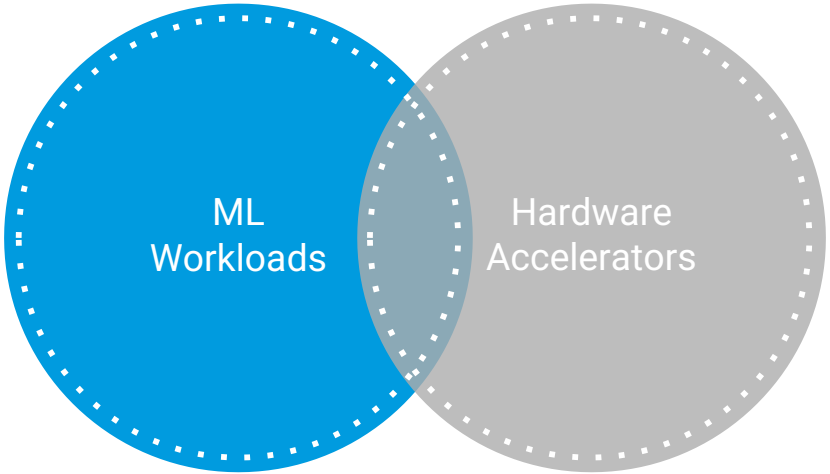
Data Scientists Consume **Massive Computing Power for Days** and Then are Idle for Weeks

GPU allocation, different users, 24 days



Challenges

Machine Learning Workflows and Hardware Accelerators Are Highly Coupled!



Under-utilized GPUs

Workflow Limitations

Constant hassles

Kubernetes, the “De-facto” Standard for Container Orchestration

Lacks the following capabilities:

- Multiple queues
- Automatic queueing/de-queueing
- Advanced priorities & policies
- Advanced scheduling algorithms
- Affinity-aware scheduling
- Efficient management of distributed workloads
- Fractional GPU’s

Gain Visibility and Control over AI Workloads to Increase GPU Utilization

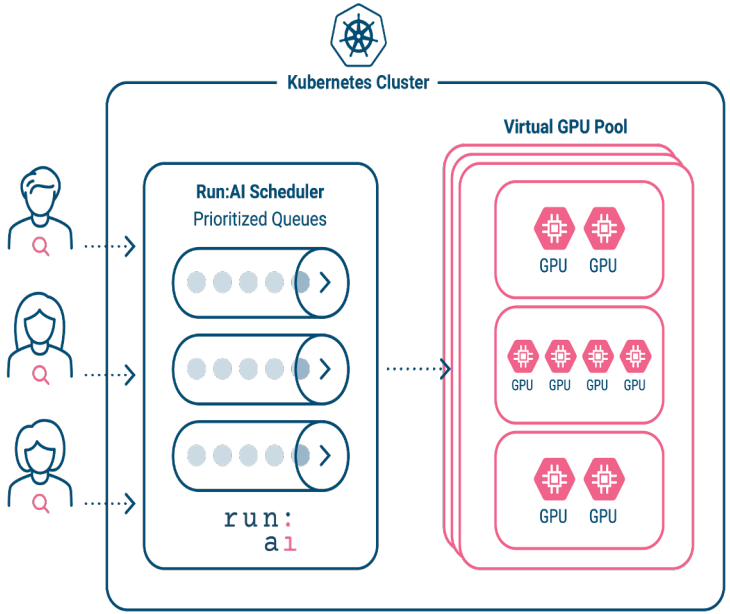
- Improved Cluster utilization
- Faster time to market
- Simplified Infrastructure management

Run:AI at-a-Glance

- Innovation & IP
  - Cloud-native high performance scheduler for AI/HPC
  - Fractional GPUs
  - Dynamic MIG allocation
  - Simple control, orchestration, management & monitoring interface

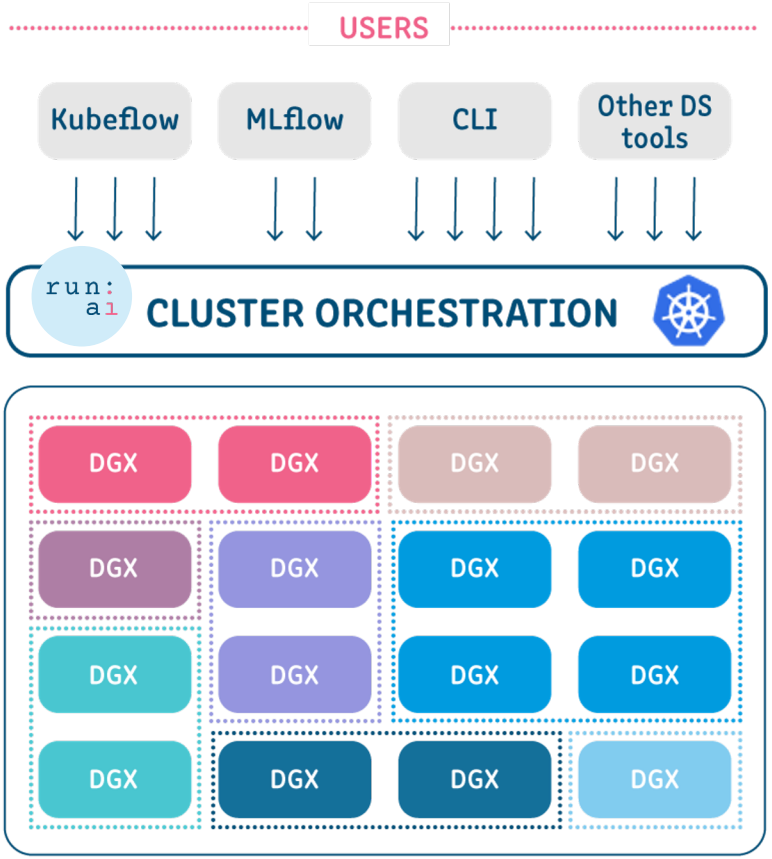
Run:AI Super Scheduler

- Guaranteed quotas:
  - Over quota system
  - Multiple queues
  - Automatic preemption
  - Presets, Priorities, policies & fairness
- Gang scheduling
- Topology-aware scheduling



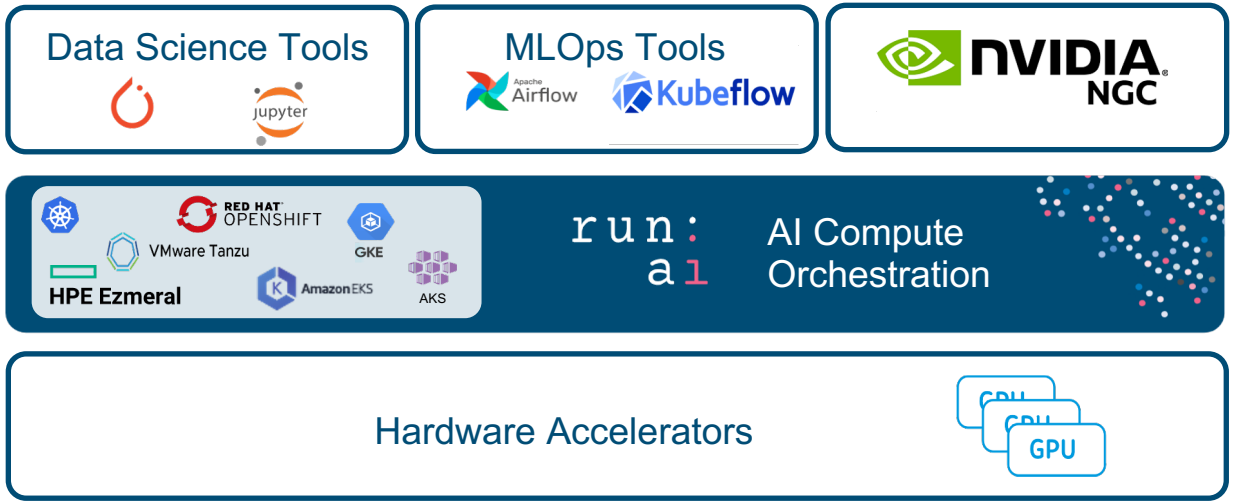
AI - Dedicated infrastructure requires centralized high performance GPU orchestration

- Efficient resource sharing between multiple data science tools & between multiple users – no static allocations
- The right allocation of resources for every user based on pre-set policies
- Centralized monitoring and control



Implemented as a Kubernetes Plug-in, Easily Integrating with Your Existing AI Stack

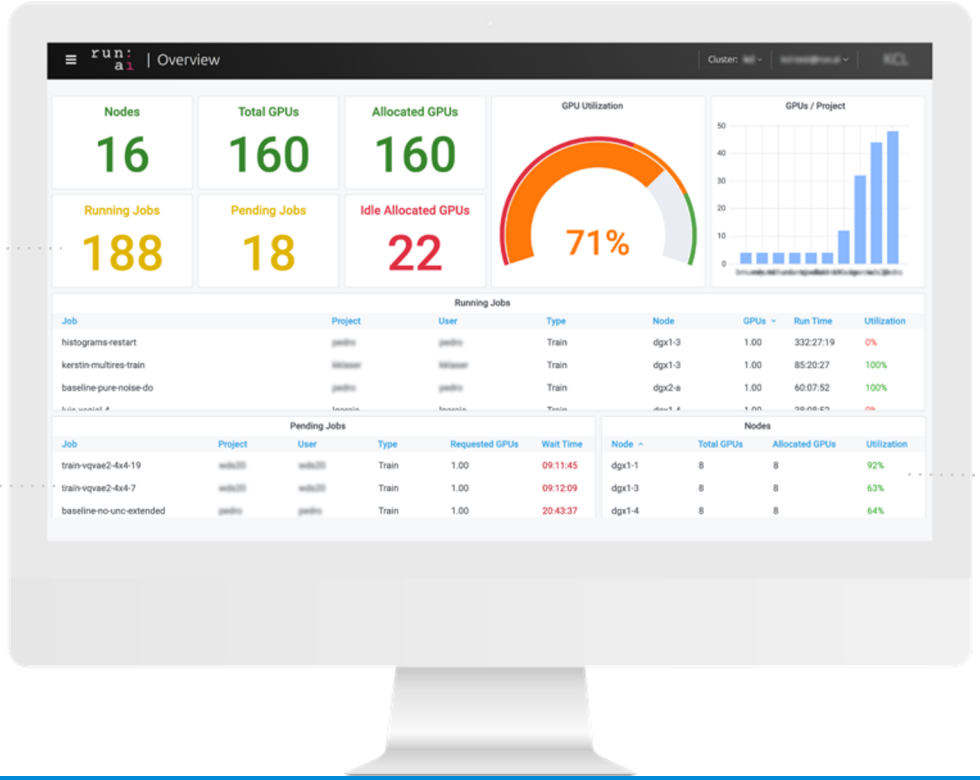
- Super Scheduler
- Fractional GPUs
- Dynamic MIG allocation
- Control, management, and monitoring



Management Tools: Control & Visibility Across Clusters

See Jobs, Projects and Resources

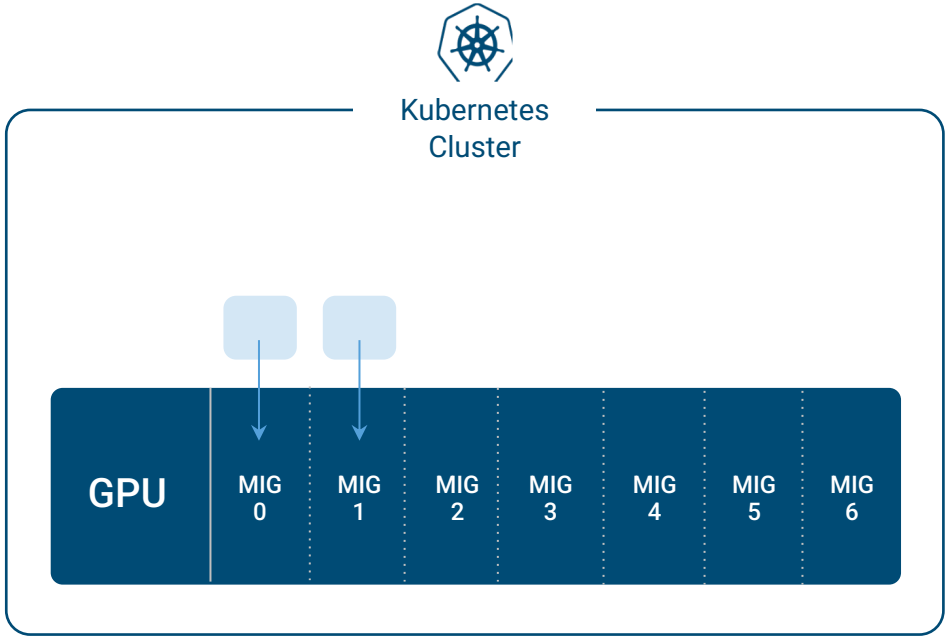
Queued Jobs



View Entire Cluster

Multi-Instance GPU (MIG)

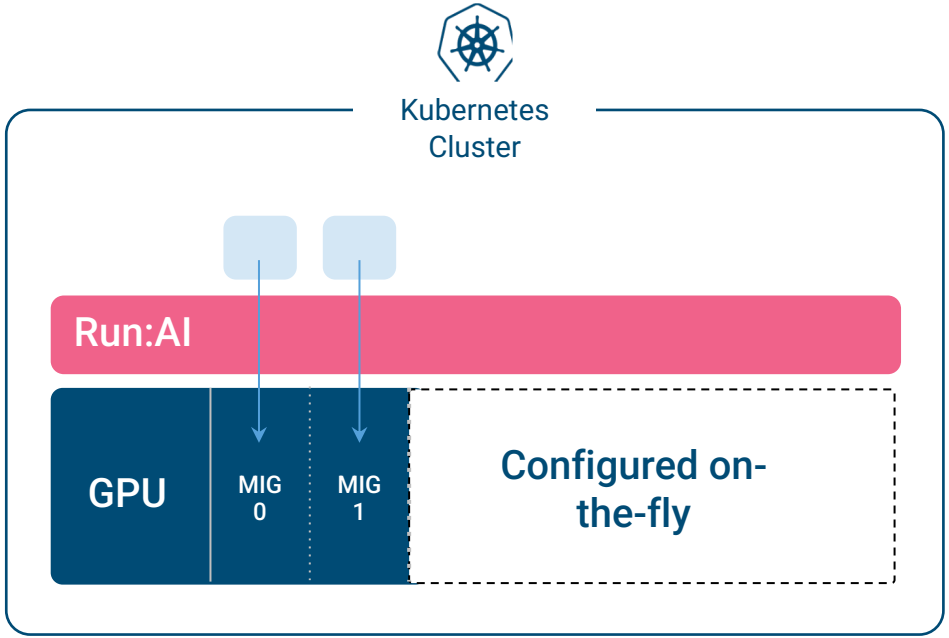
- **Available on A100 and A30**
- **Limited to up to 7 slices**
- **Configured manually and statically:**
  - Configured on node provisioning, based on demand forecasting
  - Users need to be aware of the available MIG slices
  - Changes require admin permissions and draining all running workloads



Run:AI's Dynamic MIG Allocations

MIG slices are configured automatically according to demand, on job submission, rather than on node provisioning

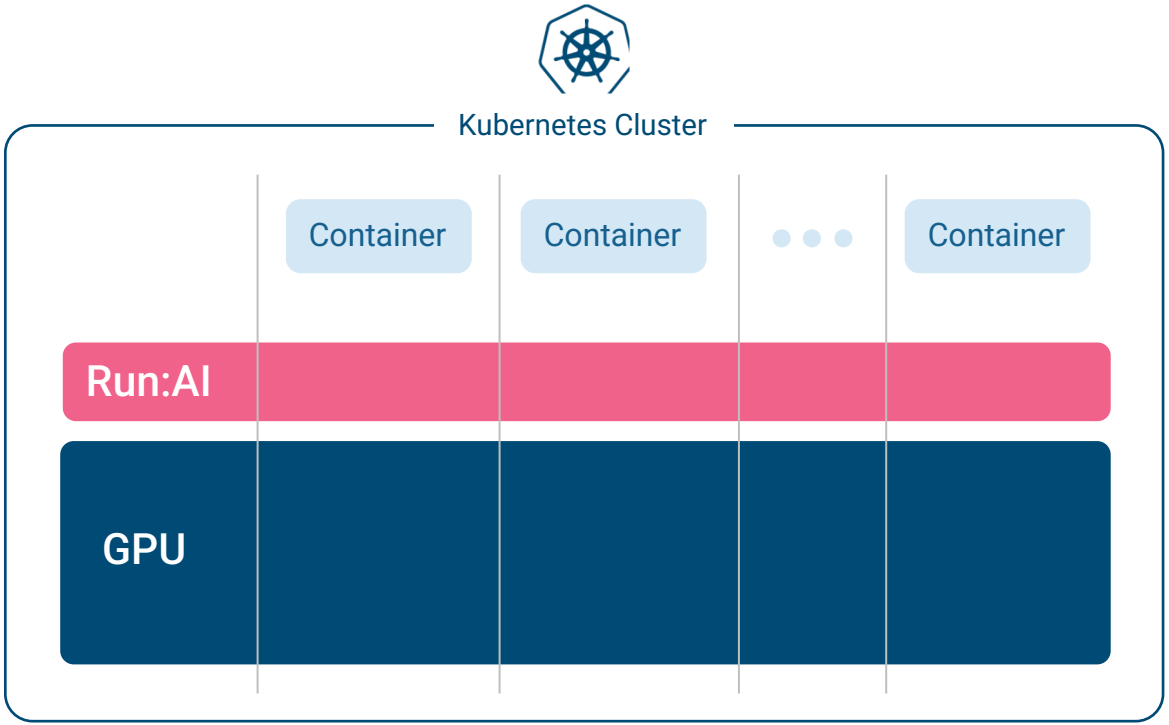
- Simpler infrastructure management
- Increased cluster utilization and decreased GPU fragmentation
- More flexibility to developers



Fractional GPUs

Multiple containers share a single GPU

- Supported by the Run:AI Kubernetes scheduler
- Intercepting CUDA API calls to manage memory and compute performance

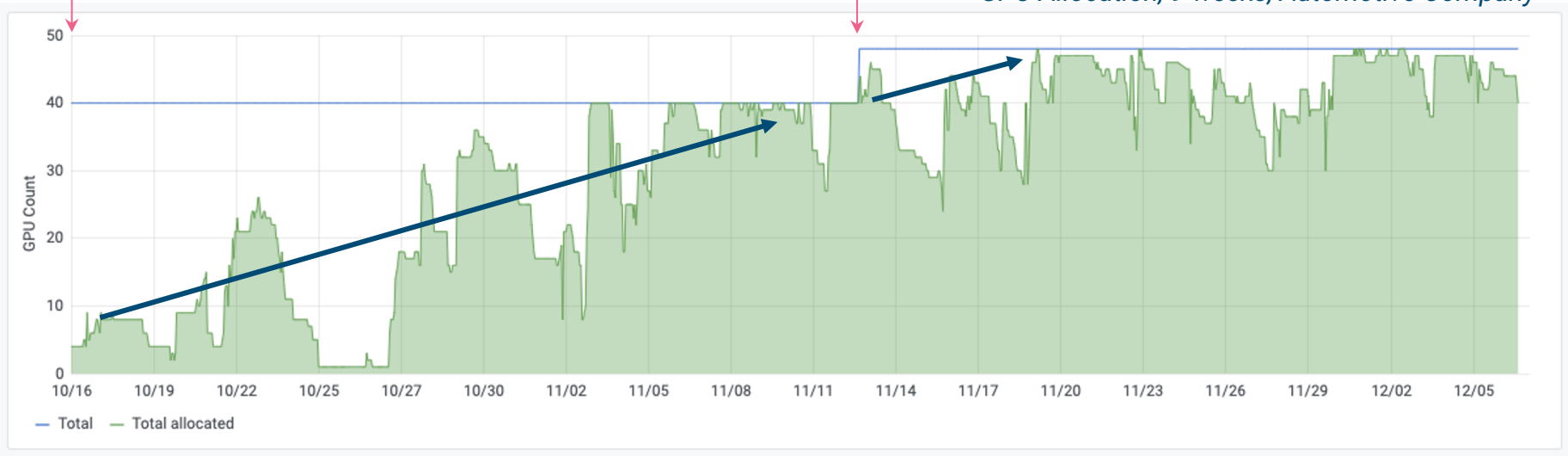


Data Scientists Consume As Much GPU Power As They Can

Run:AI onboarding

Adding another server

GPU Allocation, 9 weeks, Automotive Company



Data Scientists LOVE Run:AI !

- Accelerated AI Development
- Improved cluster utilization
- End-to-end visibility & control

> “With Run:AI we’ve seen great improvements in speed of experimentation and GPU hardware utilization. Average experimentation time reduced from 49 days to 2 days”
>
> ***Dr. M Jorge Cardoso, London AI Center for Healthcare, CTO***

KING’S College LONDON

IDC Data Award for building state-of-the-art AI platform

Centralized Compute Management for the AI Era

- Fully automate GPU provisioning and allocation
- GPU compute you need, precisely when you need it
- IT gains visibility and maximizes GPU utilization
- Enable one-click distributed computing, as well as fractional GPU allocation
- Support air-gapped, On-Prem, Hybrid and Cloud infrastructure.

IMPROVED UTILIZATION

with batch scheduling and GPU virtualization

FASTER TIME-TO-VALUE

with distributed training

FULL CONTROL & VISIBILITY

with automated policy enforcement & monitoring

With Run:AI

Maximize GPU infrastructure efficiency

Accelerate the evolution to cloud native infrastructure through hybrid clouds solution

Speed up experimentation process and time to market

www.run.ai

run:ai

run:ai

Thank you