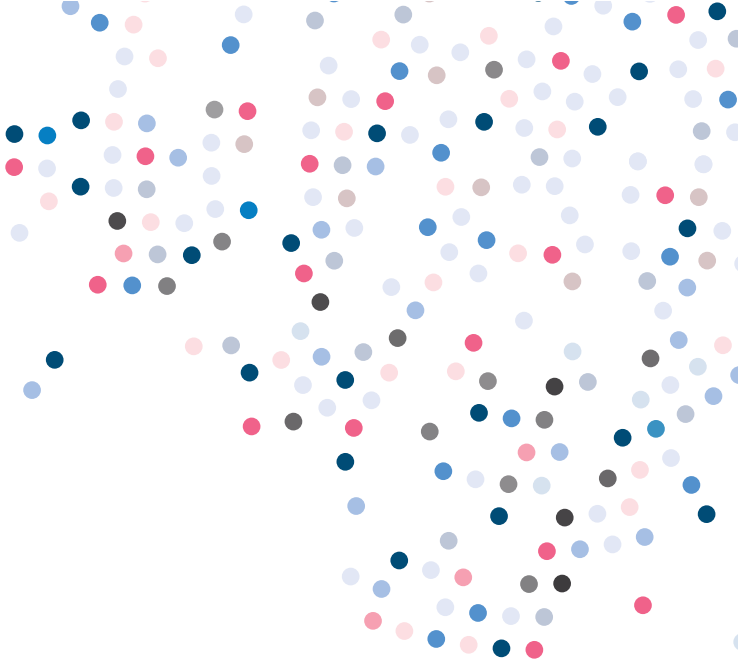
run:ai

Can New Approaches to GPU Machine Management Speed Delivery of Deep Learning Projects?

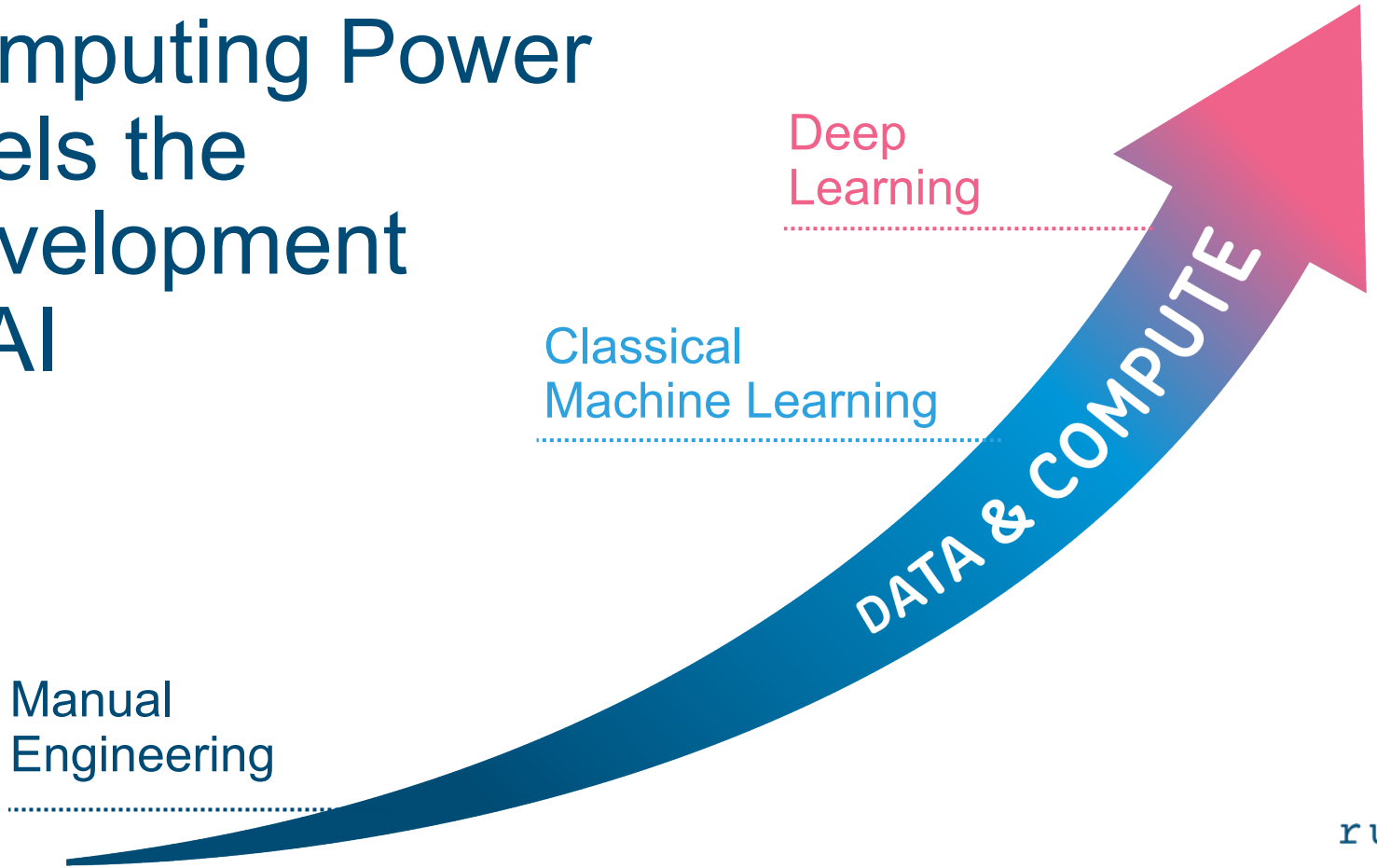
AI Summit 2019

Outline

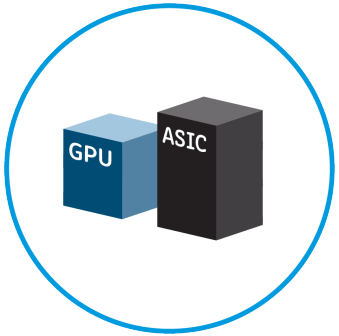
- Introduction
- Today’s Challenges
- Building a better compute workflow
- Use case



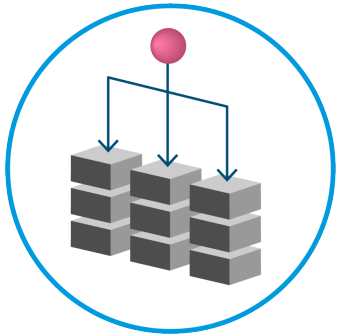
Computing Power Fuels the Development of AI



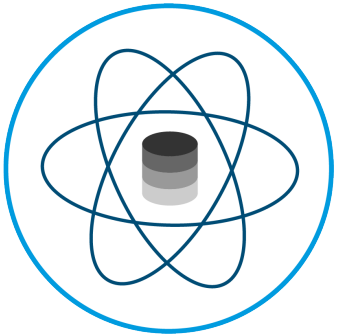
Artificial Intelligence is a Completely Different Ballgame



New accelerators

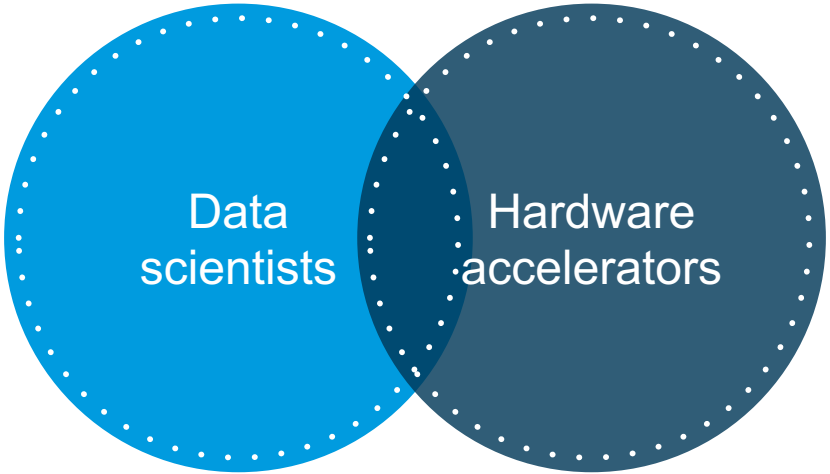


Distributed computing



Data science

Data Science Workflows and Hardware Accelerators are Highly Coupled

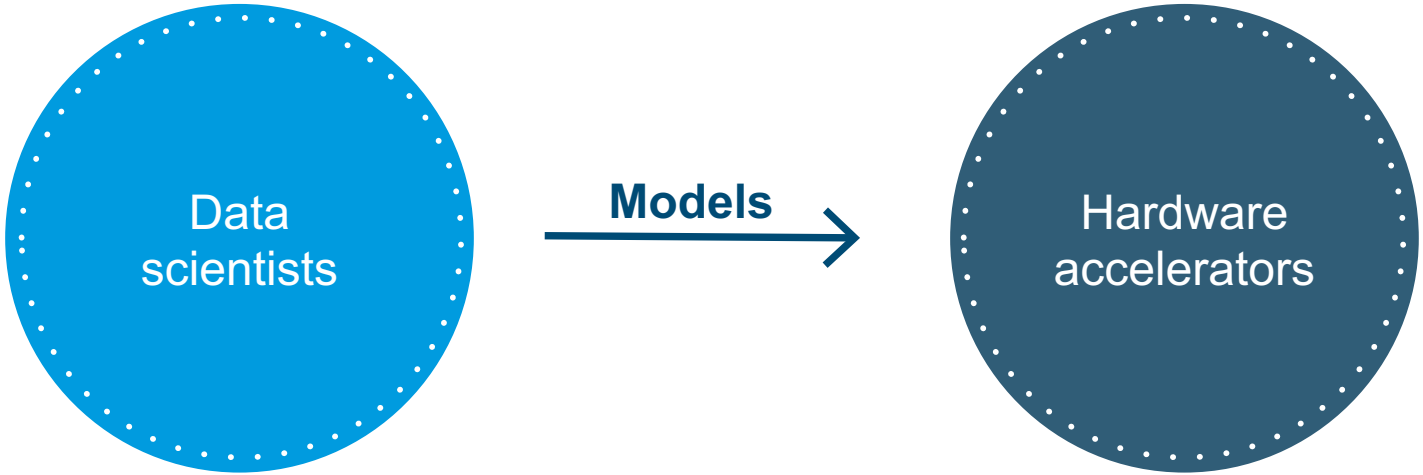


Constant hassles

Workflow Limitations

Under-utilized GPUs

The Run:Ai Vision – Full Hardware Abstraction

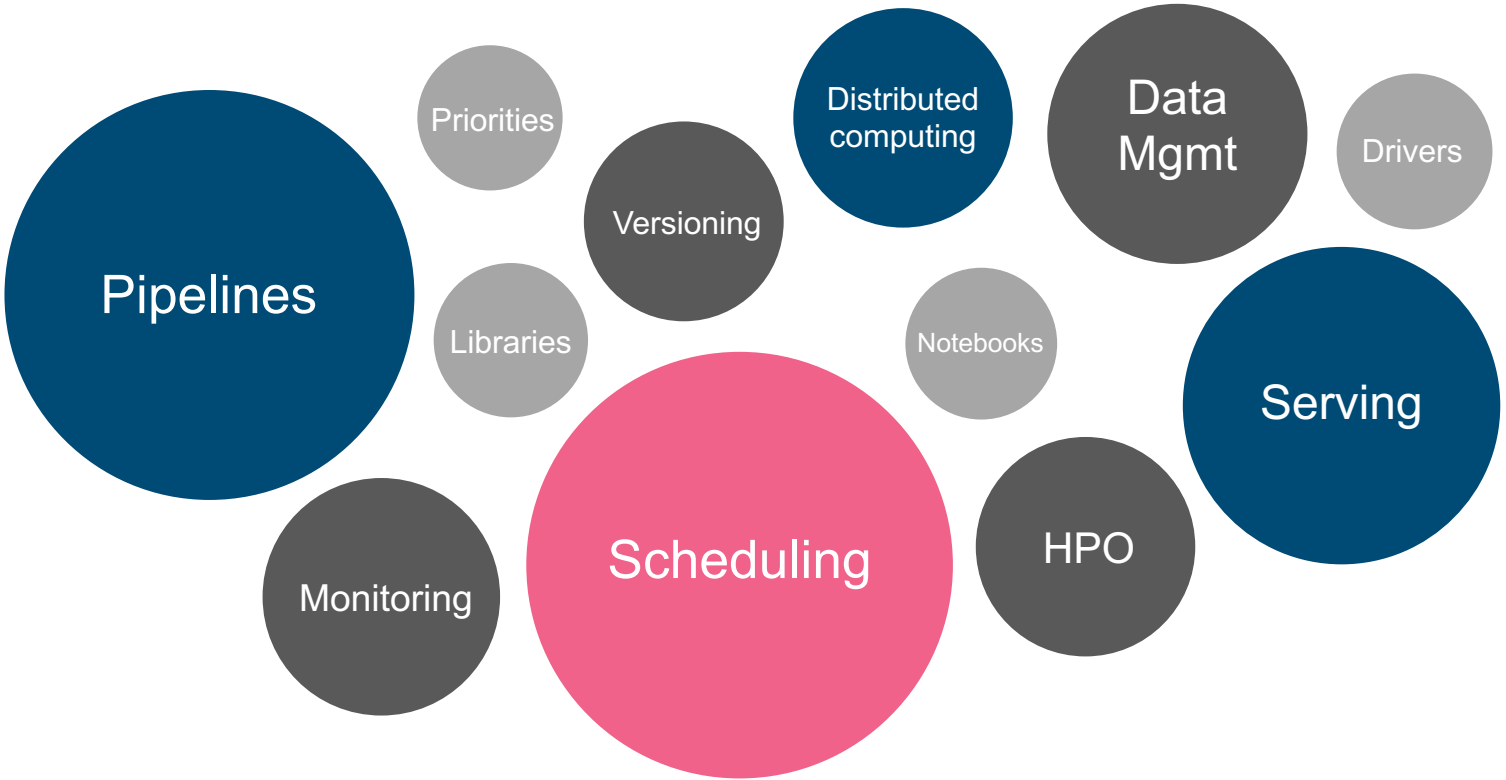


The Ideal AI Infrastructure

- Elastic
- Dynamic allocations
- Large-scale distributed computing
- Automated per business goals
- Run anywhere, anytime
- Virtualized

run:ai

But in real life …

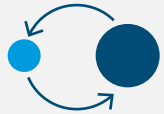


Kubernetes, the “De-facto” Standard for Container Orchestration

Lacks the following capabilities:

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

Distinguishing Between Build and Training Workflows

Build 

- **Development & debugging**
- **Interactive** sessions
- **Short** cycles
- Performance is **less** important
- **Low** GPU utilization

Training 

- **Training & HPO**
- **Remote** execution
- **Long** workloads
- **Throughput** is highly important
- **High** GPU utilization

Fixed vs. Guaranteed Quotas

Fixed quotas

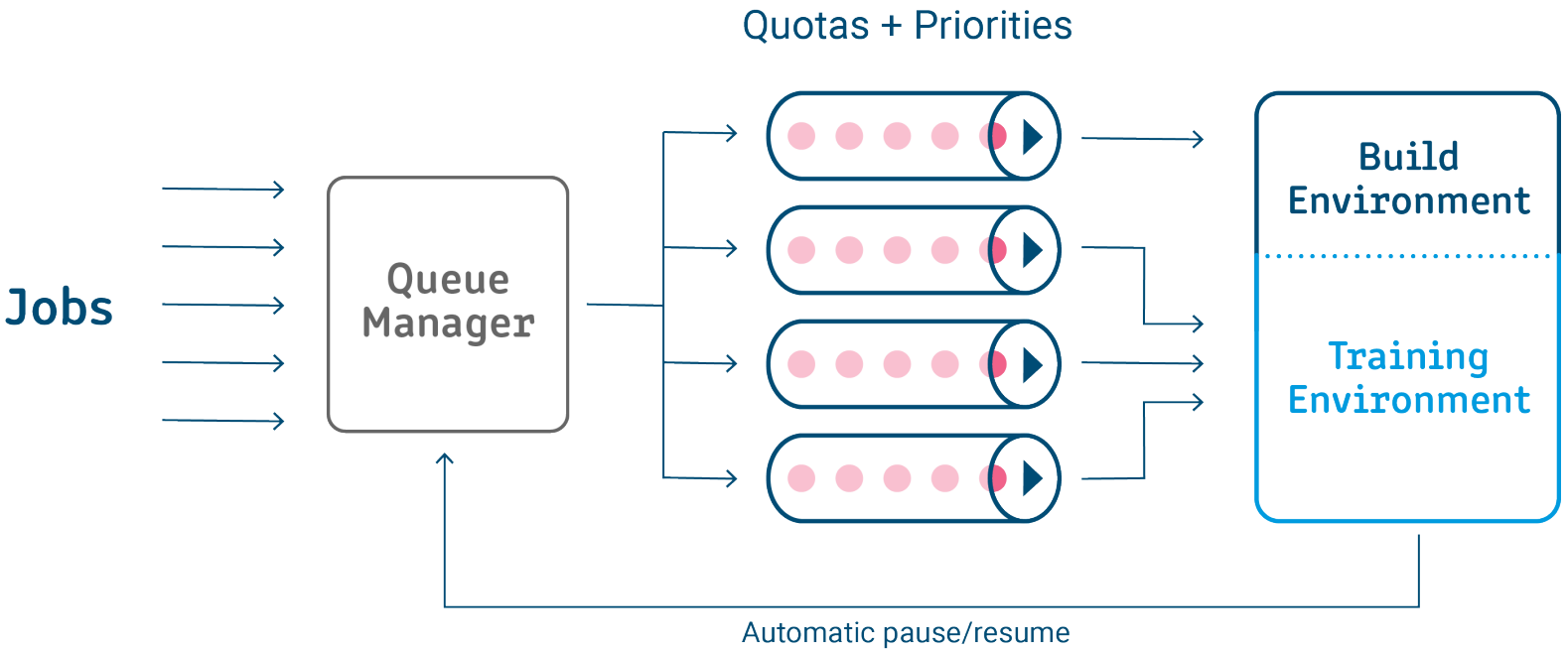
- Fits build workloads
- GPUs are always available

Guaranteed quotas

- Fits training workflows
- Users can go over quota

- More concurrent experiments
- More multi-GPU training

Queueing Management Mechanism



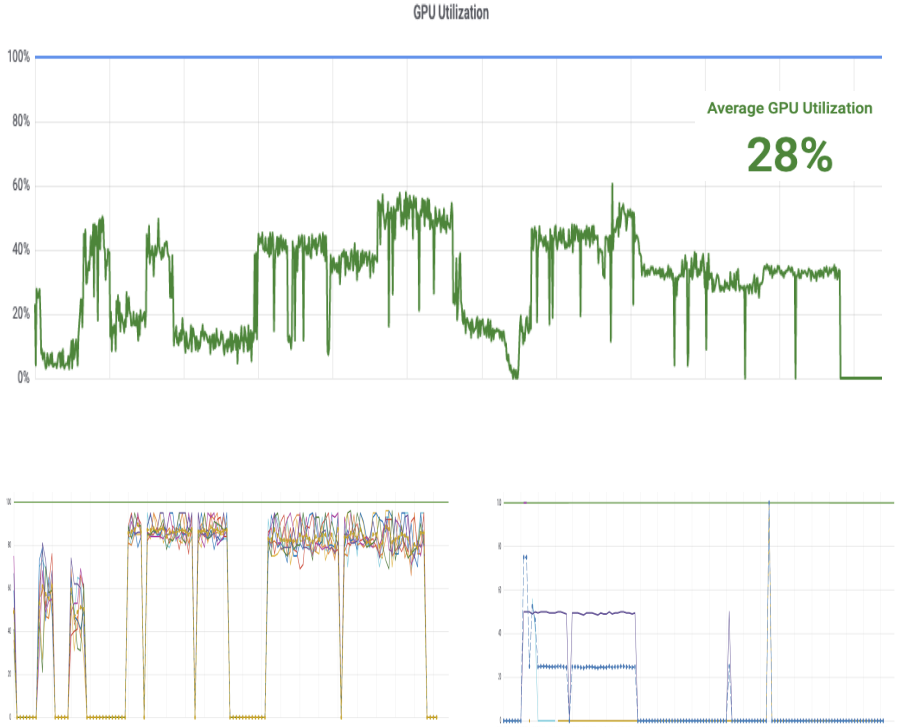
Use Case: Technology Enterprise

- Team of 20 deep learning researchers
- Dozens of GPUs on premises:
  - High-end DGX servers
  - Low-end GPU servers
  - Workstations
- Static allocations

Problem:

Unscalable system, constantly buying additional GPUs

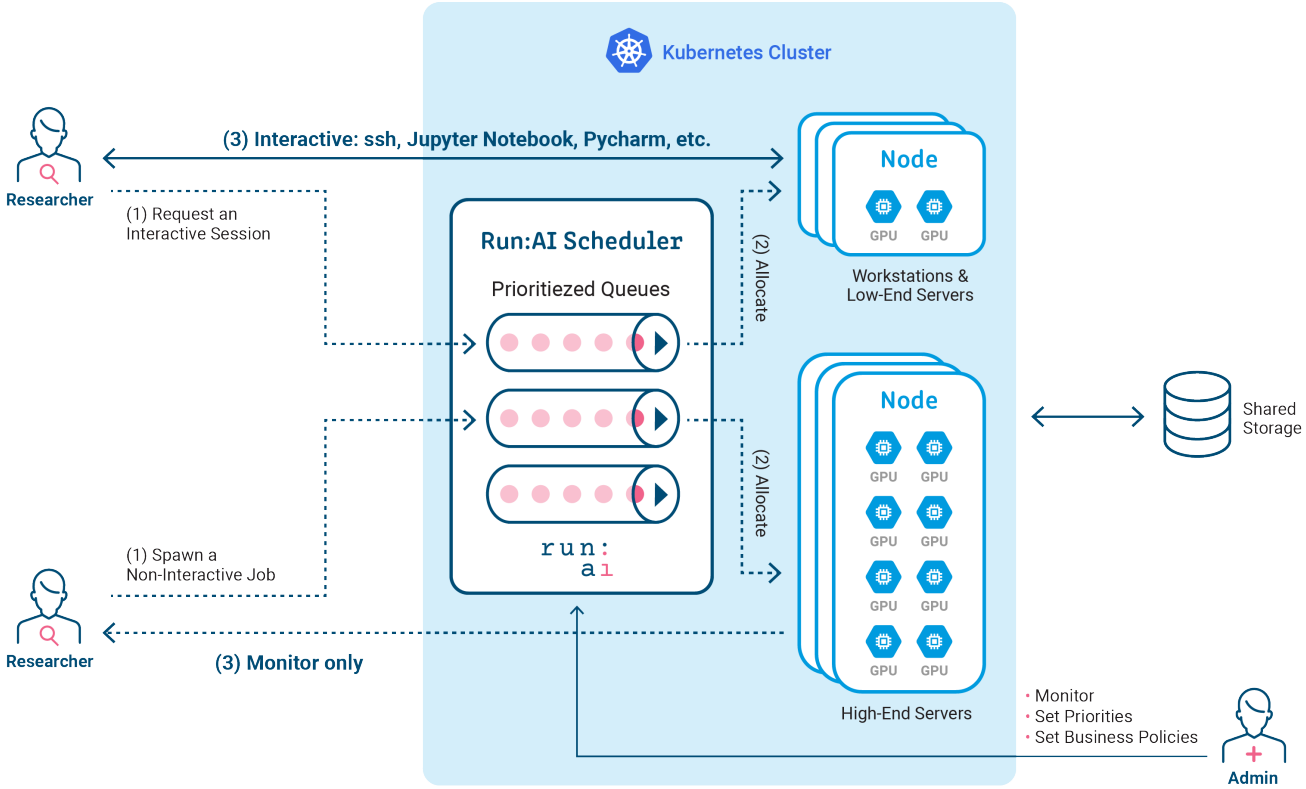
Use Case: What We Discovered



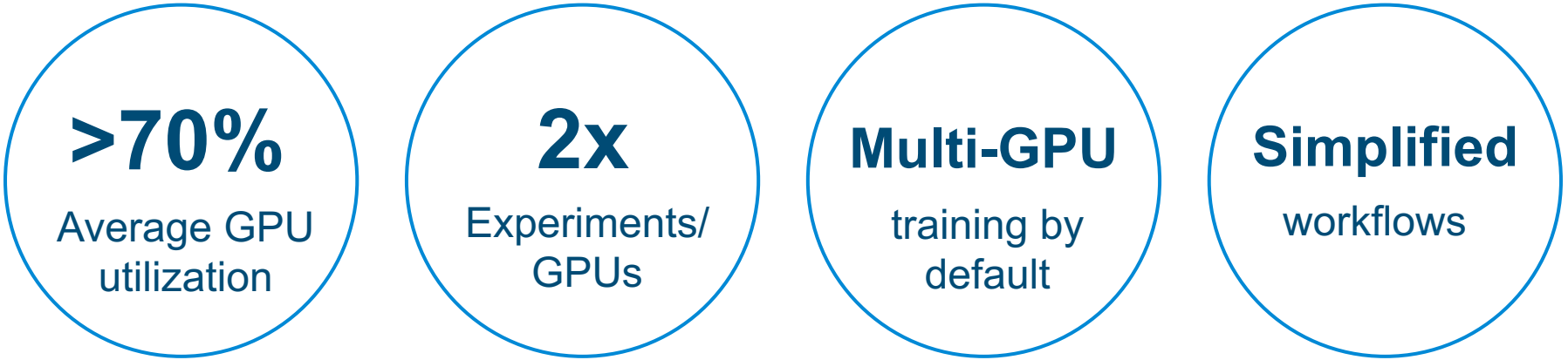
LOW GPU UTILIZATION
Some peaks, but mostly inefficient system and unused resources

DIFFERENT USAGE PROFILES
'Build' 'Train' 'Retrain' with very different needs

New Architecture: Virtual Pool of GPU Machines



4 Months Later



run:ai

Thank you

Contact: omri@run.ai