Accelerating Inferencing Rag Pipeline With The WEKA Data Platform

WHAT'S CHANGING?

FROM

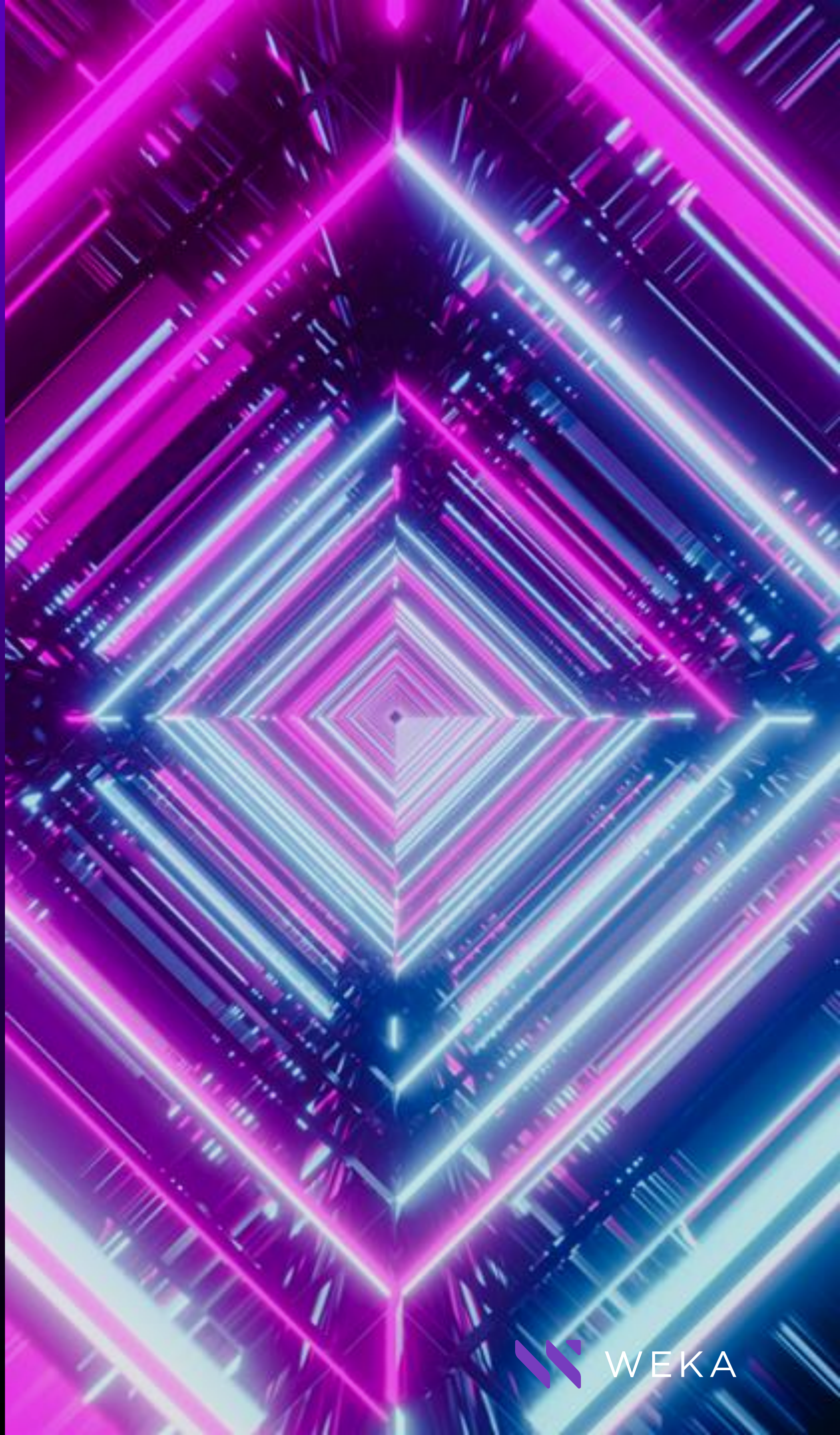
- AI Science Projects
- Practitioner Grade AI
- Petabyte Scale
- Foundational Models

TO

- Core Operational Assets
- Enterprise Grade AI
- Exascale
- Pre-trained Models with Domain-specific tuning

WHAT'S STAYING THE SAME?

- Rackspace & Power Continue to be Limiting Factors
- Our Appetite for Processing Power Continues to Increase
- Access to AI Hardware Remains a Limiting Factor
- On-premises/Cloud/Hybrid Capable Infrastructure is Dominant
- Ongoing Demand for Operational Simplicity



HOW DO WE MEET THE NEEDS OF TOMORROW AND TODAY?

INDUSTRY-LEADING PERFORMANCE & EFFICIENCY

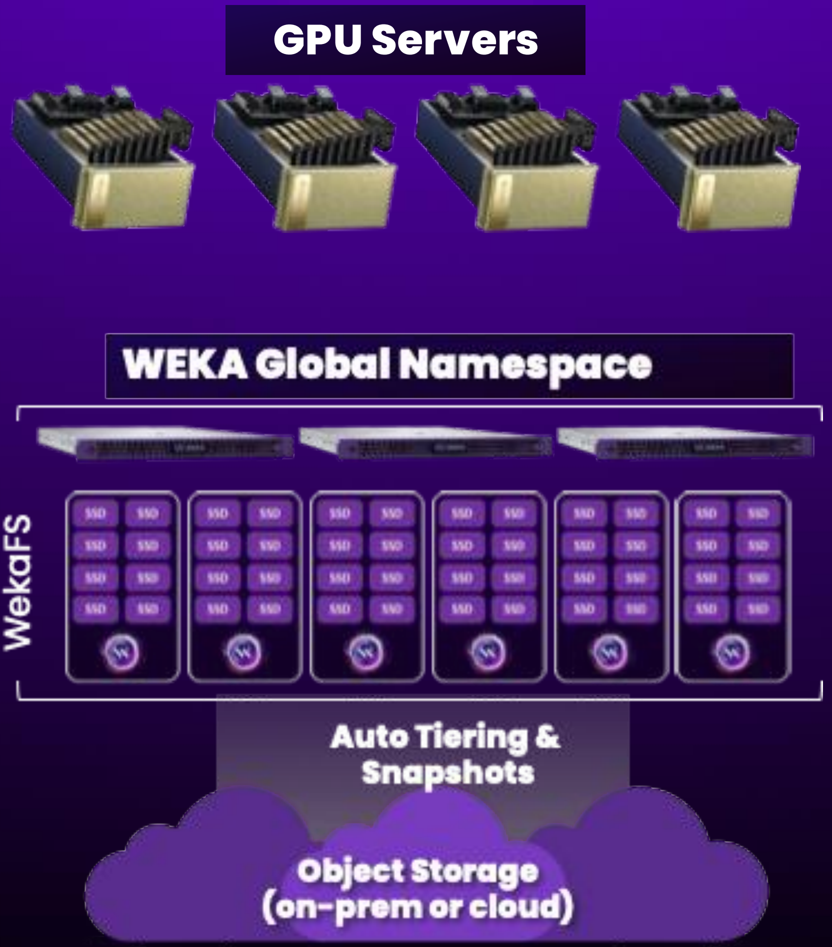
MULTI-DIMENSIONAL PLATFORM SUPPORT

DESIGNED FOR AI

WEKA IS DESIGNED FOR AI

Training

- Best Performance density = less power & rack space
- Ingest once, zero-copy architecture increases GPU efficiency up to 20x
- Low latency checkpointing reduces risk AND delivers MORE flexibility

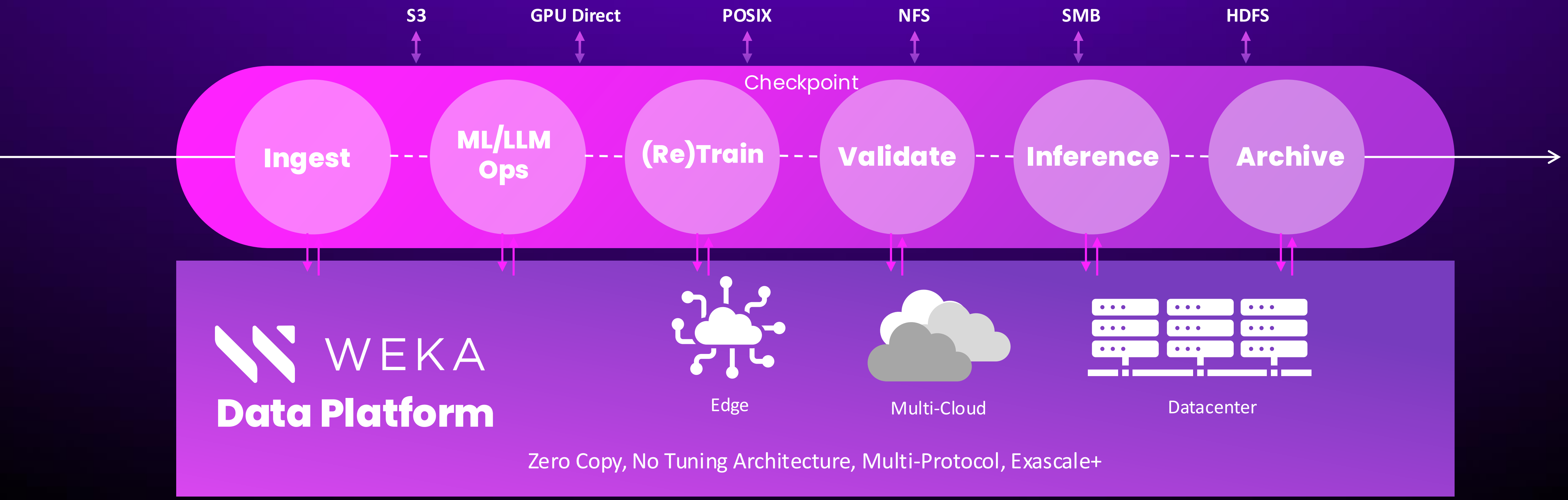


Inference

- Rapid model load times reduce latency from minutes to seconds
- WEKA enables more frequent embeddings keeping outcomes are up to date with the latest data
- WEKA's snapshot & replication capabilities means models, data & embeddings can be seamlessly shared across multiple environments

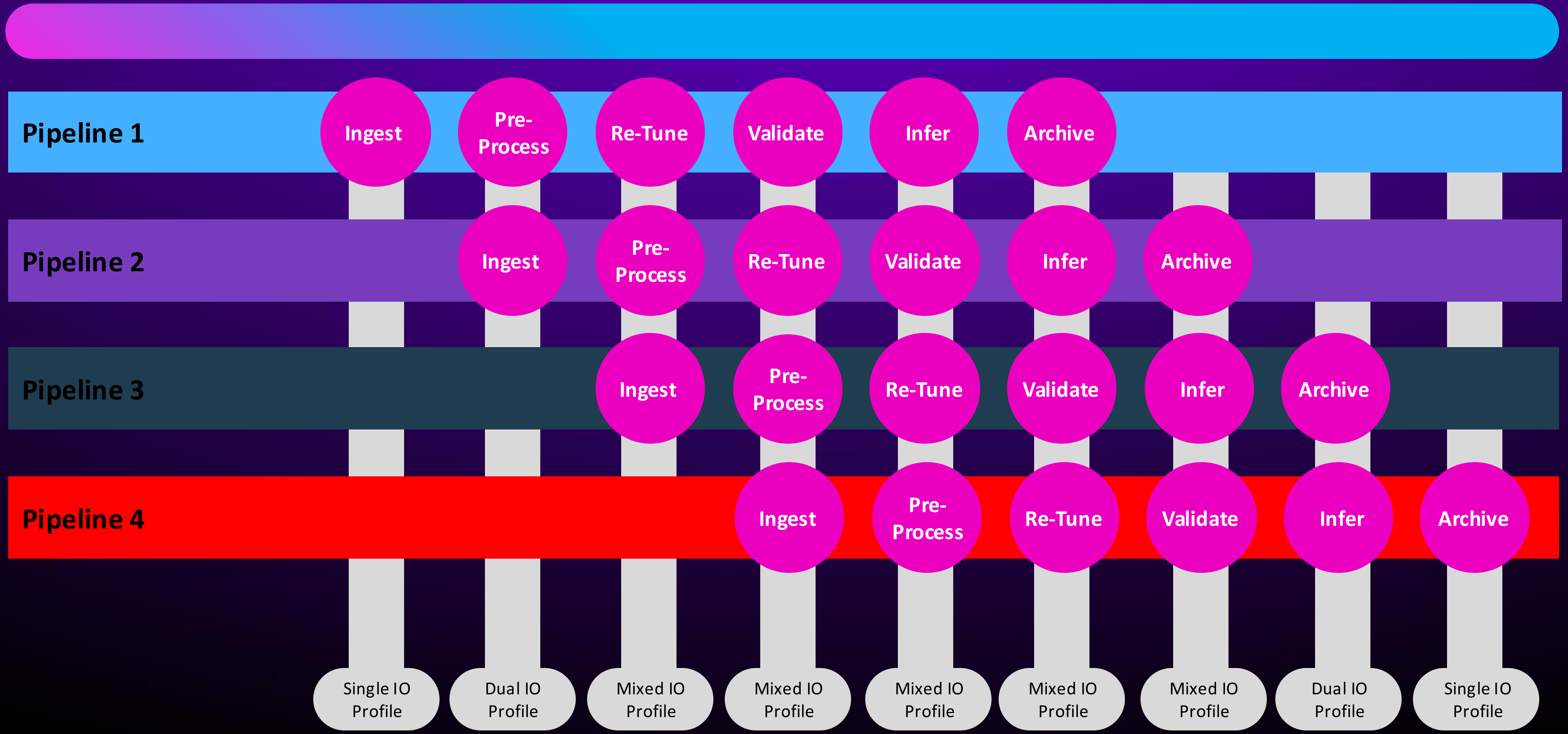
THE WEKA DATA PLATFORM

THE PLATFORM OF CHOICE FOR AI



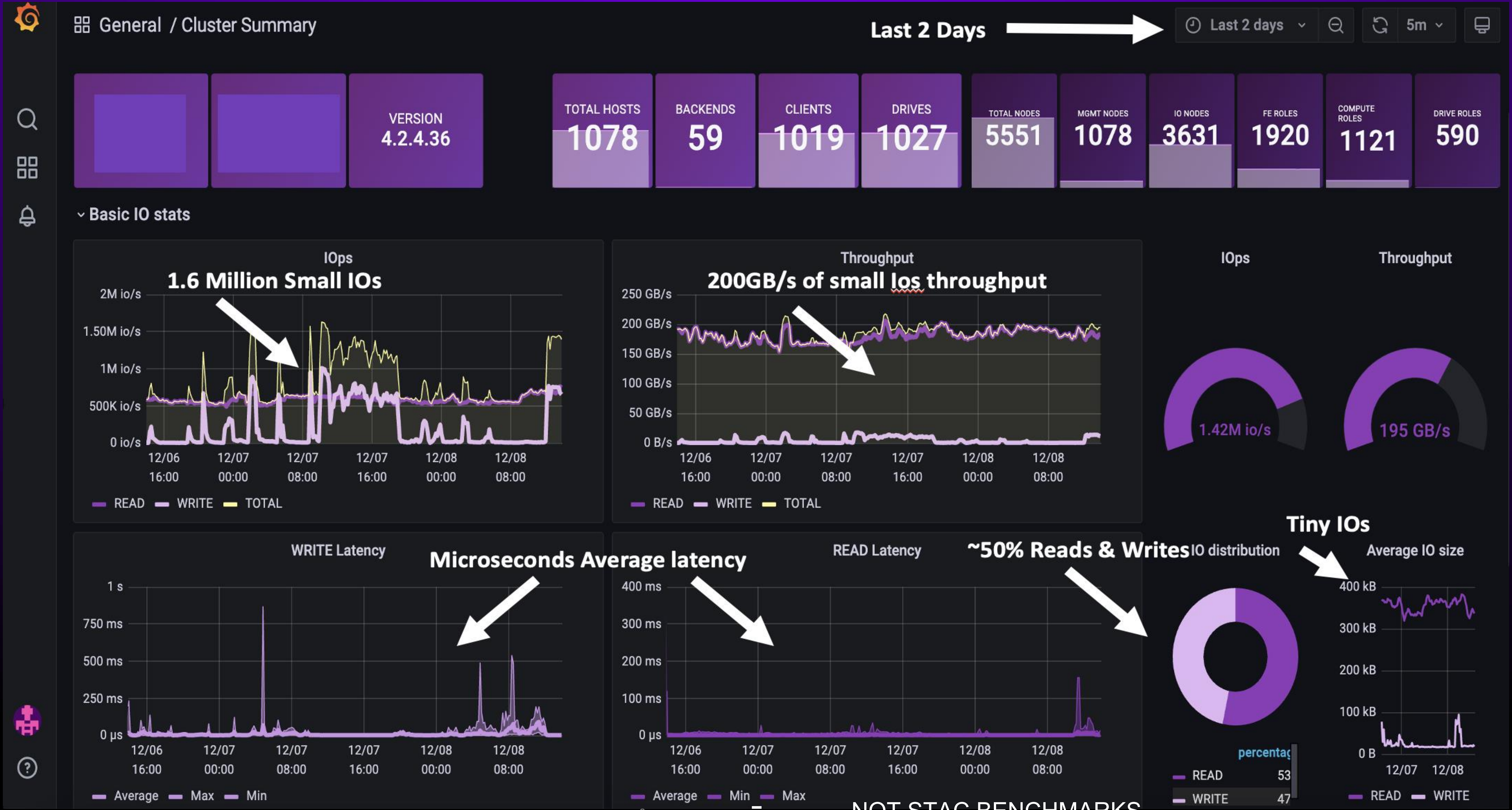
WHY WEKA FOR MODEL TRAINING?

AI Data Pipeline: multiple pipelines heating strorage



Pipeline 1	Ingest	Pre-Process	Re-Tune	Validate	Infer	Archive			
Pipeline 2		Ingest	Pre-Process	Re-Tune	Validate	Infer	Archive		
Pipeline 3			Ingest	Pre-Process	Re-Tune	Validate	Infer	Archive	
Pipeline 4				Ingest	Pre-Process	Re-Tune	Validate	Infer	Archive
	Single IO Profile	Dual IO Profile	Mixed IO Profile	Mixed IO Profile	Mixed IO Profile	Mixed IO Profile	Mixed IO Profile	Dual IO Profile	Single IO Profile

AI customer #1 IO Pattern – Millions of Tiny IOs Reads / Writes



General / Cluster Summary

Last 2 days · 5m

VERSION 4.2.4.36

TOTAL HOSTS 1078

BACKENDS 59

CLIENTS 1019

DRIVES 1027

TOTAL NODES 5551

MGMT NODES 1078

IO NODES 3631

FE ROLES 1920

COMPUTE ROLES 1121

DRIVE ROLES 590

Basic IO stats

IO distribution

percentag
READ	53
WRITE	47

Last 2 Days

1.6 Million Small IOs

200GB/s of small Ios throughput

Microseconds Average latency

~50% Reads & Writes

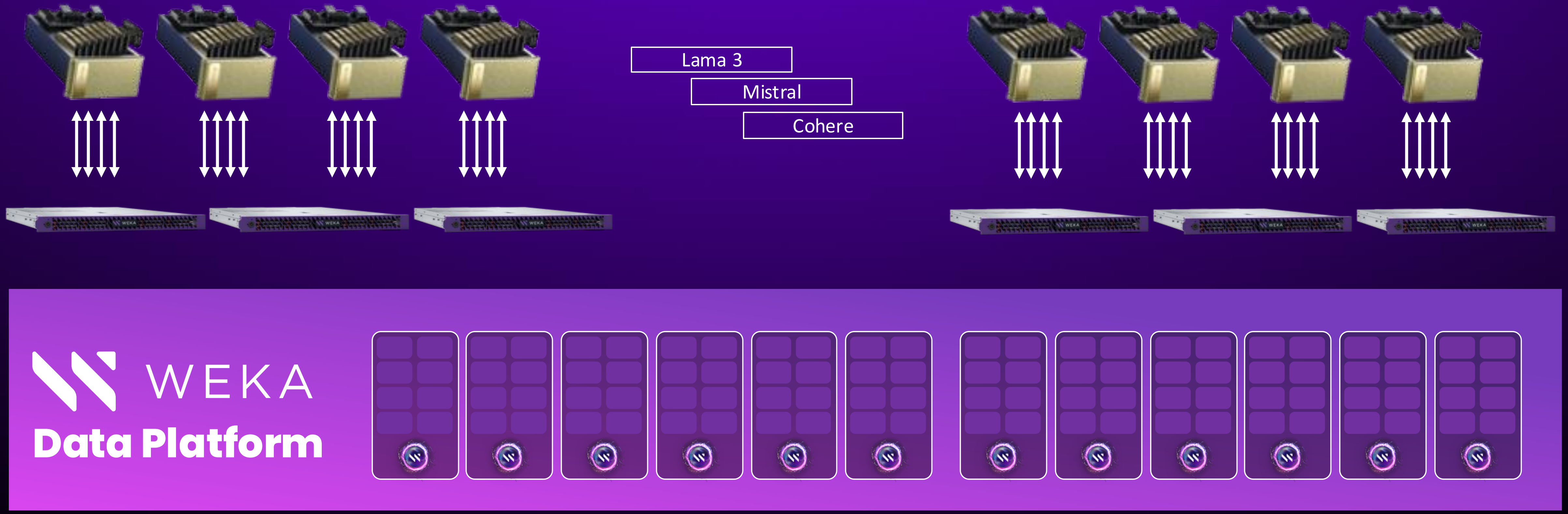
Tiny IOs

NOT STAC BENCHMARKS

WHY WEKA FOR INFERENCE?

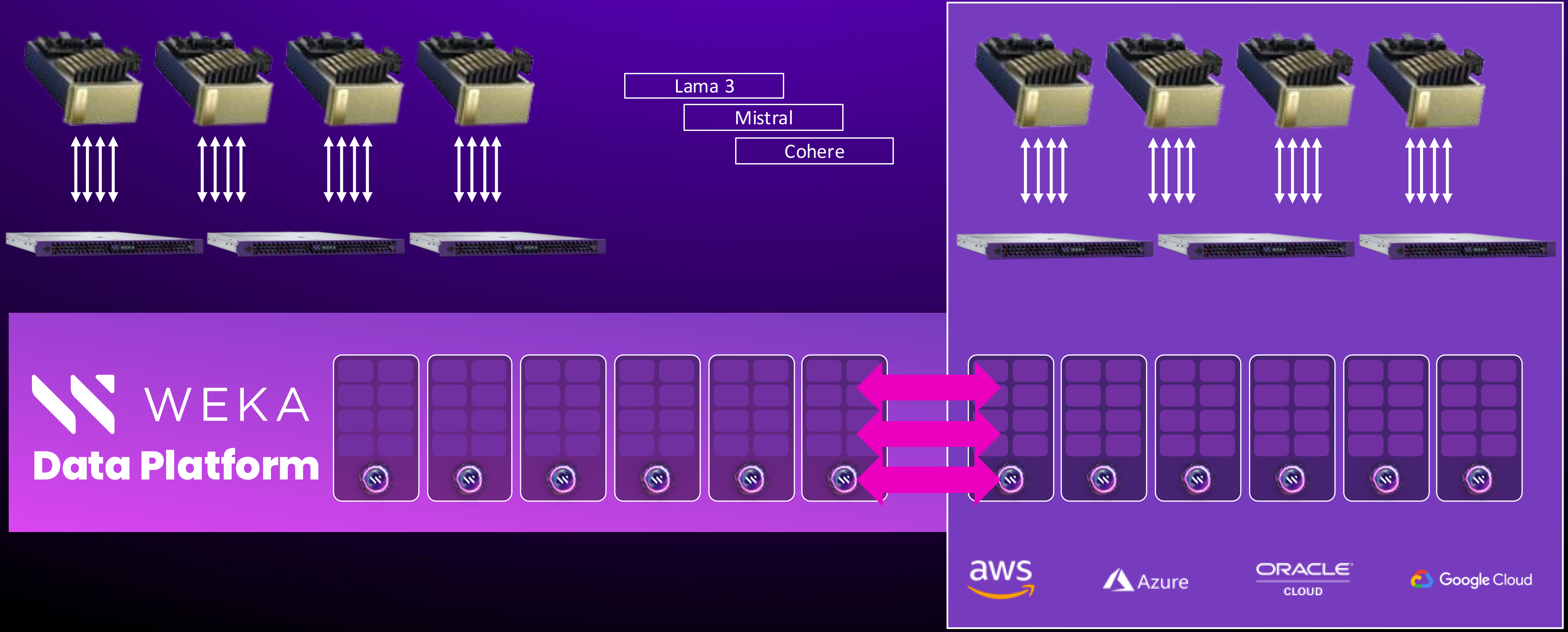
DYNAMIC INFERENCE FARMS

Avoid Over-provisioning of GPU Infrastructure



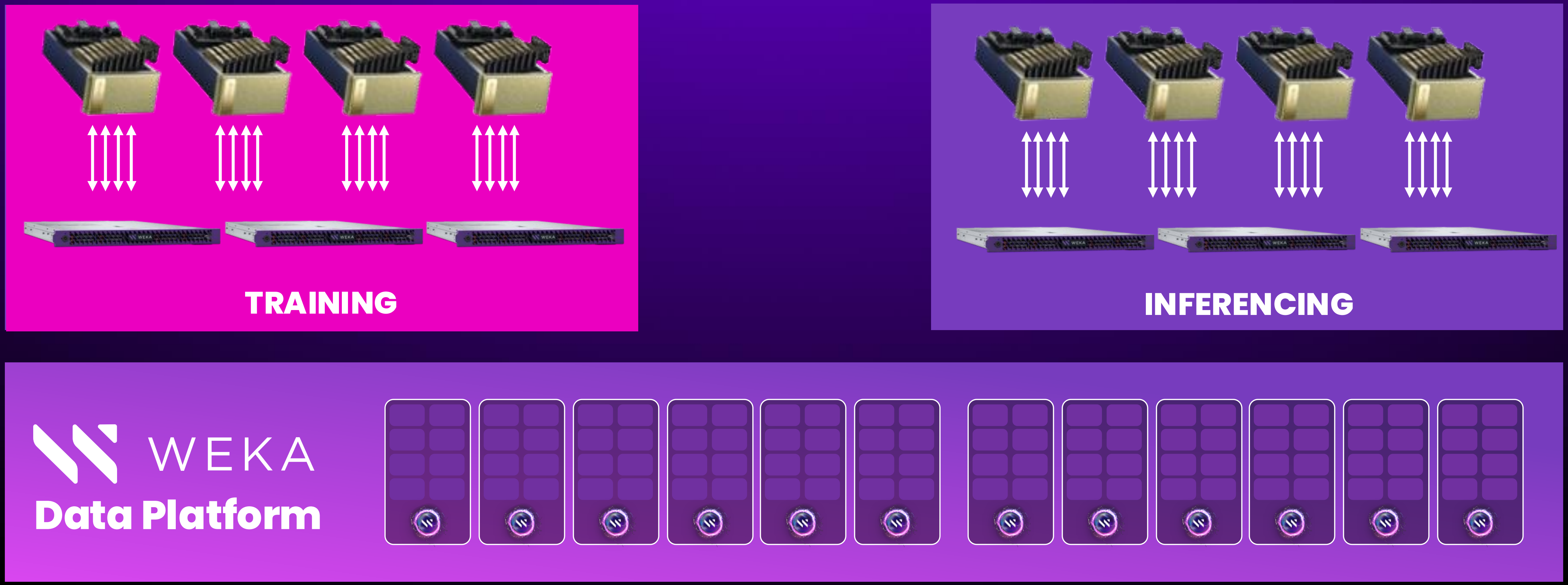
BURST INFERENCE FARMS TO THE CLOUD

On-demand Delivery of Data to Additional Resources



ELASTIC TRAINING & INFERENCING ENVIRONMENT

Utilize Your Most Expensive Assets in a Dynamic Fashion



TRAINING

INFERENCING

WEKA Data Platform

INFERENCING AT SCALE

Developer of large language model (LLM) solutions for businesses.

LLMs are trained to learn statistical relationships between words & phrases.

100+GB

Models

>10x

faster load times

Enables Dynamic Spinning Up Of GPU Inferencing Instances With Seamless Sharing Of Models Between Different Clouds

AUTONOMOUS VEHICLE TRAINING

Revolutionary autonomous vehicle manufacturer using AI for model training

Software has become the main profitability growth driver for automakers

12
days of training
(used to take 1 year)

75
years of innovation

3.5
years equivalent to
75 years

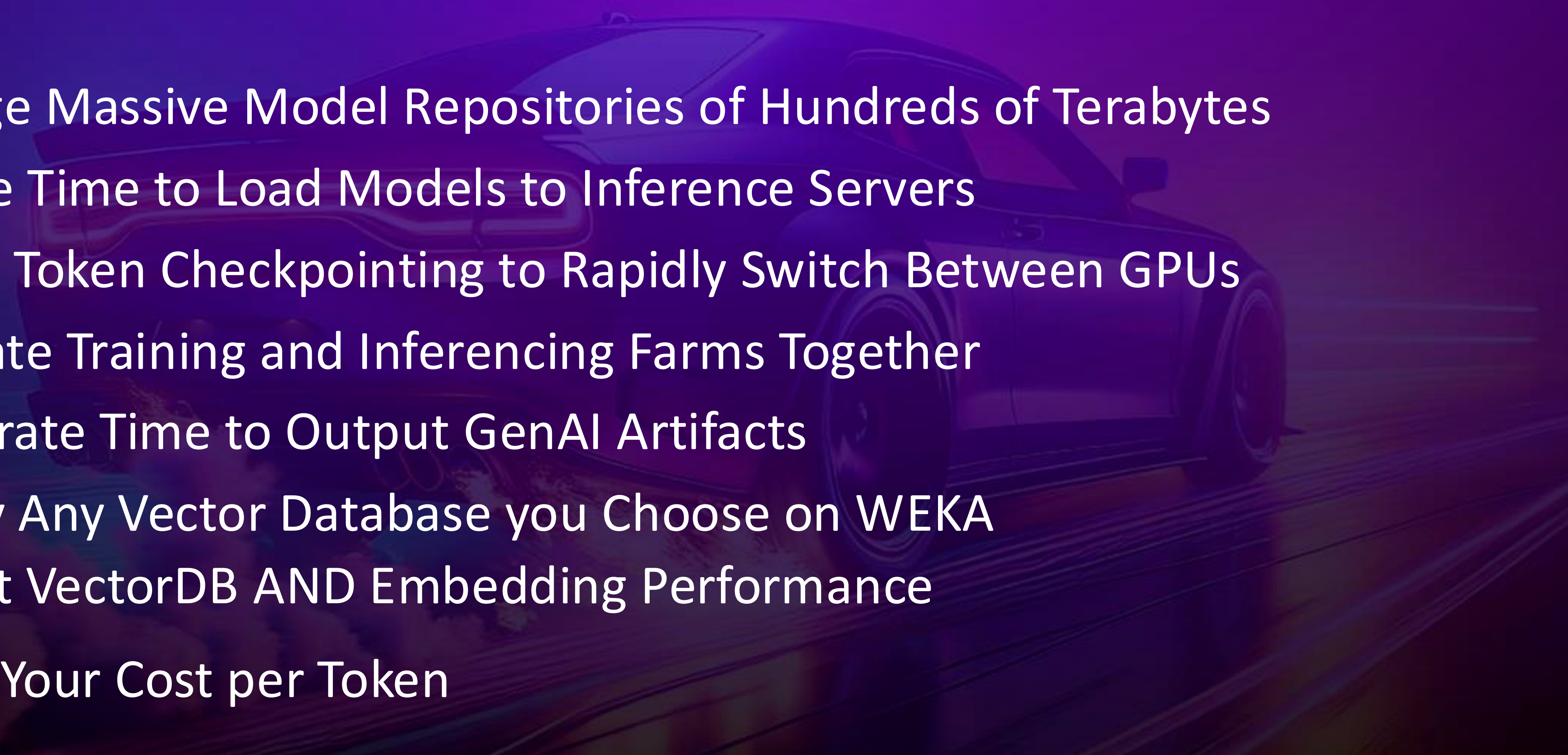
Driven by More Efficient GPU Utilization, Zero Copy & Faster Checkpointing

AI is Disrupting the Automotive Industry

ACCELERATED AI
with the WEKA Data Platform

- Easily Scale To 100k+ GPUs & Multi-EB Environments
- Run Your GPU Pipelines 20x Faster
- Bring Your Data to Your Compute, Wherever It Is…
  - On-premises, Cloud or Hybrid
- Reduced Datacenter Footprint & Power Consumption
- Greatly Reduced Operational Complexity & Cost
- Simple Path To Latest AI Software Innovations

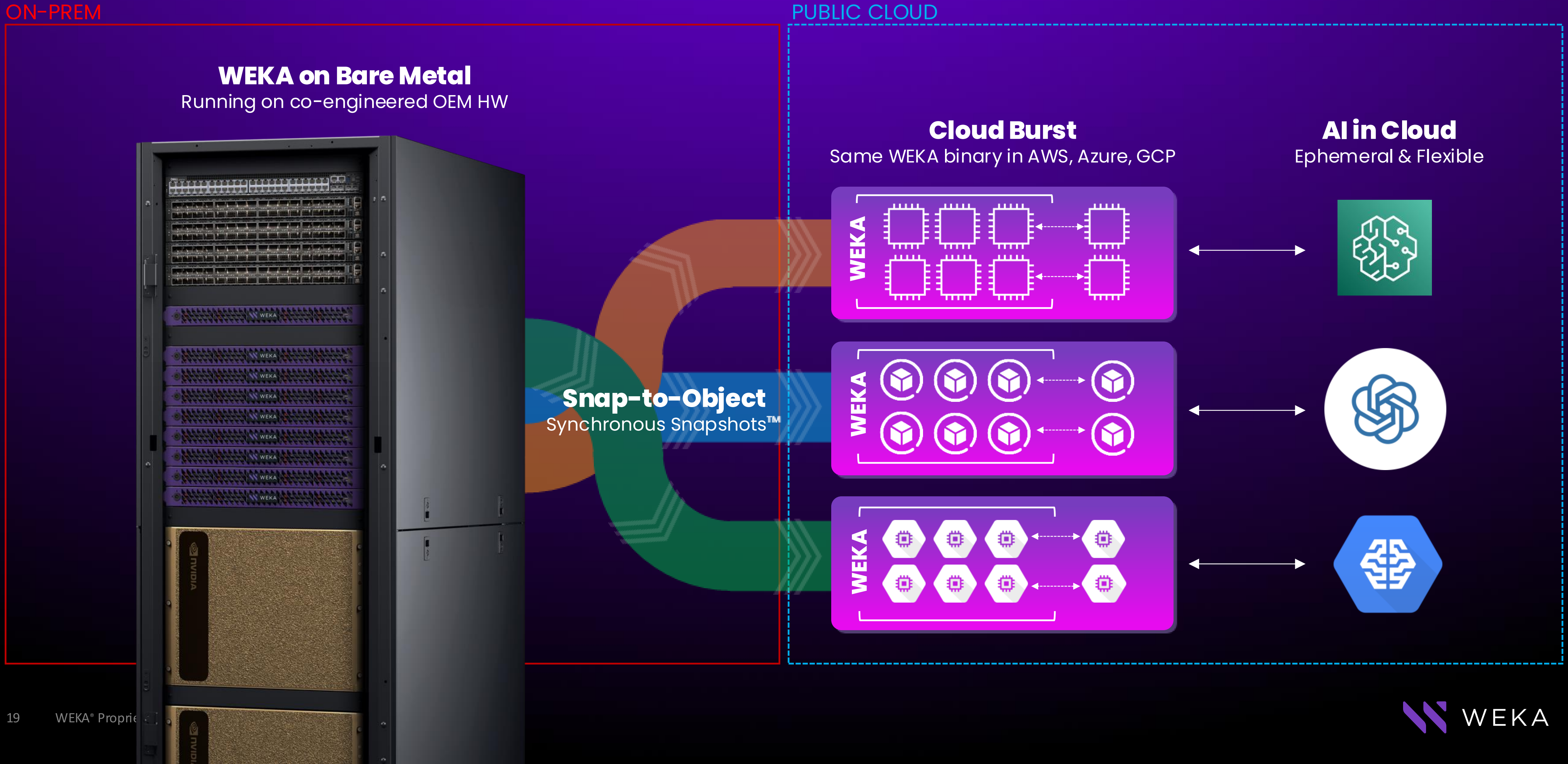
ACCELERATED AI
with the WEKA Data Platform



- Manage Massive Model Repositories of Hundreds of Terabytes
- Reduce Time to Load Models to Inference Servers
- Enable Token Checkpointing to Rapidly Switch Between GPUs
- Integrate Training and Inferencing Farms Together
- Accelerate Time to Output GenAI Artifacts
- Deploy Any Vector Database you Choose on WEKA
  - Boost VectorDB AND Embedding Performance
- Lower Your Cost per Token

WEKA

BUILT FOR HYBRID ENVIRONMENTS



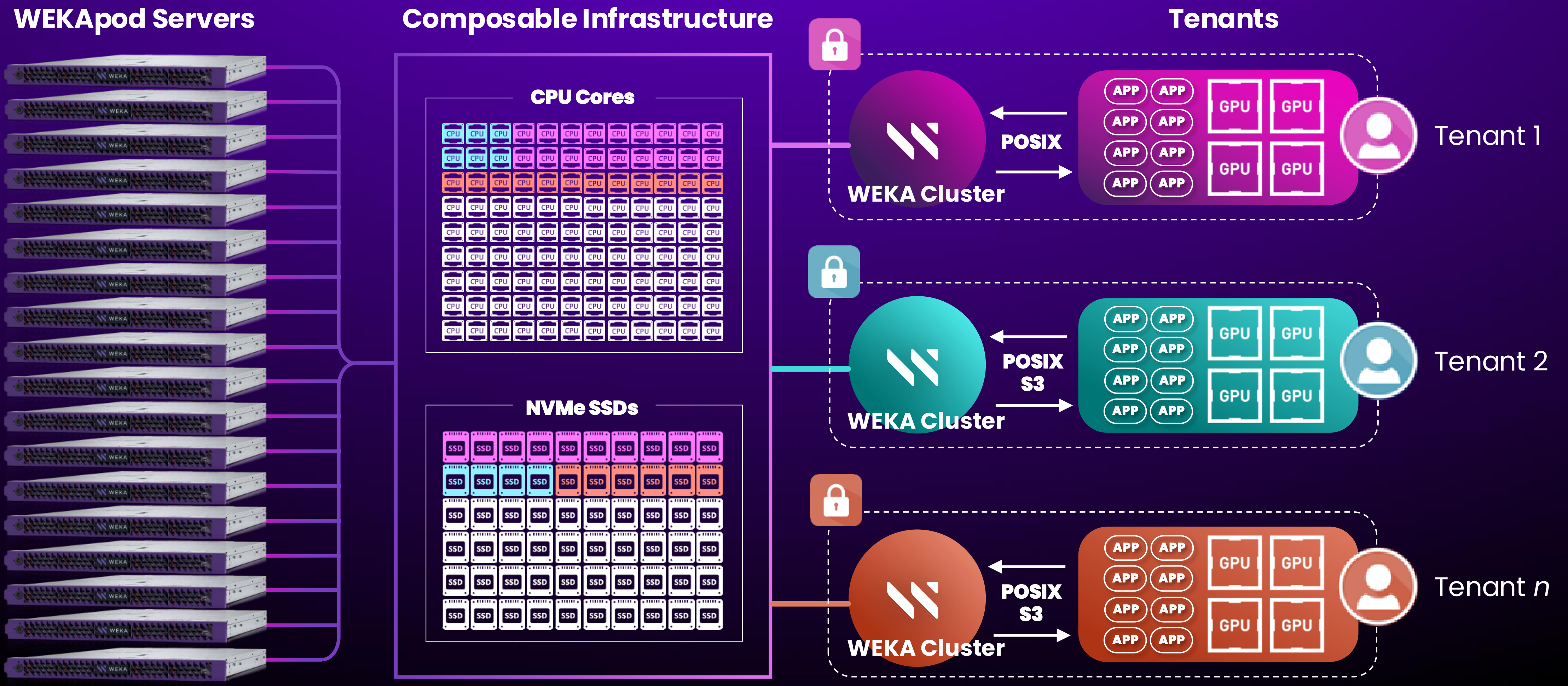
WEKA on Bare Metal
Running on co-engineered OEM HW

Snap-to-Object
Synchronous Snapshots™

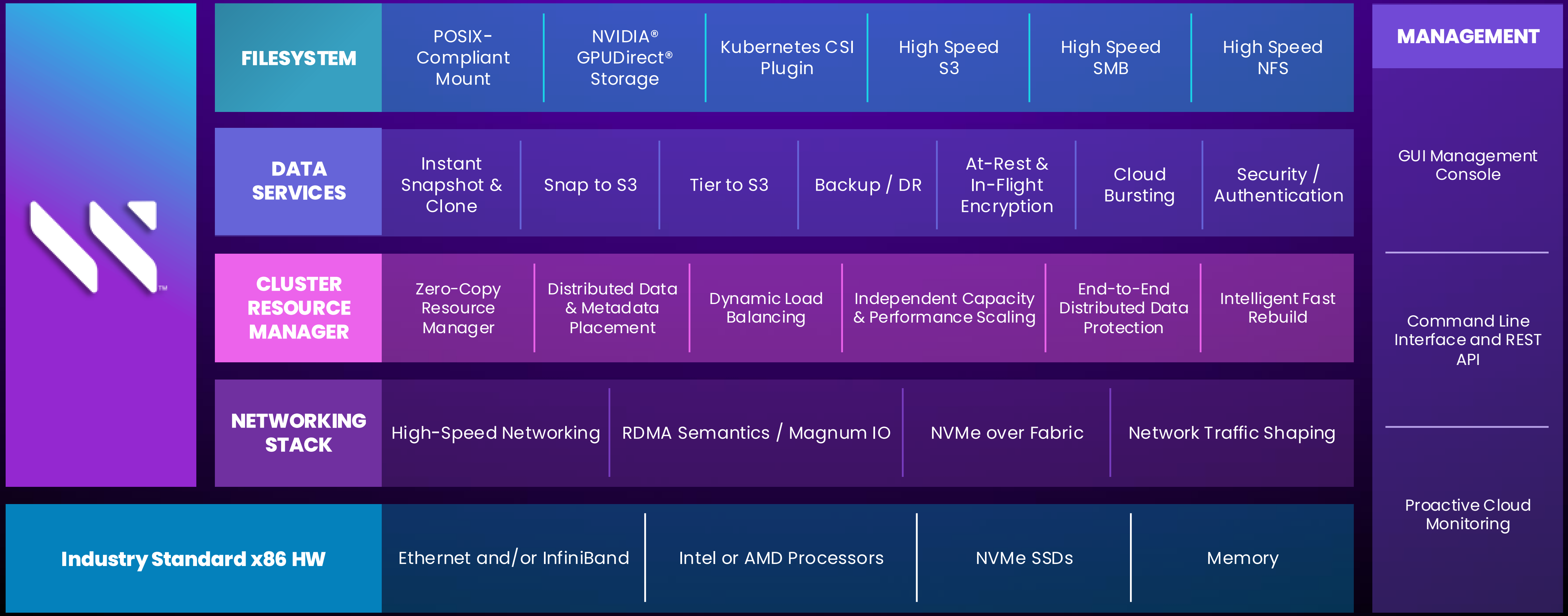
Cloud Burst
Same WEKA binary in AWS, Azure, GCP

AI in Cloud
Ephemeral & Flexible

COMPOSABLE MULTI-TENANCY



WEKA ENTERPRISE FEATURES



FILESYSTEM
- POSIX-Compliant Mount
- NVIDIA® GPUDirect® Storage
- Kubernetes CSI Plugin
- High Speed S3
- High Speed SMB
- High Speed NFS

DATA SERVICES
- Instant Snapshot & Clone
- Snap to S3
- Tier to S3
- Backup / DR
- At-Rest & In-Flight Encryption
- Cloud Bursting
- Security / Authentication

CLUSTER RESOURCE MANAGER
- Zero-Copy Resource Manager
- Distributed Data & Metadata Placement
- Dynamic Load Balancing
- Independent Capacity & Performance Scaling
- End-to-End Distributed Data Protection
- Intelligent Fast Rebuild

NETWORKING STACK
- High-Speed Networking
- RDMA Semantics / Magnum IO
- NVMe over Fabric
- Network Traffic Shaping

Industry Standard x86 HW
- Ethernet and/or InfiniBand
- Intel or AMD Processors
- NVMe SSDs
- Memory

MANAGEMENT
- GUI Management Console
- Command Line Interface and REST API
- Proactive Cloud Monitoring