The Data Platform for the
Cloud & AI Era

WEKA Korea
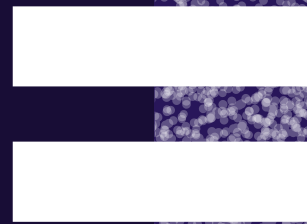김 승 훈 지사장 (chris.kim@weka.io)
Jun 2023

Data, Lots of Data

DNN Models

Compute Power

=

Cambrian Explosion of AI

"

"By 2025, AI will be the top category driving infrastructure decisions, due to the maturation of the AI market, resulting in a tenfold growth in infrastructure requirements"

– Gartner, December 2021

What Is
Artificial Intelligence?

Artificial intelligence (AI) applies advanced analysis and logic-based techniques, including machine learning, to interpret events, support and automate decisions and take action.

How does artificial intelligence work?

Here at Gartner, we define artificial intelligence (AI) as applying advanced analysis and logic-based techniques, including machine learning (ML), to interpret events, support and automate decisions and to take actions. This definition is consistent with the current and emerging state of AI technologies and capabilities, and it acknowledges that AI now generally involves probabilistic analysis (combining probability and logic to assign a value to uncertainty).

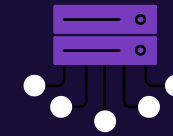Other organizations and individuals may use different definitions. There

The Problem

Exponential growth in modern AI/ML, Life Sciences, Financial Services and cloud applications is fueling unheard-of data growth

Organizations desire competitive advantage through digitalization by **10x faster processing** of 10x bigger datasets
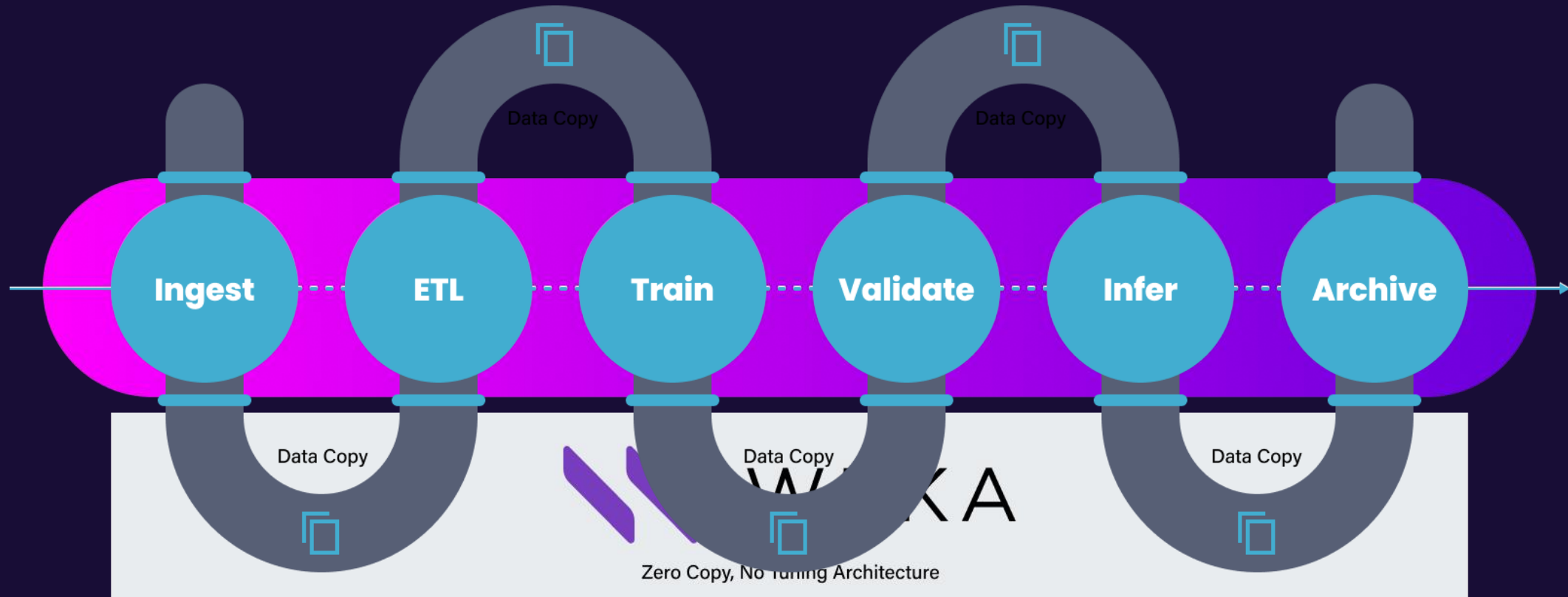
Traditional hardware and cloud storage cannot meet rapidly escalating requirements for performance AND scale across core, multi-cloud and edge data centers

...A NEW APPROACH IS NEEDED.

WEKA

Infrastructure is Slowing Down AI Data Pipelines

70% of an epoch happens before the data ever reached the GPU*

Data Copy

Data Copy

Ingest   **ETL**   **Train**   **Validate**   **Infer**   **Archive**

Data Copy

Data Copy

Data Copy

Zero Copy, No Tuning Architecture

WEKA

WELCOME TO WEKA

FOUNDED IN 2014

ENGINEERING HQ IN TEL AVIV
CORPORATE HQ IN
CALIFORNIA

GOALS

ADDRESS SCALABILITY &
PERFORMANCE CHALLENGES
OF UNSTRUCTURED DATA

FINANCE

MORE REVENUE IN Q3 FY22
THAN ALL OF FY2021.(635%
GROWTH)
SERIES D RAISED $135M IN
2022

CUSTOMERS

8 OF THE FORTUNE 50
FINANCE, HEALTHCARE,
MANUFACTURING, OIL & GAS,
GOVERNMENT, DEFENSE,
ACADEMIA/RESEARCH

PARTNERS

AMD, AWS, HITACHI
VANTARA, HPE, LENOVO,
NVIDIA, SUPERMICRO, DELL
TECHNOLOGIES,
QUANTIUM, SCALITY, IBM

외부 조사기관 평가(Gartner Magic Quadrant)

Figure 1: Magic Quadrant for Distributed File Systems and Object Storage
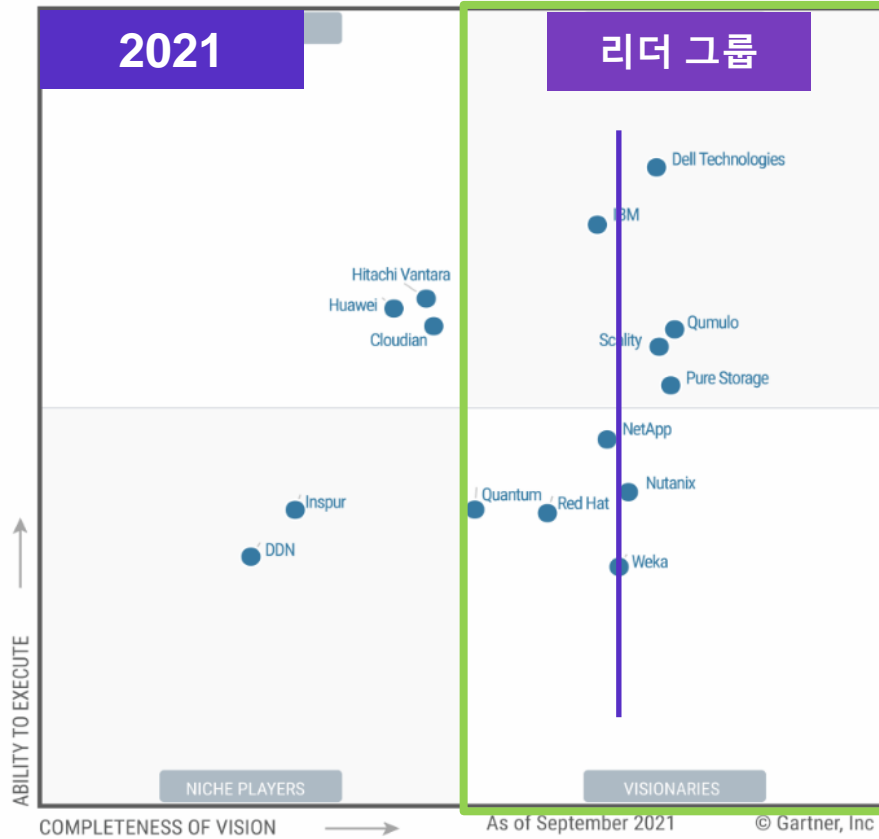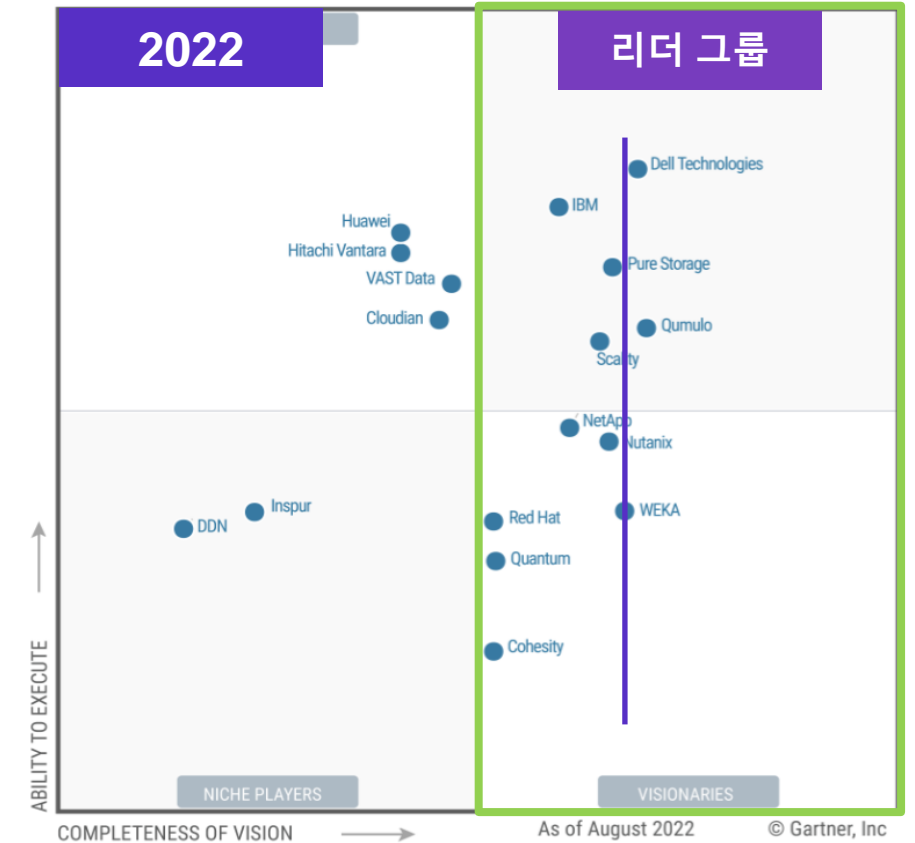
2021 — 리더 그룹

Dell Technologies, IBM, Hitachi Vantara, Huawei, Cloudian, Scality, Qumulo, Pure Storage, NetApp, Quantum, Red Hat, Nutanix, Weka, Inspur, DDN

ABILITY TO EXECUTE / COMPLETENESS OF VISION / NICHE PLAYERS / VISIONARIES

As of September 2021 © Gartner, Inc

Figure 1: Magic Quadrant for Distributed File Systems and Object Storage

2022 — 리더 그룹

Dell Technologies, IBM, Huawei, Hitachi Vantara, VAST Data, Pure Storage, Cloudian, Qumulo, Scality, NetApp, Nutanix, WEKA, Red Hat, Quantum, DDN, Inspur, Cohesity

ABILITY TO EXECUTE / COMPLETENESS OF VISION / NICHE PLAYERS / VISIONARIES

As of August 2022 © Gartner, Inc

고성능 병렬 파일시스템 벤더 중 유일한 **리더 그룹,** 리더 그룹에서도 유일한 **포지션이 약진한 고속 성장 벤더**

WEKA

WEKA's Play ground

AI/ML & DEEP LEARNING

LIFE SCIENCES

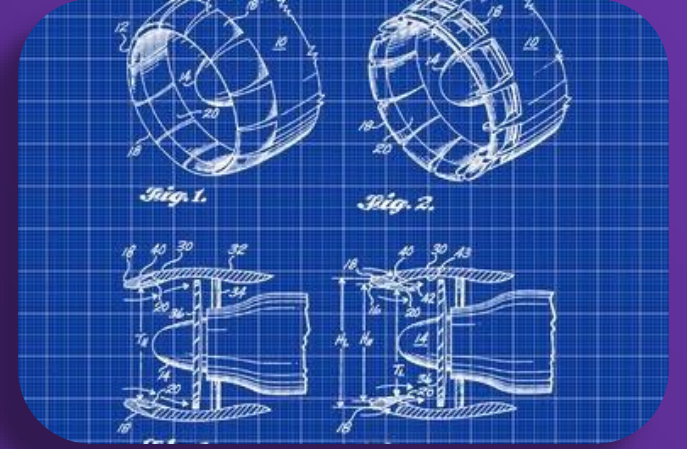
TRADITIONAL HPC

MEDIA & ENTERTAINMENT
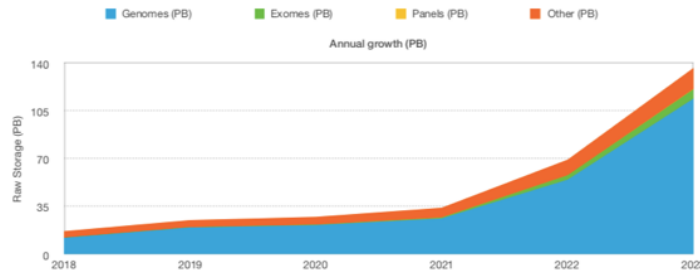

FINANCIAL SERVICES

MANUFACTURING

...And Their Common Pain Points

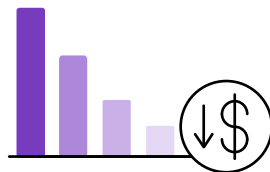## IT Budgets Are Being Squeezed

DO MORE WITH LESS!!!

Fewer Personnel

Budgets Squeezed

Data Is Growing Exponentially

Source: Data Age 2025, sponsored by Seagate with data from IDC Global DataSphere, Nov 2018

Real-time Analytics

Financial Processing & Analysis

Fraud Detection
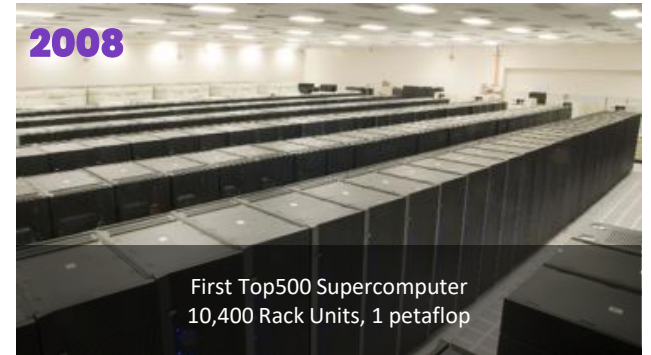
Geospatial Research

HPC

Microscopy

Life Sciences

Imaging Device

Compute Densification

2008

First Top500 Supercomputer
10,400 Rack Units, 1 petaflop

5X Performance

0.1% of the Space
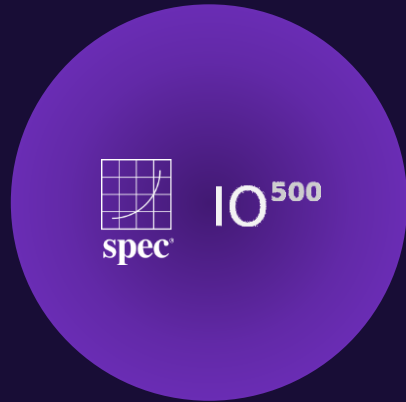
I/O per Client is off the charts

2020

NVIDIA DGX-A100
6 Rack Units, 5 Petaflop

WEKA

Uniquely Designed for Next-Gen Workloads
Software Data Platforms *Replace* Traditional *Hardware Storage Systems*

Zero Tuning for World-Record Performance

Exabyte++ **Scalability**

Cloud, Edge & On-Prem

Modern Data **Management**

Class Leading **Economics**

WEKA

One Architecture Delivers on Three Promises

File semantics

Simple to install

Simple to manage

Cloud integration
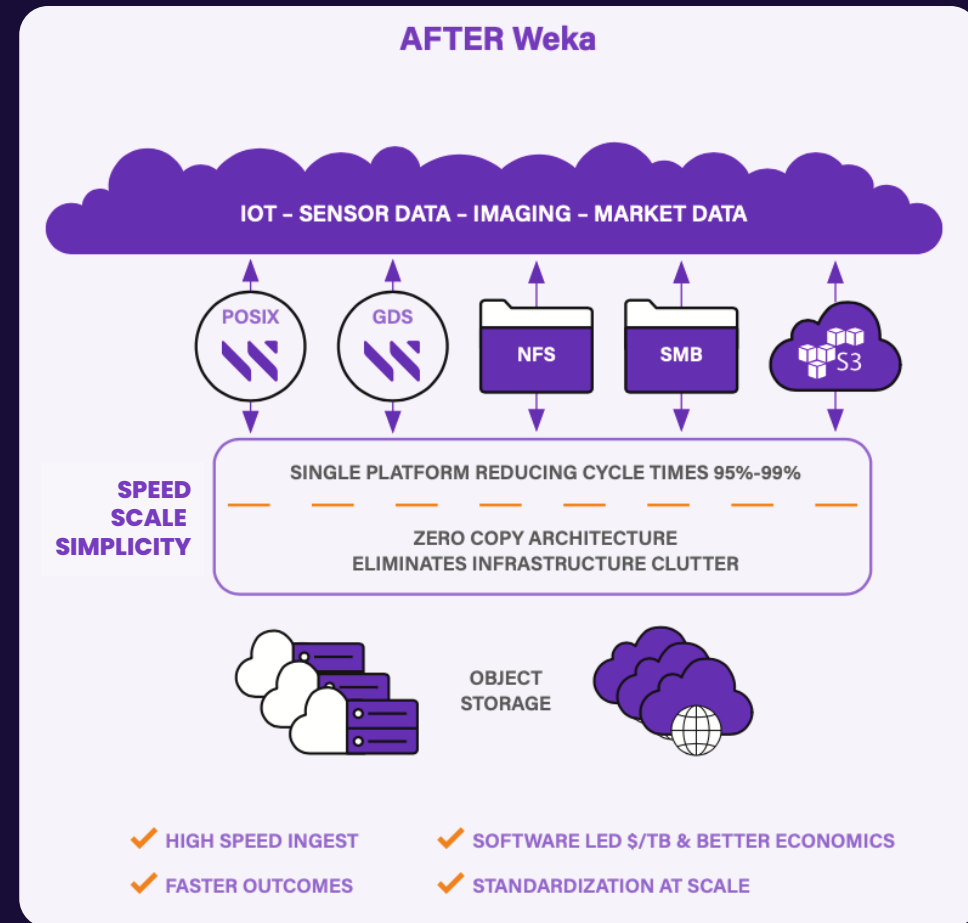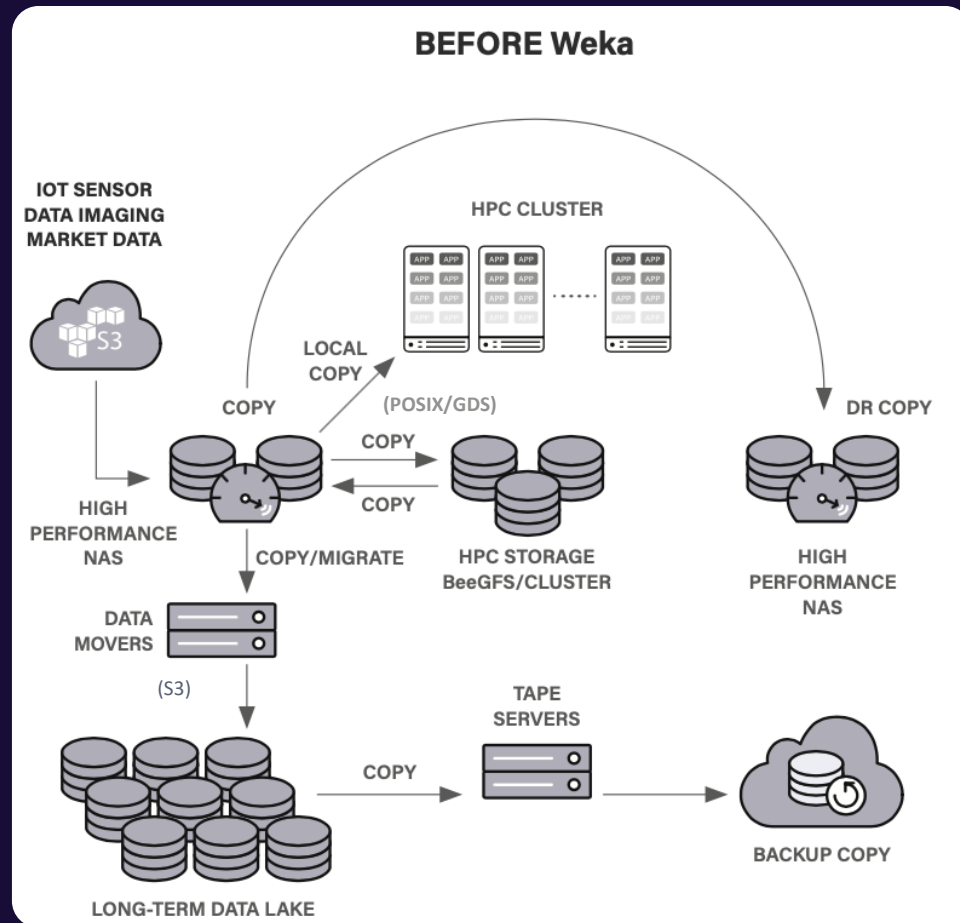
Simplicity

Speed

Lowest latency

Faster than a local filesystem
or All Flash array

Scale

Massive scale

Capacity

Throughput

Metadata

Weka® © 2022

WEKA

Eliminate A Mish-Mash Of Storage Systems

BEFORE Weka

IOT SENSOR DATA IMAGING MARKET DATA

S3

HPC CLUSTER

LOCAL COPY

COPY

(POSIX/GDS)

COPY

COPY

HIGH PERFORMANCE NAS

COPY/MIGRATE

HPC STORAGE BeeGFS/CLUSTER

DR COPY

HIGH PERFORMANCE NAS

DATA MOVERS

(S3)

TAPE SERVERS

COPY

COPY

BACKUP COPY

LONG-TERM DATA LAKE

AFTER Weka

IOT – SENSOR DATA – IMAGING – MARKET DATA

POSIX

GDS

NFS

SMB

S3

SPEED SCALE SIMPLICITY

SINGLE PLATFORM REDUCING CYCLE TIMES 95%-99%

ZERO COPY ARCHITECTURE ELIMINATES INFRASTRUCTURE CLUTTER

OBJECT STORAGE

✓ HIGH SPEED INGEST

✓ FASTER OUTCOMES

✓ SOFTWARE LED $/TB & BETTER ECONOMICS

✓ STANDARDIZATION AT SCALE

WEKA

Business Benefits

Increase productivity

up to

90%

Improvement in deep learning epoch time

Improve resource ROI

75%

Reduction in workload wall clock time

Lower storage costs

65%

Savings over traditional high-performance computing

WEKA

One Data Platform That Does It All

⊘ **1 Tier of Storage**

⊘ **1 File System**

⊘ **1 Point of Management**

⊘ **1 Copy of Data**

2 times performance guarantee program !!!

WEKA

WEKA on Cloud

Eliminate Data Copies

Auto-scale Performance

Move Data to Cloud

Automatically Tier & Restore Data

Access Data via Any Protocol

Easily Manage at Scale

WEKA

Atomwise Reduced their Epoch Time by 92%

Feed billions of tiny files into expensive GPU instances

Improve AI Epoch times

**Eliminate copies across multiple EBS instances**

Integrate with Kubernetes pipelines

Dev Boxes
(jump host / local environment)

Cluster Client
- Submit job
- Job status
- Delete job

AWS Batch (legacy)

Elastic Kubernetes Service Control Plane

EC2 instances with Weka POSIX Client

EC2 instances with Weka POSIX Client

0 - 10,000
Spot instances per day

WEKA CLUSTER
8 i3en.xlarge EC2 instances
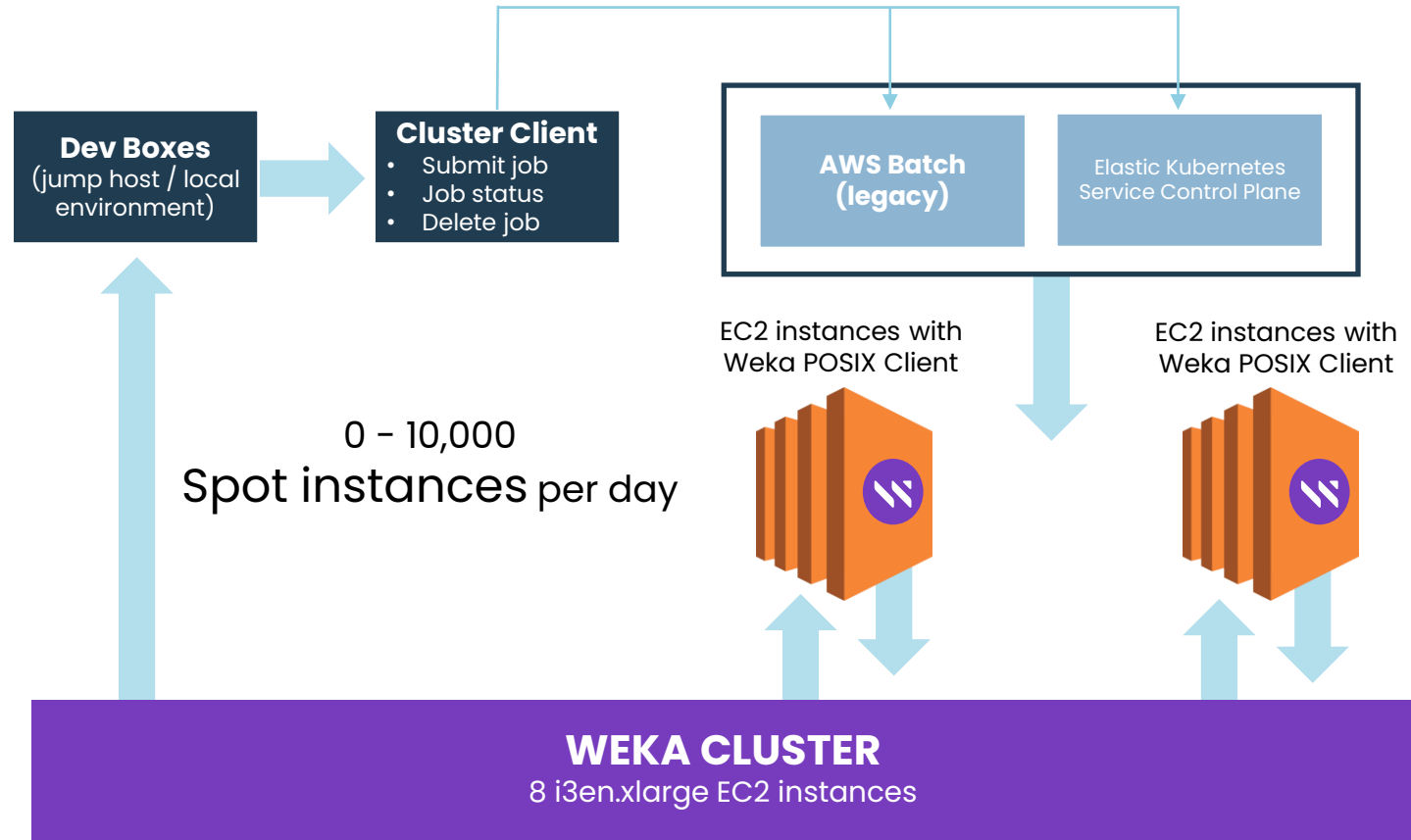
Migrated from native services (EBS) to reduce
Epoch time from 3 months to 1 week with 100%+ savings

Let's Talk About Performance That Matters

Vendors Emphasize Throughput, Which is Wrong
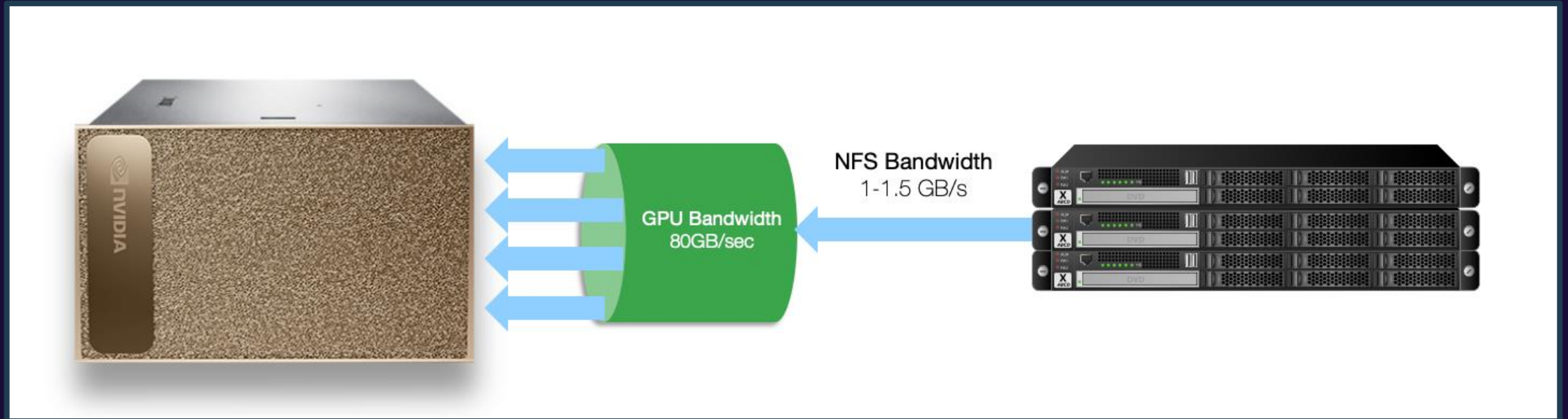
Workloads Need Bandwidth, IOPS, Latency, and Metadata Performance

Benchmarks on NVIDIA DGX Systems

WEKA Data Platform for GPUDirect Storage

162GB/s
~1 Million IOPS

113GB/s
~1 Million IOPS

DGX-A100 Benchmarks

DGX-2 Benchmarks

Solves the AI Storage Challenges

Improves productivity and faster time to market and value

Accelerate large scale data pipelines

- Reduced epoch times
- Fastest inferencing
- Highest images / secs benchmarks

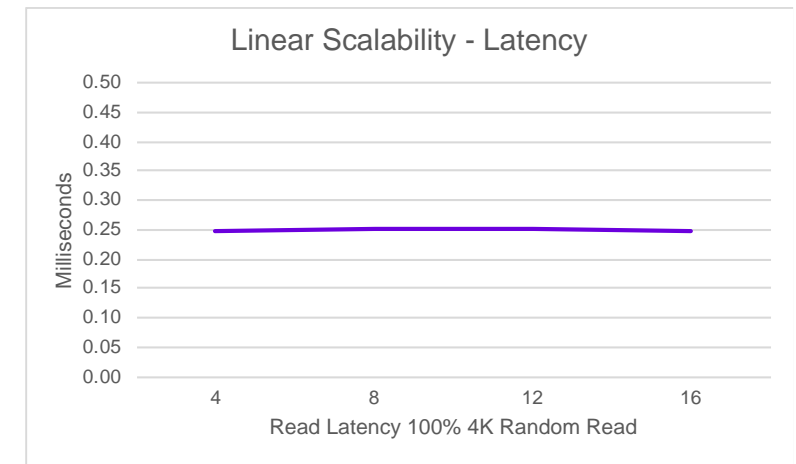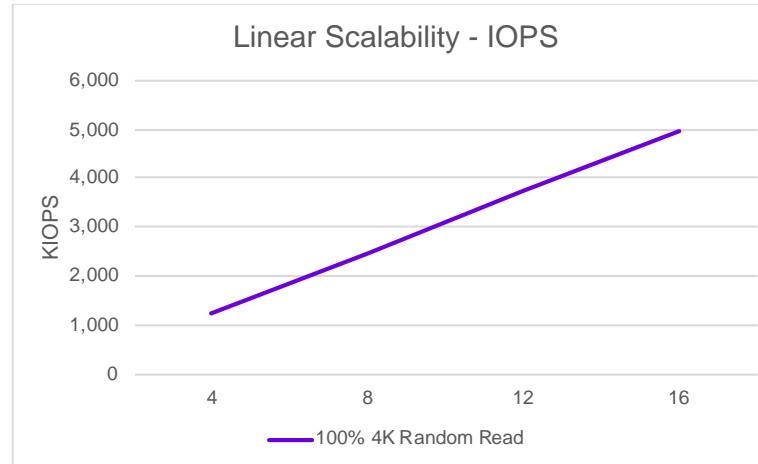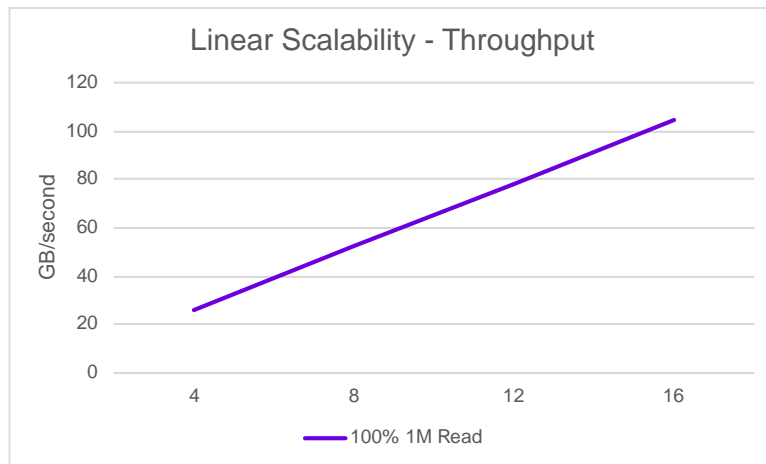Run entire pipeline on the same storage backend

Multi Protocol Support

Faster than local storage

84X Faster Training

4 Hours

2 Weeks

File System Scales Linearly with Cluster Size

WekaFS performance in AWS

Linear Scalability - Throughput

GB/second vs cluster size (4, 8, 12, 16)
— 100% 1M Read

Linear Scalability - IOPS

KIOPS vs cluster size (4, 8, 12, 16)
— 100% 4K Random Read

Linear Scalability - Latency

Milliseconds vs cluster size (4, 8, 12, 16)
Read Latency 100% 4K Random Read

- 308K OPS/AWS *i* Instance
- 6.6GB/sec/AWS *i* Instance
- <250 microsecond *i* latency

Test Environment – 16 I3en.24xlarge cluster.

Today, the world's data centers consume **roughly 3%** of the global energy supply

WEKA

Without intervention, it will grow to 8% by 2025

WEKA

That's about **40 megatons of CO_2e worldwide...**

WEKA

Reforestation is critical to Reduce Carbon

WEKA

Our Vision
for
Sustainable AI

WEKA