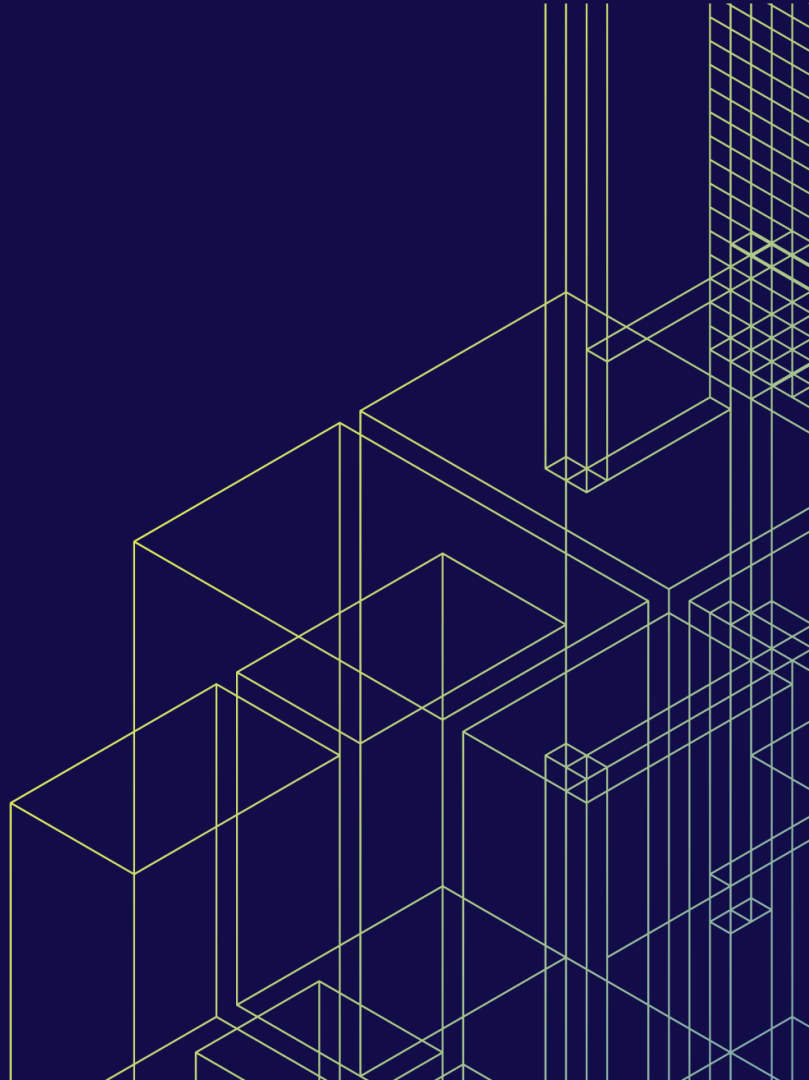
SDC 19

September 23-26, 2019
Santa Clara, CA

IO for GPU-Accelerated Machine Learning

Andy Watson, CTO
watson@weka.io / @the_andywatson
WekaIO, Inc.



Data Access Challenges Impeding AI/ML Research 1/2

- Increasingly Huge Quantities of Data
  - Available from more sources
  - Hundreds of Terabytes growing to Petabytes (or Exabytes)
- Object Storage can scale capacity, but …
  - Performance is generally only good for large-object throughput
- Also, historically most applications have been written for *file* access
  - Arguably, even new apps continue to be file-oriented
  - Many “cloud native” apps are architected for object storage

2019 Storage Developer Conference. © WekaIO, Inc. All Rights Reserved.

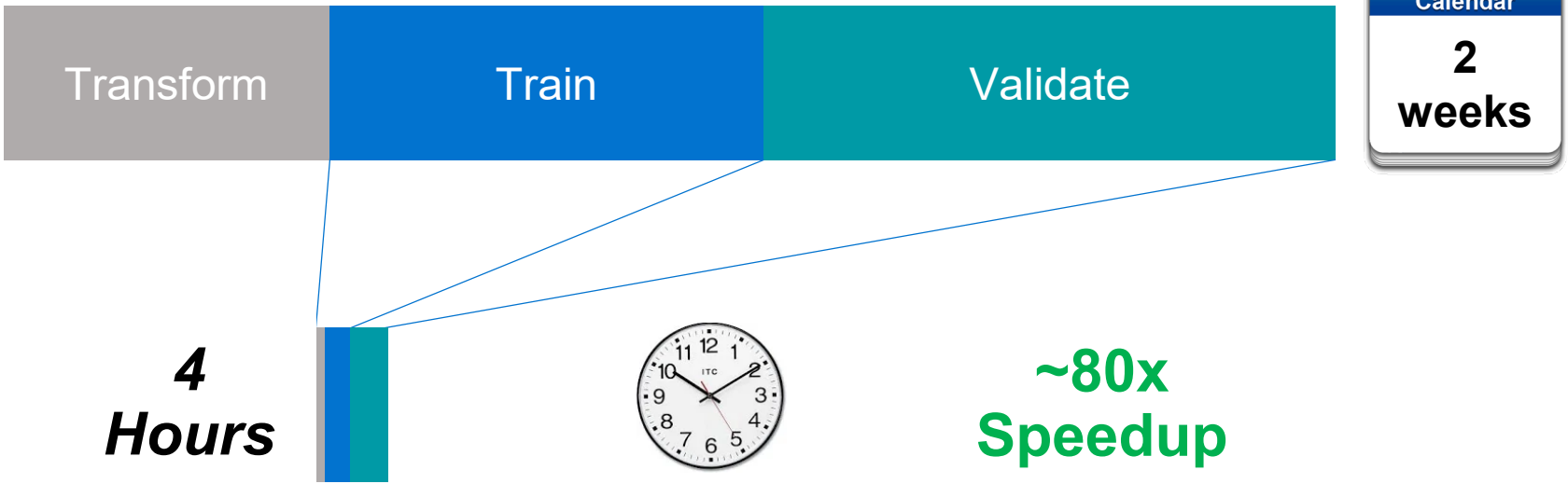
Data Access Challenges Impeding AI/ML Research 2/2

- Some Filesystems can scale capacity, but …
  - Diminishing performance with scaling is typical
  - Instability with huge numbers of *files per directory* reflects metadata issues independent of capacity
  - Single-mount-point performance is often insufficient
- GPU-Accelerated AI/ML Training & Validation
  - Some GPU tasks are compute-bound, but …
  - Training & Validation are usually IO-constrained

2019 Storage Developer Conference. © WekaIO, Inc. All Rights Reserved.

Dramatic Reduction in Training Epoch Cycle Times

at an actual AI/ML customer site (autonomous vehicle software development)



2019 Storage Developer Conference. © WekaIO, Inc. All Rights Reserved.

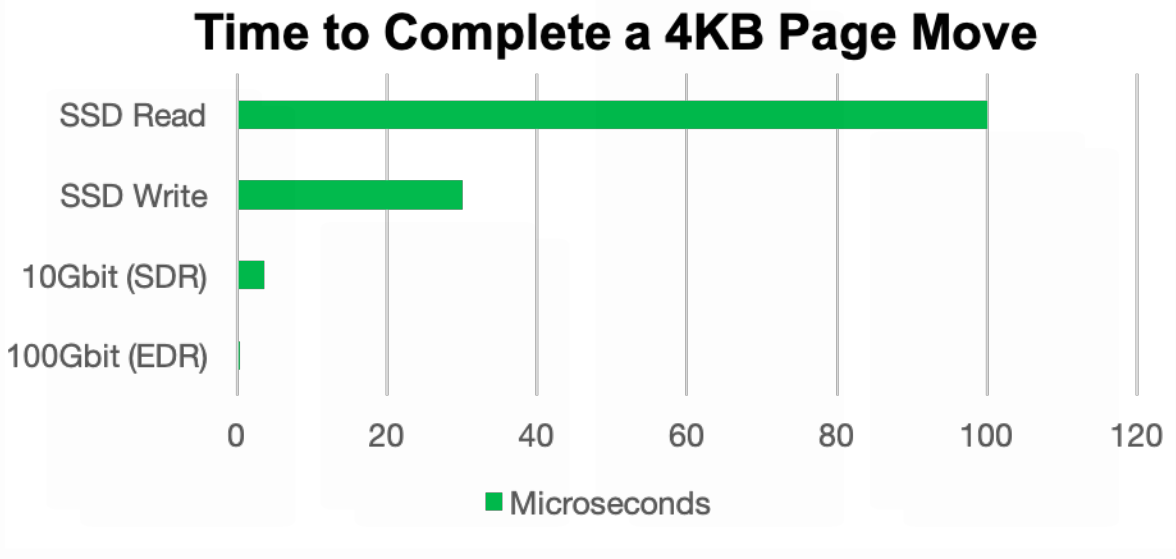
What Happened There? 1/2

- Previously they'd been copying data to local flash
  - Seemed "fast" — but …
    - Burdened by wall-clock time for copying data to local SSD
    - Extra steps in the workflow are opportunities for errors
  - Better to have shared data access — if it's fast enough
- But by using the network to get to WekaIO …
  - Instead of reading from only a small # of local flash devices, dozens or hundreds are accessed in parallel
  - And WekaIO's Matrix™ filesystem is flash-native, with no historical baggage based on HDD — intrinsically faster

2019 Storage Developer Conference. © WekaIO, Inc. All Rights Reserved.

What Happened There? 2/2

- And why not traverse the network?
  - Latency contribution of a high-speed network is trivial



2019 Storage Developer Conference. © WekaIO, Inc. All Rights Reserved.

Aggregate Performance Isn't Enough

- These are not your grandfather's compute servers
  - Faster CPU's? Yes.
  - Faster NIC's, with offload processors? Yes.
- Adding GPU's to accelerating the AI/ML workload …
  - Significantly elevates the demand at each mount point

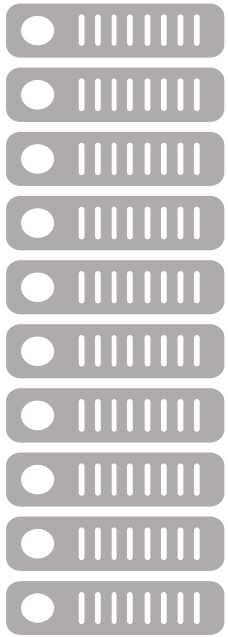
2019 Storage Developer Conference. © WekaIO, Inc. All Rights Reserved.

GPU-Acceleration Is A Game-Changer

- Each GPU is a sophisticated Array Processor with extremely demanding IO requirements
  - Each individual NVIDIA Tesla V100 GPU's ~4 GBytes/s
  - An x86 compute server can have multiple GPU's installed, maxing out at 8
  - NVIDIA's DGX-2 platform has 16 Tesla V100's — with 8 100-gbit/s network links
- So: What should data storage infrastructure look like for GPU-accelerated compute platforms?

2019 Storage Developer Conference. © WekaIO, Inc. All Rights Reserved.

GPU-Acceleration Is A Game-Changer



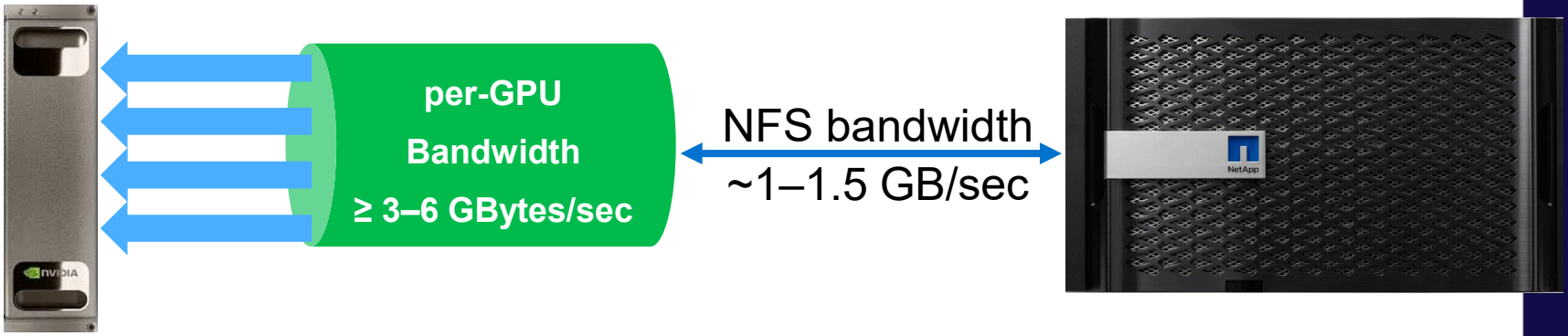
10
CPU-only
Servers

1
GPU-Accelerated
Server

adding GPUs *concentrates* compute infrastructure's associated IO density (by ~10x in this example)

2019 Storage Developer Conference. © WekaIO, Inc. All Rights Reserved.

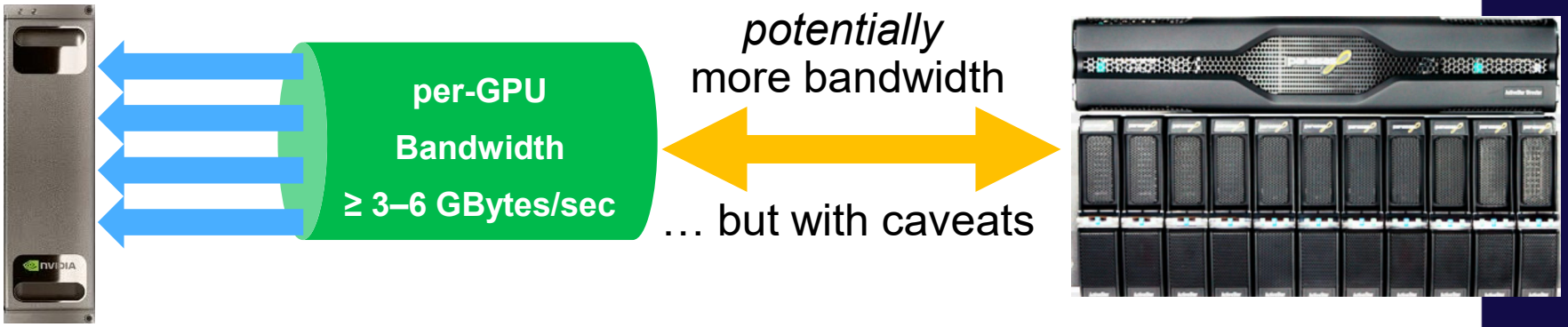
Shared File Storage via NFS



- NFSv2, v3, v4 — only ~1–1.5 GB/s per client
  - NFSv4's pNFS & Session Trunking *potentially* better, but not yet performing in production
  - Multiple GPU's per compute server exacerbate this per-mount-point limitation

2019 Storage Developer Conference. © WekaIO, Inc. All Rights Reserved.

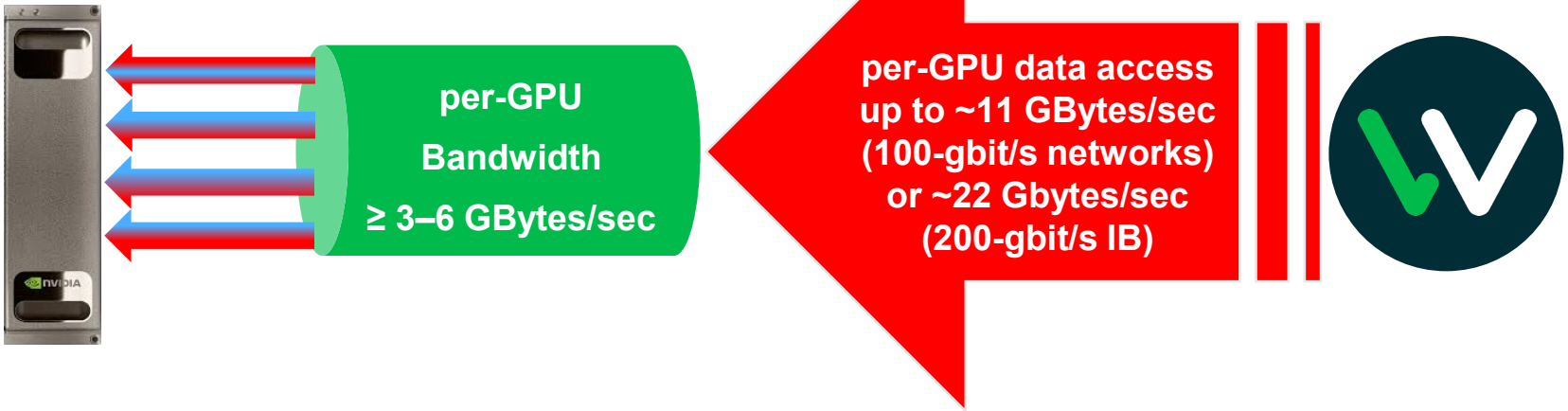
Shared File Storage via Distributed Parallel Filesystems



- Lustre, IBM Spectrum Scale (GPFS), Panasas
  - Best results typically for large-file sequential access
  - > 100,000 files per dir stressful for metadata servers
  - Complex to admin; designed for HDD

2019 Storage Developer Conference. © WekaIO, Inc. All Rights Reserved.

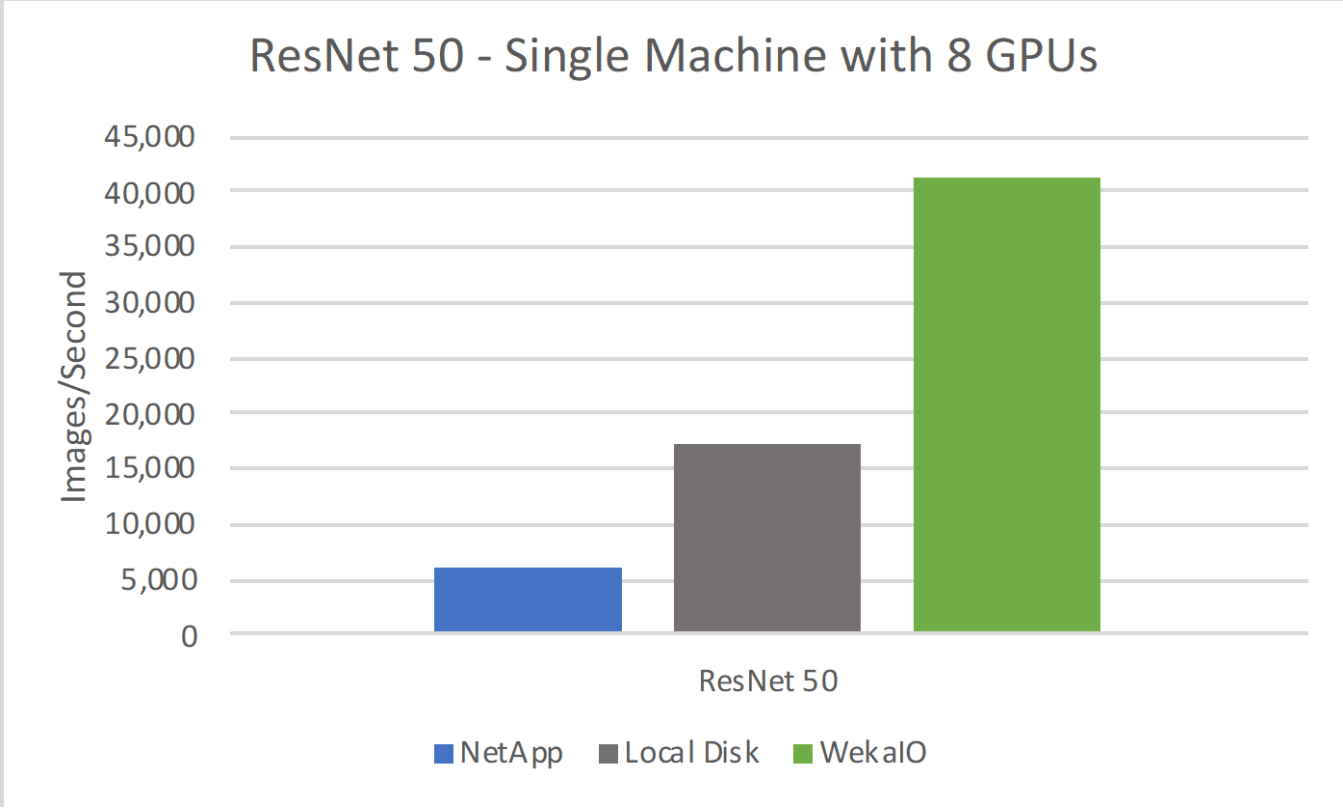
Shared File Storage via WekaIO Matrix™ Filesystem



- WekaIO's Matrix™ is a shared parallel fs
  - Designed exclusively for flash (not for HDD)
  - Optimized across WekaIO's own NVMe Fabric
  - InfiniBand or 100-GbE preferred (10-GbE minimum)

2019 Storage Developer Conference. © WekaIO, Inc. All Rights Reserved.

ResNet 50 Inference Benchmark

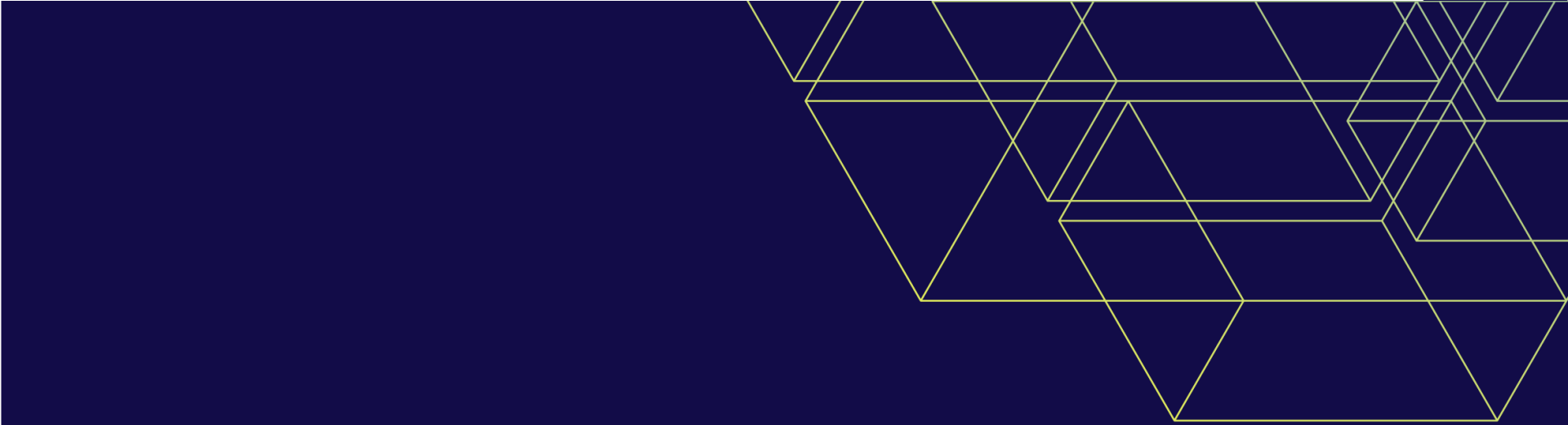


2019 Storage Developer Conference. © WekaIO, Inc. All Rights Reserved.

Linear Scaling — Resnet50 Training

2019 Storage Developer Conference. © WekaIO, Inc. All Rights Reserved.

THANKS



Please take a moment to rate this session.

Your feedback matters to us.