Flash Memory Summit

Next-Generation NVMe-Native Parallel Filesystem for Accelerating HPC Workloads

Liran Zvibel

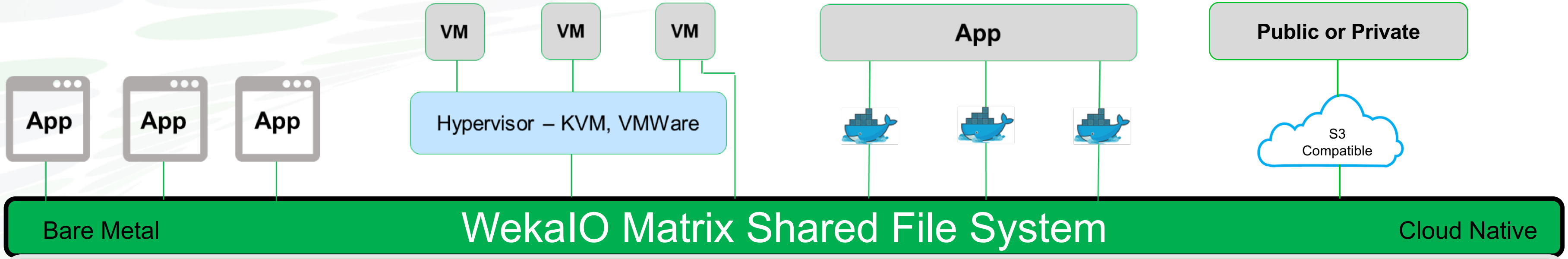
CEO, Co-founder WekaIO

@liranzvibel

Flash Memory Summit 2018
Santa Clara, CA

WEKA.IO

WekaIO Matrix: Full-featured and Flexible



Fully Coherent POSIX File System That is Faster than a Local FS

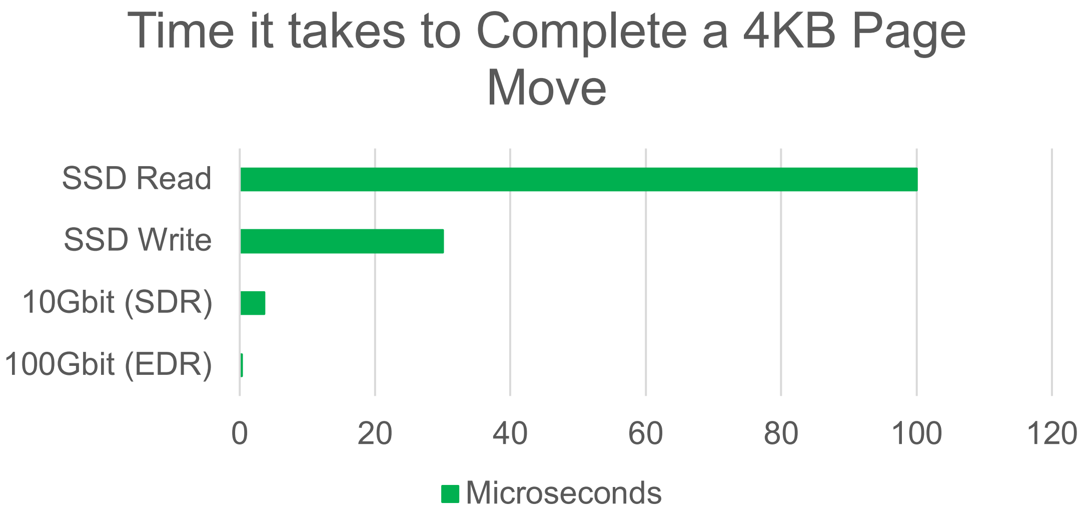
Distributed Coding, Scale-out metadata, Fast Rebuilds, End-to-End DP

Instantaneous Snaps, Clones, Performance Tiering to S3, DR, Backup

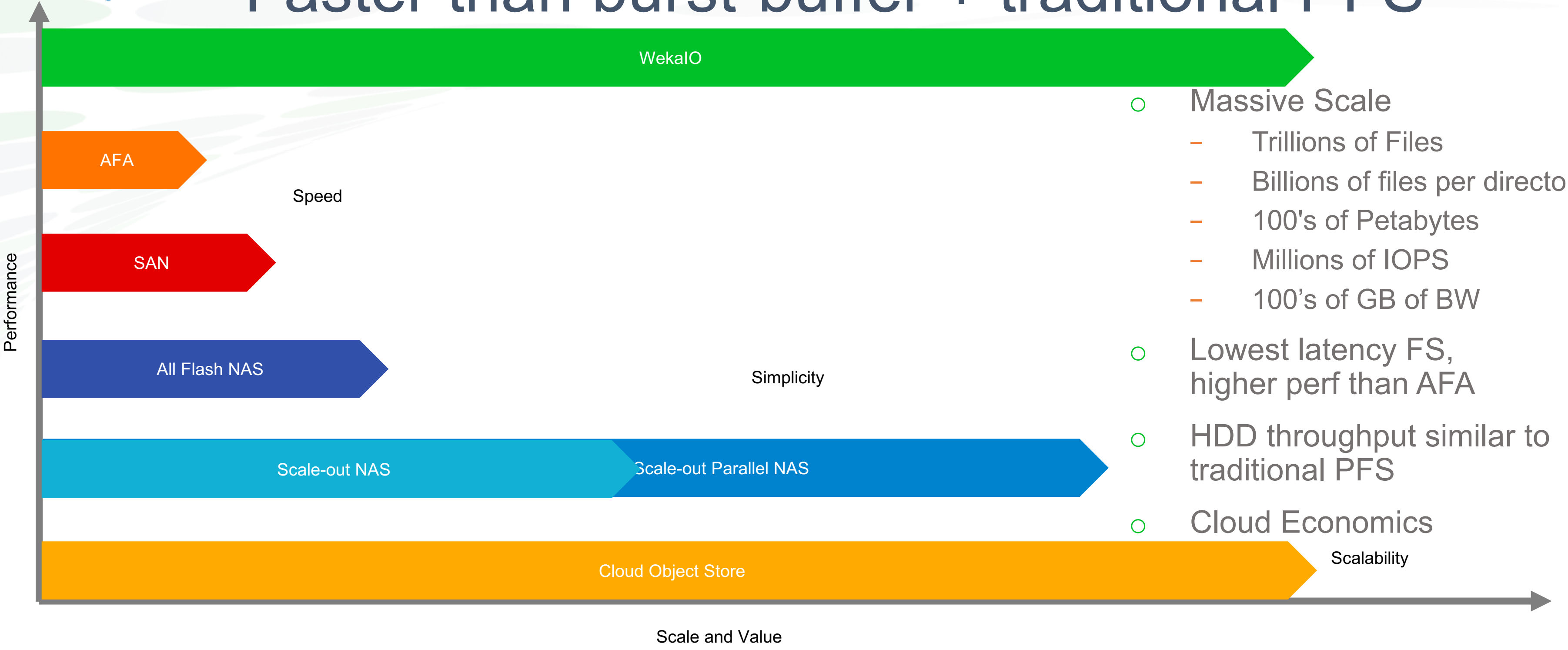
InfiniBand or Ethernet, Hyperconverged or Dedicated Storage Server

Why NVMe-oF parallel FS?

- Local copy architectures were developed when 1GbitE and HDDs were standard
- Modern networks on 100Gbit are 100x faster than SSD
- It is much easier to create distributed algorithms when locality is not important
- 4KB IOs latency similar to local FS, bigger IOs parallelize, so even lower latency



Only PFS for NVMe-oF, PFS over S3
Faster than burst-buffer + traditional PFS

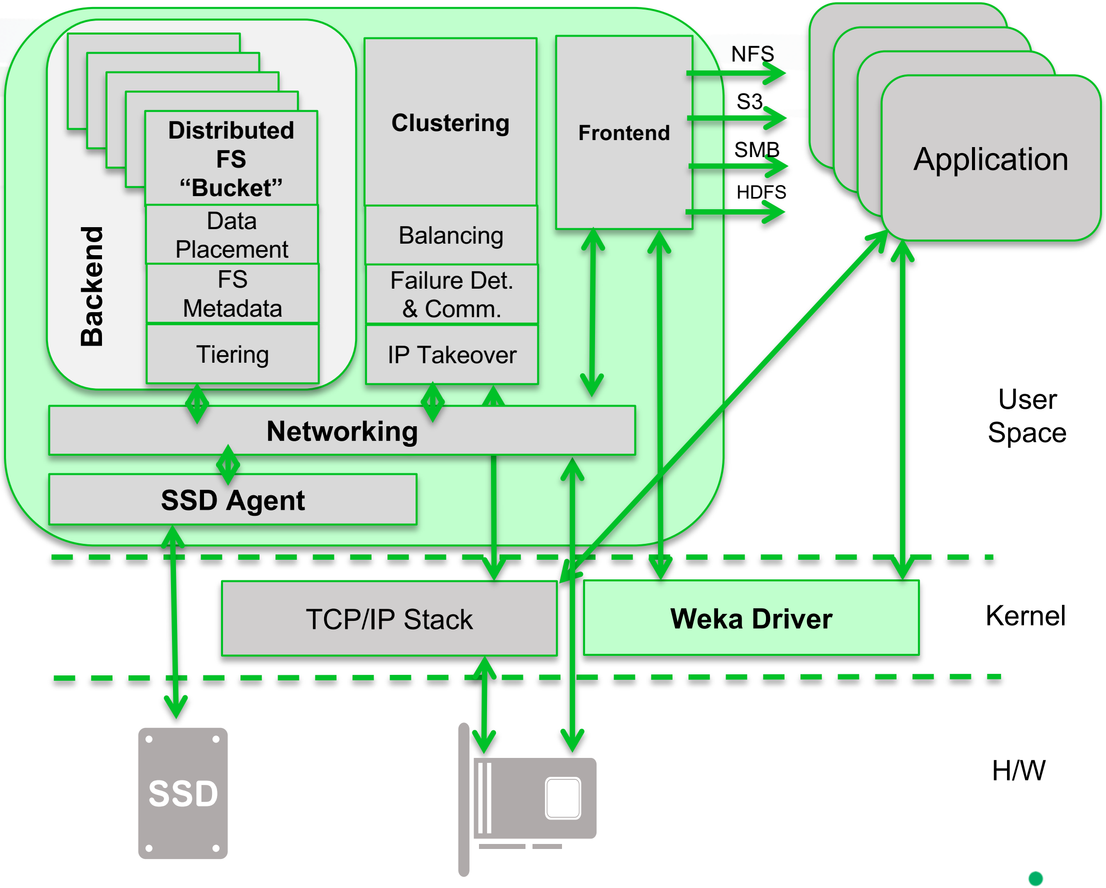


- Massive Scale
  - Trillions of Files
  - Billions of files per directory
  - 100's of Petabytes
  - Millions of IOPS
  - 100’s of GB of BW
- Lowest latency FS, higher perf than AFA
- HDD throughput similar to traditional PFS
- Cloud Economics

Flash Memory Summit

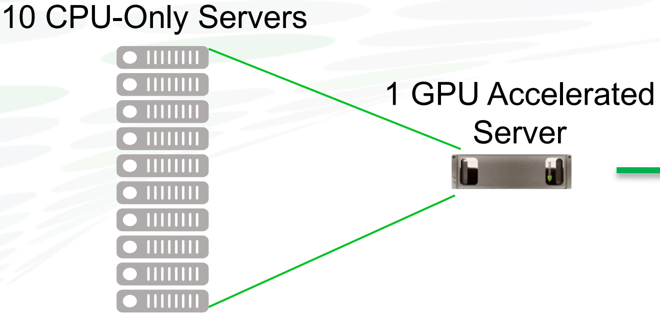
Software Architecture – Keep out of kernel

- Runs inside LXC container for isolation
- SR-IOV to run network stack and NVMe in user space
- Provides POSIX VFS through lockless queues to WekaIO driver
- I/O stack bypasses kernel
- Scheduling and memory management also bypass kernel
- Metadata split into many Buckets – Buckets quickly migrate ➔ no hot spots
- Support, bare metal, container & hypervisor

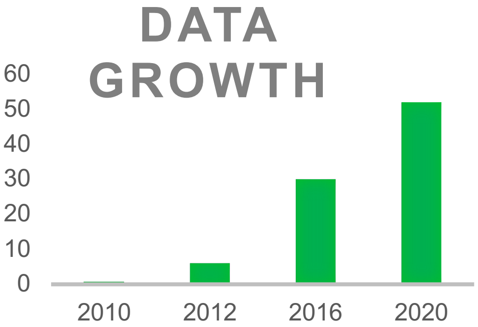


WEKA.IO

Processing Has Shrunk while Data Sets Explode



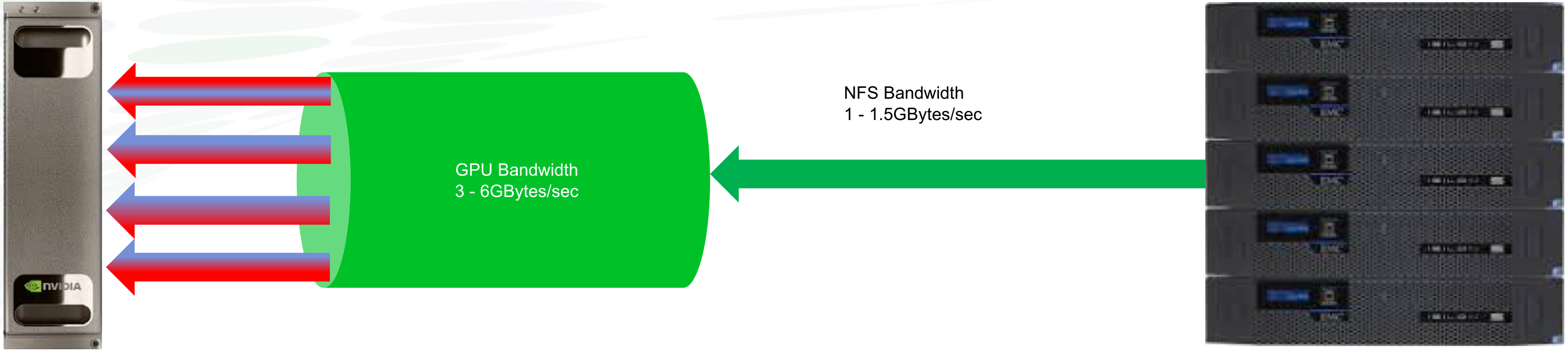
GPUs have shrunk compute infrastructure by 10x



But the data that needs processing has grown 50x

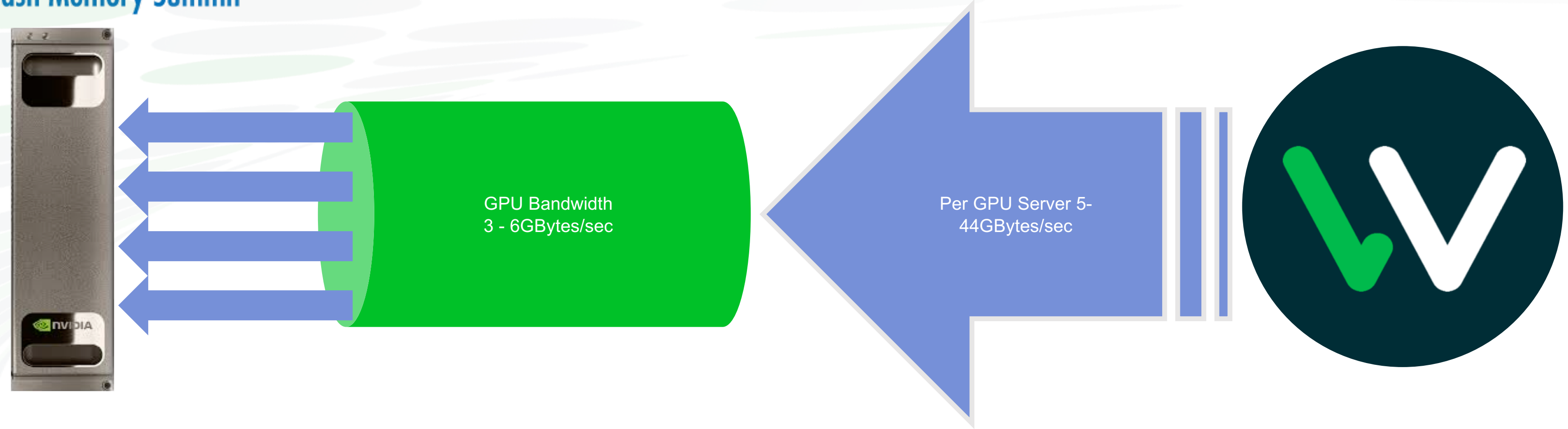
Industry is cornered into an I\O nightmare

NFS = Not For Speed



- A protocol developed in 1984 trying to solve a 2018 problem
- pNFS tried to fix NFS but failed when metadata workloads exploded
- Legacy parallel file systems like Lustre and GPFS cannot handle billions of small files
  - And they require a PhD to operate

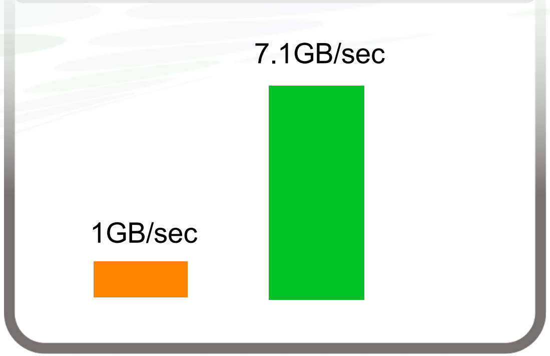
WekaIO Solves the Data Accessibility Problem



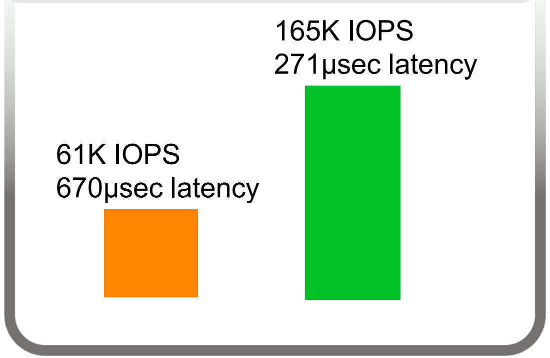
- Shared, Parallel file system written for NVMe
- POSIX Client runs on GPU Servers
- Cluster of servers provide high performance file services from NVMe
- Low latency networking on InfiniBand or Ethernet
- Training “data lake” stored on low cost object storage for best cost

Actual Results from Bake-off vs All flash filer

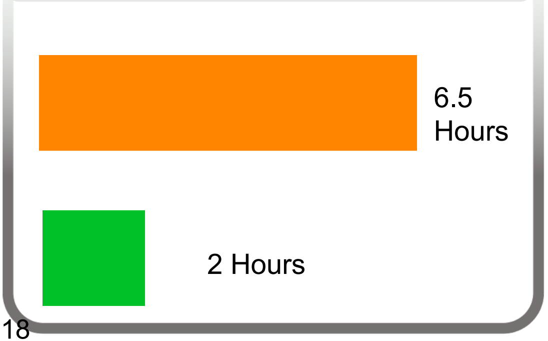
7x Faster 1MB Throughput



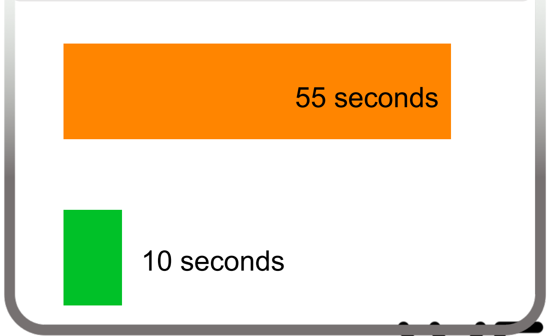
2.6x Better 4KB IOPS/Latency



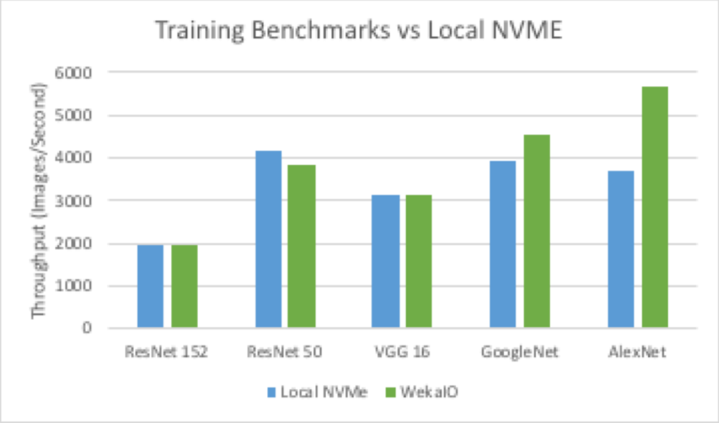
3.3x Better FS Traverse (Find)

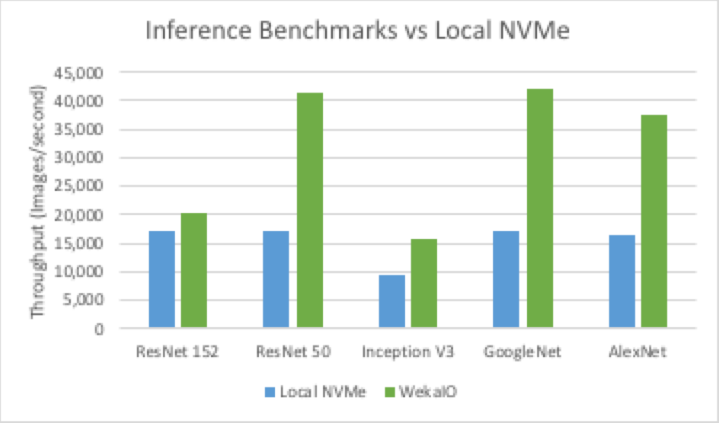


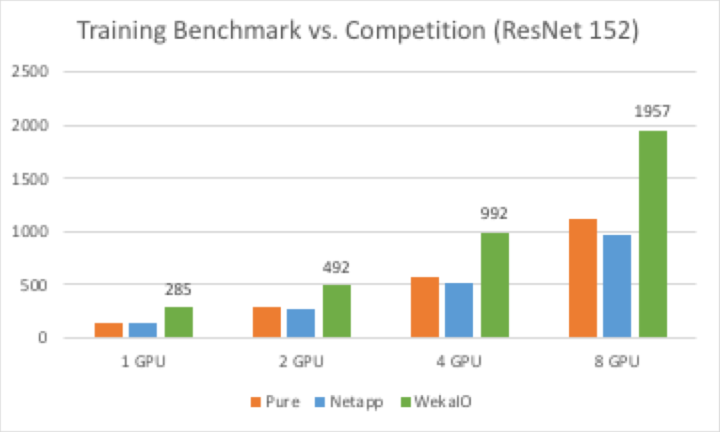
5.5x Better 'ls" Directory

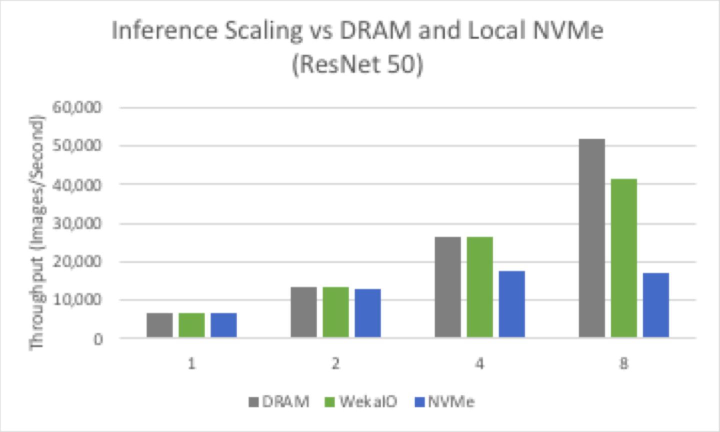


GPU Performance vs. Alternatives


Training Benchmarks vs Local NVME


Inference Benchmarks vs Local NVMe


Training Benchmark vs. Competition (ResNet 152)


Inference Scaling vs DRAM and Local NVMe (ResNet 50)

https://www.theregister.co.uk/2018/06/07/pure_beats_netapp_at_ai/

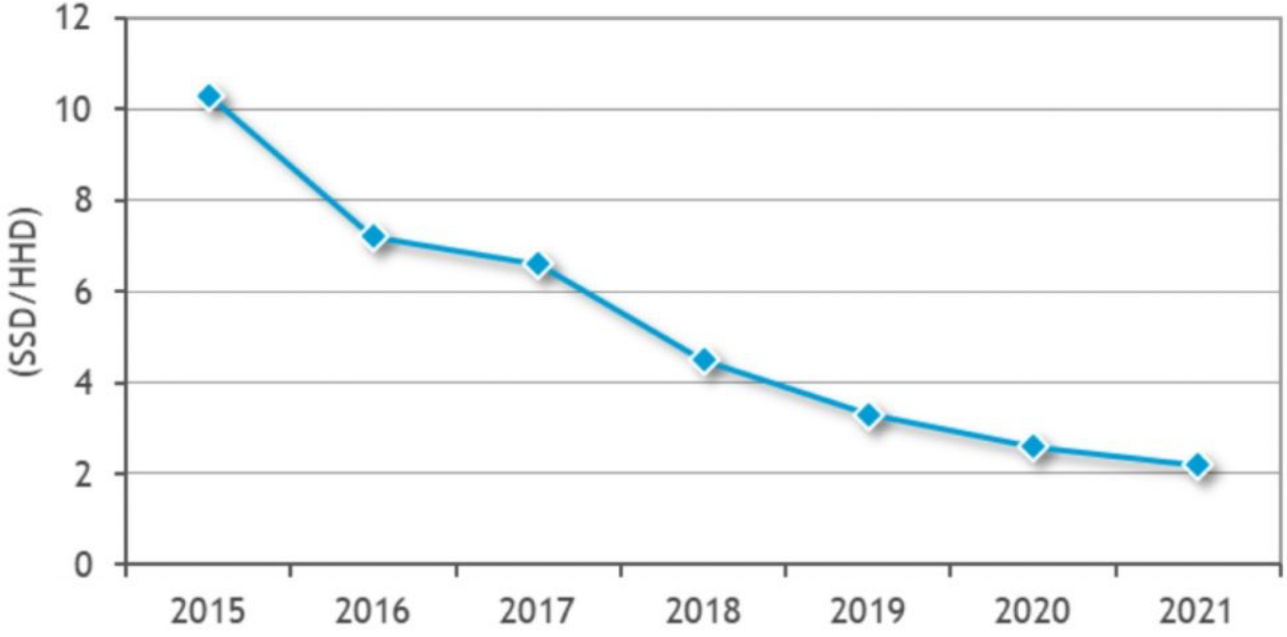
http://dlpg.labs.hpe.com

Deep Learning Requirements

- Actually very close to HPC problems…
- Store a vast amount of data
  - Effectively “**stage**” working set **back on fast storage**, for efficient access
- High bandwidth, **low latency**
- Very good **metadata performance**, traverse files quickly
  - Billions of files per directory, huge namespaces
- Very high **single host** performance
- Support **multiprotocol** (S3, HDFS, SMB, NFS)

Flash Memory Summit

SSD vs HDD pricing (per gb ratio)



Source: Hyperion research
https://www.storagenewsletter.com/2018/08/07/flash-storage-trends-and-impacts/

WEKA.IO

HPC only cares about throughput, right?

- NAND is cheaper for IOPS (and obviously latency) for several years now
- HDD stats: 160MB/sec ; $0.02/GB capacity for 10TB devices
- 3.84TB TLC devices read at 1700MB/sec ; so faster than 10 HDDs
  - Total HDD cost needed to read at 1700MB/sec → $2000; avg per NAND device $0.52/GB
  - Already cheaper today!
- 7.68TB QLC devices coming next year writing at 1000MB/sec; 6 HDDs needed
  - Total HDD cost needed to read at 1000MB/sec → $1200; avg per NAND device $0.16/GB
  - Next year QLC will be cheaper for write throughput
  - Endurance will probably not hold for checkpointing; but anywyas small capacity that TLC makes sense for

Future of HPC storage is NAND FLASH

- Currently HDDs still make sense for some workloads
- In a year (and obviously later) HPC storage should steer towards NAND FLASH technologies
- Parallel FS for NVMe-oF require different data structure, and algorithms based on modern workloads (scaling metadata, small IOPS, etc)
- HPC applications should consider NAND FLASH only for active workload; other media (tape;optics; etc) for archival capacity

Flash Memory Summit

WEKA.IO