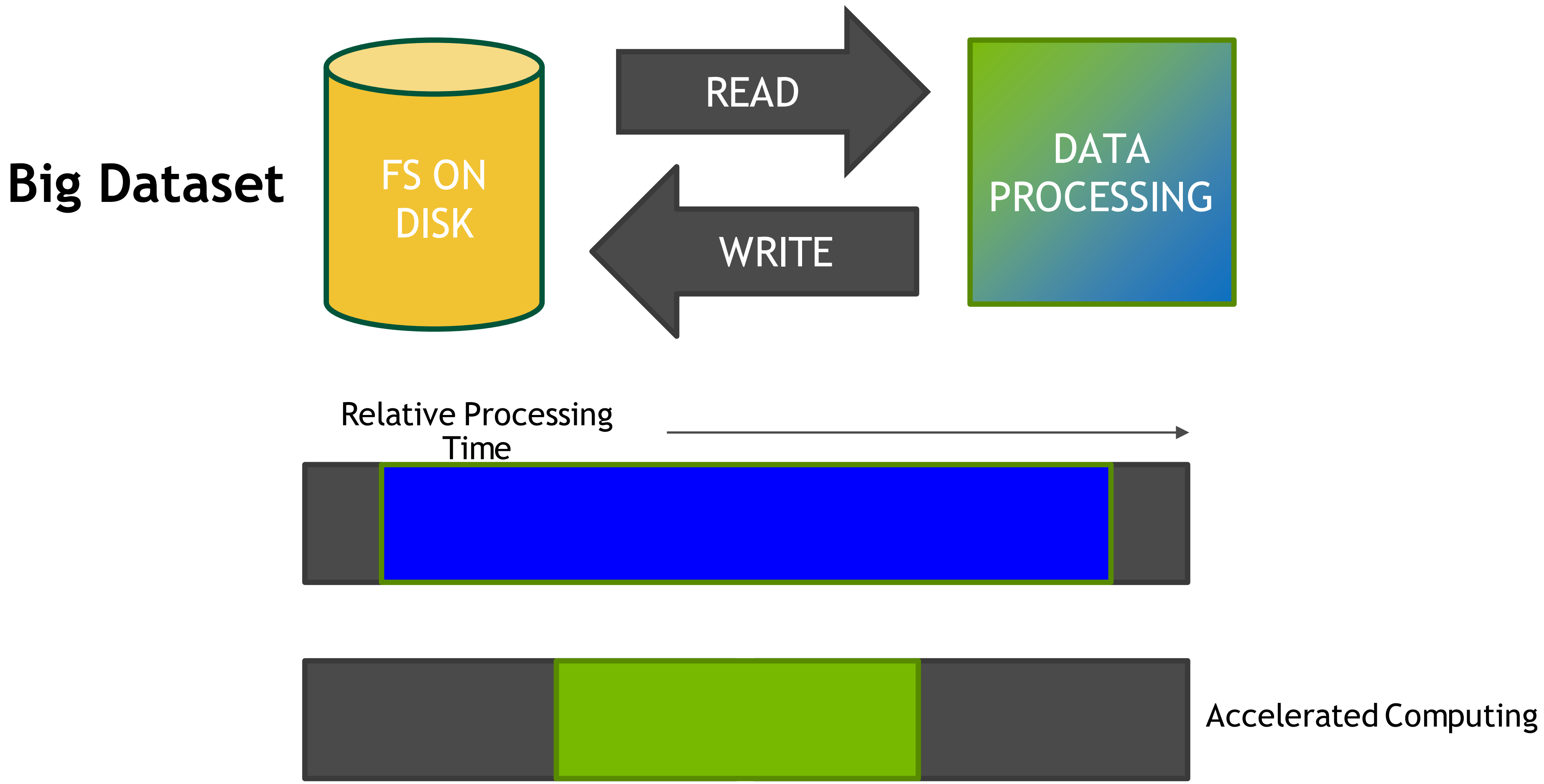
GPUDIRECT STORAGE: A DIRECT GPU-STORAGE DATA PATH

CJ Newburn, Kiran Modukuri, Akilesh Kailash, Saptarshi Sen, Sandeep Joshi

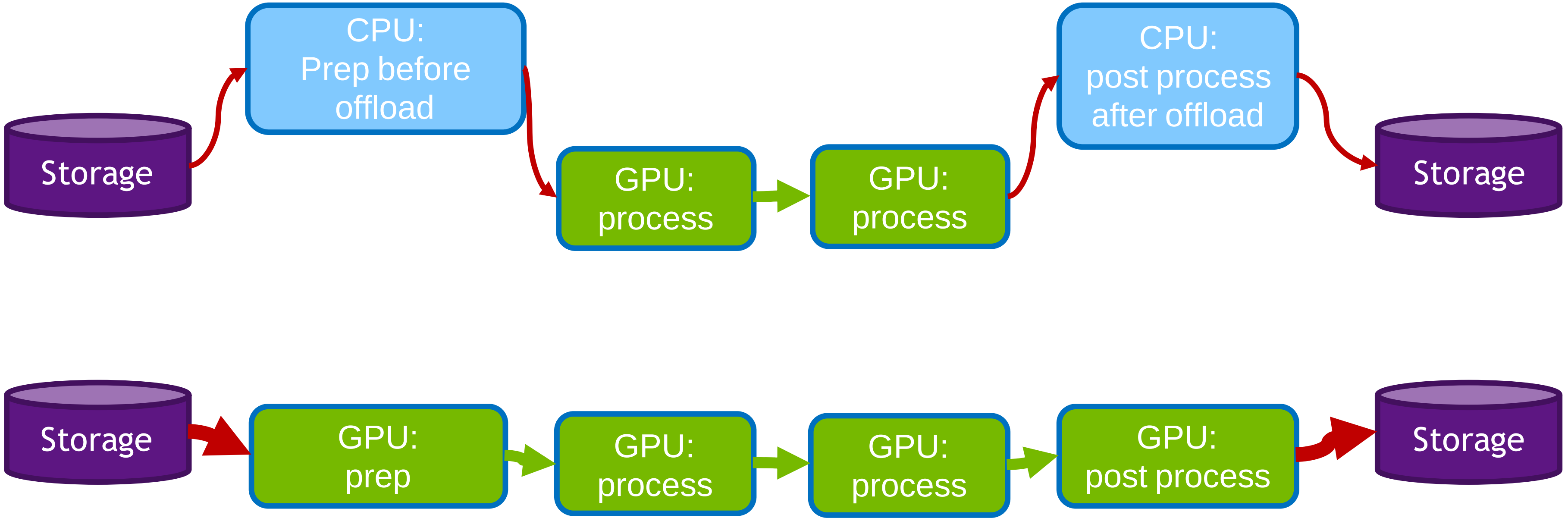
Shiva Shankar, Barton Fiske

THE CHALLENGE



SHIFTING TO THE GPU

IO acceleration to GPU is a force multiplier for compute acceleration



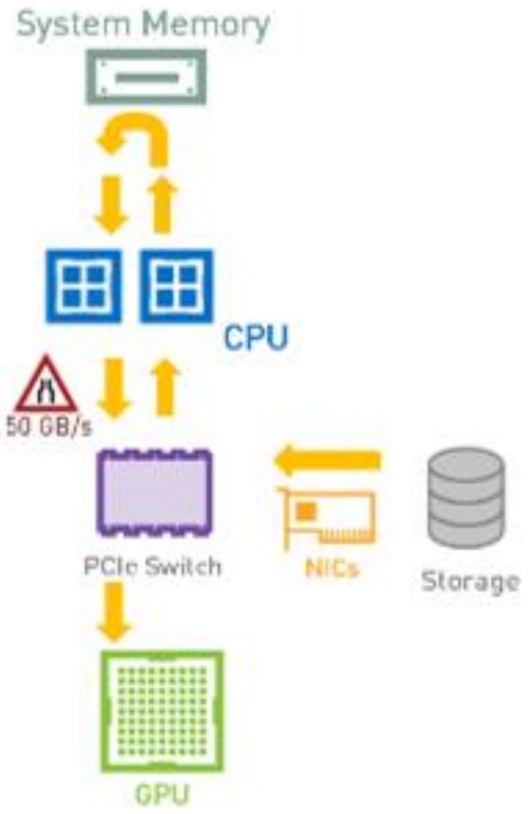
GETTING DATA TO GPUS: IO

CPUs introduce bottlenecks on data path between storage and GPUs

SysMem - GPU: **48-50** GB/s with 4 PCIe trees, **1TB**

Bottlenecks

- Copying into bounce buffer in SysMem
- Bandwidth limitations into CPU and back
- Load interference on CPU, GPU
- cudaMemcpy overhead



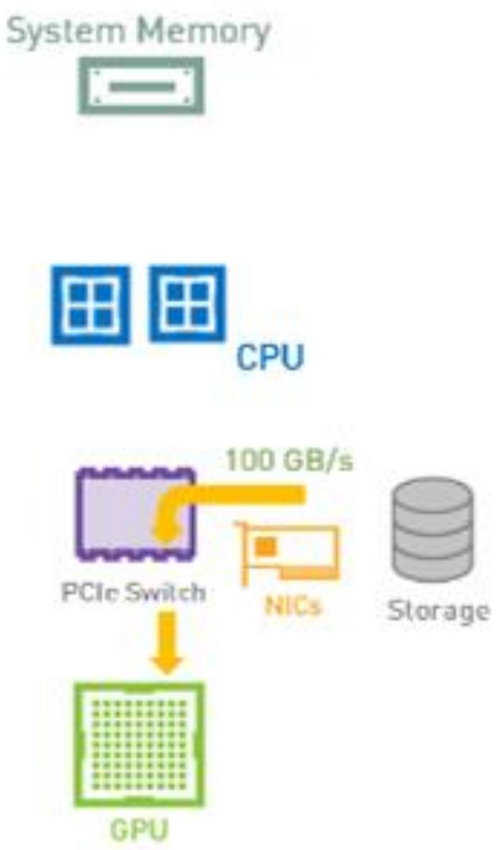
SYSMEM vs. STORAGE

IO bandwidth to GPUs exceeds SysMem bandwidth to GPUs

More capacity and bandwidth (DGX-2)

SysMem - GPU: **48-50** GB/s with 4 PCIe trees, **1TB**

Local storage - GPU **53.3** GB/s with 16 drives, **O(100TB)**



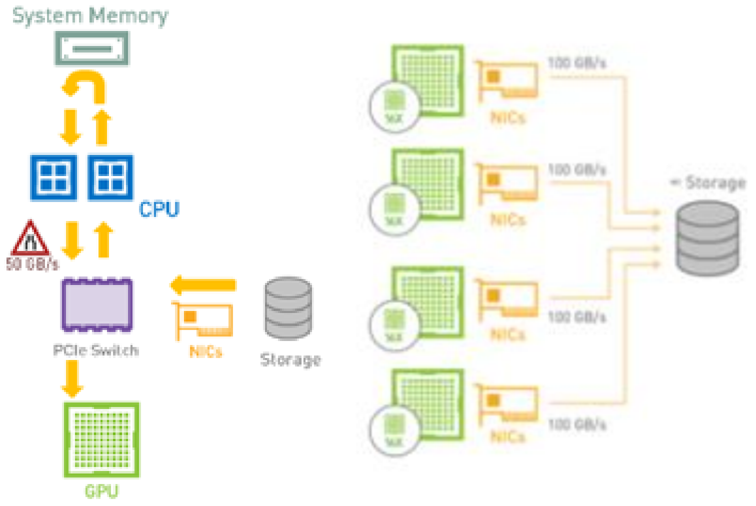
NVIDIA.

SYSMEM vs. STORAGE

IO bandwidth to GPUs exceeds SysMem bandwidth to GPUs
More capacity and bandwidth (DGX-2)

SysMem - GPU: **48-50** GB/s with 4 PCIe trees, **O(1TB)**

Local storage - GPU **53.3** GB/s with 16 drives, **O(100TB)**



Storage outside enclosure

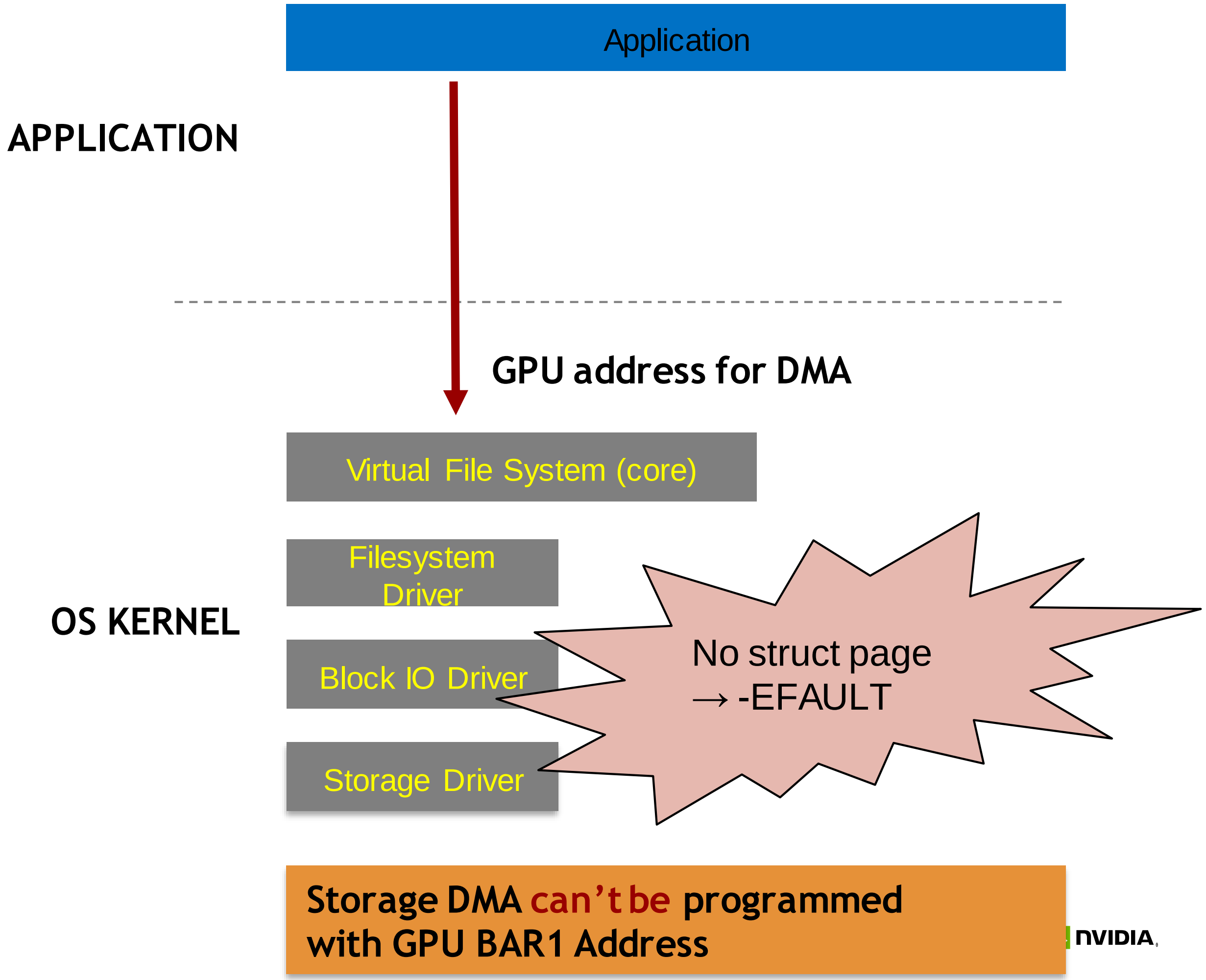
- 8 RAID cards @ 8*14=112* GB/s, **O(100TB)**
- 8 NICs, e.g. IB NVMe-oF, @ 8*10.5=**84** GB/s, **O(PB)**

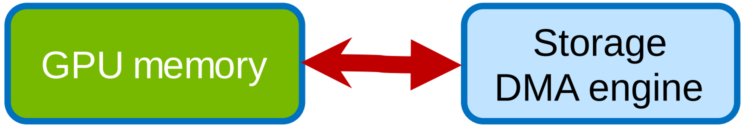
**Measured by MicroChip*

GPUDIRECT STORAGE SW ARCHITECTURE

User and kernel components

Linux is not enabled to handle
GPU Virtual Addresses
needed for DMA





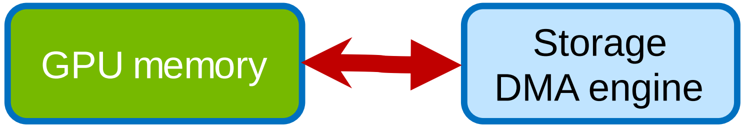
GPUDIRECT STORAGE SW ARCHITECTURE

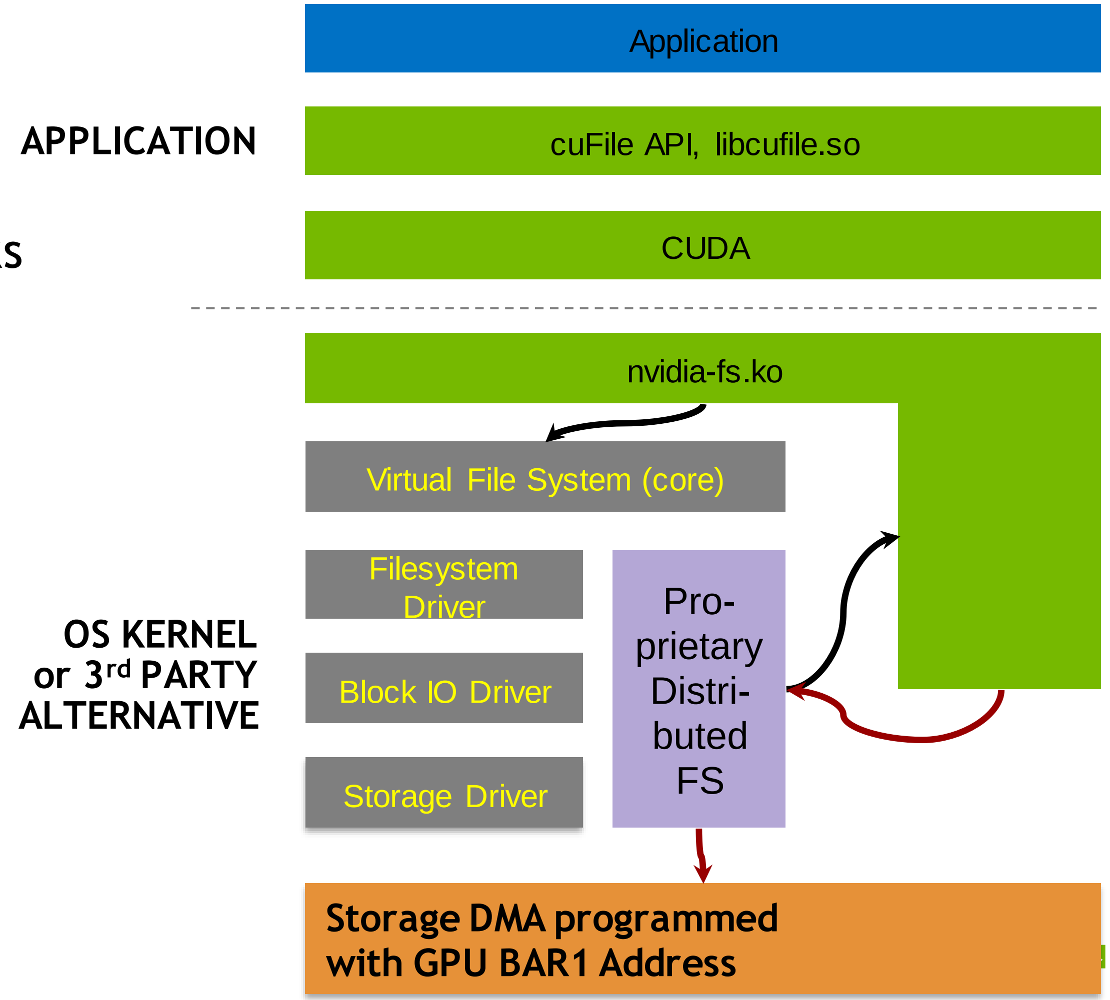
User and kernel components

cuFile API
Enduring API for applications and frameworks

nvidia-fs Driver API
For filesystem and block IO drivers
Vendor-proprietary solutions: no patching
avoid lack of Linux enabling

NVIDIA is actively working with the community on upstream first to enable Linux to handle GPU VAs for DMA





RELEASE TIMELINE

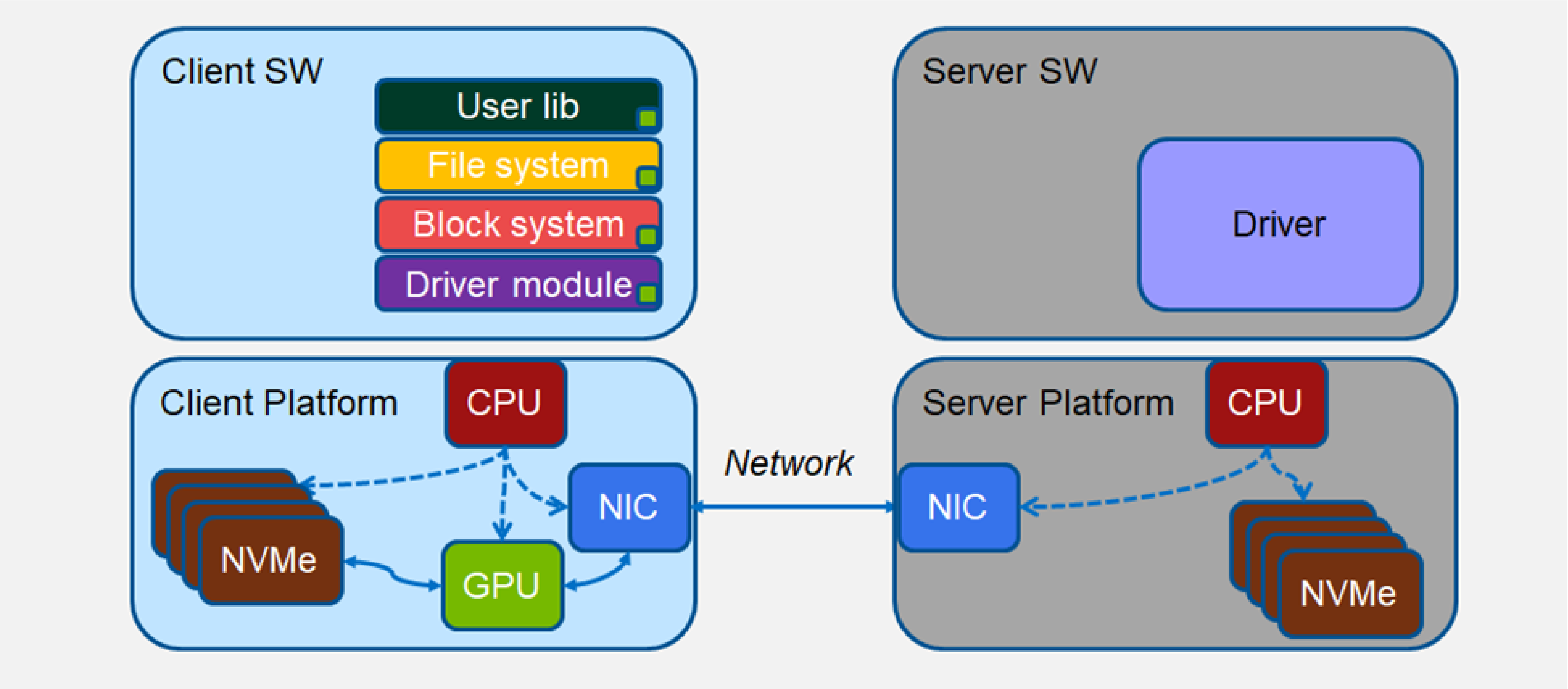
Q3'19
Closed Alpha
Vet design
Establish partners

SC'19 Launch
Launch with partners
Enhance design
E2E showcase solns

Early '20
Open Beta
Enhance design
Launch with partners
Augment E2E solns

Mid '20
GA
Incorporate into CUDA
Broaden partners
Broaden E2E solns

A DISTRIBUTED FILE SYSTEM



MAGNUM IO PARTNERS

A broad and growing ecosystem

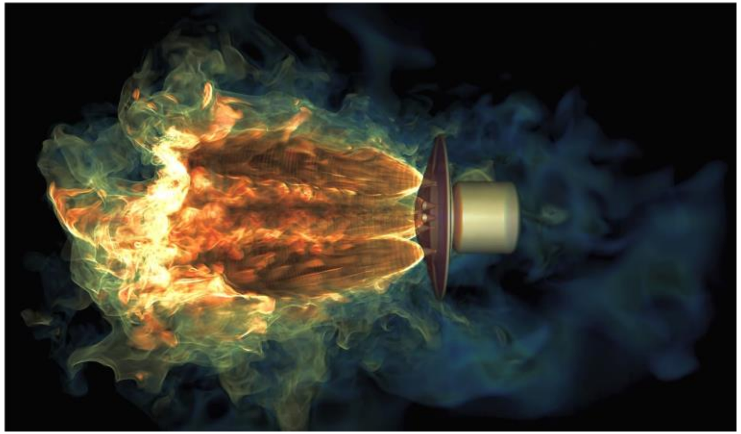
GDS-enabled

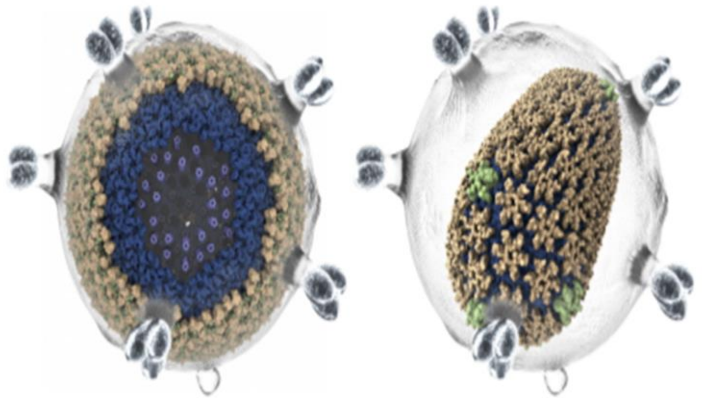
Evaluation

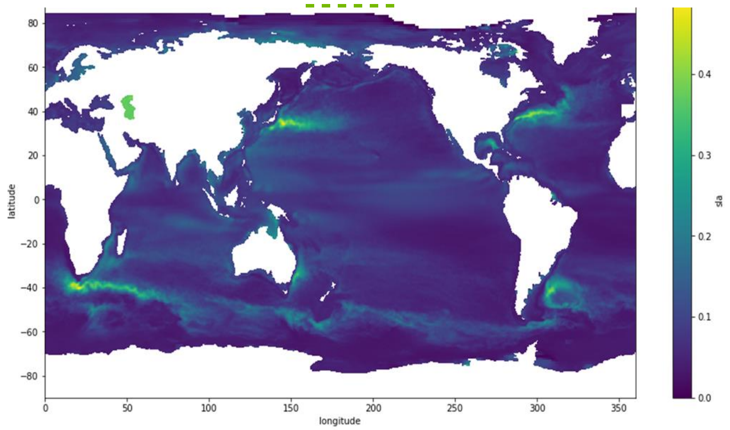
Ecosystem

OEMs

GPUDIRECT STORAGE – USE CASES IN NV BOOTH







<u>NASA Mars Lander</u>

Simulation → visualization
128TB data, must stream in from remote
Part render, part IO; not quite linear in IO

5 GB/s (1 DGX-2) → 160-180 GB/s (4 DGX-2s) with GPUDirect Storage

<u>Molecular Dynamics</u>

Simulation → analytics → visualization
30TB data, can be remote + local
O(N^2) IO problem to build dissimilarity matrix of macromolecule poses across time steps so can find stable configs

7 GB/s → 22 GB/s local, 60 GB/s remote
3x from GDS vs. heroic effort, 4x threads

<u>Pangeo Earth Science</u>

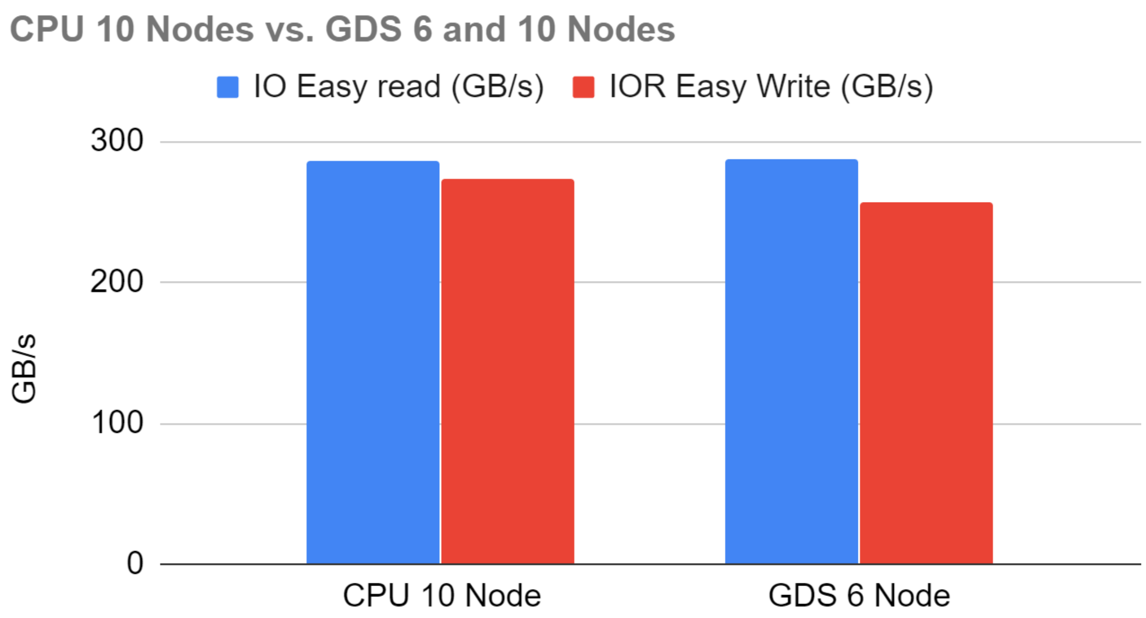
Simulation → DA, DL → visualization
100 TB–PB data, streamed from remote
Coming to GPUs because of faster IO
Increasing richness: DA, DL

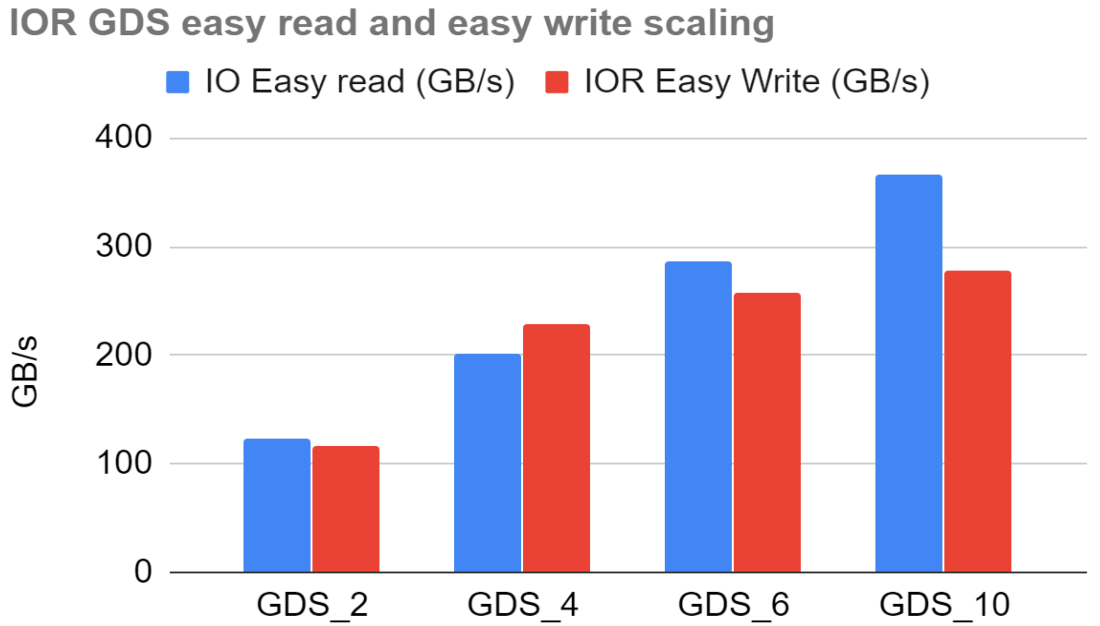
Moving from 1 per day to 2-3 per day
What ifs vs. safe bets

IO500

- IO to compute dominates
- DGX-2: 16 GPUs, 8 NICs, 2 CPUs
- Move data directly to GPUs
- Relieve CPU bottleneck

6 or 10 DGX-2s,
10 DDN EXA5 on A3I AI400X
GPUDIrect Storage used for IOR easy



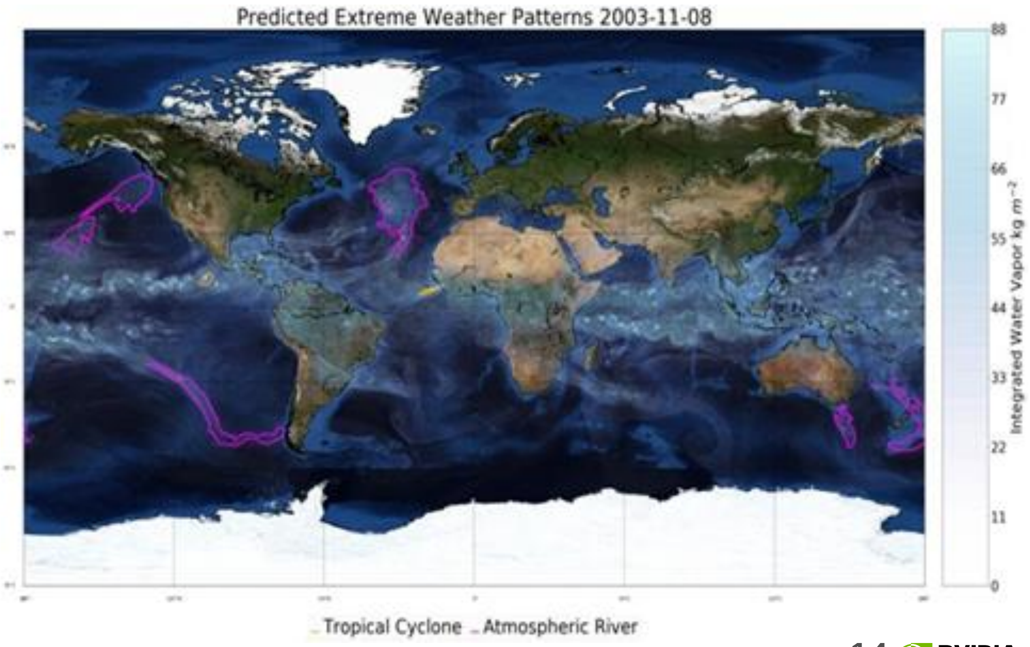


GPUDIRECT STORAGE UNDER PYTORCH

Proof of concept for deep learning with DeepLabv3+/Tirumisu

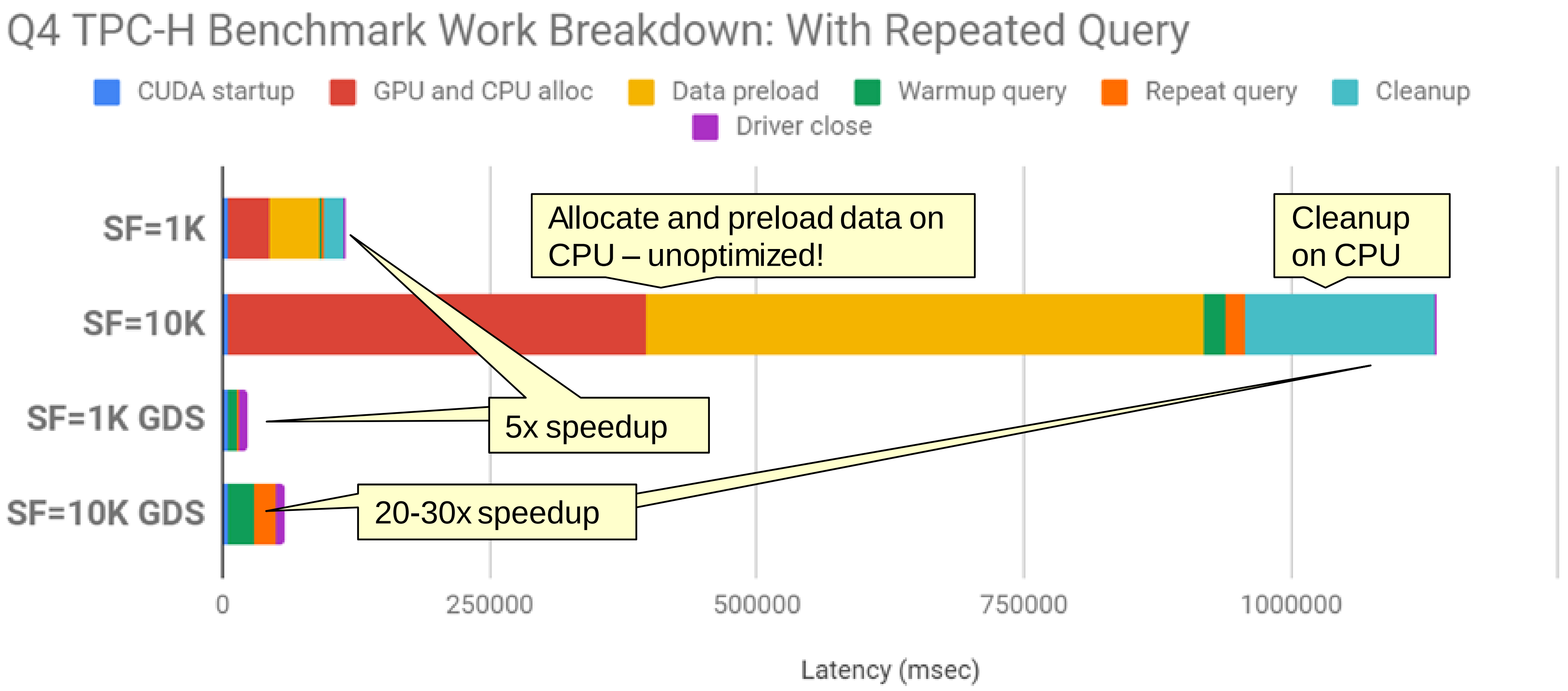
- Prototype: PyTorch for .npy
- 2x perf vs. hand-optimized Python multiprocess-based input pipeline
  - CAM5 climate science dataset (Gordon Bell Prize 2018)
- Other readers
  - Adotable for TFRecord, LMDB
  - NVIDIA DALI planned for near future
    - → TensorFlow, MXNet, PaddlePaddle
  - HDF5 for scientific workloads is possible

Courtesy of Thorsten Kurth



TPC-H (real but extreme case)

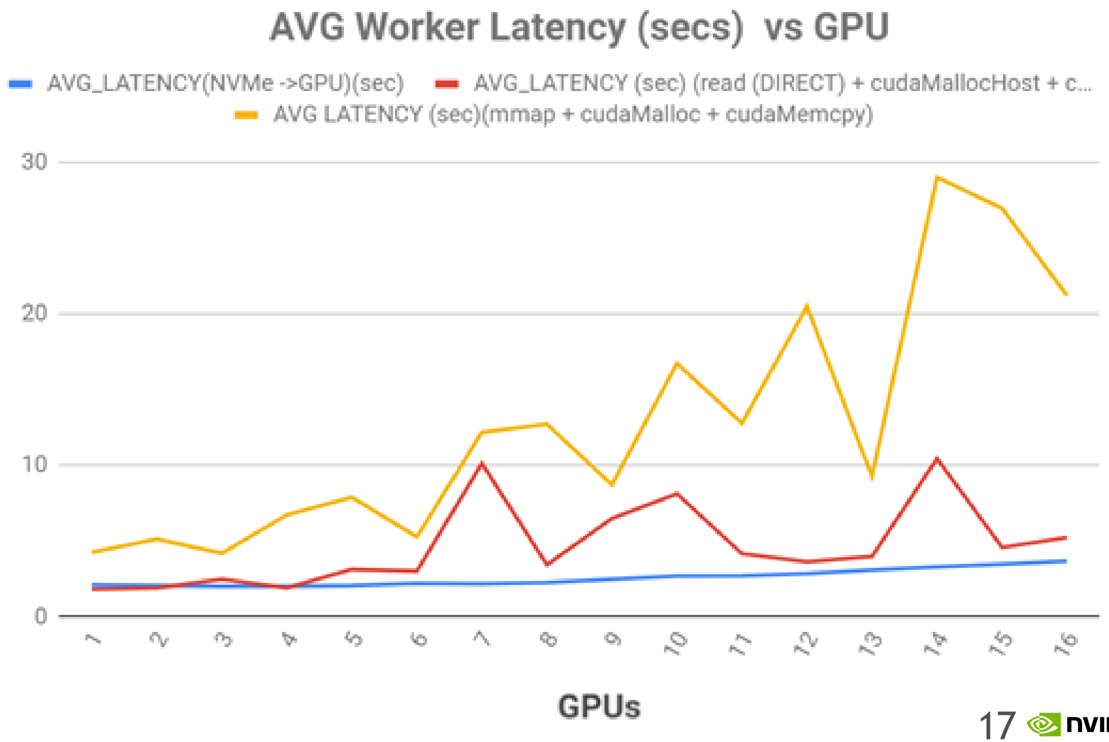
Speedups from both IO and savings in CPU memory management



LATENCY COMPARISON FOR cuDF

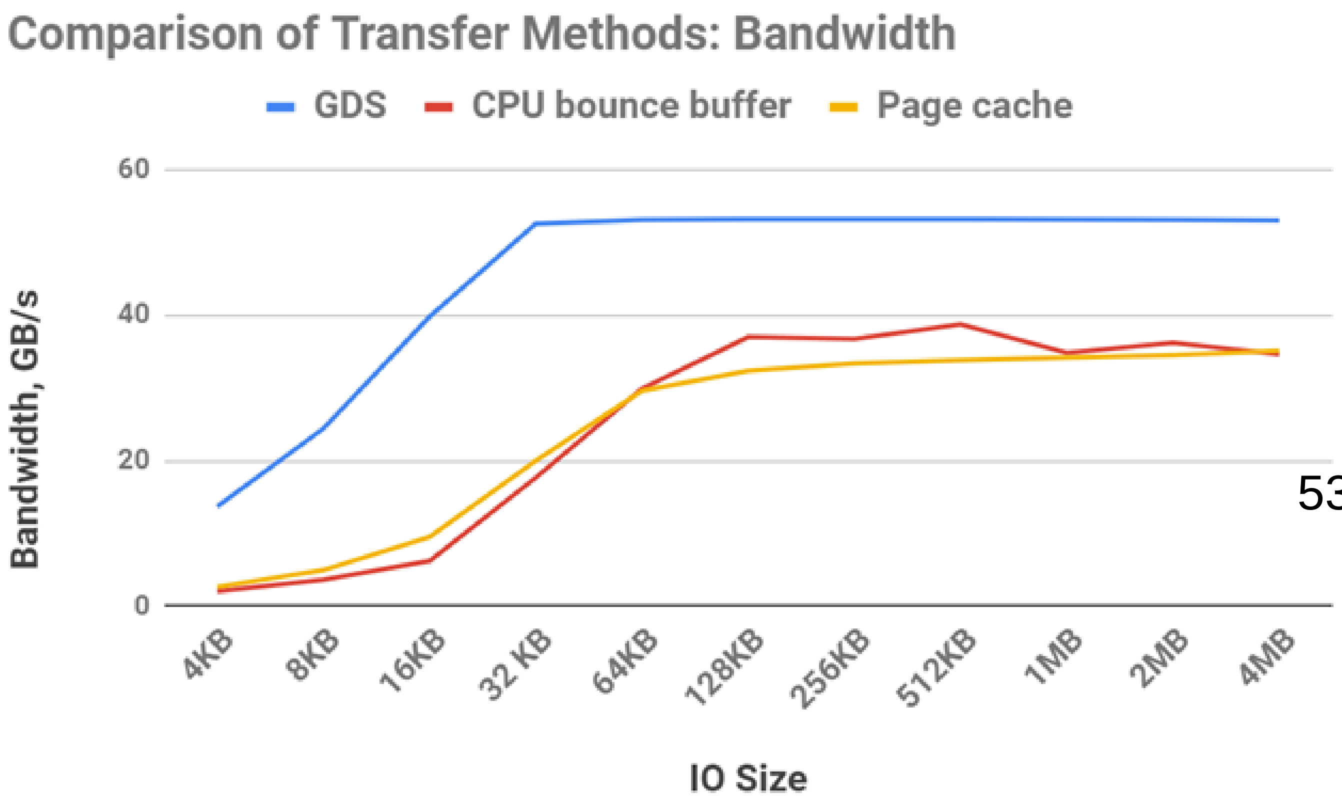
Direct has better scaling and stability

- Progression of latency improvements
  - Original, with 2 faults: slow, jittery
  - Best without direct: better
  - Direct: fairly flat, smooth
- Jitter
  - CPU interference leads to variation
  - Less predictable



HIGHER BANDWIDTH

Direct path leads to more throughput



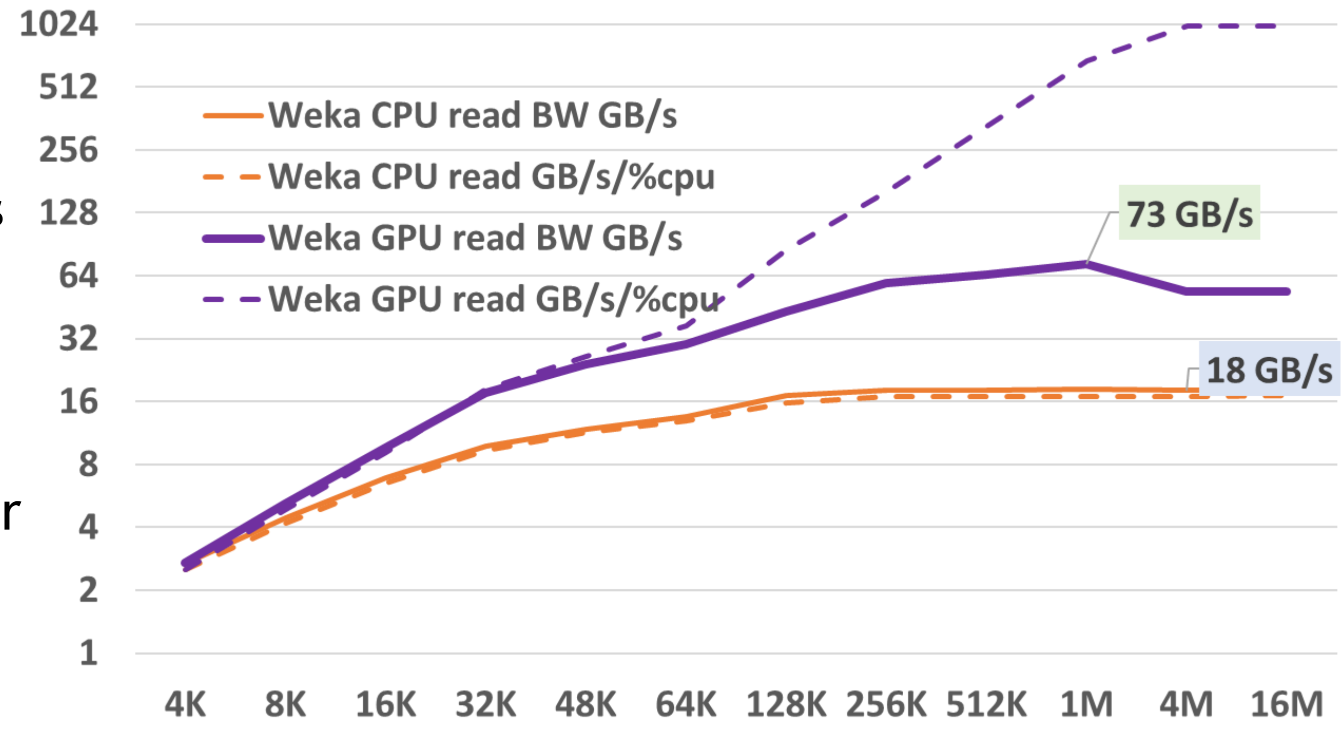
DGX-2 with 16 drives, 53.3 GB/s, would be more with remote IO

BANDWIDTH EFFICIENCY

Reducing the impact on the CPU

- Remote IO: 73 GB/s
  - One DGX-2, 8 NICs
  - Weka.io: 2 Supermicro servers

- CPU for control path
- GDS avoid CPU's data mov't
- GDS beats CPU on 1GB reads
- User-level systems offer greater CPU efficiency: 1000 vs. 17



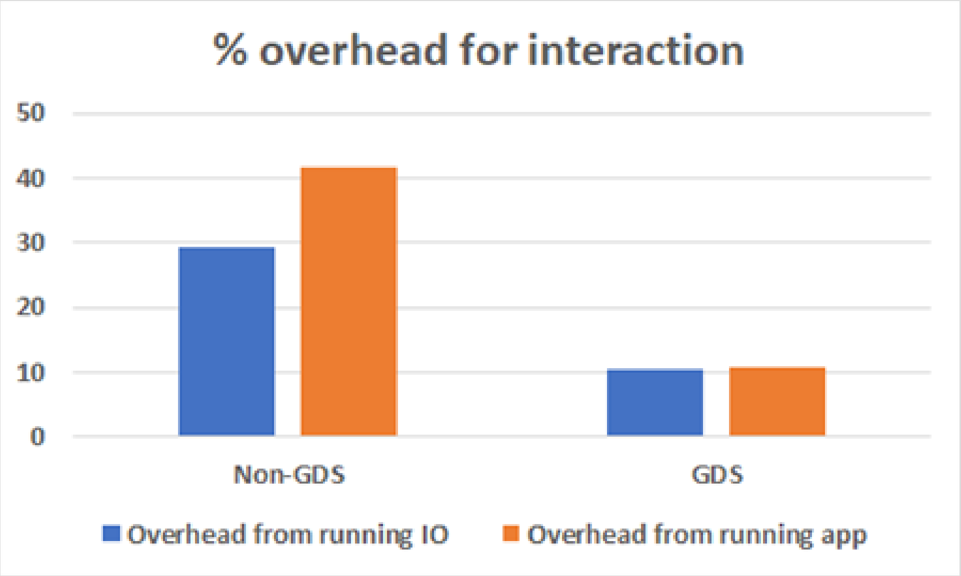
DIRECTED OVERHEAD TESTS

Less interference with and by the CPU

McCaplin Stream running alongside IO

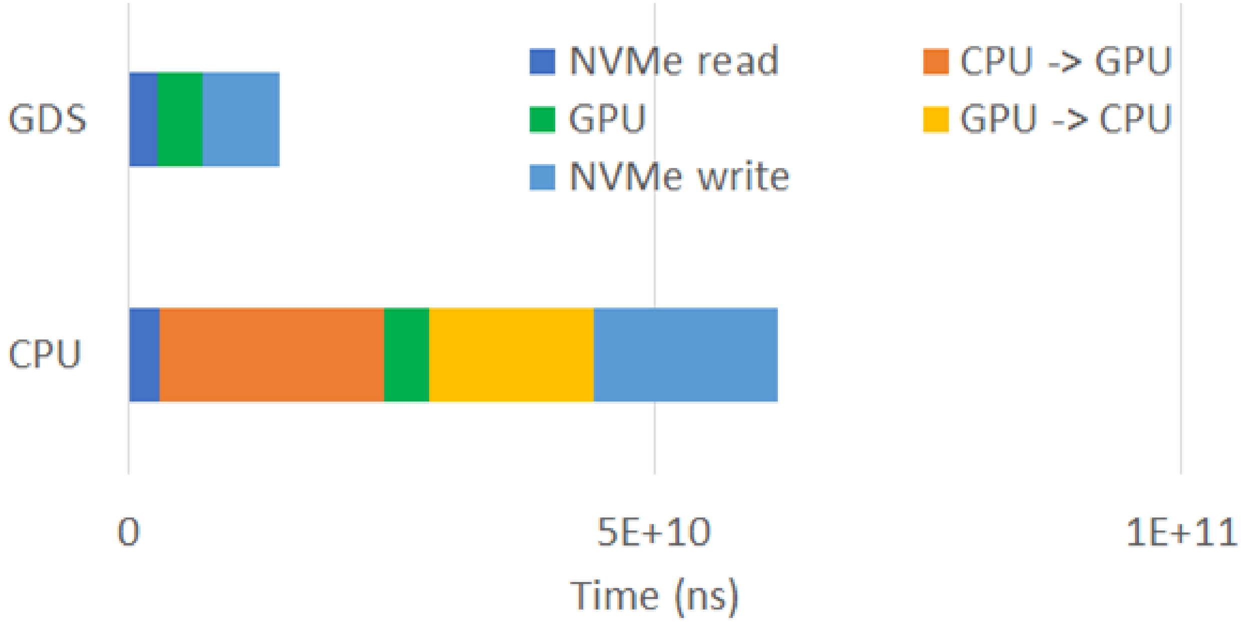
GPUDirect Storage

- Stays out of the way 10% vs 29%
- Less affected by CPU utilization 10% vs. 42%



Read, compute, write

4.3x speedup on just 4 drives in one DGX-2



- Courtesy of Matthew Nicely. 4 NVMe drives, 30GB transfers

TAKEAWAYS

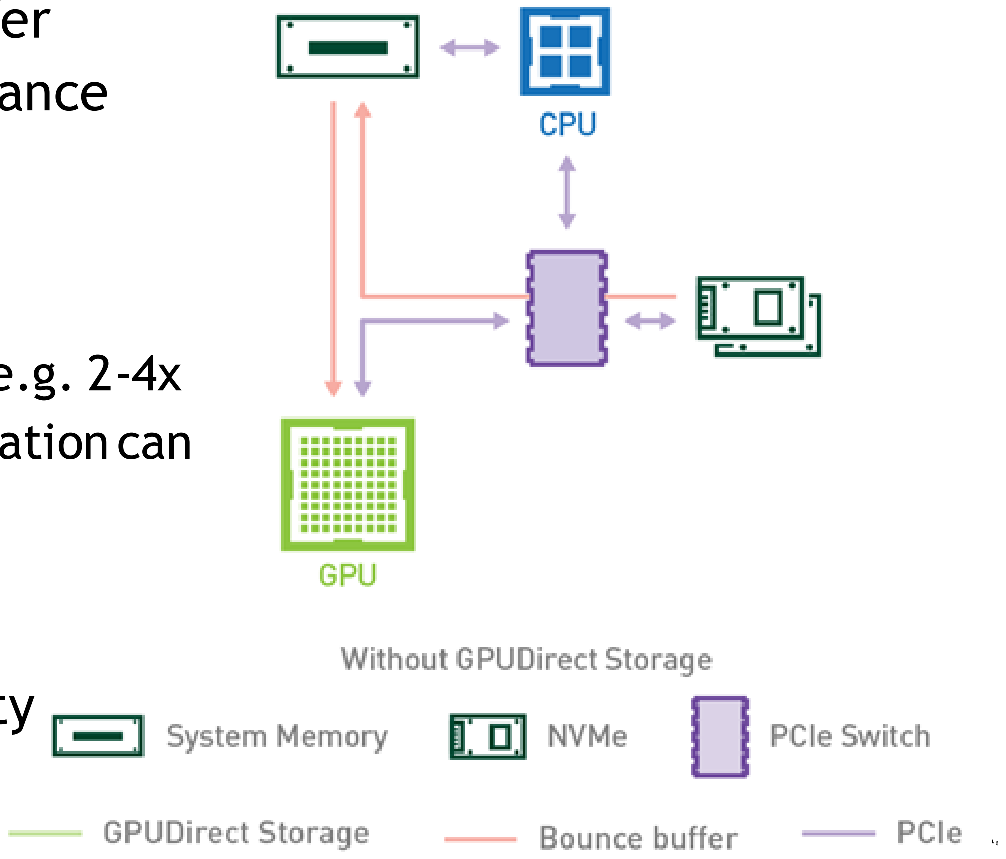
- GPU is now the fastest computing element with the fastest IO
- The easiest way to get IO performance
    - Higher bandwidth, lower latency
    - Varies by platform; we’ve seen 2-4x in several cases
- Broad ecosystem interest, active enabling
- Gathering end to end use cases → more readers (Zarr, HDF5, TFRecord)
- Working with the broader Linux community
- Coming to a CUDA near you

One-slide summary

GPUDIRECT STORAGE

The easiest way to get IO performance with a direct IO path to storage

- Avoids copying through a CPU bounce buffer
- cuFile APIs: easiest way to get IO performance
  - Bolster bandwidth, lower latency
  - Avoid CPU, GPU utilization burden
- Performance
  - Raw IO bw difference varies by platform, e.g. 2-4x
  - Savings in memory management and utilization can be a force multiplier on top of that
  - Varies by platform
- Broad ecosystem interest, active enabling
- Enabling with the broader Linux community
- Coming to a CUDA near you



SAMPLE USE CASES

Breadth of applicability

Demonstrations

- Synthetic file IO
- cuDF/cuIO
- TPC-H, Query 1 has no joins; Query 4 does joins
- Visualizing simulations
- DA for earth science with 3D data
- Molecular dynamics trajectory analysis

Additional cases

- Data lake
- Graph analytics
- Deep learning
- Checkpointing