G2M RESEARCH

# G2M Research Multi-Vendor Webinar: Utilizing HPC-Scale Storage and AI for Business Intelligence

- Tuesday May 19, 2020

# Webinar Agenda

| | |
|---|---|
| **9:00-9:05** | Ground Rules and Webinar Topic Introduction (G2M Research) |
| **9:06-9:41** | Sponsoring Vendor presentations on topic (9 minute each) |
| **9:42-9:51** | Key Question 1 (2-minute question; 2 minutes response per vendor) |
| **9:52-9:52** | Audience Survey 1 (1 minutes) |
| **9:53-10:02** | Key Question 2 (2-minute question; 2 minutes response per vendor) |
| **10:03-10:03** | Audience Survey 2 (2 minutes) |
| **10:04-10:13** | Key Question 3 (2-minute question; 2 minutes response per vendor) |
| **10:14-10:23** | Audience Q&A (10 minutes) |
| **10:24-10:25** | Wrap-Up |

# G2M Research Introduction and Ground Rules

Mike Heumann

Managing Partner, G2M Research

# Panelists

NVIDIA / Mellanox TECHNOLOGIES

Rob Davis
VP, Storage Technology
www.mellanox.com

SAMSUNG

Young Paik
Sr. Director, Product Planning
www.samsung.com

WEKA

Shailesh Manjrekar
Head of AI, Strategic Alliances
www.weka.io

DATYRA

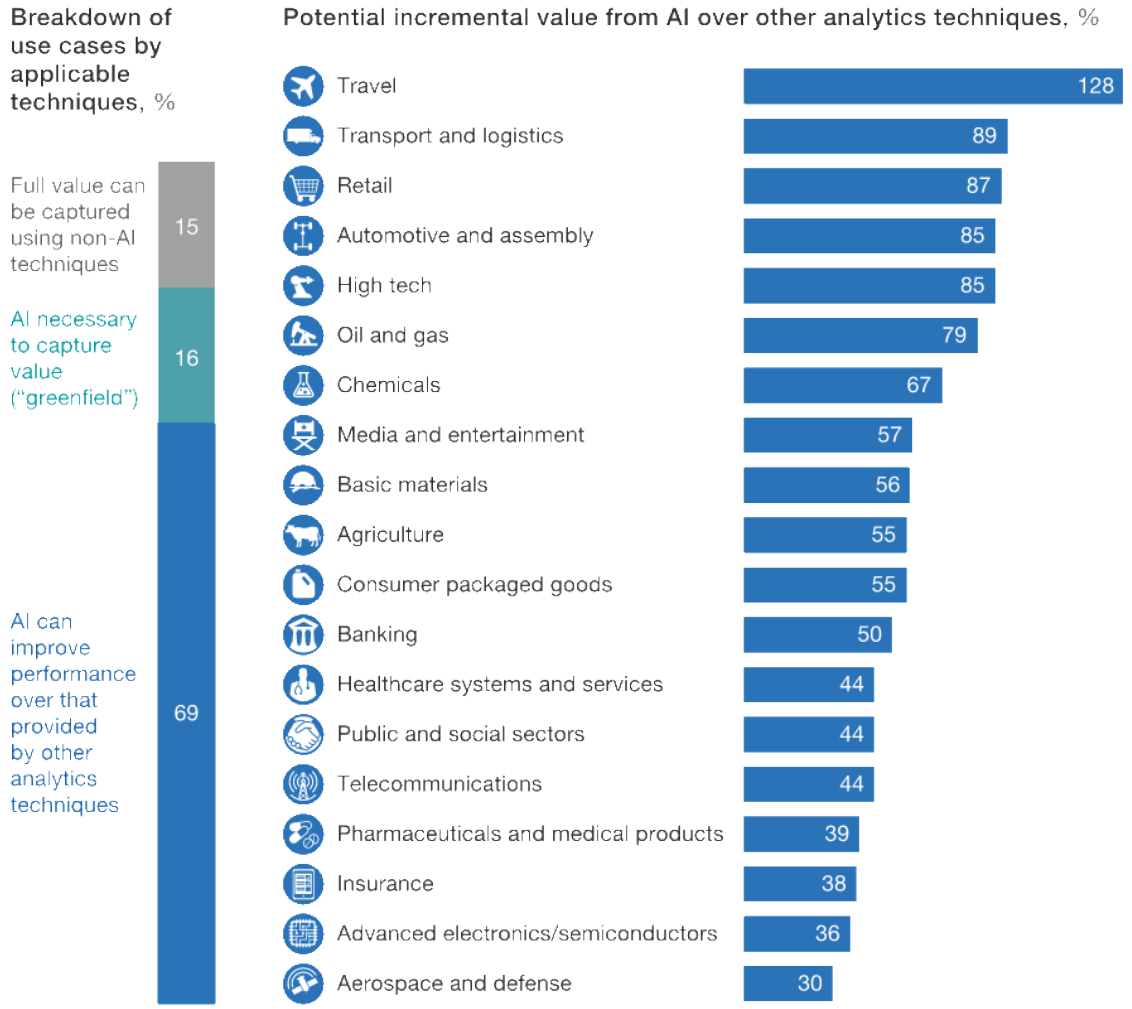
Keith Klarer
Chief Executive Officer
www.datyra.com

G2M RESEARCH

Mike Heumann
Managing Partner
www.g2minc.com

# AI Can Have a Profound Impact on Business

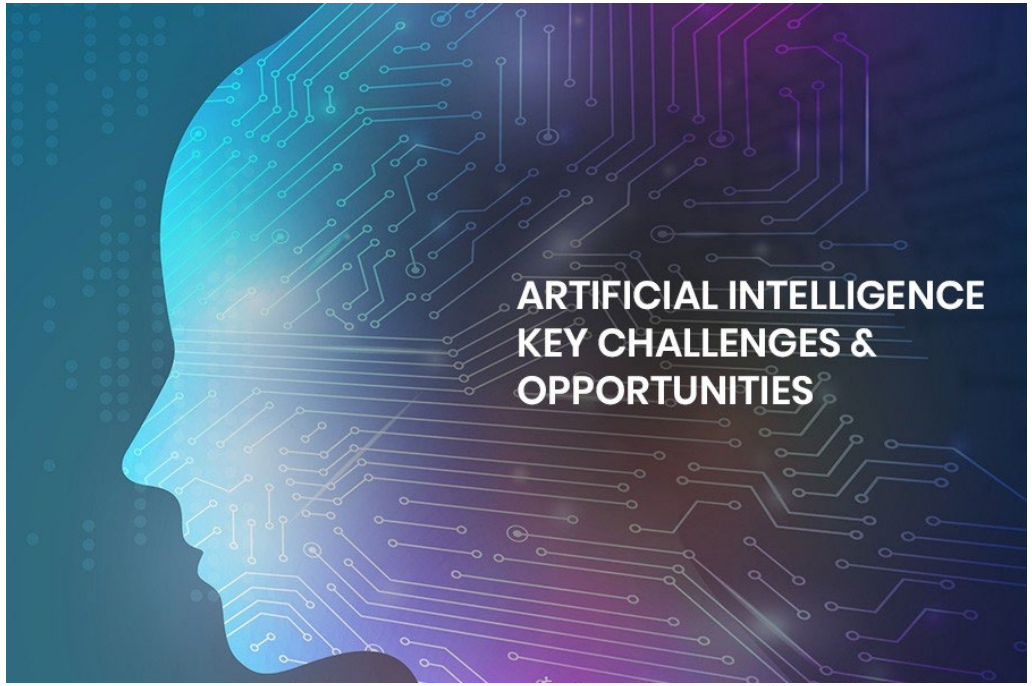
- AI has clear value for business
  - Improved outcomes when compared to other analytic techniques
  - Some problems can only be solved with AI
- However, enterprise adoption is very nascent and uneven
  - Half of large enterprises have at least one instance of AI in their business processes, BUT
  - Only 3% of large enterprises have integrated AI across their entire workflow




G2M RESEARCH

# Why Is AI (Perceived As) Being “Hard to Do”?

- Lack of an overall corporate AI strategy
- Construction/optimization of deep neural networks is an “art”
- Building training, validation data sets
- **ADOPTION HURDLES AND COSTS**
  - Data Acquisition
  - Data Storage Scale and Performance
  - Data Networking Performance
  - Lack of Computational Speed




G2M RESEARCH

G2M RESEARCH


# NVIDIA/Mellanox

Rob Davis

Vice President Storage Technology, Nvidia Worldwide Networking Business Unit

www.nvidia.com

Forbes Billionaires Innovation Leadership Money Business Small Business Lifestyle

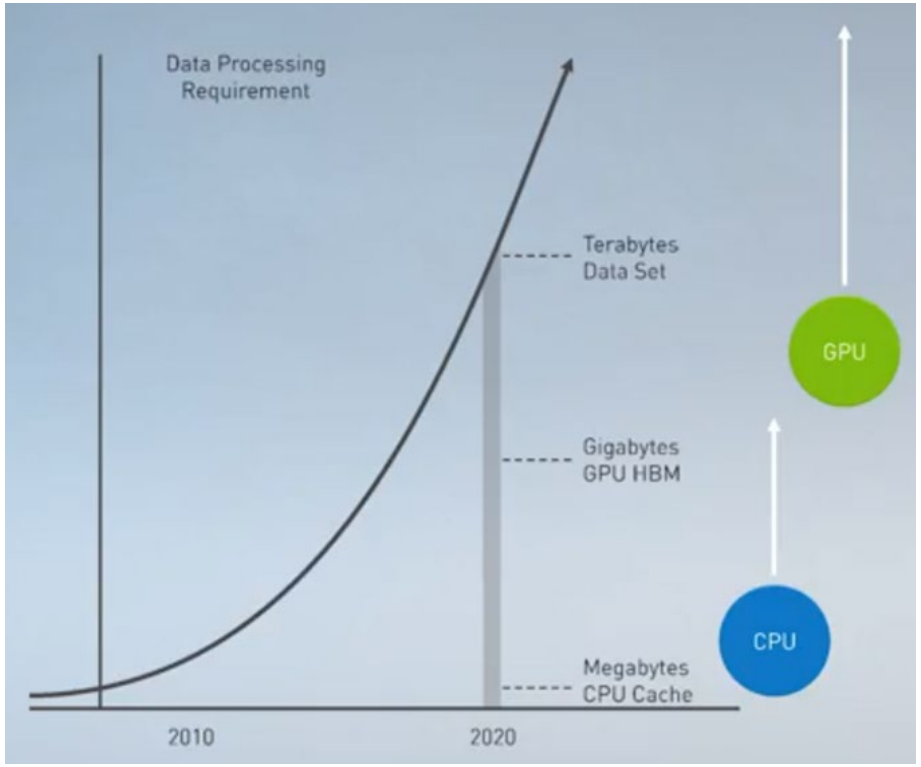
# Don't Forget About Storage When Planning For AI And ML

**Patrick Moorhead** Contributor ⓘ

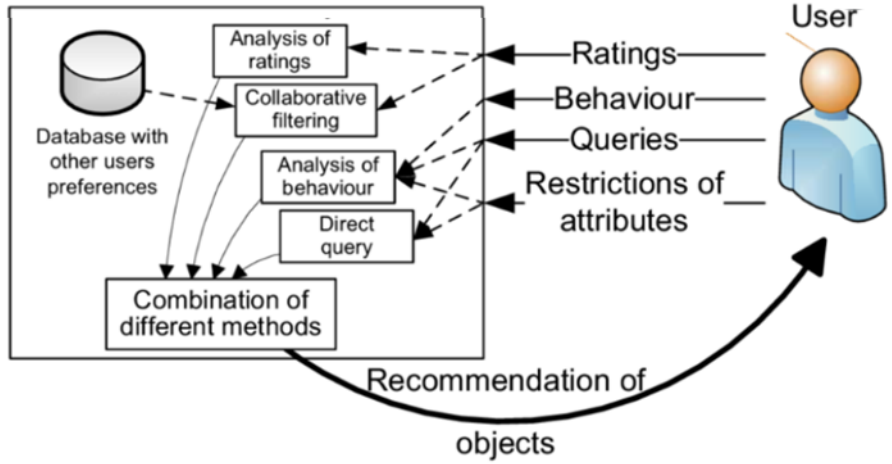
Enterprise & Cloud

*I write about disruptive companies, technologies and usage models.*


“It’s not who has the best algorithm that wins.
It’s who has the most data.”
Andrew Ng



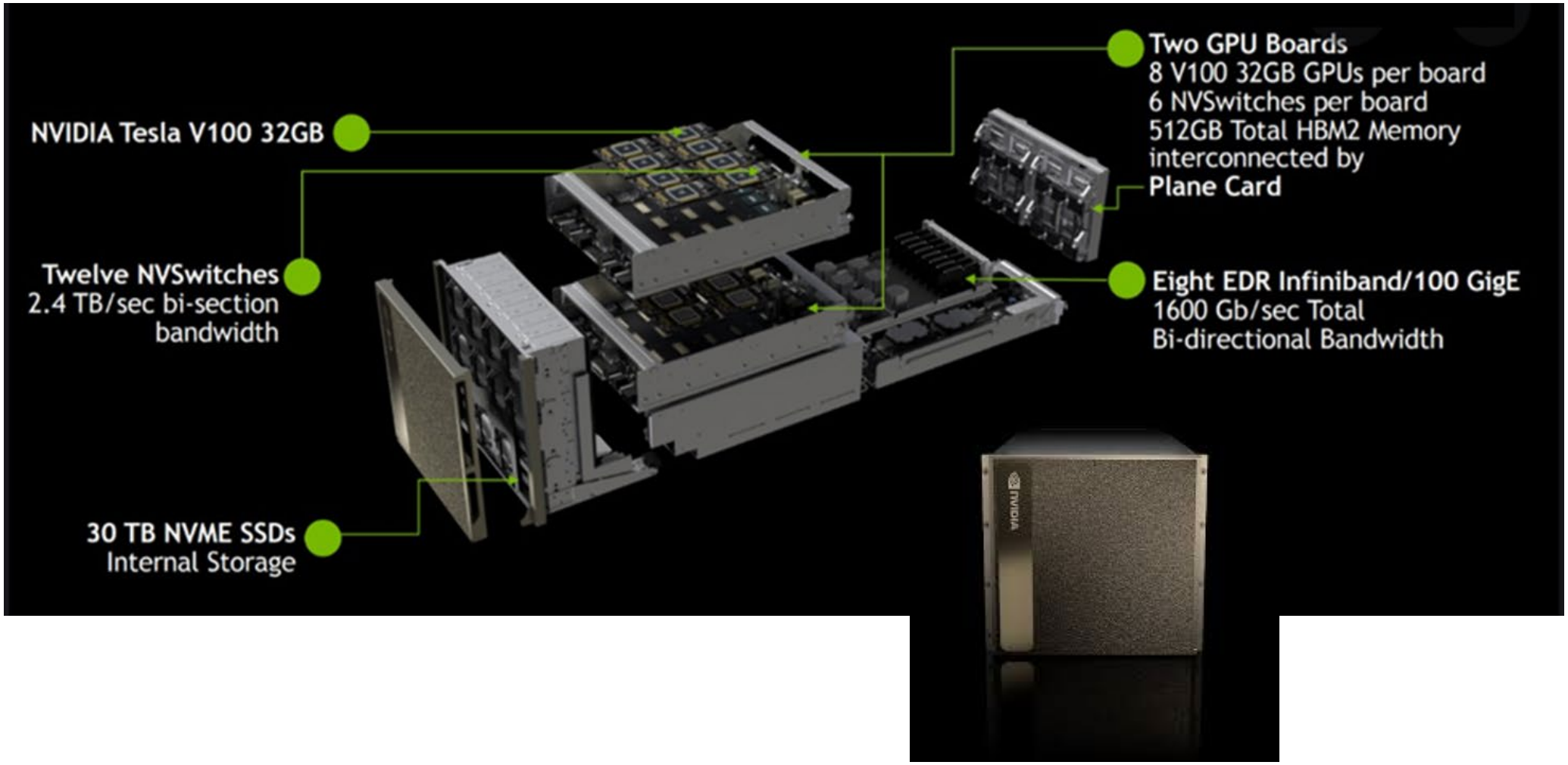

Structure of a recommender system



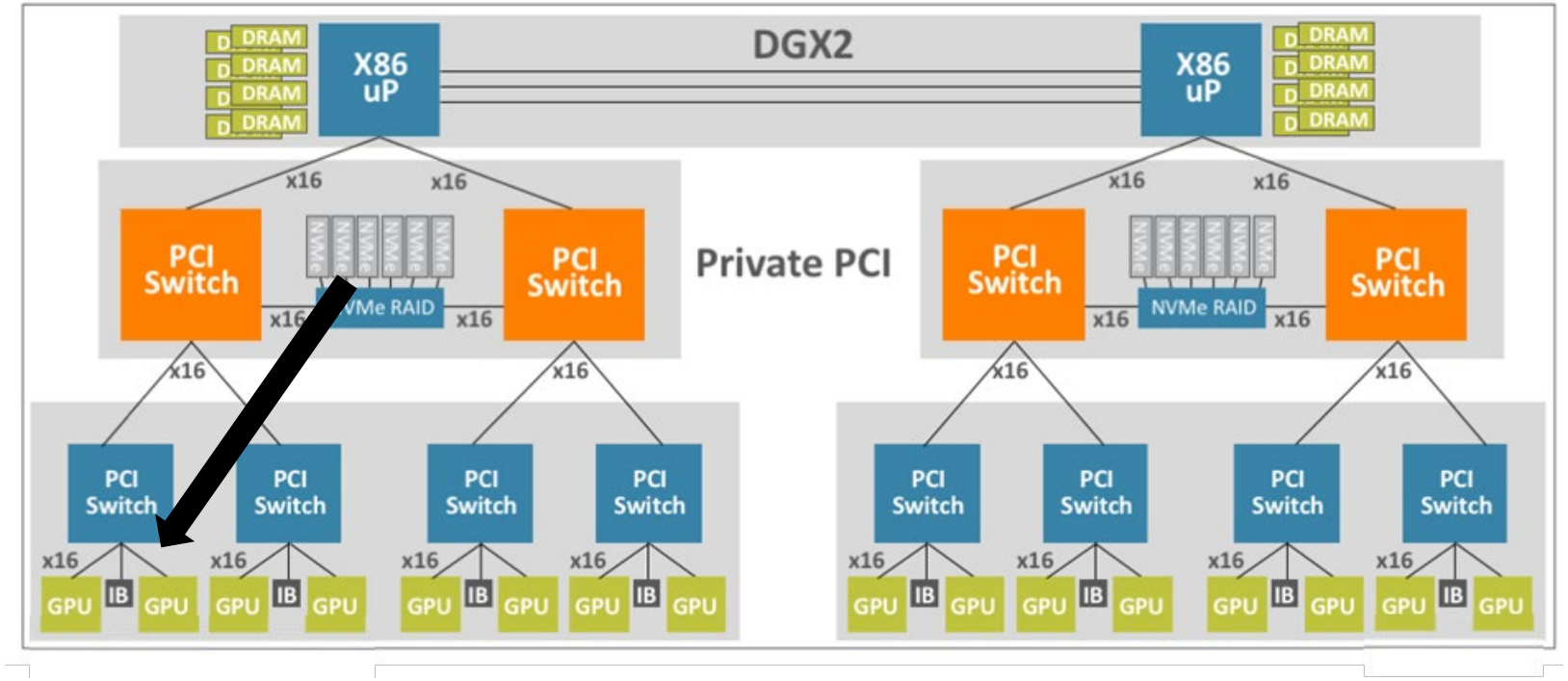


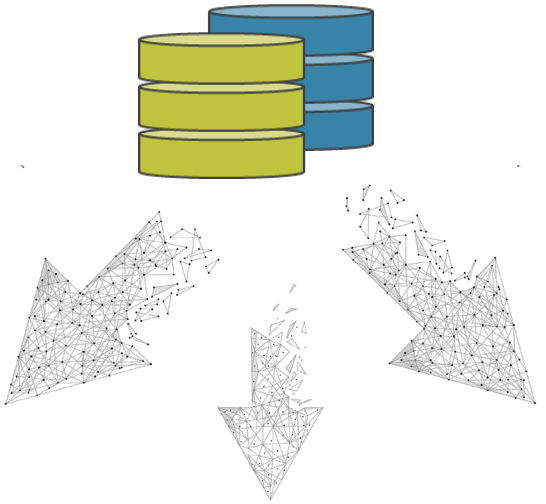

NVIDIA.

# NVIDIA DGX-2

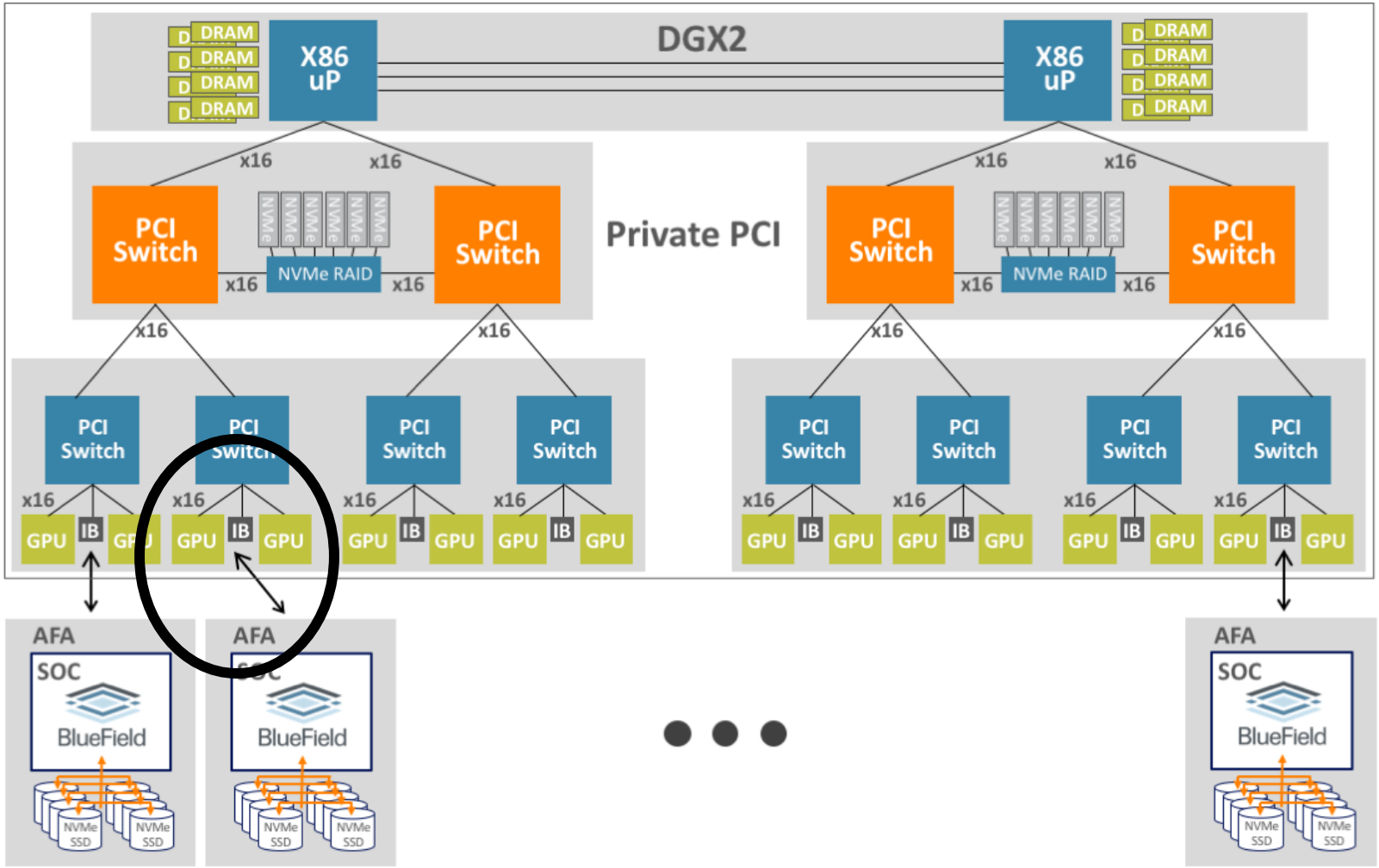

NVIDIA Tesla V100 32GB
Twelve NVSwitches
2.4 TB/sec bi-section
bandwidth
30 TB NVME SSDs
Internal Storage
Two GPU Boards
8 V100 32GB GPUs per board
6 NVSwitches per board
512GB Total HBM2 Memory
interconnected by
Plane Card
Eight EDR Infiniband/100 GigE
1600 Gb/sec Total
Bi-directional Bandwidth

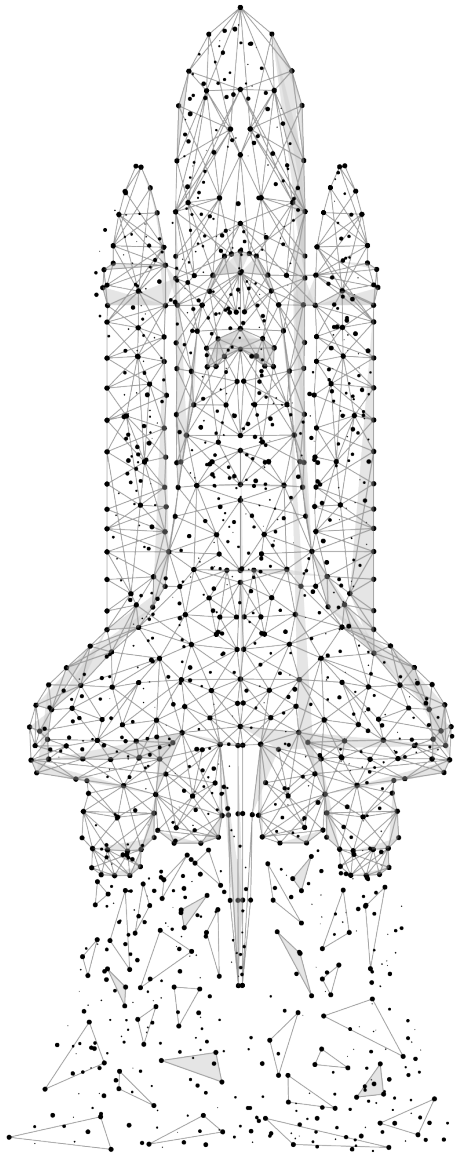
# 12.5GB/S OF INTERNAL STORAGE TO EVERY GPU PAIR



# 25GB/S TO EVERY GPU PAIR



**External IB or Ethernet attached NVMe JBOFs**

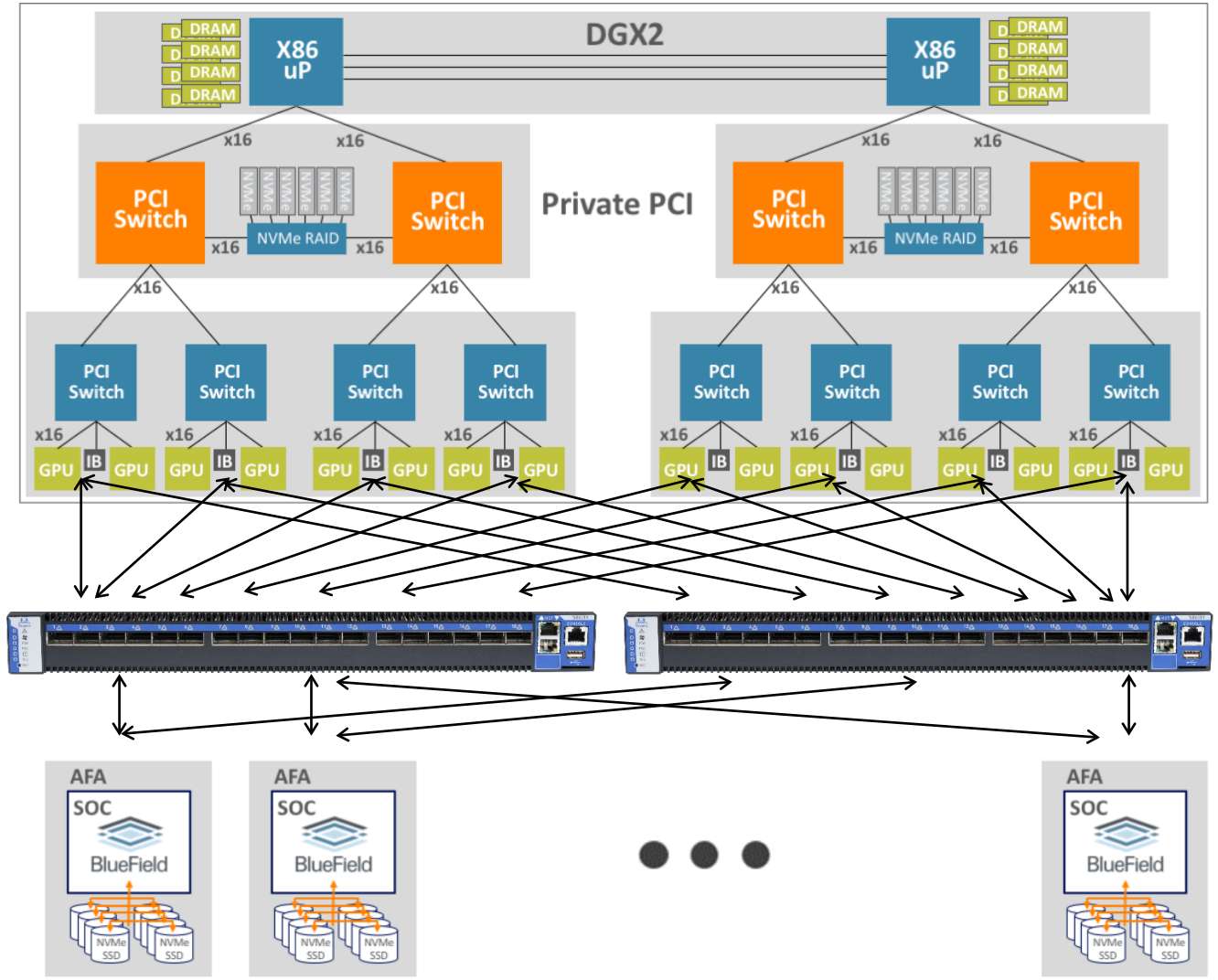
# UNLIMITED HIGH PERFORMANCE STORAGE WHEN NETWORKED

**Storage Networking**

**Advantages over**

**Local Storage for GPUs**

- Unlimited capacity
- High Availability
- Higher Utilization
- Lower TCO

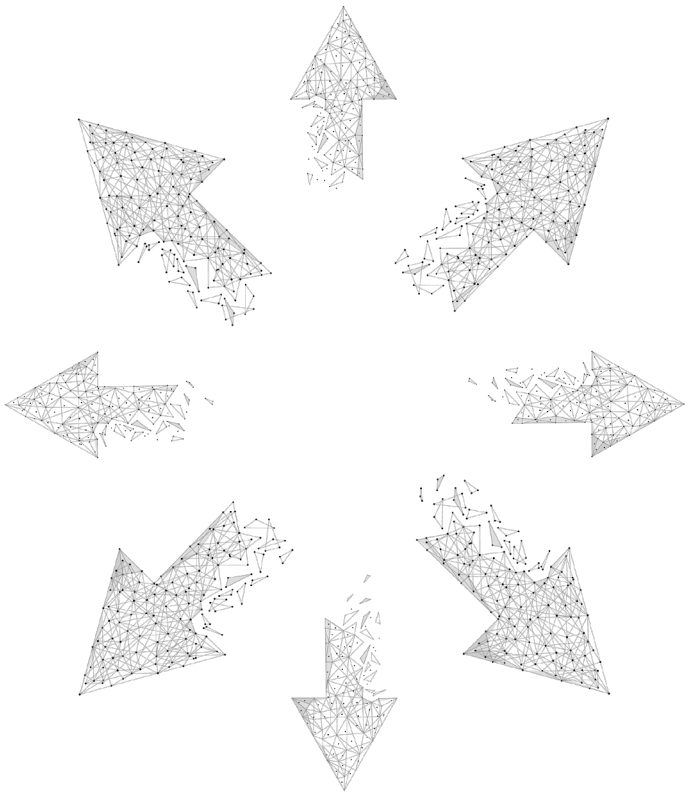


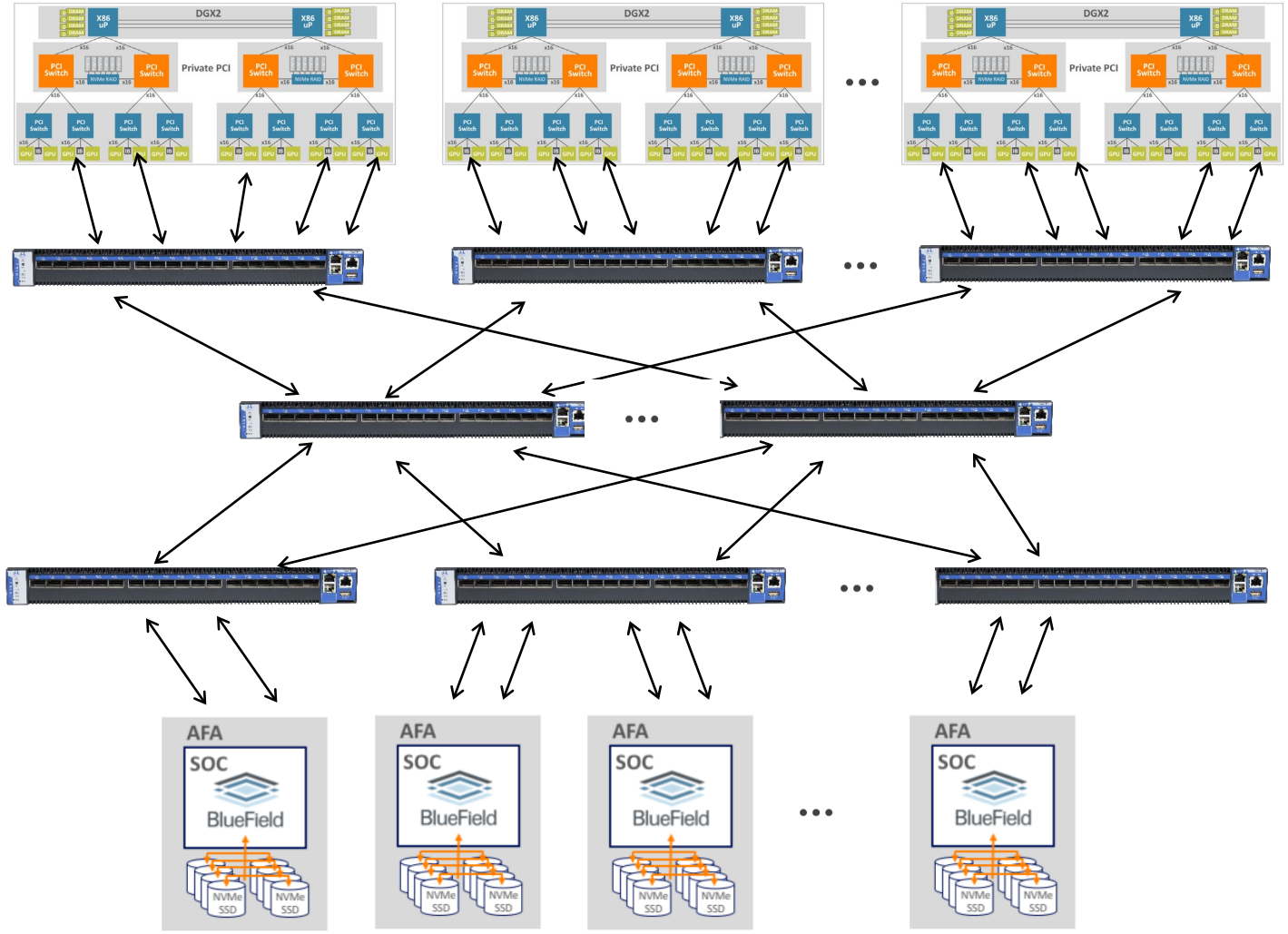

Dual port 200Gb HCA or NICs

200Gb IB or Ethernet Switches

NVMe-oF All Flash Arrays

# NOW STORAGE & GPUS CAN SCALE INDEPENDENTLY

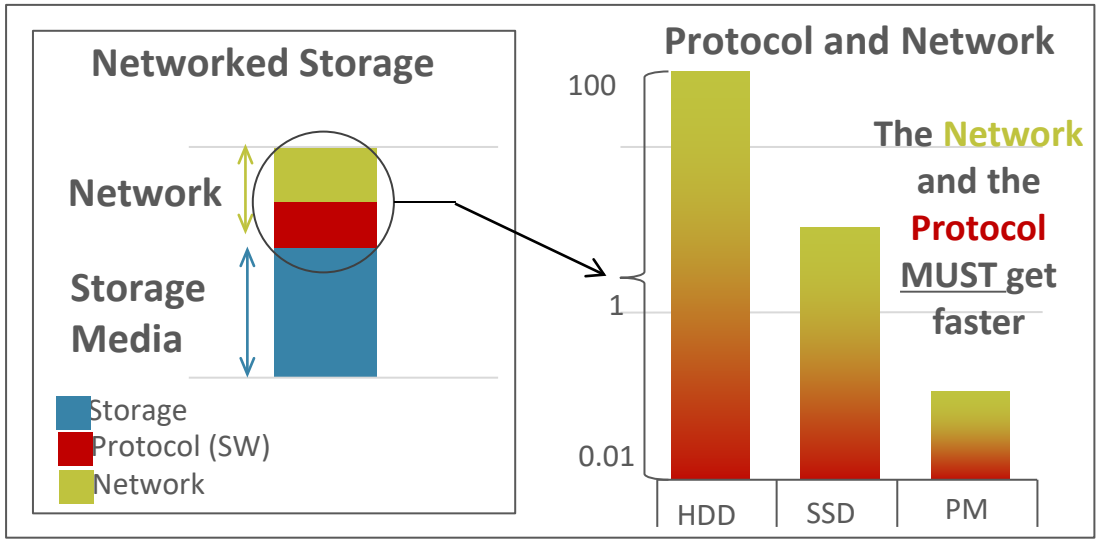


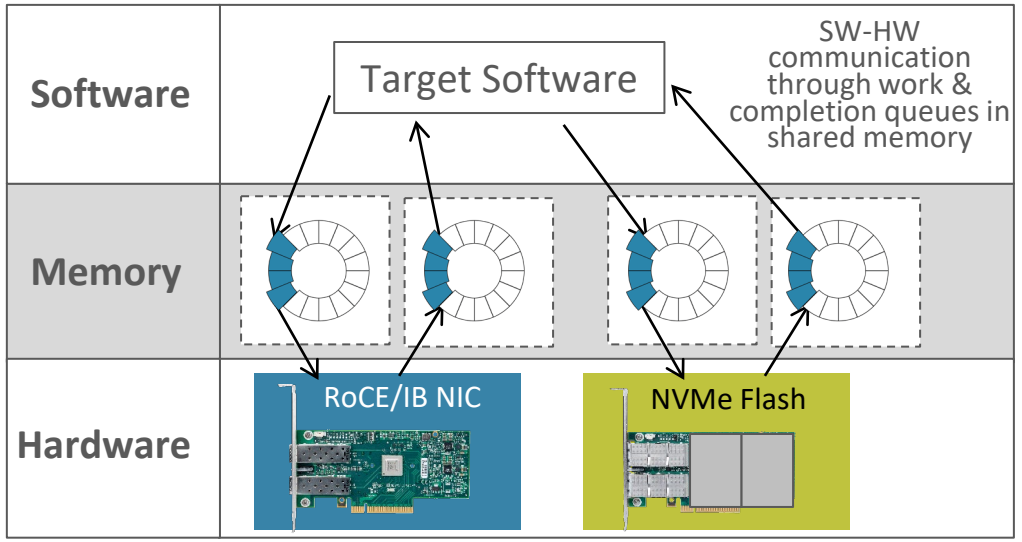


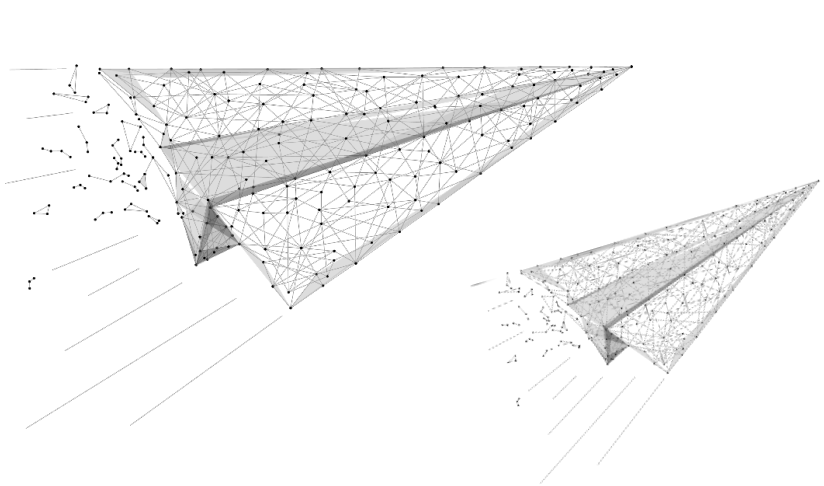


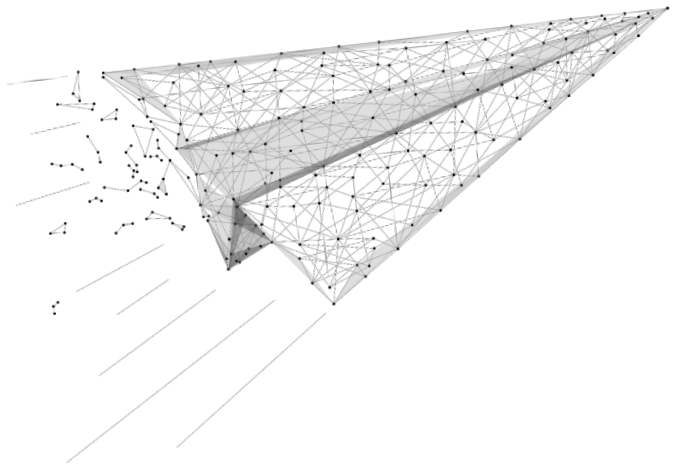

NVIDIA

# NVME, NVME-OF & RDMA PROTOCOLS

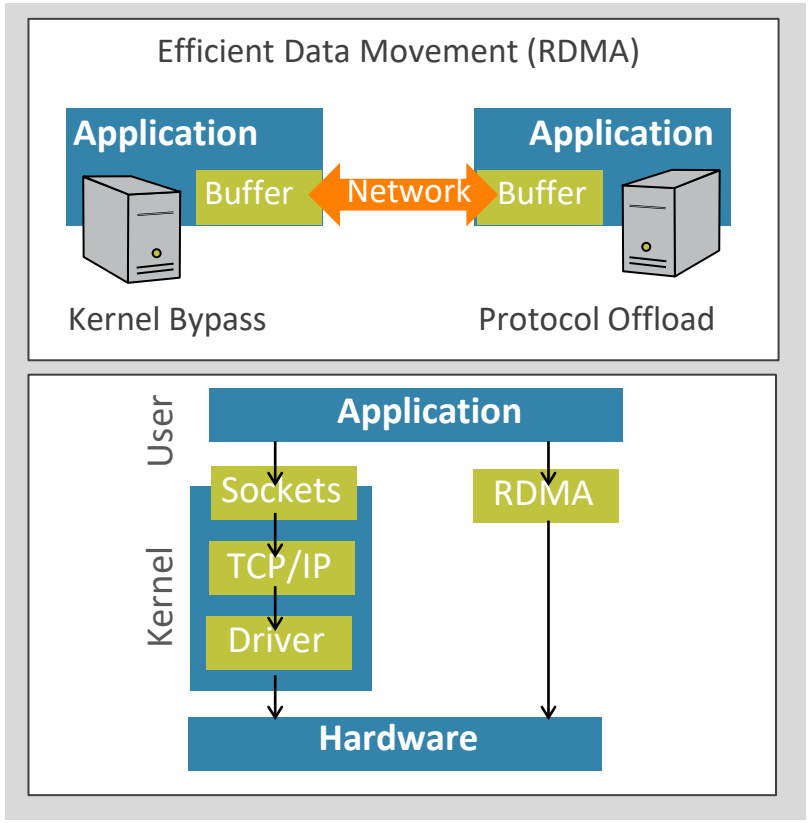


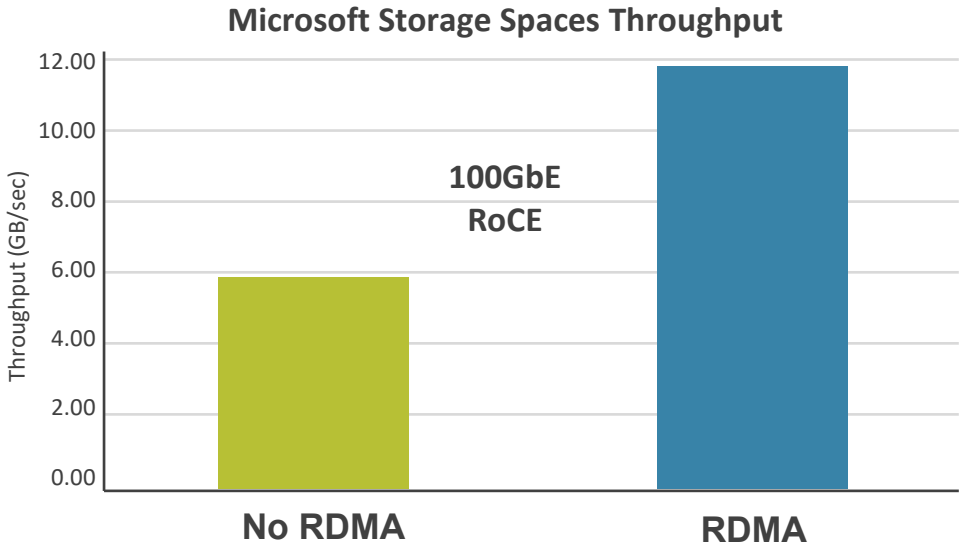





# RDMA PERFORMANCE



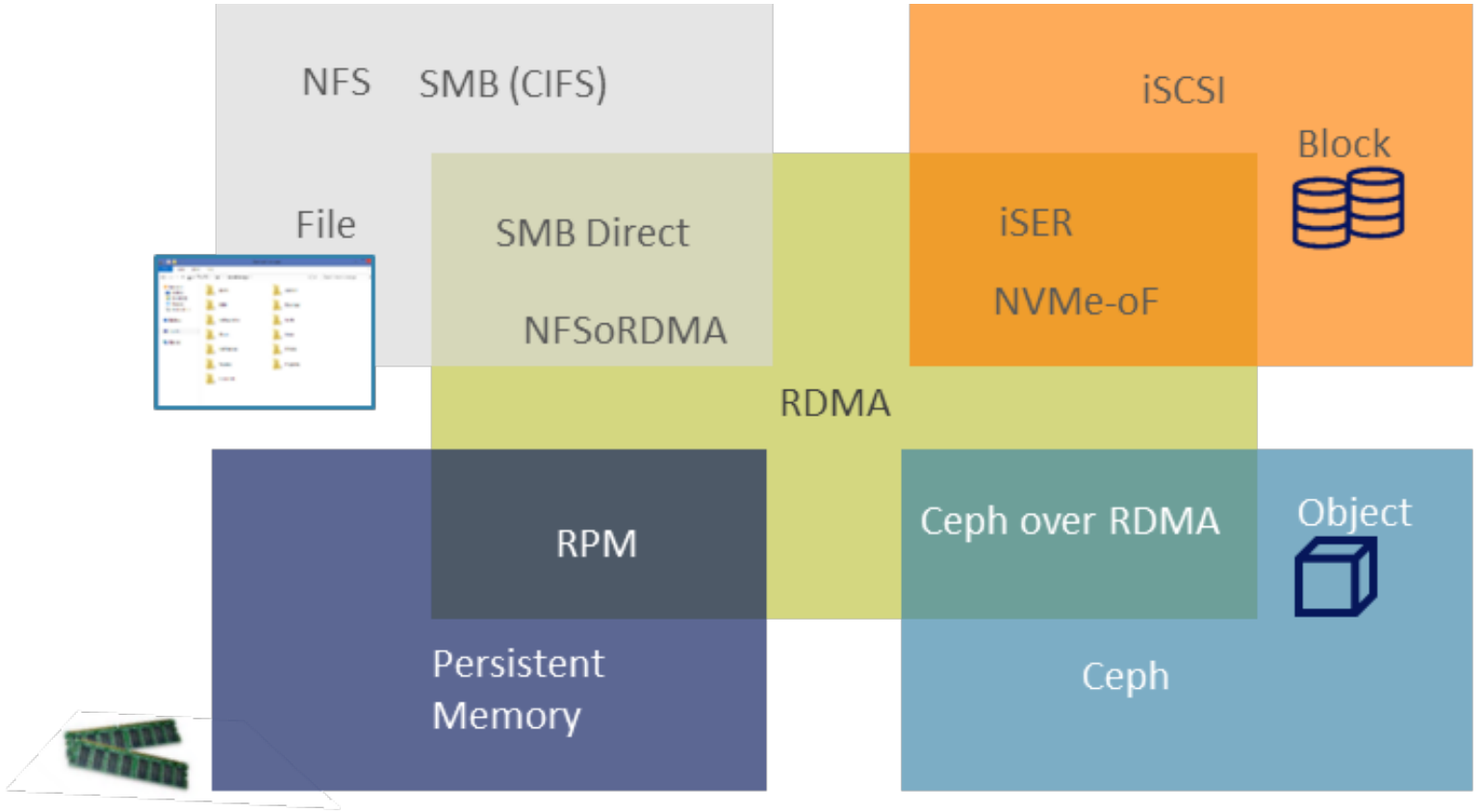




**With RDMA**

- **2x Better Bandwidth**
- **Half the Latency**
- **33% Lower CPU**

See MS demo: https://www.youtube.com/watch?v=u8ZYhUjSUoI

# INDUSTRY WIDE RDMA ADOPTION

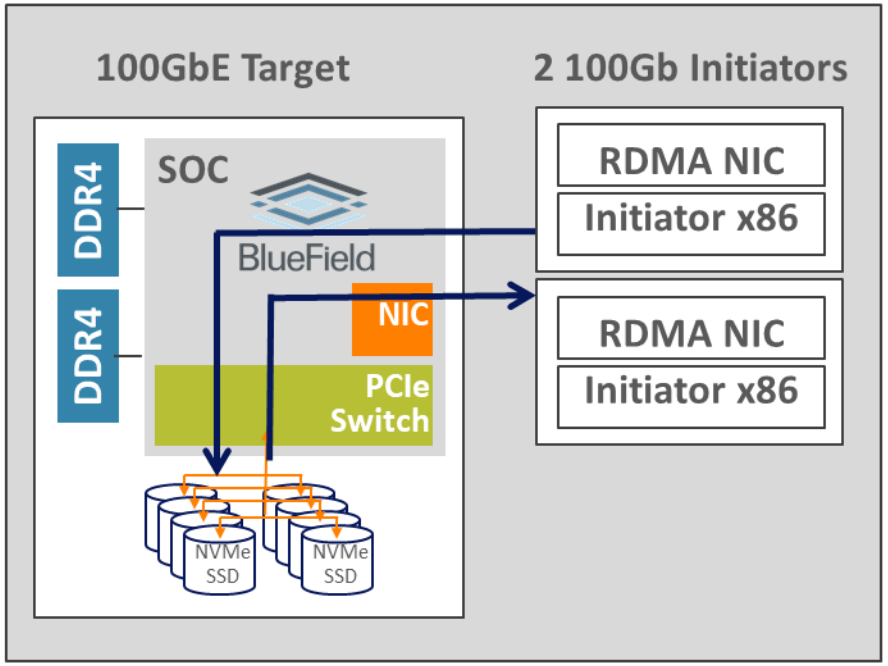


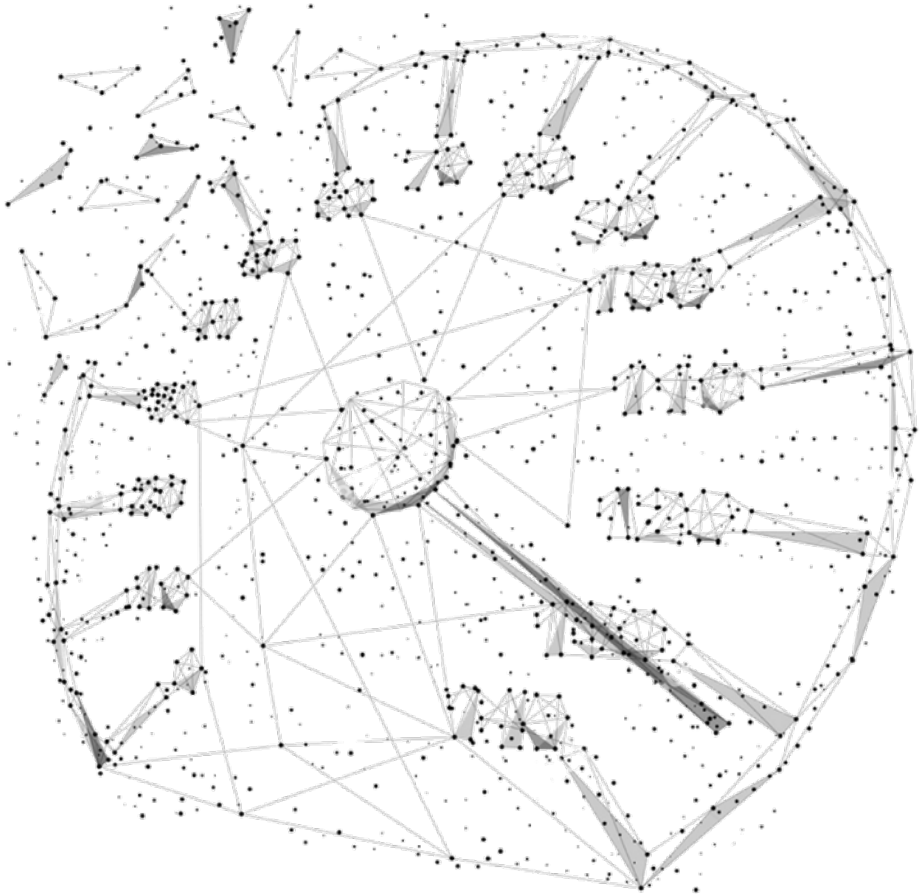




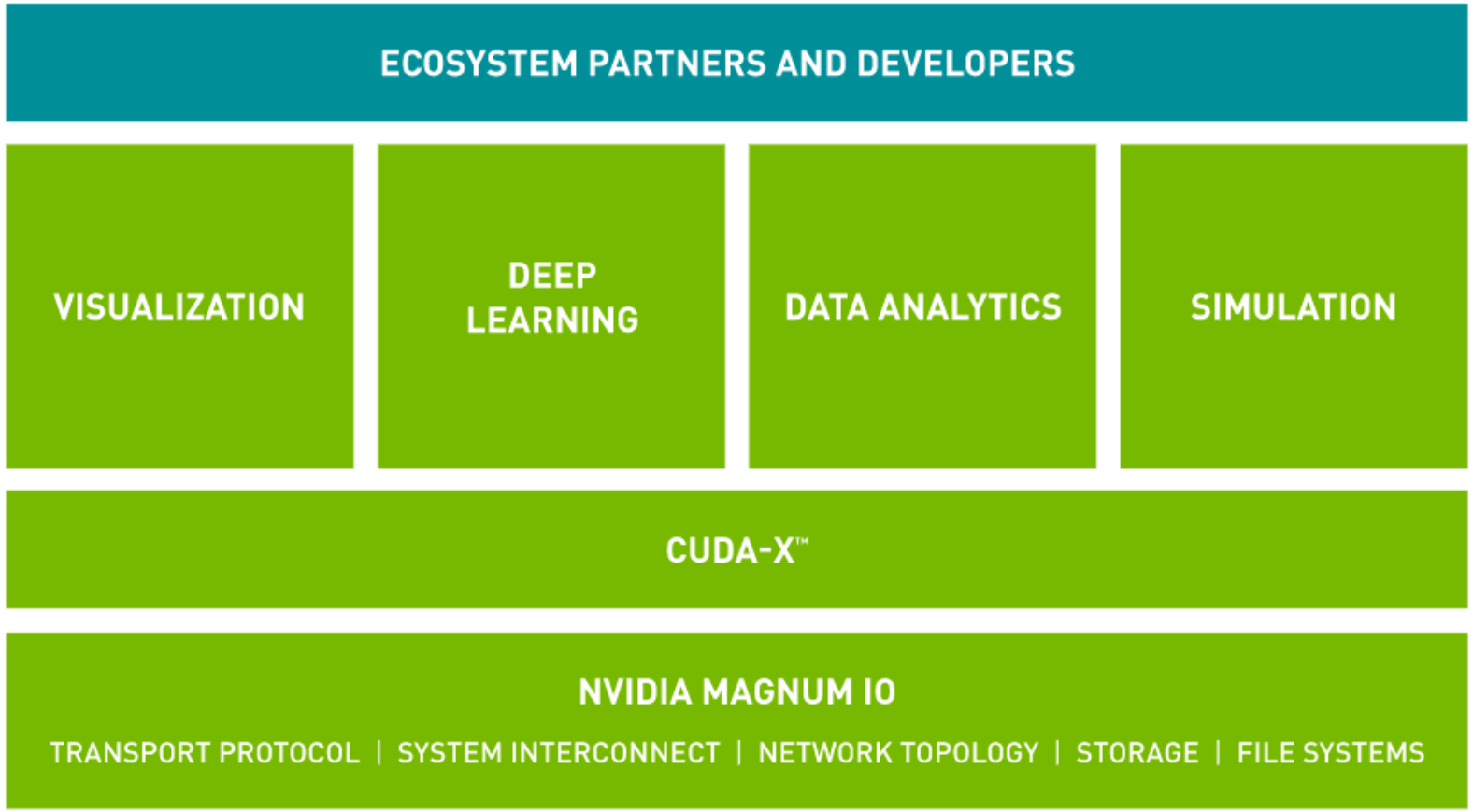
# NVME-OF PERFORMANCE WITH RDMA

- More Bandwidth ✓5M IOPs, 4K block side
- Less Latency ✓~3usec latency
- Less CPU ✓0.01% CPU utilization



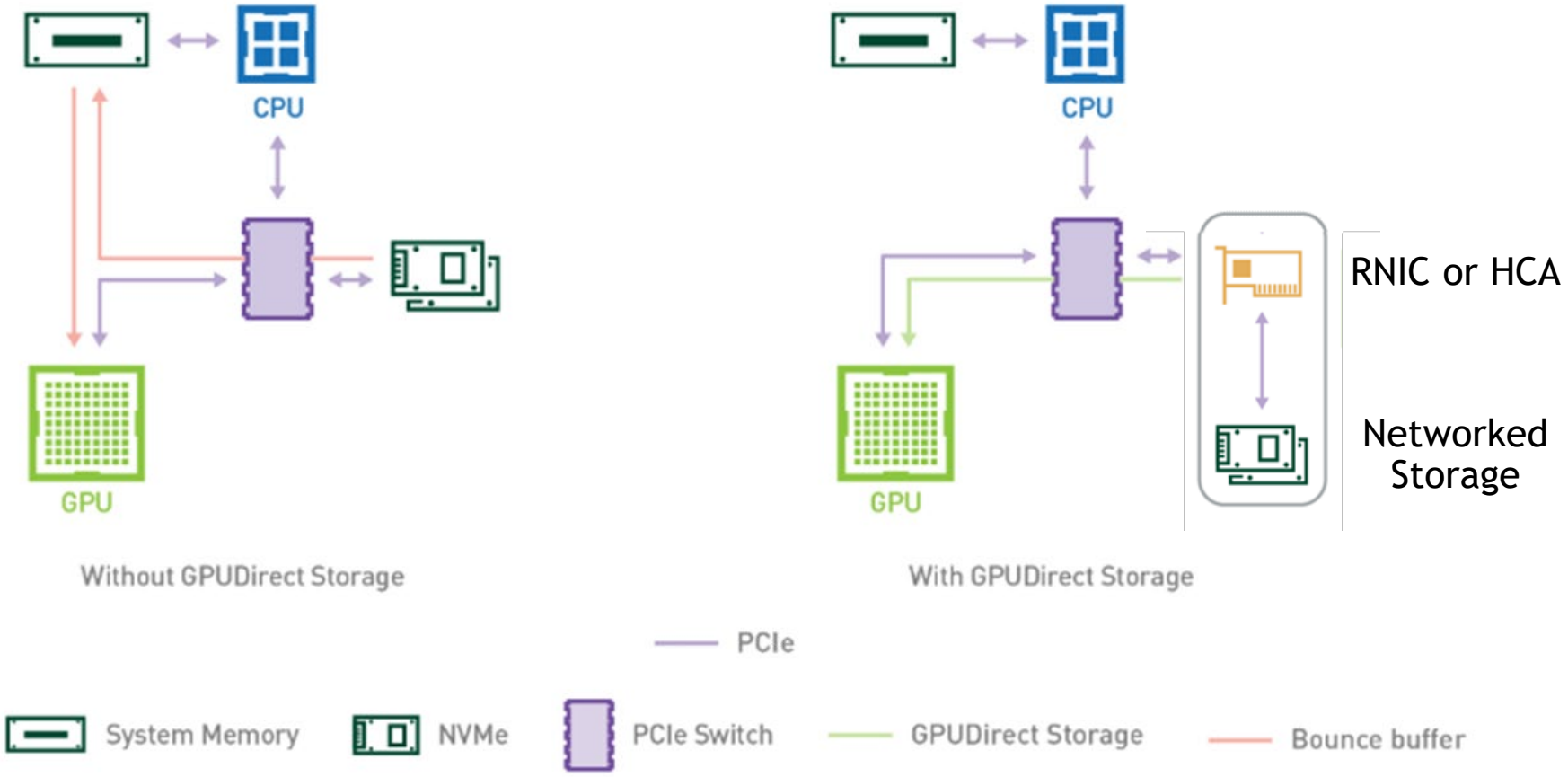



# MAGNUM IO




https://www.nvidia.com/en-us/data-center/magnum-io/

# GPUDIRECT STORAGE

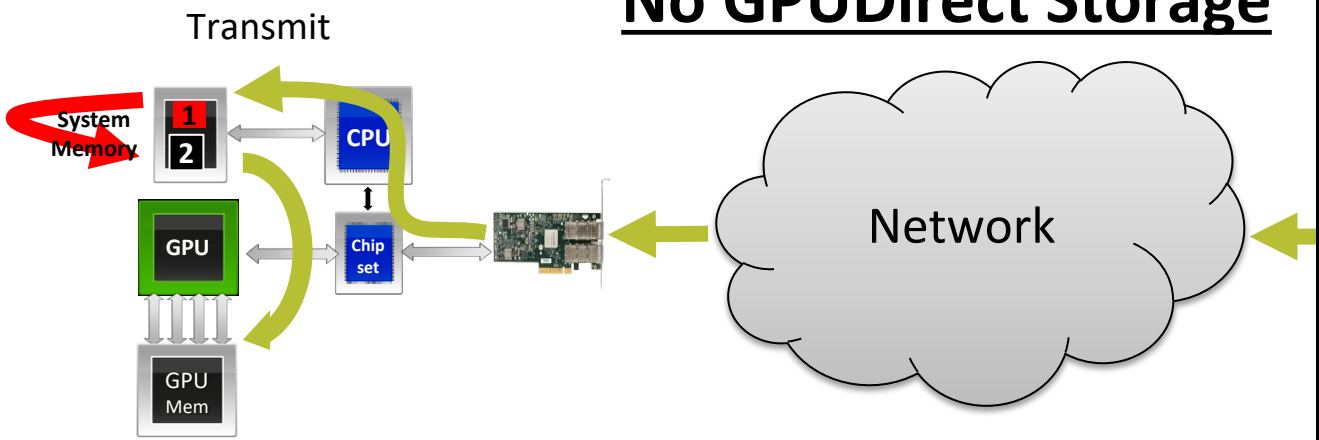


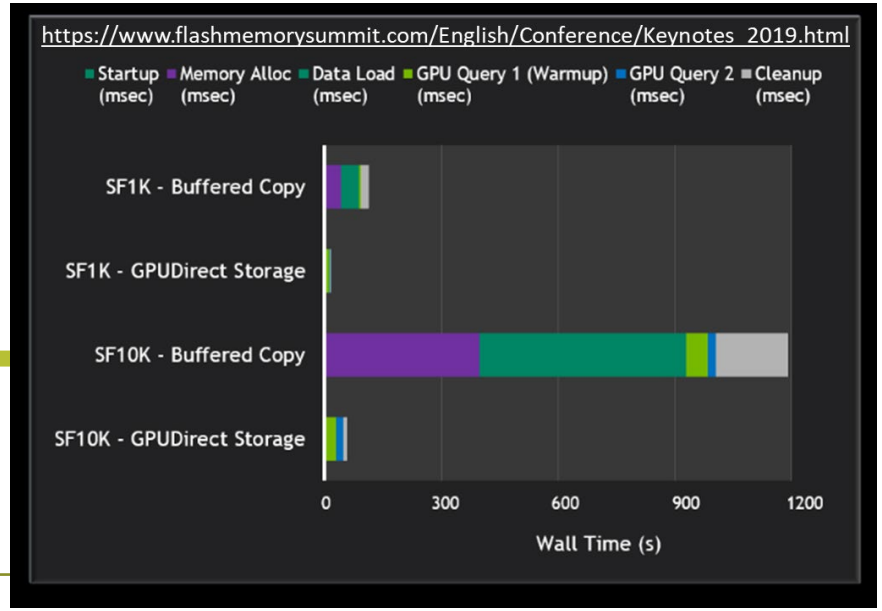

https://developer.nvidia.com/gpudirect

https://devblogs.nvidia.com/gpudirect-storage/

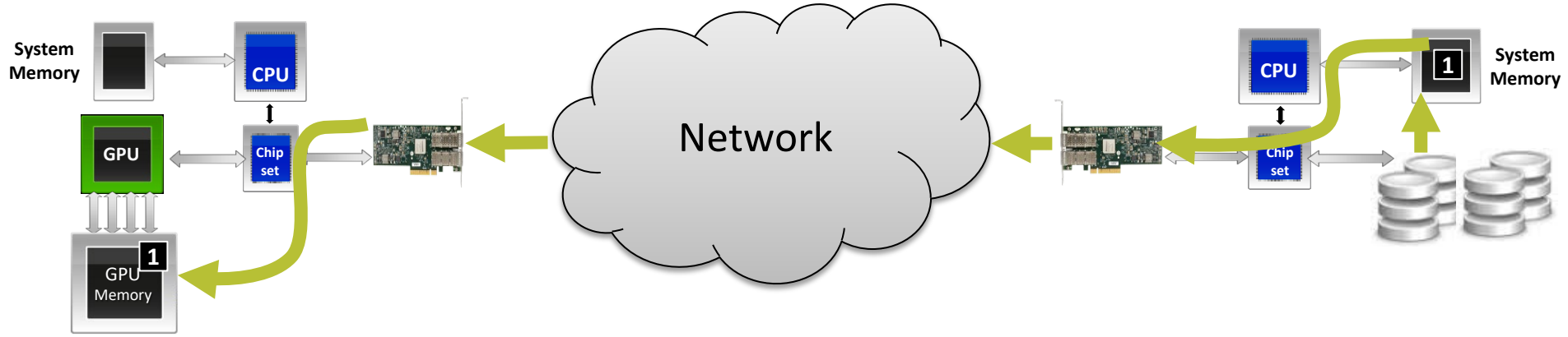
# GPUDIRECT STORAGE EXAMPLE

## No GPUDirect Storage



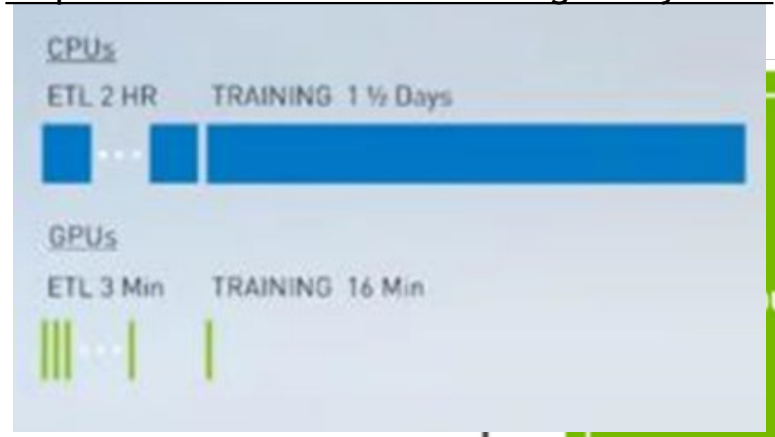



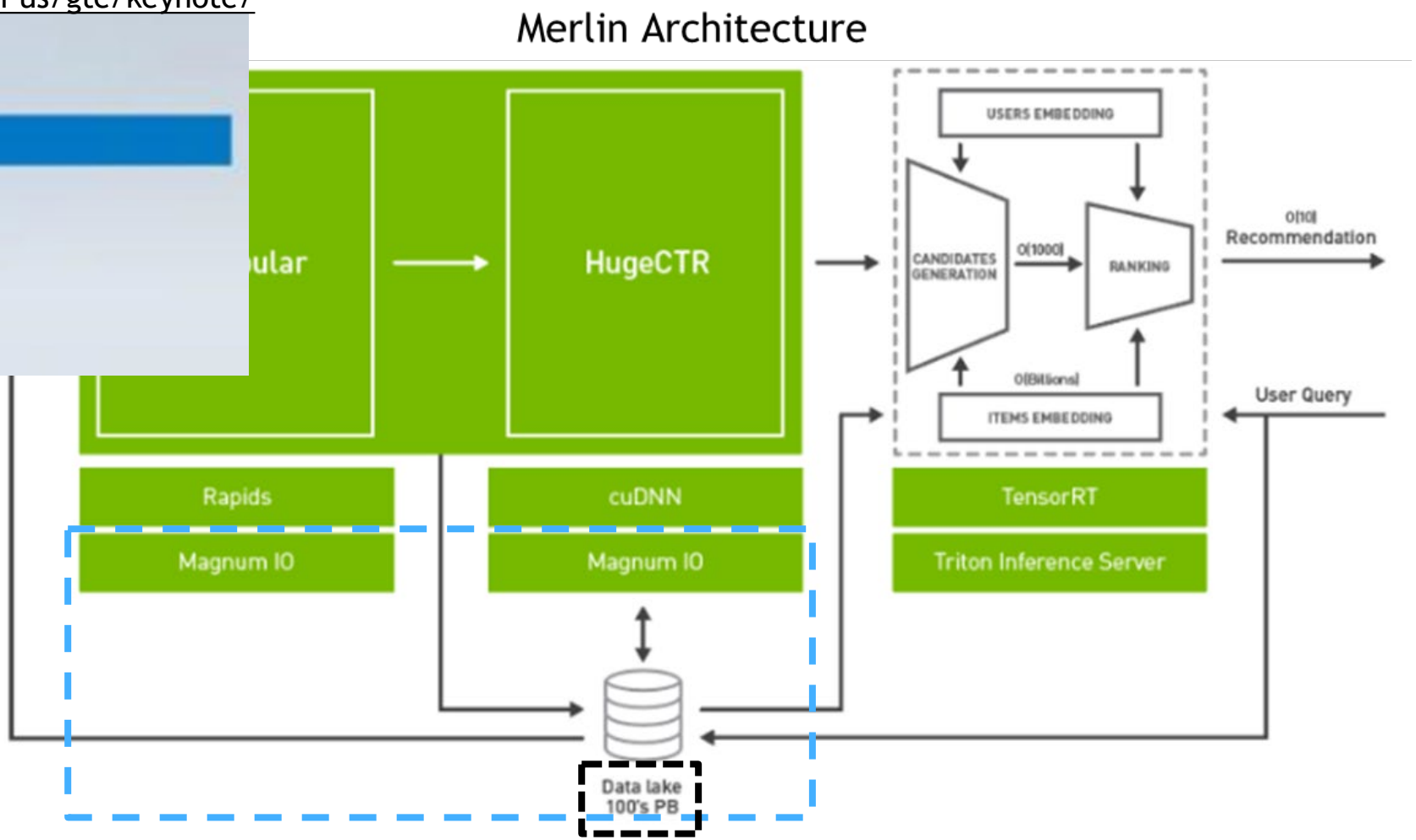

## With GPUDirect Storage



# MERLIN – RECOMMENDER APPLICATION FRAMEWORK

https://www.nvidia.com/en-us/gtc/keynote/

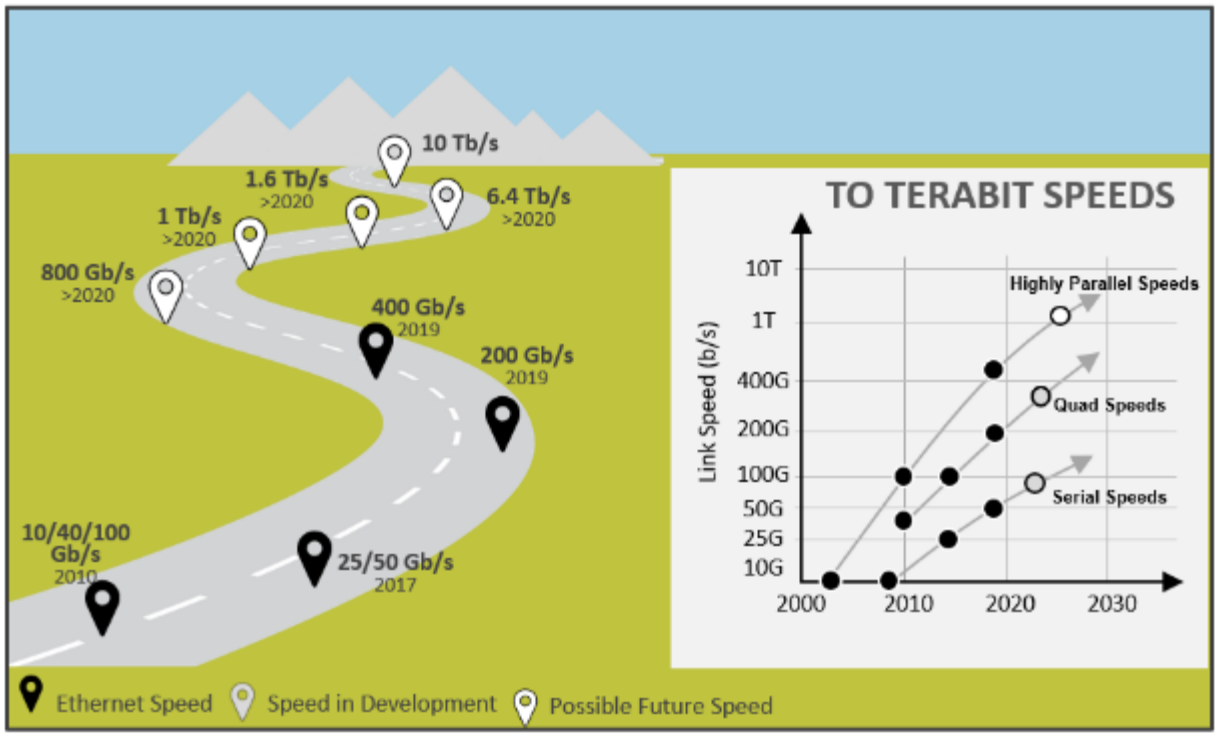


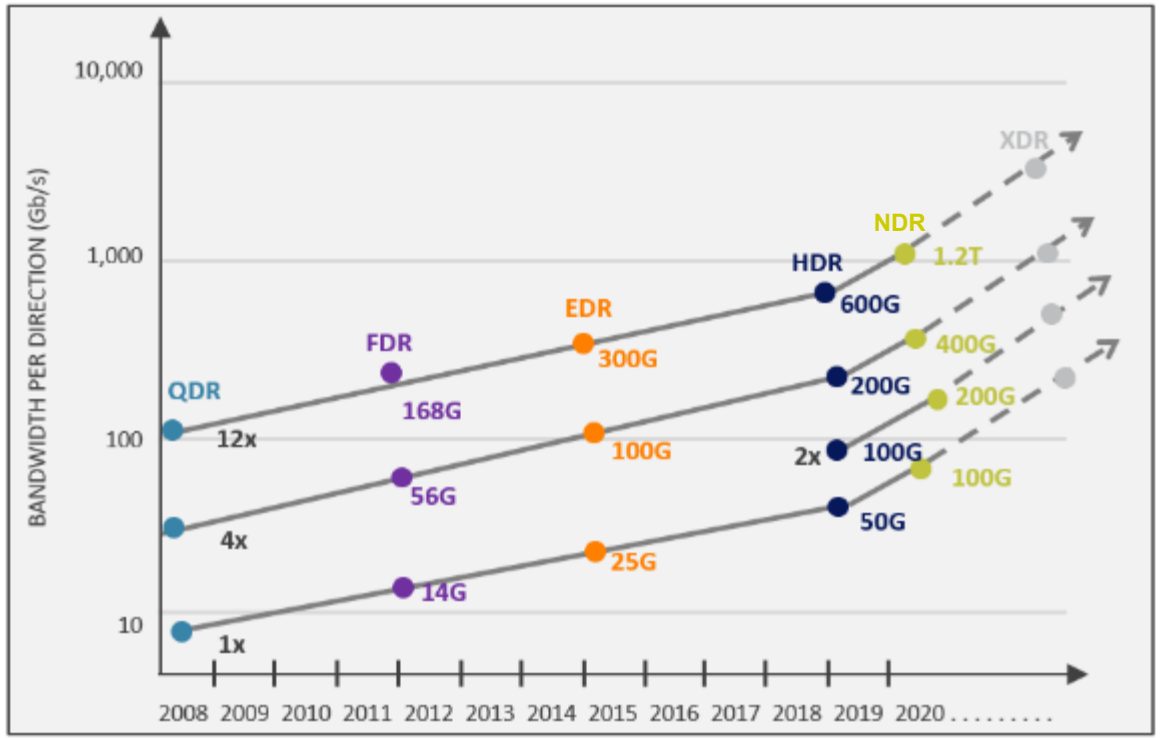

https://devblogs.nvidia.com/announcing-nvidia-merlin-application-framework-for-deep-recommender-systems/

# FUTURE PERFORMANCE

## Ethernet




## InfiniBand



G2M RESEARCH

WEKA

# Weka


- Shailesh Manjrekar
- Head of AI and Strategic Alliances
- www.weka.io

WEKA

NVIDIA PREFERRED SOLUTION ADVISOR

# WEKA ai for Accelerated DataOps

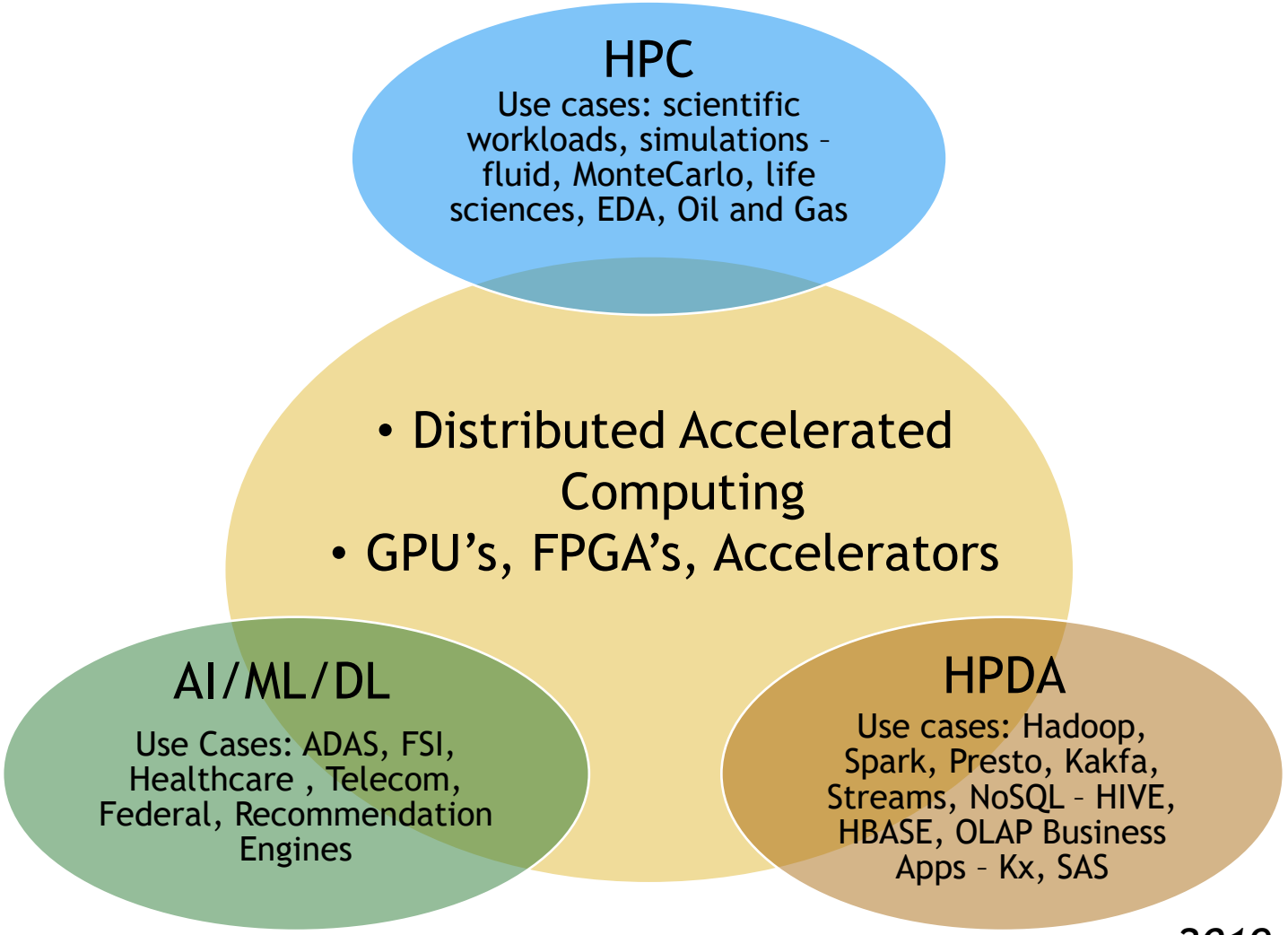
Shailesh Manjrekar,
Head of AI and Strategic Alliances

# Agenda

- New Workloads – AI/ML/DL apps are inherently different
- New Architecture – Edge to Core to Cloud
- New Approach – Fuel your Digital Transformation with Accelerated DataOps

**“Data is the new source code”**

# AI 2.0 Market Scape – Use Cases




*2019 Hyperion Research*

# Software Defined Car – ADAS data pipeline and other examples

1000+ manual labelers
20 Million+ objects
labelled / month

20+ DNN Models
actively developed

ADAS Eco-system Partners

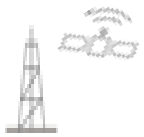
Labeling UI

ML Metrics/ UI

Hardware-In-Loop

5000+ GPU cluster -> 625 PFLOPS
Kubernetes Orchestration

Edge



Data Ingestion | Data Selection | Feature Engg./Dataprep | Model tuning / Test | Hardware-In-Loop tests | Deployment, Monitoring | Lifecycle Mngmt.

EGX / Edge Aggregation

DGX / Core

HGX - Start.weka.io SAAS

Ingested Datasets

WEKA

S3 API's

Selected Datasets | Labeled Datasets | Metrics and Logs | Trained Models | Synthetic data / Replay data

WEKA

S3 API's

Test/Dev datasets

WEKA

S3 API's

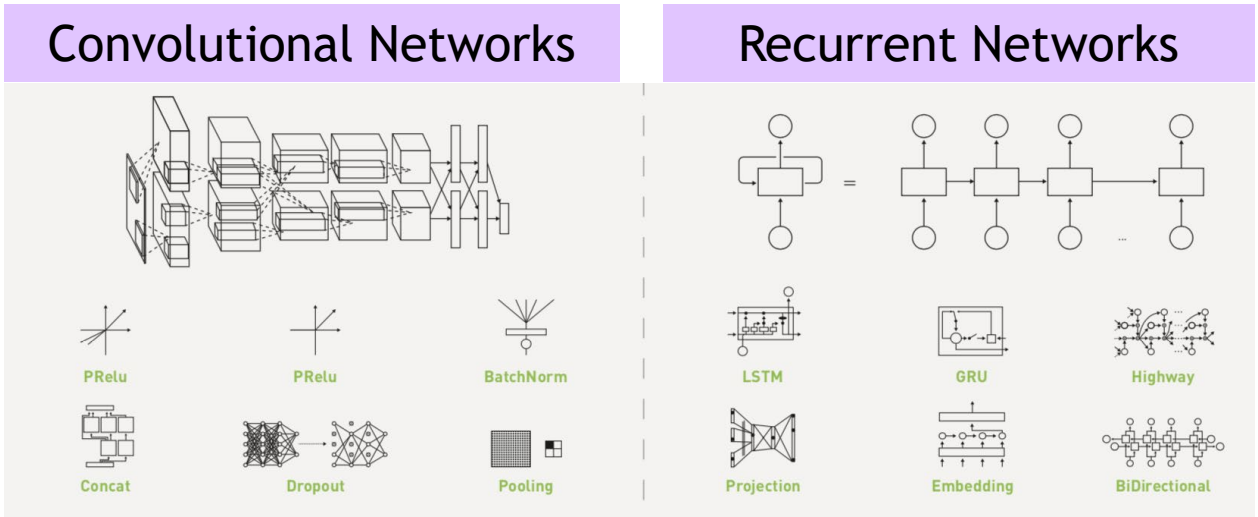

WEKA

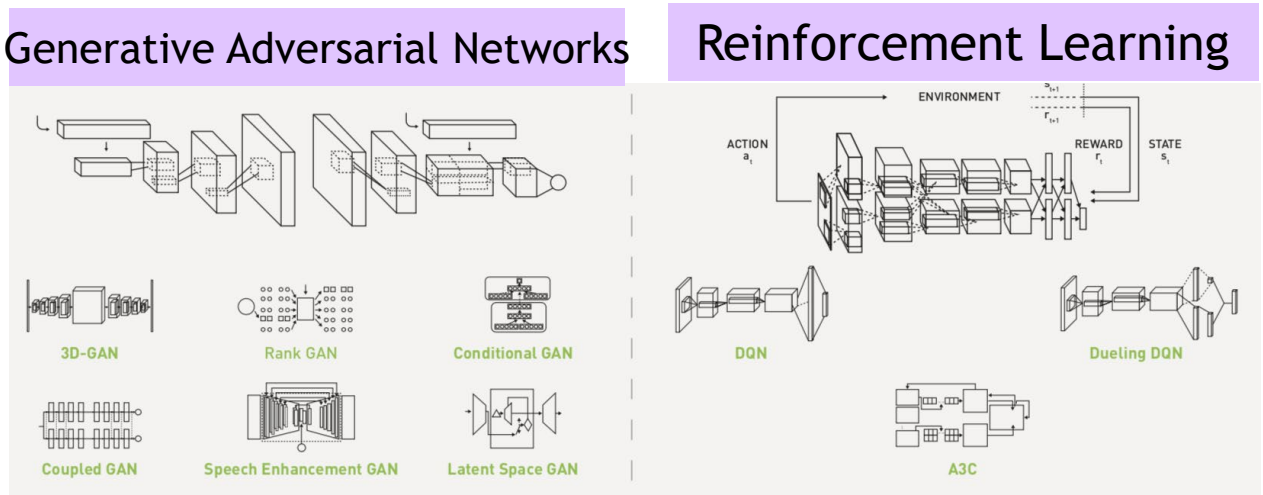
# DNN's Becoming More Complex With Several Billion Hyperparameters

Use cases moving from Computer Vision to NLP/NLU and multi-modal

Advances in Deep Learning Methodologies:

- Deep Learning
- Transfer Learning
- Federated Learning
- Active Learning



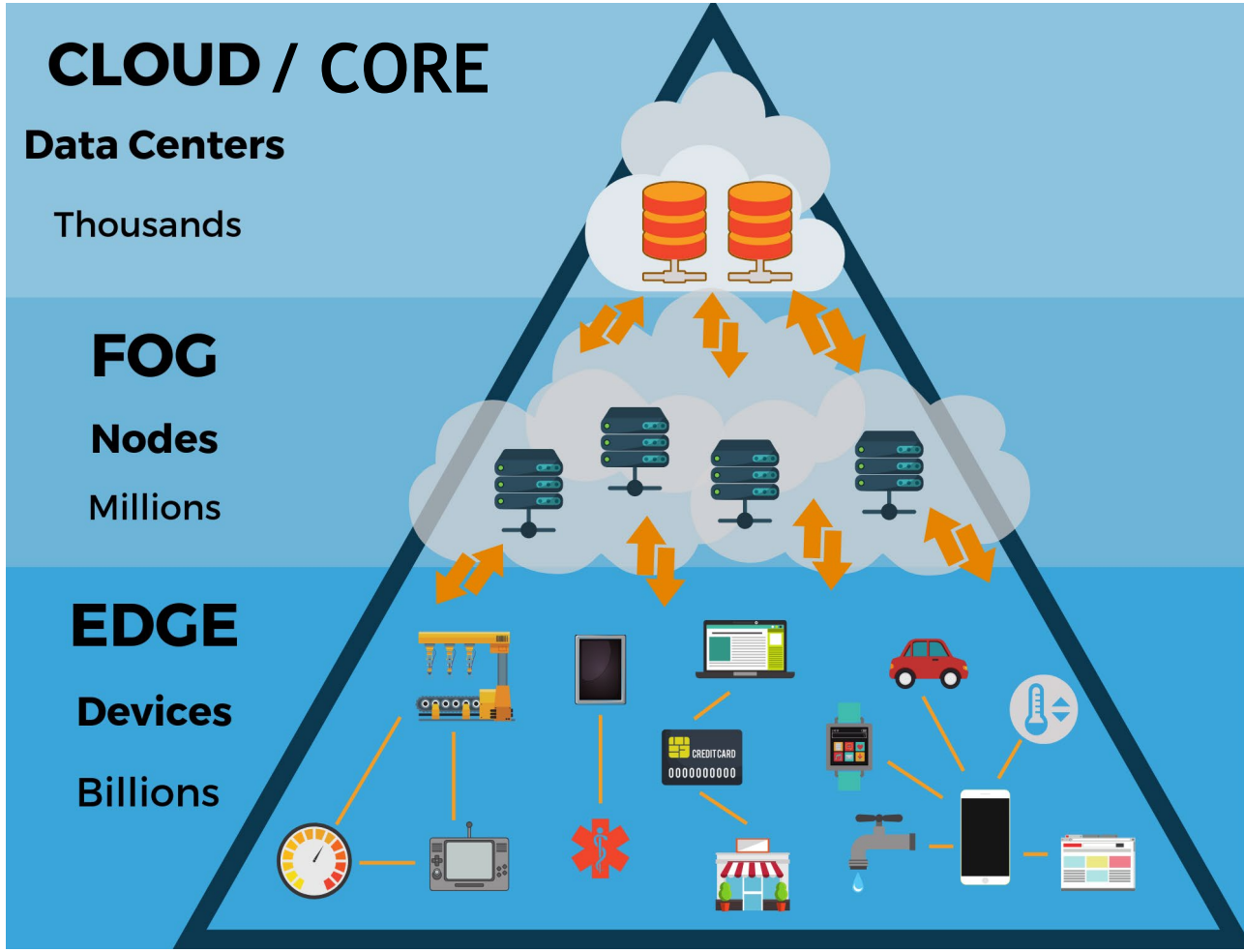



# Data Anywhere

- Training / Inferencing testing
- Application specific processing
- High cost

- Edge Aggregation
- Tagging / High Ingest

- Intelligent 5G Edge – Bigger than cloud
- Inferencing
- Time sensitive
- Task specific processing
- Low cost



# GPUs Have “Densified” Compute into a Single Server Creating a Huge Data Bottleneck

CPU-Only Servers

- 100’s of servers with CPUs
- 100’s of low bandwidth network connections
- No one server was particularly demanding on storage

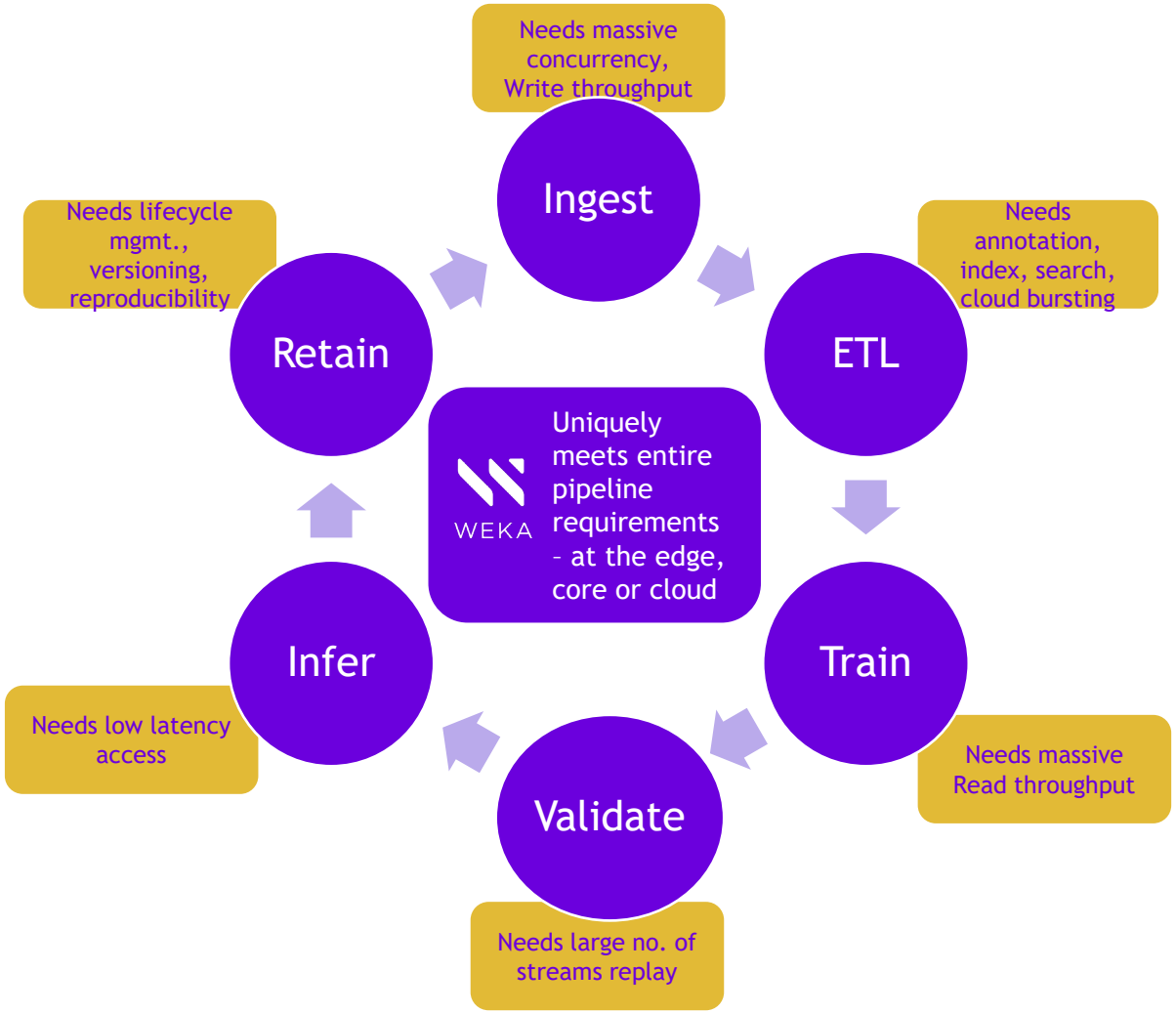
100x More Compute
40x more network

GPU Accelerated Server

Current NAS solutions cannot feed these machines with enough data

# Storage has become the last mile problem

## Results in storage silos and delayed time-to-value



# WekaAI for Accelerated DataOps – Small / Medium / Large Bundles

Data Collection: Edge, Core, Cloud

Model Training, Simulation

Inferencing

Lifecycle Mgmt.

Aerial

Kubeflow

VALOHAI

Jarvis

TensorFlow

Clara

Parabricks

Metropolis

Ingest, ETL, Query

RAPIDS

Software-in-Loop, Hardware-in-Loop

Kubernetes / Container / NGC Hub

Accelerated Compute: Nvidia DGX, EGX, HPE Apollo, Graph Processors, FPGA, Accelerators, MagnumIO

WEKA on Storage Servers

S3 On-Prem Cloud

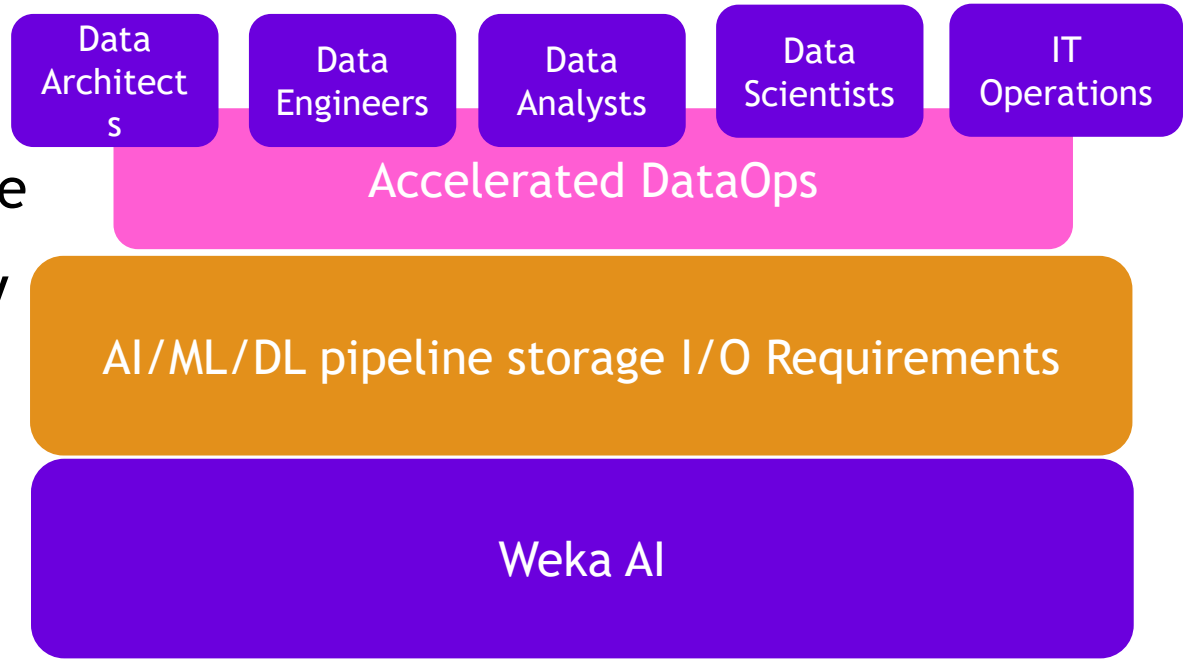
S3 Public Cloud WEKA

# Datacenter scale computing

- Multi-Workload convergence – NVIDIA DGX A100
- GPUDirect Storage enables Data Analytics, Training and Inference
- Personalized Internet with Merlin accelerated Recommender systems
- Conversational and Multi modal AI
- Clara Parabricks, Clara for Healthcare and the new Clara Guardian
- NVIDIA DRIVE for Autonomous Driver Assistance Systems (ADAS)
- NVIDIA EGX A100 with Aerial, Issac and Metropolis for Edge to Core to Cloud pipelines

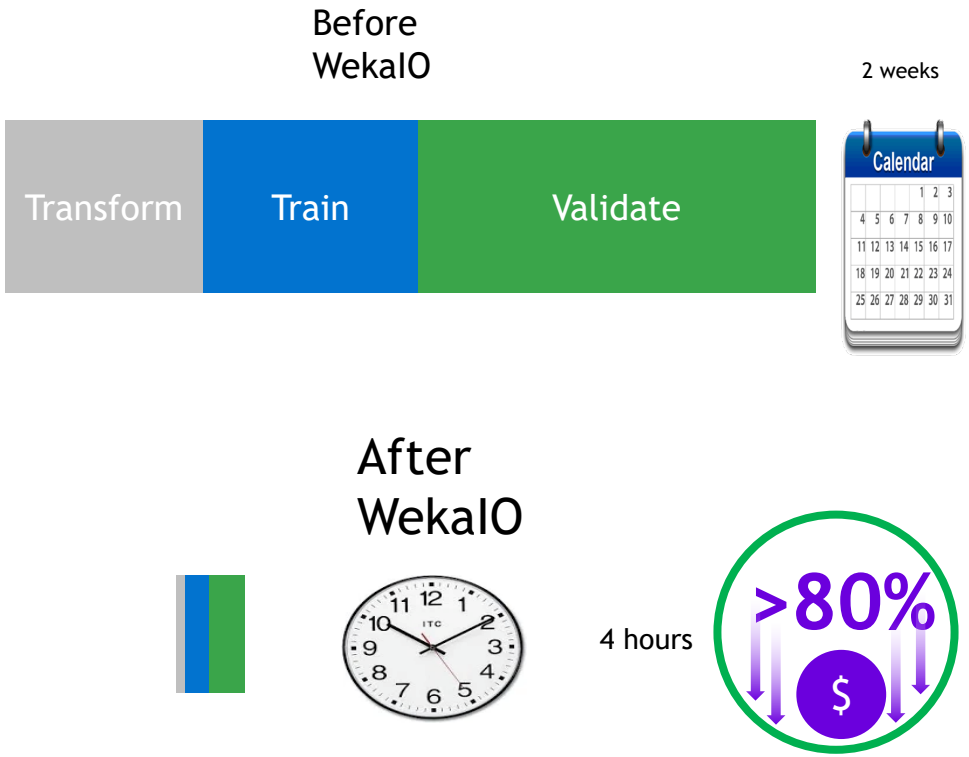
# Accelerated DataOps – Business and IT Convergence

- Accelerated DataOps for Analytics – Actionable Intelligence with BI and AI
  - Descriptive, Predictive, Prescriptive and Cognitive Analytics with same storage substrate
- Accelerated DataOps for Operational Agility – Improve productivity, reduce TCO
  - Data new source code - versioning, B&R, test/dev
  - Data Anywhere - Edge to Core to Cloud pipelines
  - Cloudstore - manage performance and capacity tiers as single namespace
- Accelerated DataOps for Governance
  - In-line encryption, virtual filesystems



# WekaAI for Datascientists, CDO's and CAO's

- Improve productivity and faster time to market and value
  - accelerate large scale data pipelines with reduced epoch times, fastest inferencing and highest images / secs benchmarks
  - run entire pipeline on the same storage backend
  - Faster than local storage

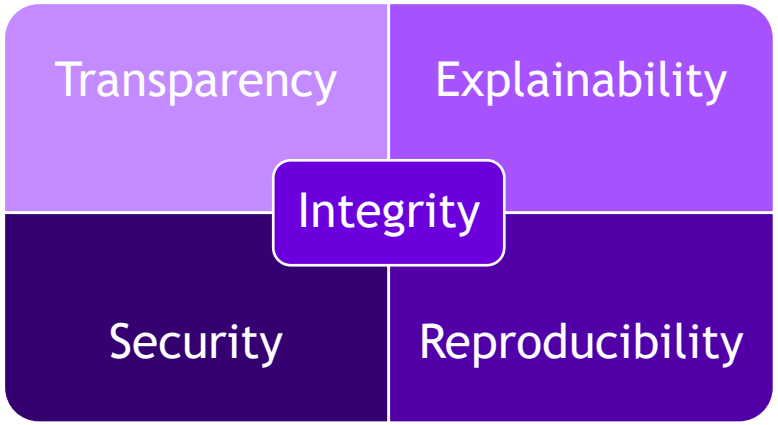



*30% better utilization results into $1.13M in savings for 10 node GPU cluster with 3 Data scientists, over 3 years*

# WekaAI for Data scientists

- Data compliance and security
  - in-line encryption support enables compliance
- Explain-ability and Reproducibility for experiments
  - instant space efficient snapshots make it easy to maintain versions
  - Snap2object retains versions for reproducibility and explain-ability
- Hybrid workflows
  - Dev and Test experiments in the public cloud, data mobility and rehydration on-premise for production

Pillars of AI Trust



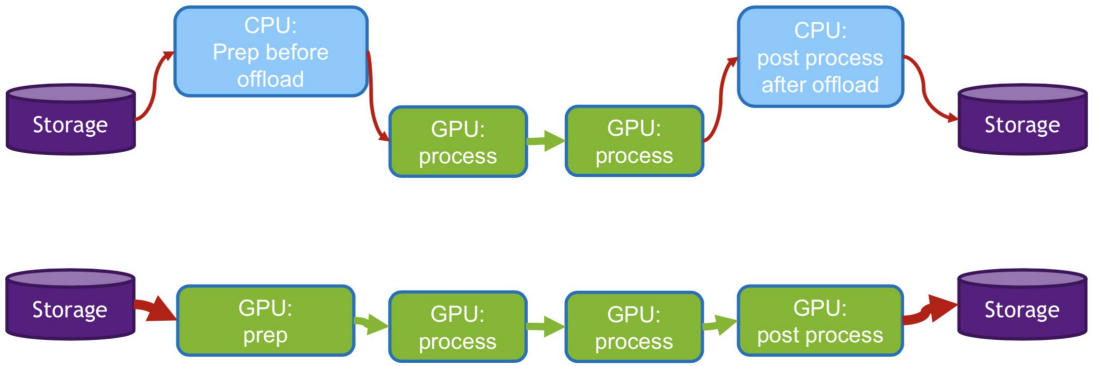
# WekaAI for Data engineers – GPU Accelerated Storage

## GPU Direct / Accelerated Libraries

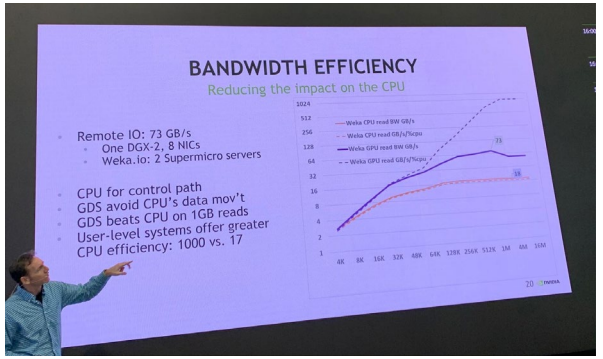
GPU Direct over NVMe over Fabrics Ready

RAPIDS | IndeX | DALI

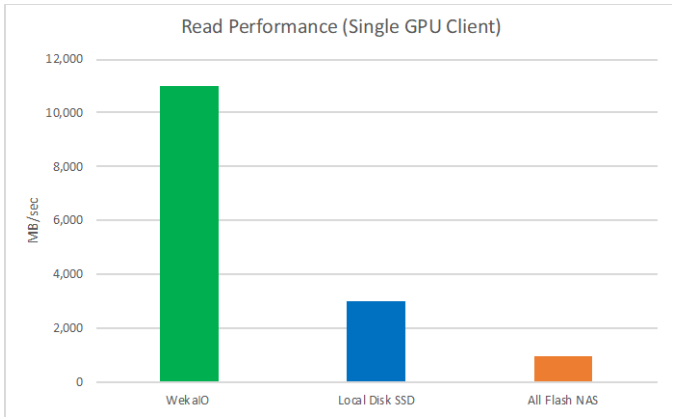
Accelerated Libraries




NVIDIA demonstrated Weka performance over 80 GB/s to a single DGX-2



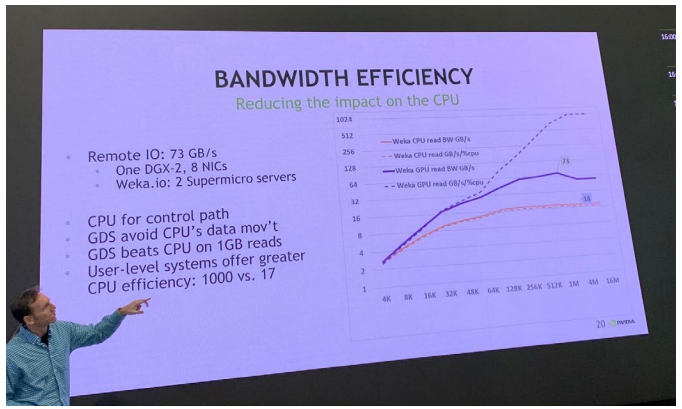
# Proof points with Accelerated Compute




- Fully saturate 100Gbit Network link
- 3x faster than local drive Storage
- 10x faster than all flash NAS


- Perfect linear scaling as cluster expands
- NVIDIA validated reference architecture

- NVIDIA demonstrated Weka performance over 73GB/second to a single DGX-2




## IO500

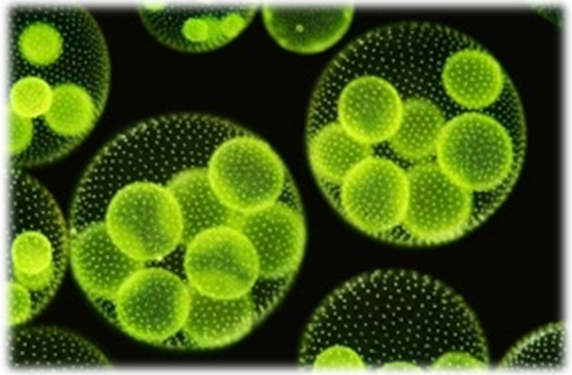
This is the official list from Supercomputing 2019. The list shows the best result for a given combination of system/institution/filesystem.

Please see also the 10 node challenge ranked list.

| # | information | | | | | | | | io500 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | list id | institution | system | storage vendor | filesystem type | client nodes | client total procs | data | score | bw | md |
| | | | | | | | | | | GiB/s | kIOP/s |
| 1 | sc19 | WekaIO | WekaIO on AWS | WekaIO | WekaIO Matrix | 345 | 8625 | zip | 938.95 | 174.74 | 5045.33 |
| 2 | sc19 | Intel | Wolf | Intel | DAOS | 26 | 728 | zip | 933.64 | 183.36 | 4753.79 |
| 3 | sc19 | National Supercomputing Center in Changsha | Tianhe-2E | National University of Defense Technology | Lustre | 480 | 5280 | zip | 453.68 | 209.43 | 982.78 |
| 4 | sc19 | NVIDIA | DGX-2H SuperPOD | DDN | Lustre | 10 | 400 | zip | 249.50 | 86.97 | 715.76 |
| 5 | sc19 | University of Cambridge | Data Accelerator | Dell EMC | Lustre | 128 | 2048 | zip | 229.45 | 131.25 | 401.13 |

WEKA

# Focused on High Performance Use Cases

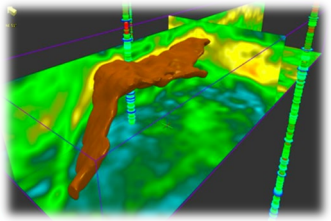


Genomics and Life Sciences

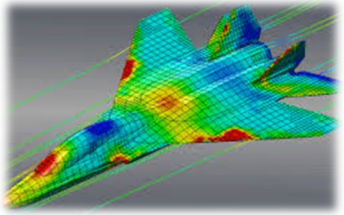
AI/Machine Learning

Financial Analytics

Secondary

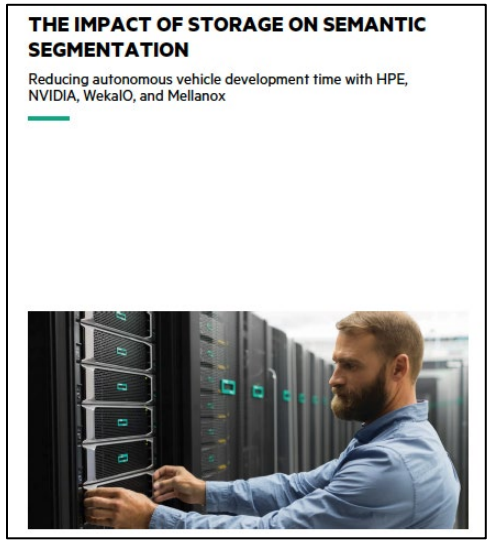


Traditional HPC

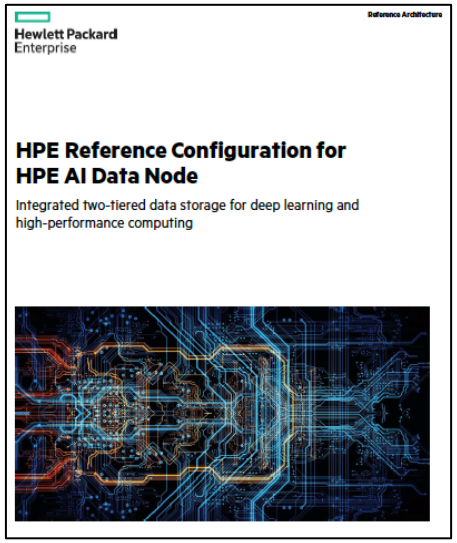


Manufacturing/ Engineering

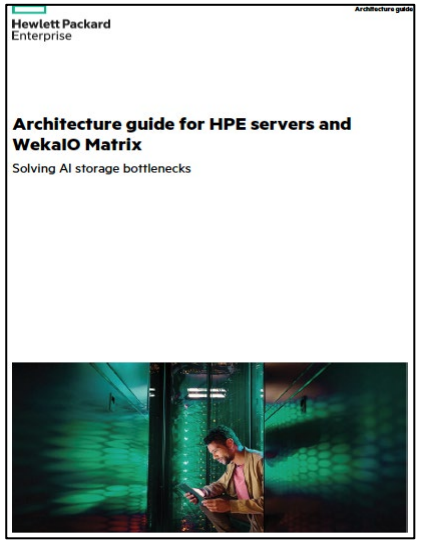
# HPE-Weka Solutions portfolio and assets



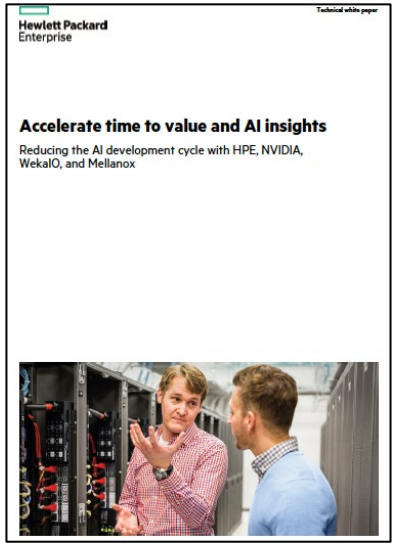

ADAS – Autonomous Driving



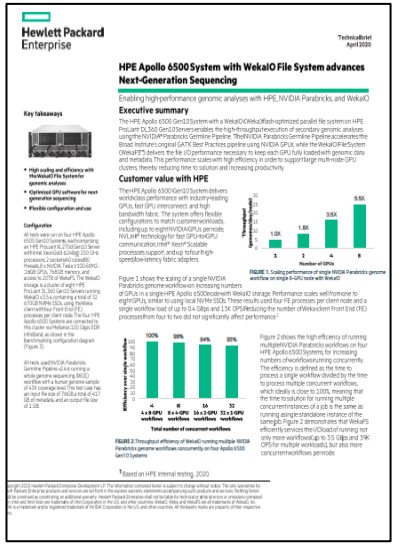

HPE AI Data Node




HPE Weka RA



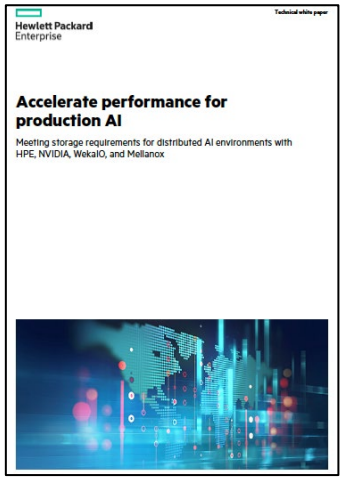

AI benchmarks




Genomics solution brief

STAC M3 benchmarks

Production AI



# Customer Profile

INNOVIZ TECHNOLOGIES


## Leading Electric Car company

tusimple

G2M RESEARCH

SAMSUNG

# Samsung


- Young Paik
- Sr. Director, Product Planning
- www.samsung.com

# Why Does Samsung Care About AI

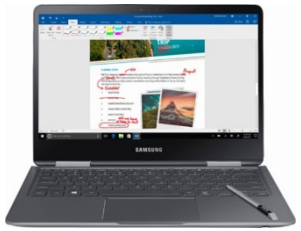
Samsung not only aids in making AI, but is a huge consumer of AI/ML


GPUs
FPGAs
ASICs


High Performance Storage Systems

Create AI

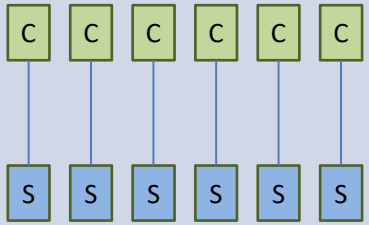
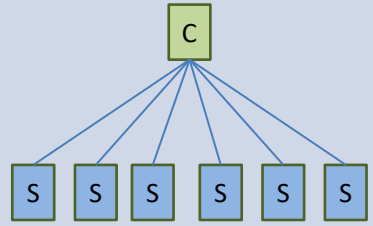




Use AI

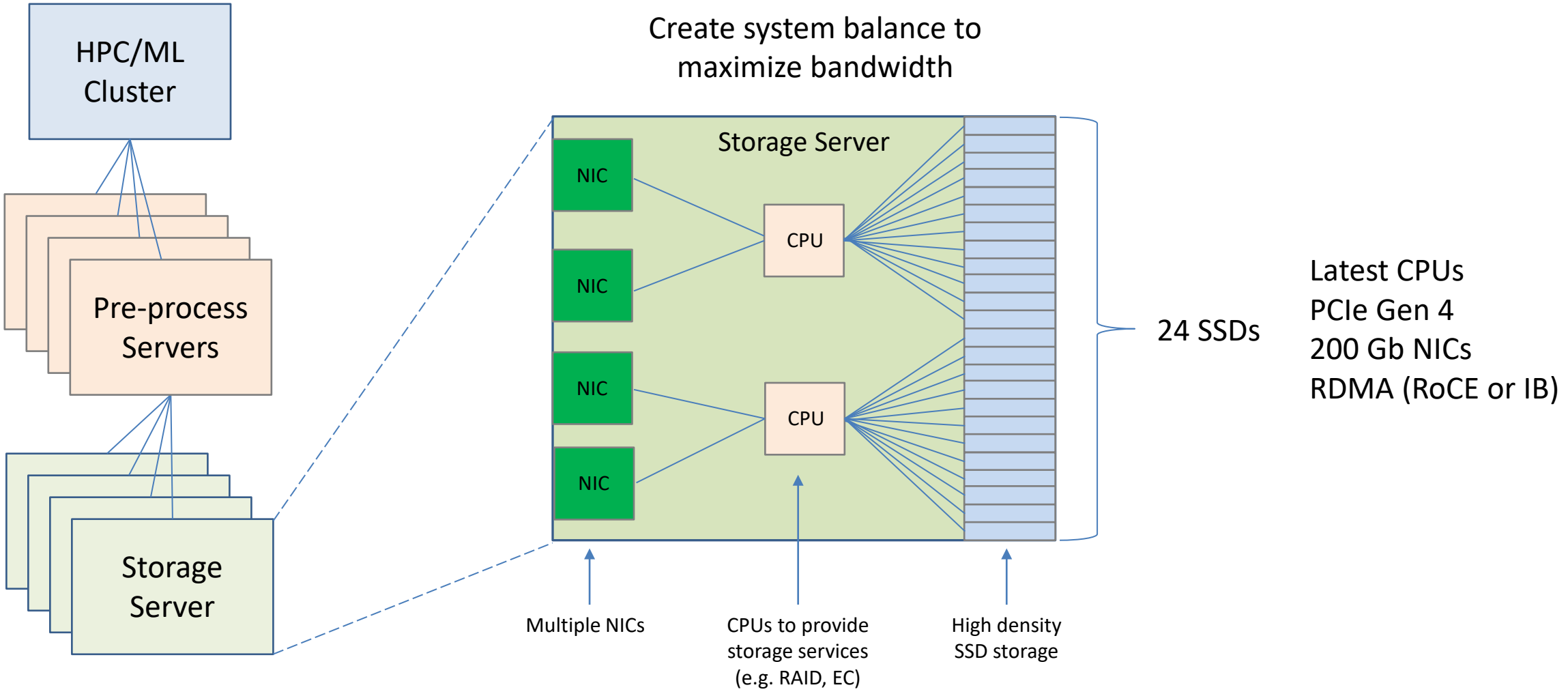
# Comparison of Big Data/ML Training/HPC

ML Training Architecture is closer to HPC than standard Enterprise Storage

| | Big Data/ML Inference | ML Training | HPC |
|---|---|---|---|
| **Data size** | 100 TB+ | 1 PB+ | 1 PB + (SSD) |
| **Workload Characteristic** | I/O bound | Compute bound | Fabric latency/bisection BW bound |
| **Scaling Strategy** | Scale Out   |   | Scale Up   |
| **Pain point** | Long query times | Large-scale data sets (100+ PB)<br>Days to weeks to train a single model once | Large-scale data sets (100+ PB)<br>Very high bandwidth reads/writes |

# HPC/Modern ML Storage Architecture

Modern HPC storage is meant to make the most of the latest storage technology



# Samsung's Most Popular SSD for HPC

The Samsung PM1733 is used by most HPC storage vendors

Shipping now!


| PM1733 | |
|---|---|
| Interface | PCIe x4 |
| PCIe | Gen 3 or 4 |
| Capacity | 1.92 – 15.36 TB |
| Read BW (GB/s) | 7.0 PCIe Gen 4<br>3.5 PCIe Gen 3 |
| Write BW (GB/s) | 3.8 PCIe Gen 4<br>3.2 PCIe Gen 3 |
| Read Latency (us) | 100 |
| Write Latency (us) | 25 |
| Dual-port | yes |

# Can We Do Better?

What will the future bring for HPC and ML training storage?

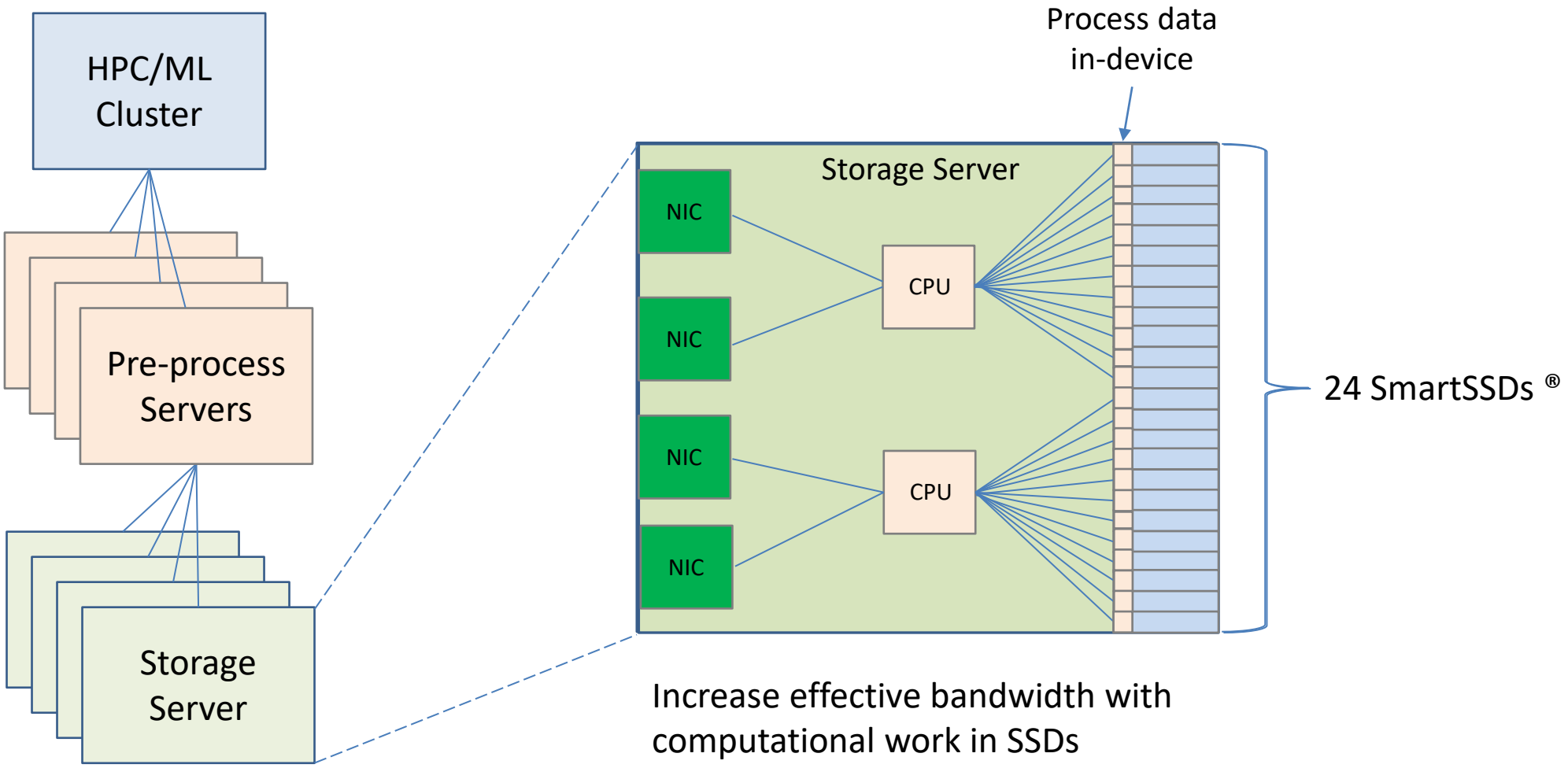
Samsung is working on new technologies to achieve even better training performance:

- Computational Storage Devices – Increases effective bandwidth by moving compute to storage
- Ethernet SSDs – Changes interface from PCIe to Ethernet

# HPC/ML Future Storage Architecture: SmartSSD®

Pushing Compute into Storage (Computational Storage Devices)

In Beta now

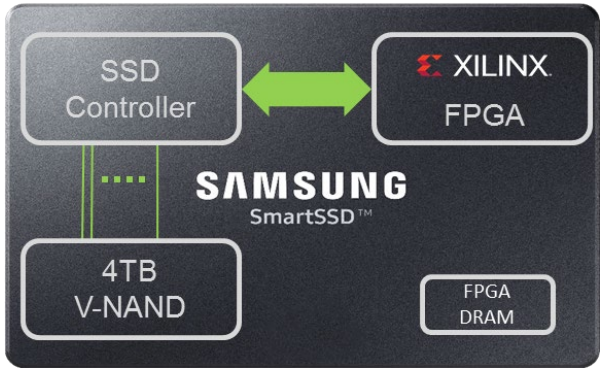



Increase effective bandwidth with computational work in SSDs

# SmartSSD® CSD Scales to Accelerate Data-Rich Workloads

SmartSSD® CSD Scales to Accelerate Data-Rich Workloads

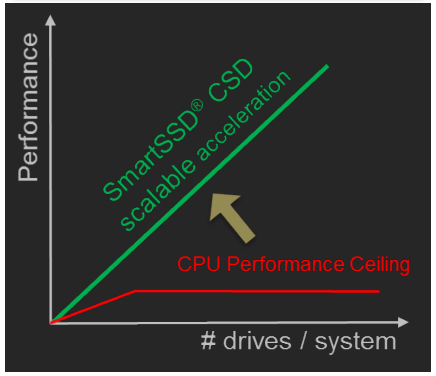
## SmartSSD U.2 Platform




### Computational Storage

- ✓ **3 & 6 GBps internal BW per device:** Minimize external data movement
- ✓ **FPGA:** Each device has 3x~10x core equivalents for offload/acceleration
- ✓ **4TB storage, 4 GB FPGA DRAM:** For Inline and Data@Rest processing

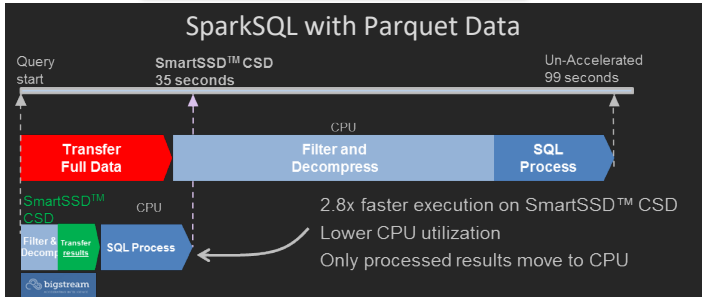
## Acceleration Concept

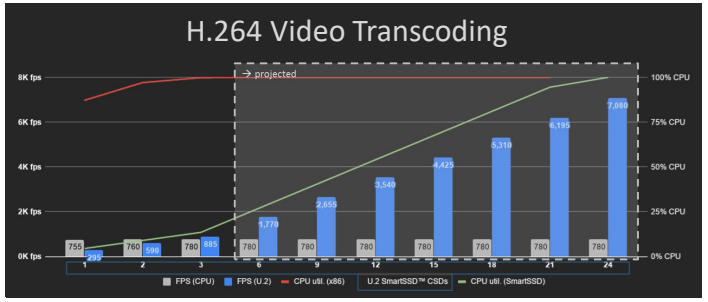


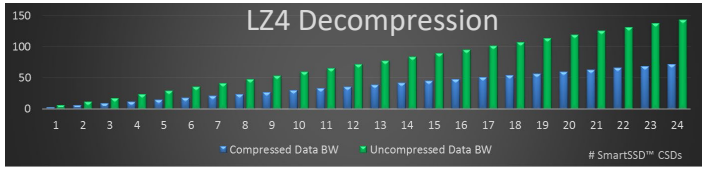

### Scalable Performance

- ✓ **Near Data Processing:** Data format conversion, Filtering, Metadata management, DB Analytics, Video processing
- ✓ **New Services:** Secure content, Edge acceleration

## Partner Solutions



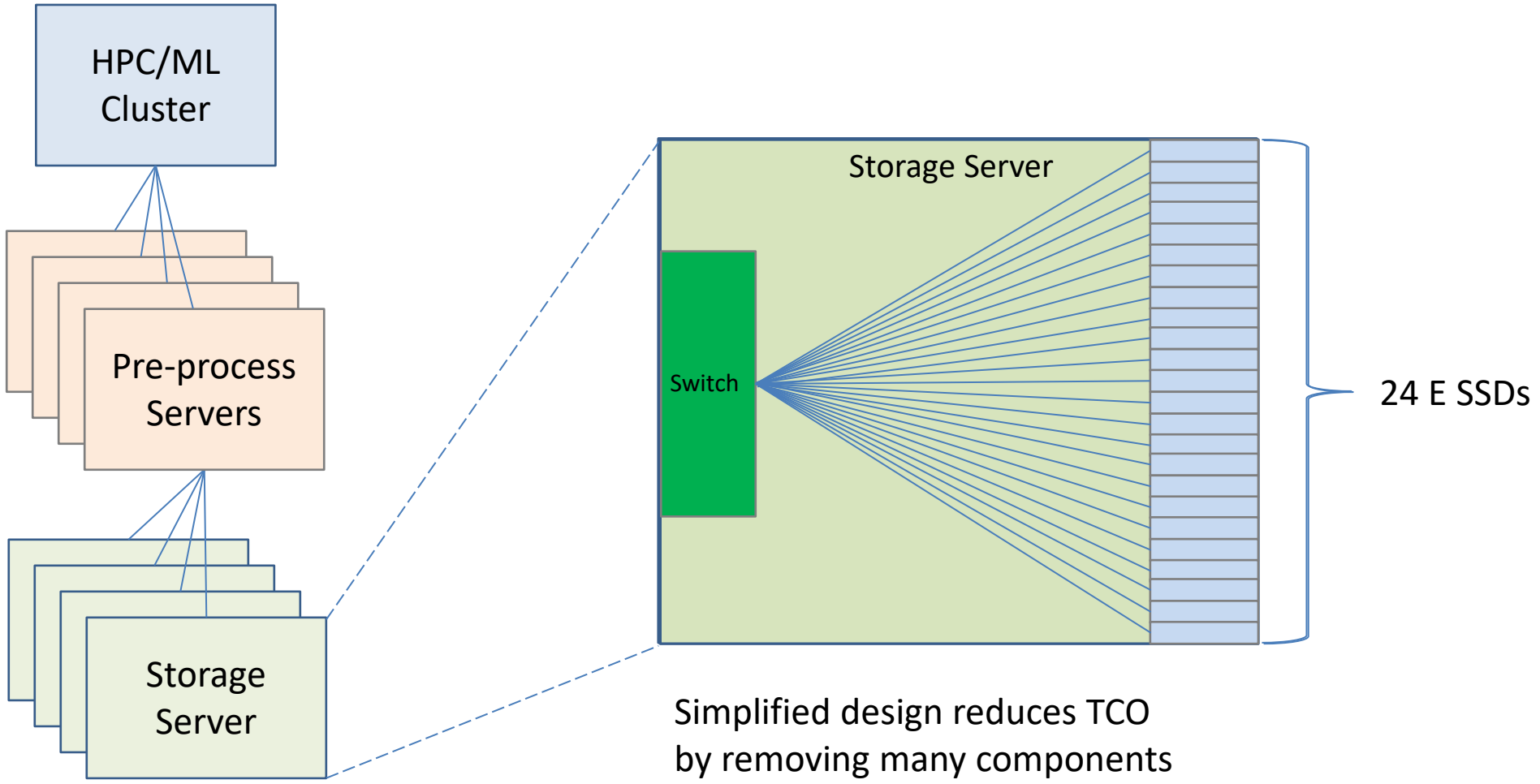



# HPC/ML Future Storage Architecture: Ethernet SSD

Simplifying deployment of storage by using Ethernet

Prototype now



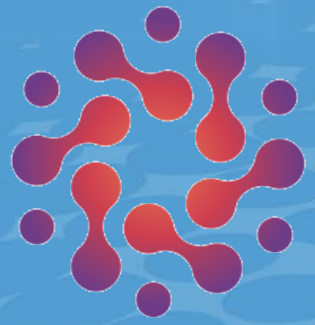

Simplified design reduces TCO
by removing many components
of legacy systems

# Thank you!

Young Paik

Young.Paik@Samsung.com

G2M RESEARCH



DATYRA

# Datyra


- Keith Klarer
- Chief Executive Officer
- www.datyra.com

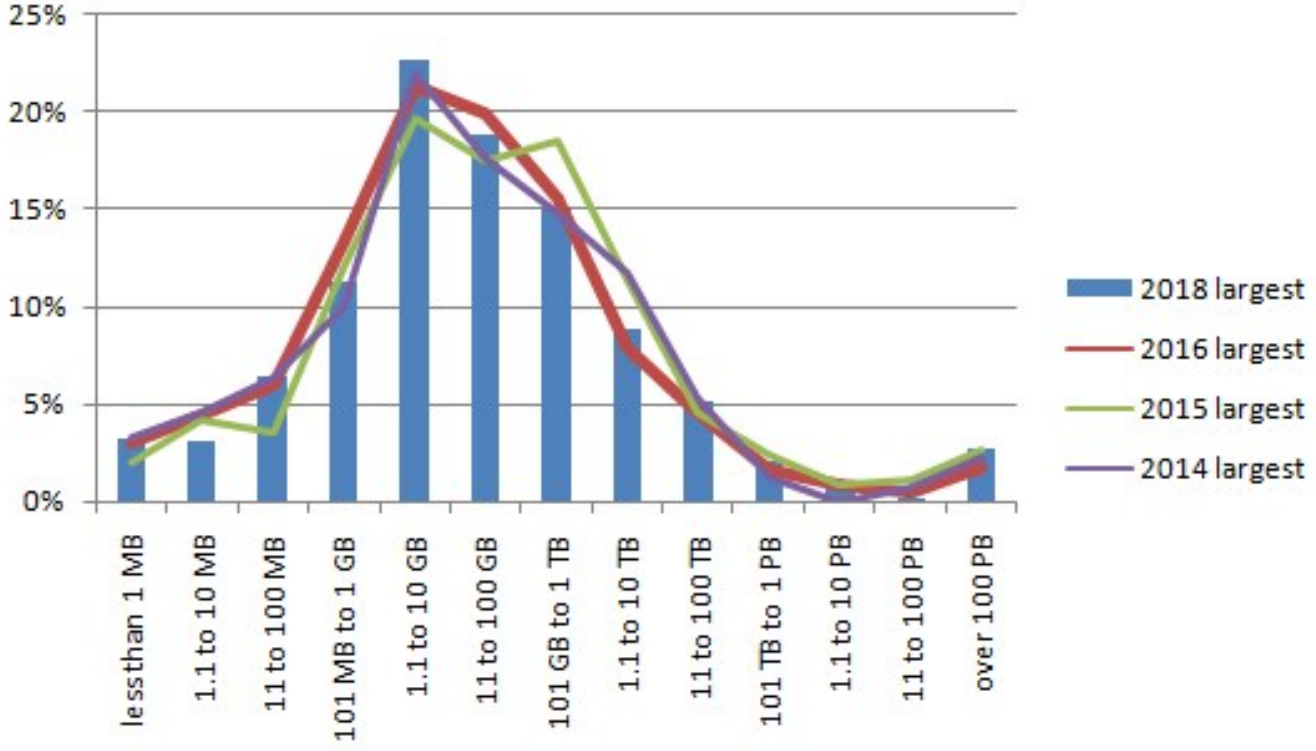
# Machine Learning Infrastructure


- What type of infrastructure do I need to for Machine Learning?
- How is going to work with my legacy infrastructure?
- How “HPC” are my requirements?
  - Is it all about size and speed?
- How can I scale up (and down)?
- How can I achieve business goals while optimizing capex, opex and development costs?

# Data Size Surveys

DATYRA

Largest Dataset Analyzed 2014-2018



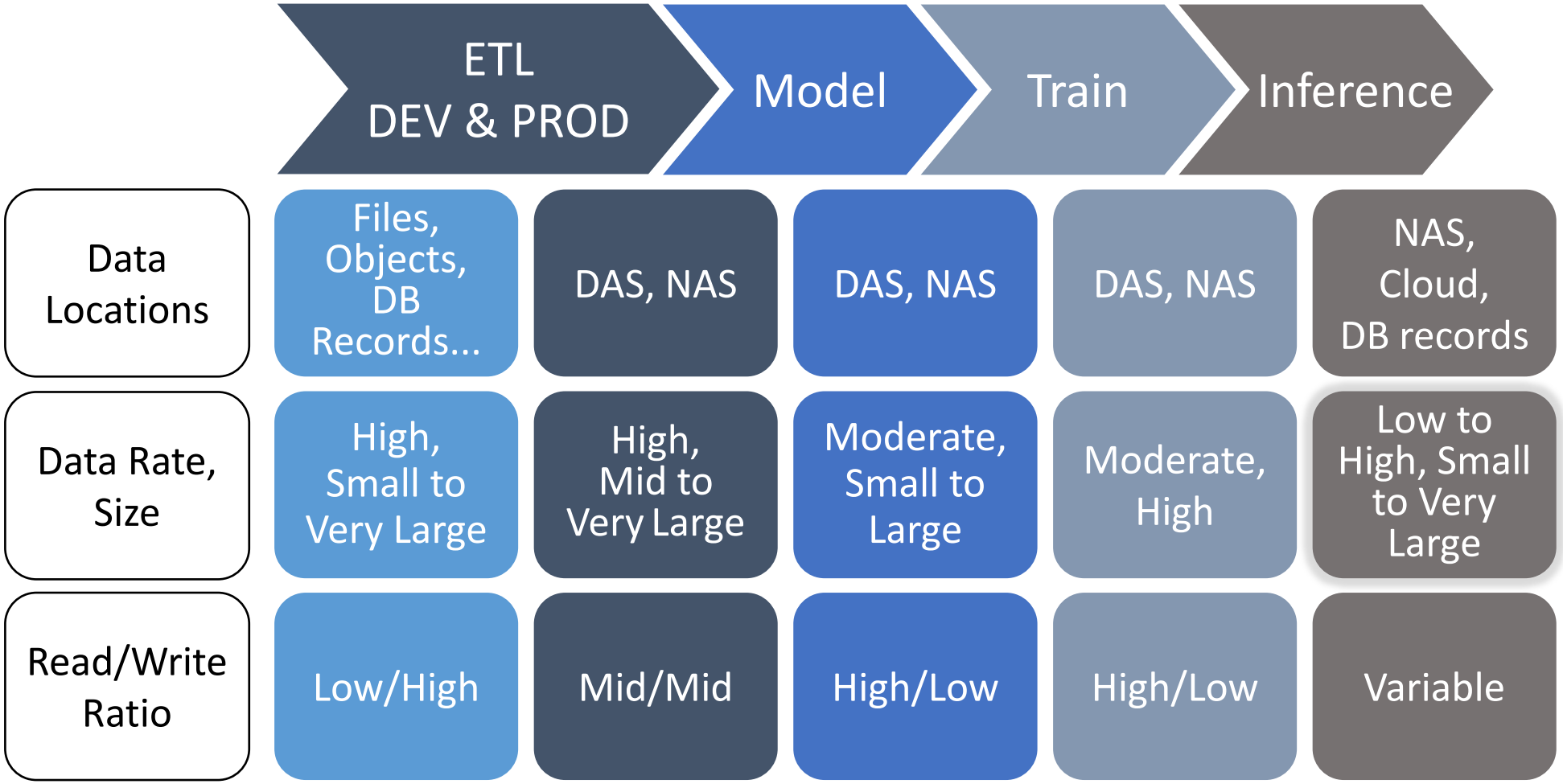

KDnuggets Poll
https://www.kdnuggets.com/2018/10/poll-results-largest-dataset-analyzed.html

ML Dataset Examples:

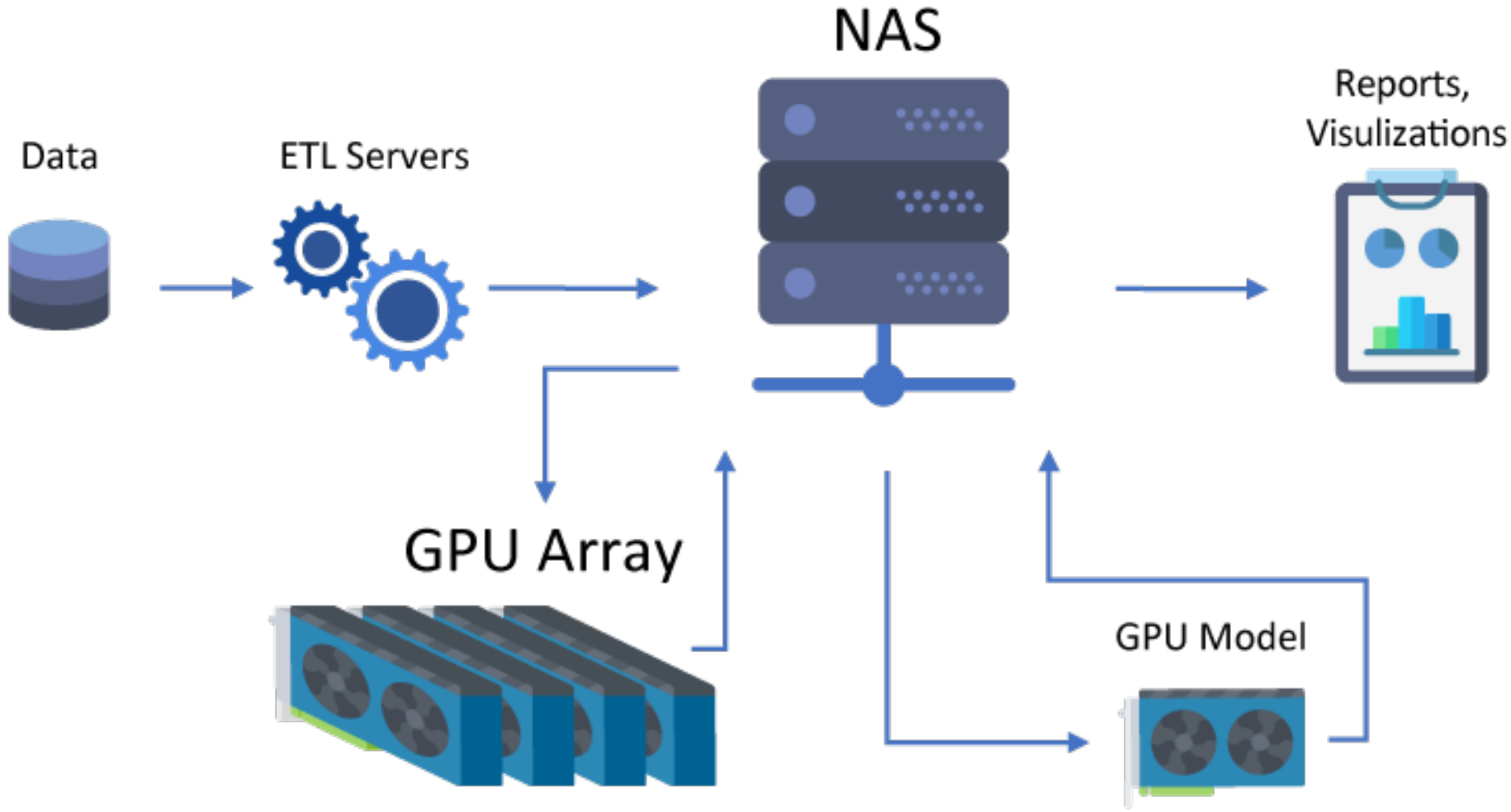
- Open Images: 9M images, 500GB
- Tencent ML: 18M images, 1TB
- Free Music Archive: 100k files, 1TB
- Million Song Dataset: 280GB
- Yelp: 2.7GB JSON, 2.9GB SQL, 7.5GB images
- Genome: 200GB per person
- Oil Exploration: 4TB per site
- Movie: 1-2PB for production
- Sumo Logic: 100PB logs daily

# Data Flow in ML Applications

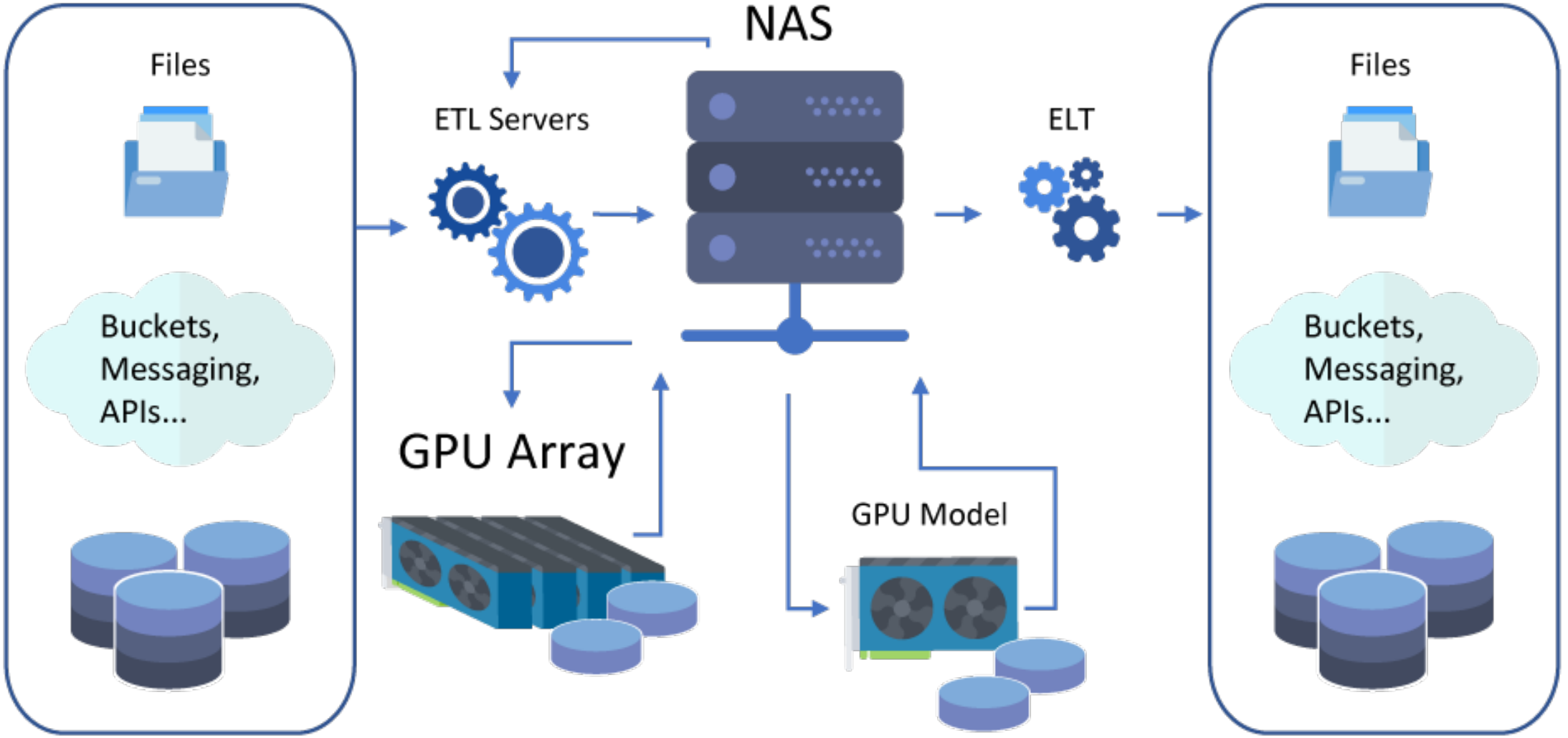
DATYRA



# ML Infrastructure – Simple View



# ML Infrastructure – Better View



# Scope the Infrastructure


- Survey your current and planned data stores
  - There are often more than originally anticipated
  - Virtually always heterogeneous
  - Understand legacy requirements at inputs and outputs
- Understand the data velocity along the data pipeline
- ML training and inference loads are hard to guestimate
  - Prototype and make sure you have scalability here
- Never underestimate ETL requirements
  - Can be greater than ML inference requirements!

# Select Infrastructure Features


- Nearly all infrastructures can benefit from some “HPC” features
- A NAS system that:
  - Easily scales capacity
  - Can provide high data velocity when needed
  - Can connect with a wide variety of other data stores
  - Can be deployed locally and in the cloud
- Aggregate servers to take advantage of high performance interconnects
- Use NVME flash devices for both NAS and DAS

# Cost Optimization


- Early on, deploy capable modeling systems to developers
  - Individual workstations have a short payback time
  - Prototype to understand training and inference requirements
- Work with legacy data owners to determine access and quality
  - Watch out for data egress costs
- A hybrid cloud model can often be more economical
  - Containers and Kubernetes are enablers
- Automate the data pipeline
  - It can save a lot of opex (and grief)
- More info: https://datyra.com/publications/

G2M RESEARCH

# Panel Questions and Audience Surveys

# Panel Question # 1

- There is a perception that successful AI implementations require “big iron” compute and storage platforms. Can meaningful AI solutions be built for less than seven figures?
  - NVIDIA/Mellanox
  - Weka
  - Samsung
  - Datyra

# Audience Survey Question #1

- Has your organization explored and/or deployed AI-based systems for business intelligence yet? (check one):
  - We have deployed AI for a variety of business applications: 38%
  - We have deployed AI for a couple of business applications : 15%
  - We are performing proof of concept evaluations on AI solutions, with the idea of deploying them in the near future: 31%
  - We are talking to vendors about potential AI solutions: 0%
  - We aren't actively exploring using AI in our organization: 15%

# Panel Question #2

- Data management is a significant issue in building and maintaining AI models. What are some best practices for managing data sets for AI?
  - Weka
  - Samsung
  - Datyra
  - NVIDIA/Mellanox

# Audience Survey Question #2

- What do you see as the greatest challenge for your organization to implement an AI solution? (check all that apply):
  - Understanding what business value we can reasonably expect from AI: 27%
  - Finding the right vendor and/or people to implement an AI solution: 20%
  - Building the right training data set: 40%
  - Affording the hardware required for a meaningful AI solution: 13%
  - Achieving the right level hardware and software performance: 40%
  - Other issues: 7%

# Panel Question # 3

- When optimizing storage performance for AI training and validation, what factors should be considered?
  - Samsung
  - Datyra
  - NVIDIA/Mellanox
  - Weka

# Audience Q&A

G2M RESEARCH

# Thank You For Attending

G2M

RESEARCH