G2M
RESEARCH

G2M Research Multi-Vendor Webinar: AI, GPUs, and Storage Use Cases in Healthcare

Tuesday Oct 20, 2020

NVIDIA | WEKA | KIOXIA | DATYRA

Webinar Agenda

9:00-9:05	Ground Rules and Webinar Topic Introduction (G2M Research)
9:06-9:37	Sponsoring Vendor presentations on topic (8 minute each)
9:38-9:46	Key Question 1 (1-minute question; 2 minutes response per vendor)
9:47-9:47	Audience Survey 1 (1 minute)
9:48-9:56	Key Question 2 (1-minute question; 2 minutes response per vendor)
9:57-9:57	Audience Survey 2 (1 minutes)
9:58-10:06	Key Question 3 (1-minute question; 2 minutes response per vendor)
10:07-10:18	Audience Q&A (12 minutes)
10:19-10:20	Wrap-Up

G2M RESEARCH

G2M RESEARCH

G2M Research Introduction and Ground Rules

Mike Heumann

Managing Partner, G2M Research

Panelists

Brad Genereaux
Medical Imaging Alliances Mgr
www.nvidia.com

Shimon Ben-David
Field Chief Technical Officer
www.weka.io

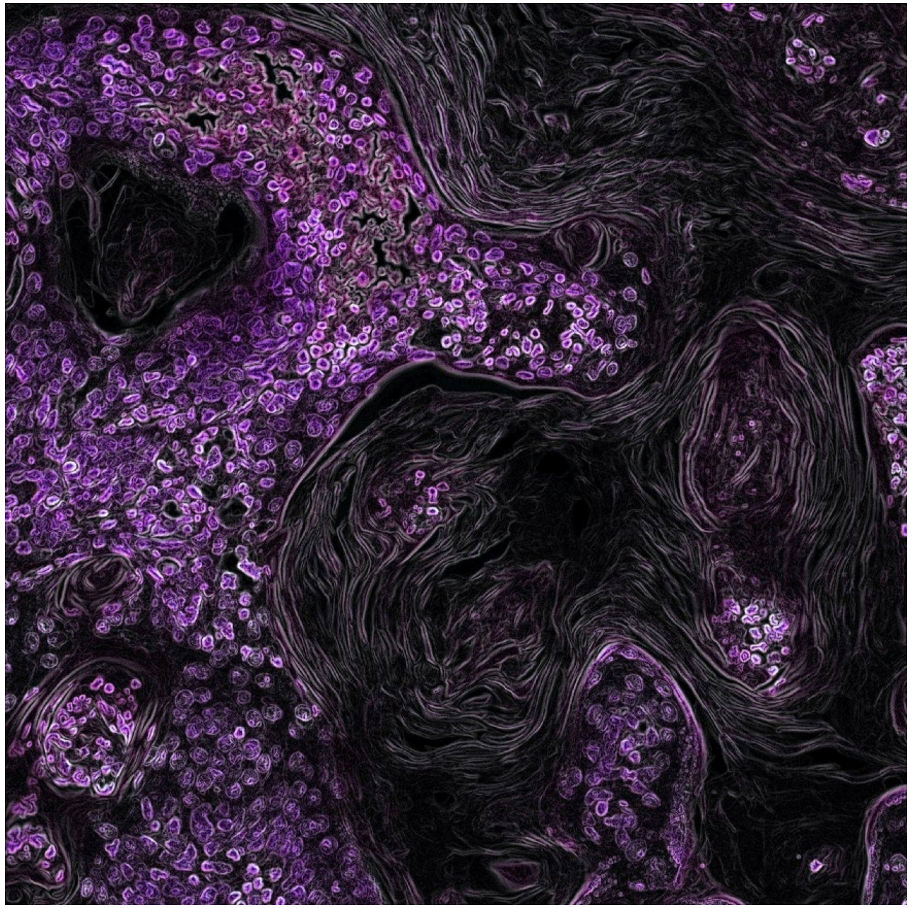
Matt Hallberg
Sr. Product Marketing Mgr
www.kioxia.com

Keith Klarer
Chief Executive Officer
www.datyra.com

Mike Heumann
Managing Partner
www.g2minc.com

AI/ML, Imaging, and Oncology

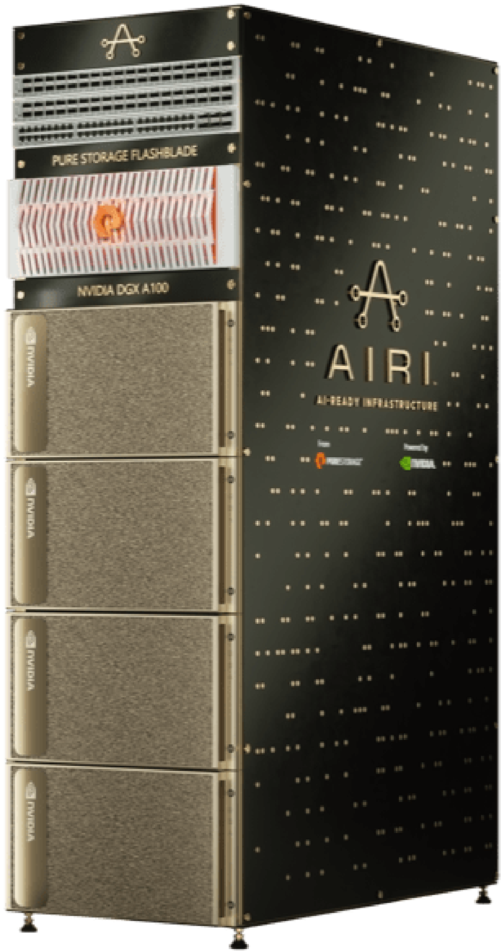
- Early ID, diagnosis of tumors is a strong predictor of positive oncology outcomes
- AI and ML analysis of large historical imaging data shows promise for building effective models for ID and diagnosis
- Platforms such as PAIGE, and the Johns Hopkins RAIL lab, are building AI-based oncology imaging models
- These models can utilize petabytes for training, and need robust storage and data management architectures



G2M RESEARCH

Data Architectures for Healthcare AI/ML

- These architectures must efficiently categorize, move, and retain disparate data types between complex storage, servers, and GPU clusters
- NVIDIA's AI Ready Infrastructure (AIRI) is a monolithic example of such an architecture
- Other architectures can include distributed nodes, resources, and data movement, or can be embedded within imaging solutions
- This webinar will explore the tools for building AI-based solutions, and the hardware behind them



G2M RESEARCH

AI, GPUS, AND STORAGE USE CASES IN HEALTHCARE

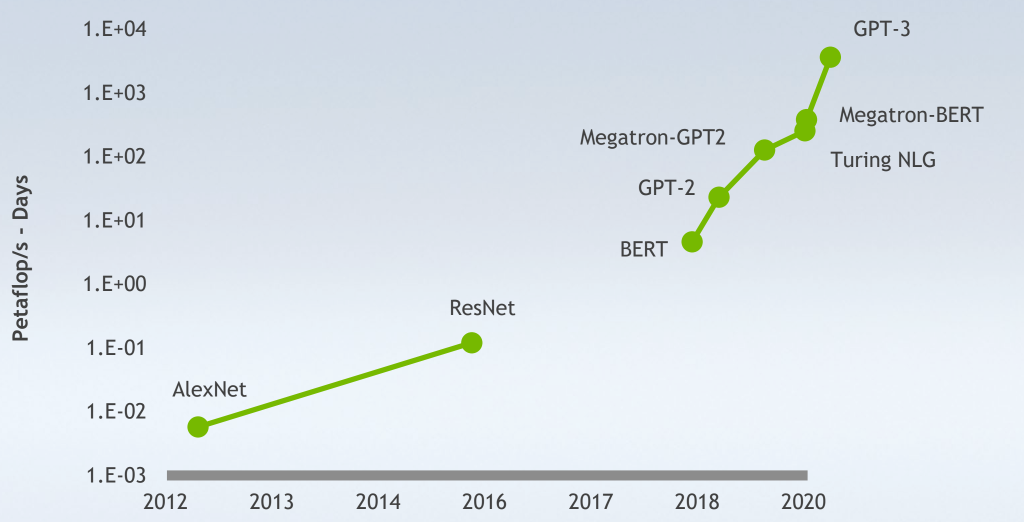
Brad Genereaux | Medical Imaging Alliance Manager

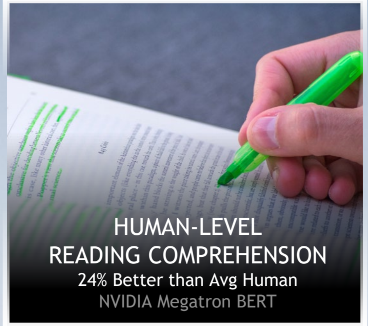
ARTIFICIAL INTELLIGENCE

The Most Powerful Technology Force of Our Time

EXPLODING MODEL COMPLEXITY

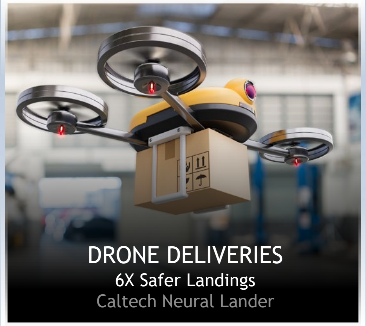
30,000X in 5 Years | Now Doubling Every 2 Months





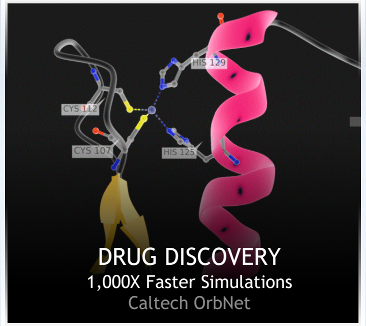
HUMAN-LEVEL READING COMPREHENSION
24% Better than Avg Human
NVIDIA Megatron BERT

HUMAN-LIKE CHATBOTS
Half of Users Preferred AI
Facebook BlenderBot



DRONE DELIVERIES
6X Safer Landings
Caltech Neural Lander

FIGHTING CORONAVIRUS
Mapping Millions of Proteins
Technical U. of Munich NLPs



DRUG DISCOVERY
1,000X Faster Simulations
Caltech OrbNet

ALL-TERRAIN WALKING ROBOT
On the Fly Surface Adaptation
NVIDIA RL Controller

WHAT ARTIFICIAL INTELLIGENCE CAN DO FOR HEALTHCARE

Shortage of Doctors
US facing shortages of up to 122,000 physicians by 2032

Imaging Data Explosion
Value in diagnostic/therapeutic workflows results in millions of images being acquired

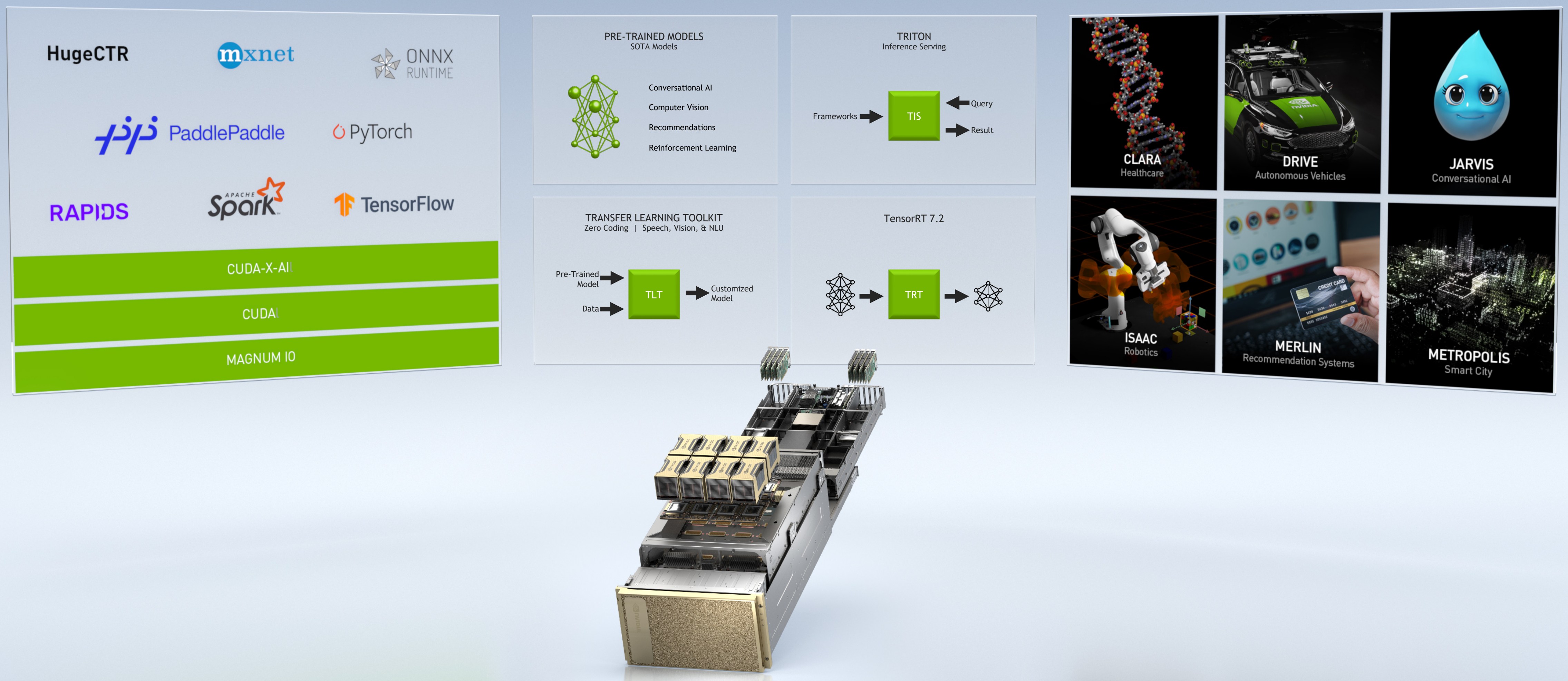
AI for Quality Care
Alleviating the 5% error rate in the system and to create new care pathways

Doctors Involved In Developing AI
Doctors, as subject matter experts, can train AI with their knowledge

AI for Smart Hospitals
Driving operational efficiency is key to reducing cost and service more patients

AI for Intelligent Instruments
Instruments are rapidly adopting AI for positioning, QA, and decision support

NVIDIA AI

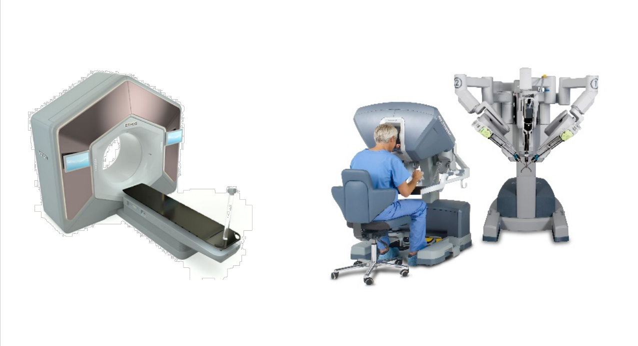

HugeCTR
mxnet
ONNX RUNTIME
PaddlePaddle
PyTorch
RAPIDS
APACHE Spark
TensorFlow
CUDA-X-AI
CUDA
MAGNUM IO
PRE-TRAINED MODELS
SOTA Models
Conversational AI
Computer Vision
Recommendations
Reinforcement Learning
TRITON
Inference Serving
Frameworks
TIS
Query
Result
TRANSFER LEARNING TOOLKIT
Zero Coding | Speech, Vision, & NLU
Pre-Trained Model
Data
TLT
Customized Model
TensorRT 7.2
TRT
CLARA
Healthcare
DRIVE
Autonomous Vehicles
JARVIS
Conversational AI
ISAAC
Robotics
MERLIN
Recommendation Systems
METROPOLIS
Smart City

CONNECTED HEALTH AND THE SMART MEDICAL DEVICE REVOLUTION

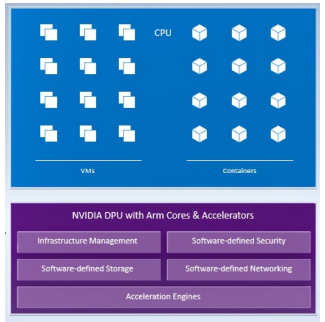
Connected health enables smart hospitals by creating a comprehensive solution set of products and services that facilitates patient engagement, collaboration among caregivers and cultivation of insights toward clinical outcomes for increased efficiencies and quality of care

Smart Sensing
Monitor, respond to, and support the patient within the hospital and beyond

Smart Imaging
Make diagnostic imaging more efficient and more portable with acceptable quality



Smart Treatment
Enabling real-time tools for physicians, surgeons, and technologists

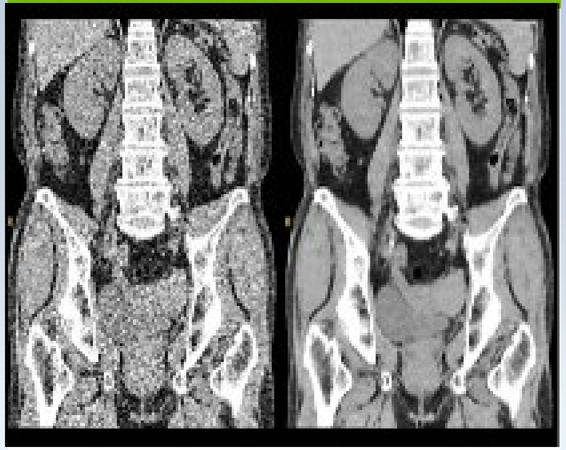


Data Processing Unit (DPU)
Software-defined Data Center Infrastructure-on-a-Chip

NVIDIA GPU COMPUTING PLATFORM

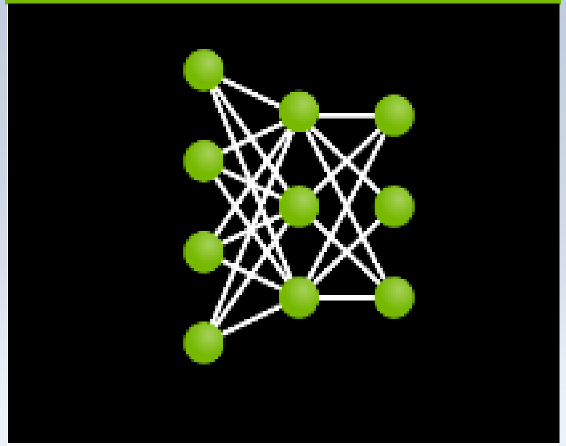
The Modern Healthcare Datacenter

GPU Accelerated Visualization



Real-time Processing

GPU Deep Learning

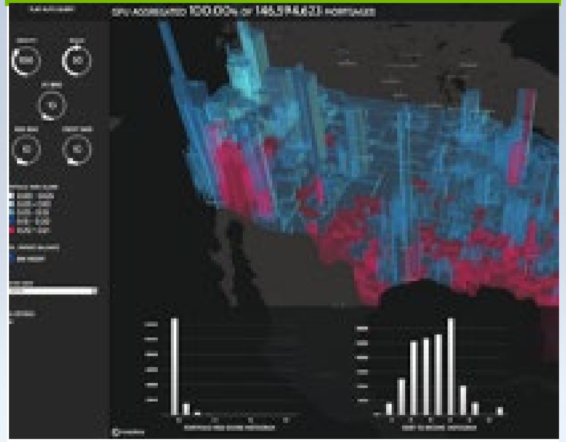


Workflow & Image Analysis

GPU Accelerated Virtualization

VDI & Virtual Rad Workstations

GPU Data Analytics



Outcomes, Prevention, Cost

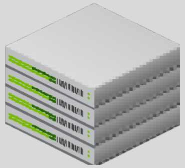
NVIDIA COMPUTING PLATFORM

SAMPLE COVID-19 AI USE CASE

Using NVIDI Clara to Encapsulate an AI Algorithm into an Integrated AI Application

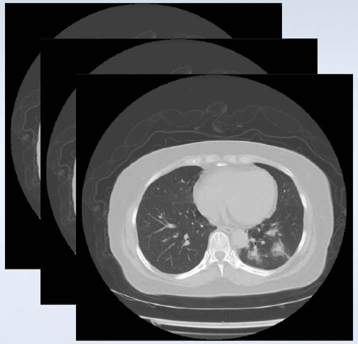
PACS → Interoperability Transport → Receive Adapter → Pre-Transforms → Inference → Post-Transforms → Send Adapter → Interoperability Transport → PACS

DICOM Input	DICOM Reader	Lung Segmentation	Chest CT Classification	DICOM-SR Writer	DICOM Output



PACS

Lung CT DICOM Datasets



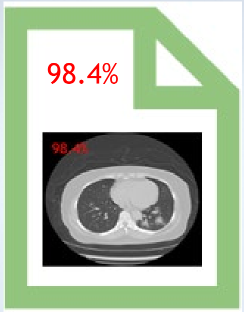
Load DICOM series
Extract pixels
Extract metadata
Export MHD

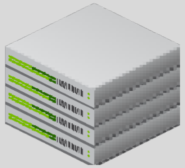
Segment Lung Regions

COVID 98.4%

No-COVID 01.6%

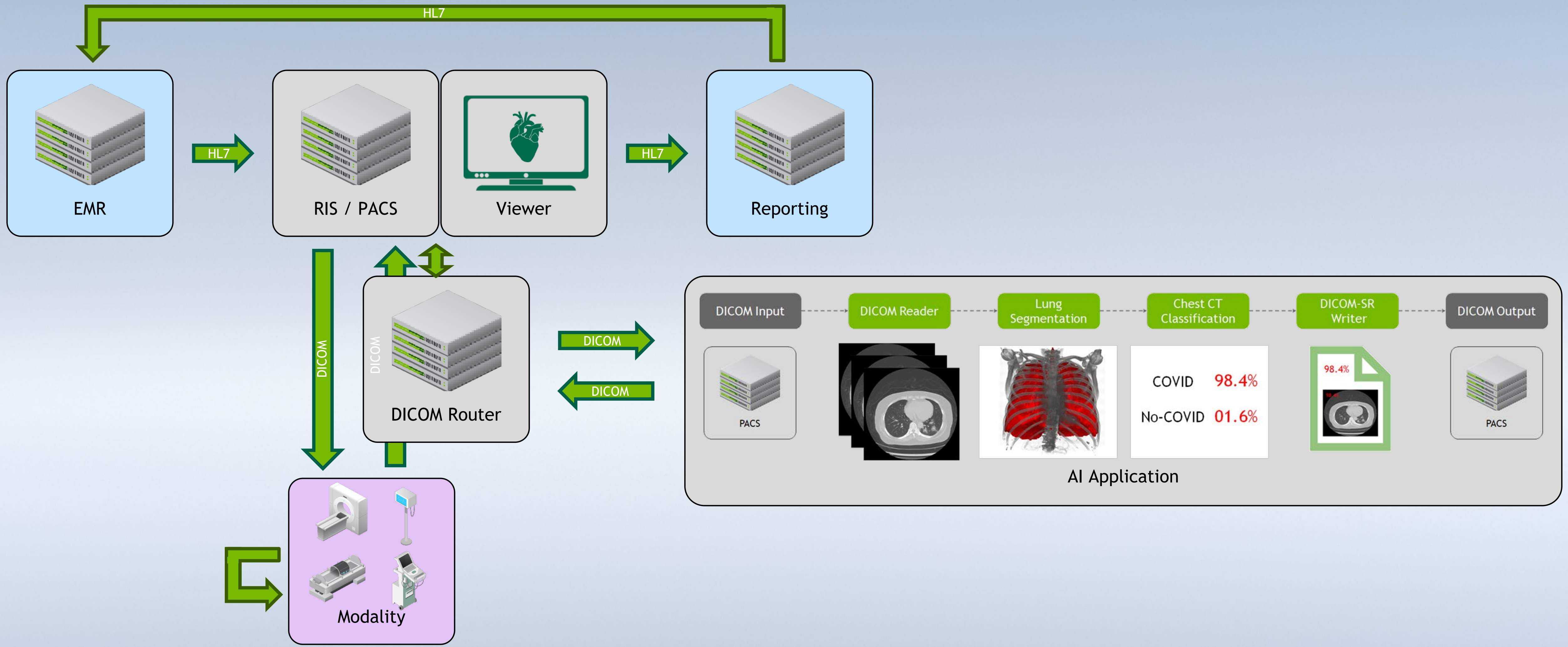
Classify 3D Dataset

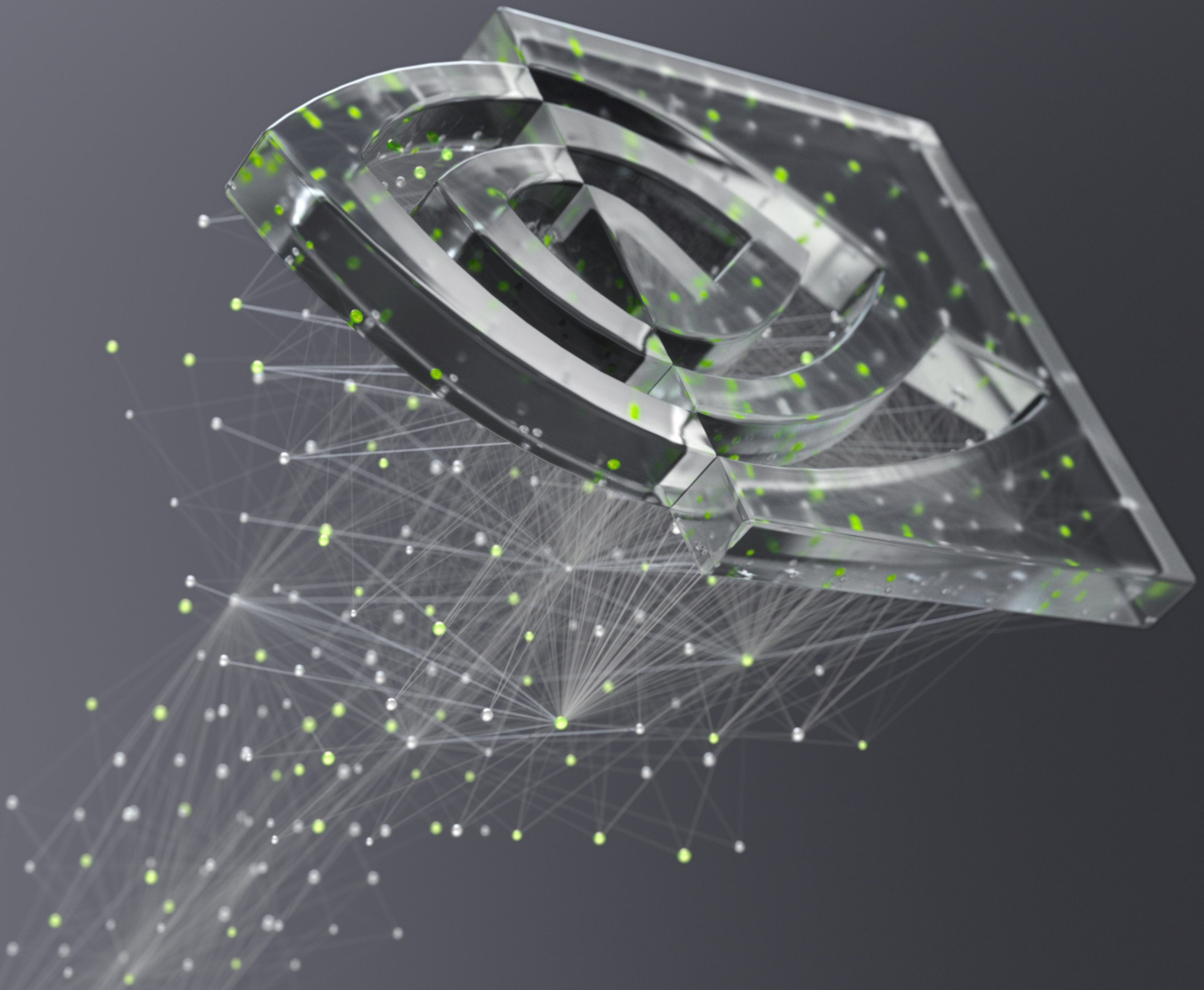




PACS

THE CONNECTED AI MEDICAL ECOSYSTEM





NVIDIA

G2M – AI, GPUs and storage in Healthcare

Shimon Ben David
Field CTO

WEKA FOR THOSE WHO SOLVE BIG PROBLEMS

Weka has built a software-defined, parallel file system that leverages NVMe and high-speed networks to unleash the value of your data

FINANCE

INCREASING PROFITABILITY FOR FINANCIAL TRADING

REDUCING FRAUD

REDUCING RISK

IMPROVING BUSINESS PERFORMANCE THROUGH PREDICTIVE/PRESCRIPTIVE ANALYTICS

LIFE SCIENCE

ACCELERATING DRUG & VACCINE DISCOVERY

ACCELRATING PRECISION MEDICINE

MORE ACCURATELY DETECTING DISEASES

AI/ML

WINNING THE RACE TO LEVEL 4 VEHICLE AUTONOMY

MORE ACCURATELY DETECTING ANOMOLIES IN MEDICAL IMAGES

INCREASING SALES & PROFITABILITY WITH RECOMMENDATION SYSTEMS

INCREASING MANUFACTURING YIELDS

HPC

INCREASING NATIONAL SECURITY

DISCOVERING & MAXIMISING PROFITABILITY FROM OIL & GAS RESERVOIRS

MANUFACTURING SAFER, GREENER & MORE PROFITABLE PRODUCTS

THE BIG PROBLEMS

WEKA CUSTOMERS SOLVING BIG PROBLEMS

Genomics england

The Genomics England cluster required a new solution to allow scaling of the company's DNA data bank in line with the anticipated five-year growth. Already at 25 petabytes, our existing solution had already ***reached its limit and performance had deteriorated****. We needed a modern solution that could scale to 100s petabytes while maintaining performance scaling, and it had to be simple to manage at that scale. With its clever combination of flash for performance and object store for scale, Weka has proven to be a great solution.*

David Ardley, Director of Technology

tu simple

WekaIO was the clear choice for our DNN training… ***standard NAS would not scale*** *and Matrix [was] the* ***most performant of all the parallel file systems*** *we evaluated…we really liked that it was hardware-independent allowing us better* ***control over our infrastructure costs****.*

Dr. Xiaodi Hou, Co-founder and CTO

---

tgen® TRANSLATIONAL GENOMICS RESEARCH INSTITUTE

Future-thinking companies like WekaIO, complement our core principle of ***accelerating research and discovery****. The ability to run more concurrent high performance genomic workloads will significantly advance our time to discovery.*

Nelson Kick, Manager of HPC Operations

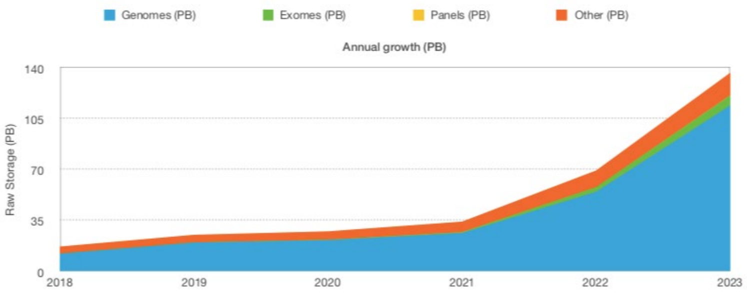
cerence™

We looked at our legacy architecture and instead of taking an evolutionary step, ***we took a revolutionary approach. Weka cost-effectively*** *enables both the use of POSIX and object storage with performance and latency that is* ***far superior to any other system.***

Bridget Collins, Chief Information Officer

THEIR COMMON CHALLENGES

IT BUDGETS ARE BEING SQUEEZED

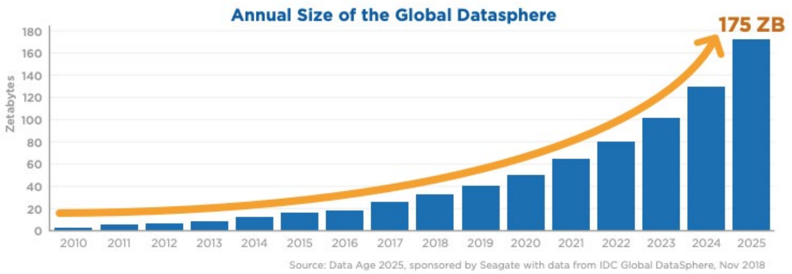


DO MORE WITH LESS!!!

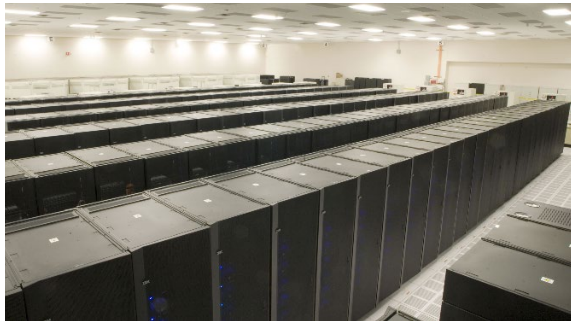
FEWER PERSONNEL

BUDGETS SQUEEZED

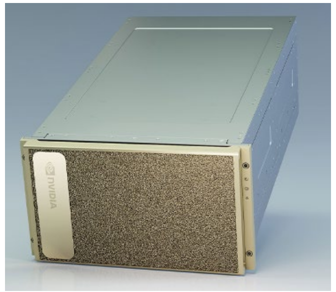
DATA IS GROWING (NOT NEW NEWS ☺)



COMPUTE PERFORMACE DENSITY IS INCREASING



ROADRUNNER AT LANL, INSTALLED 2008
PERFORMANCE = 1 petaFLOP
116,640 PROCESSOR ~CORES
SIZE = 260 RACKS

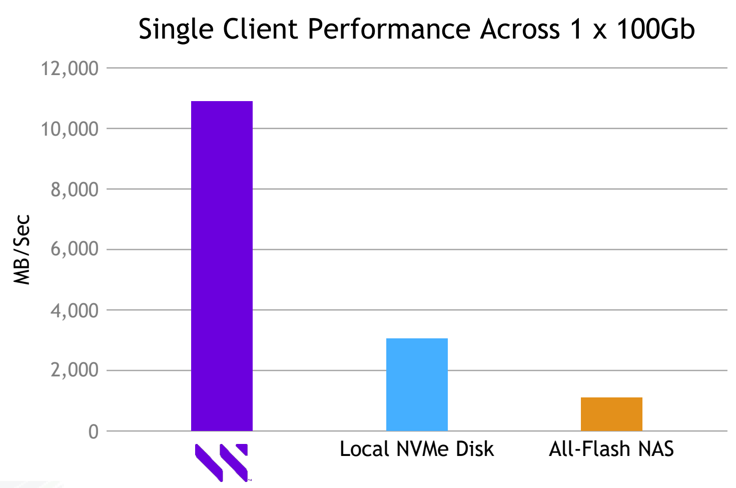


NVIDIA DGX-A100, LAUNCHED 2020
PERFORMANCE = 5 petaFLOPs
8 X NVIDIA A100 GPUs
SIZE = 6 RACKABLE UNITS

INTRODUCING WEKA FOR HIGH VALUE WORKLOADS

- Highest performance
- Massively scalable
- Easy to manage
- Software on your favorite servers

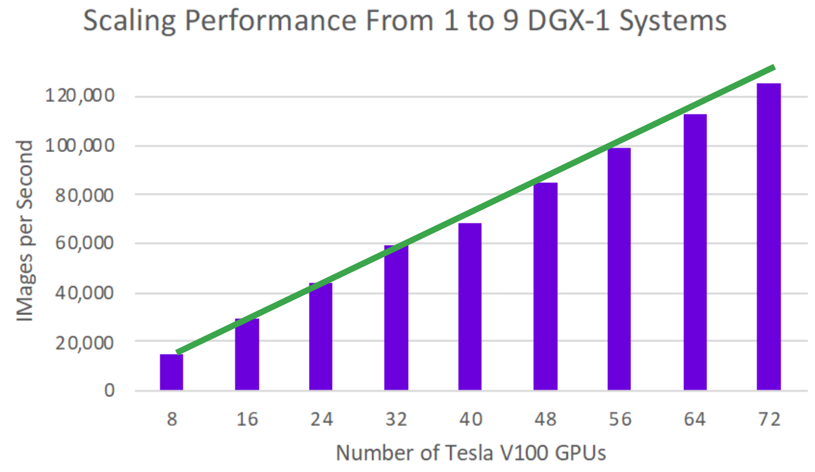
WekaFS PERFORMANCE



HIGH-SPEED NETWORK SATURATION

162GB/Sec & 970K IOPs Performance to a Single GPU client

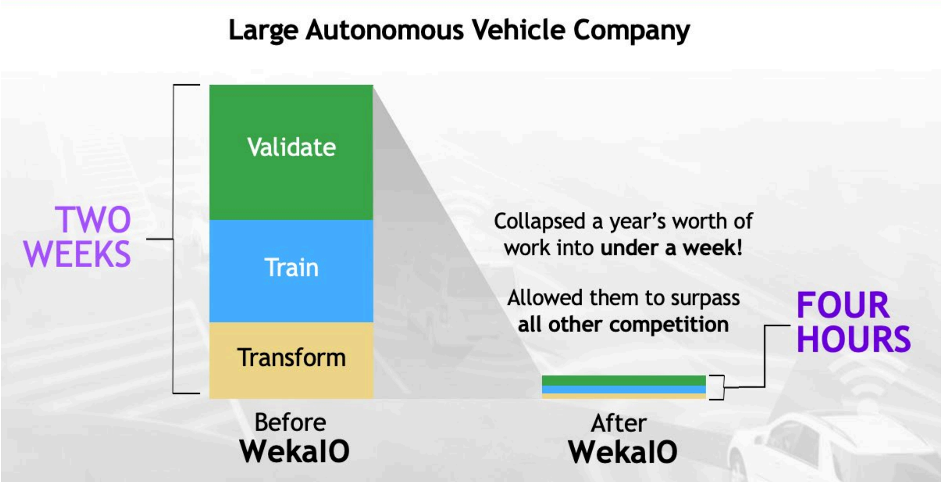
MASSIVE SINGLE CLIENT PERFORMANCE



PERFORMANCE SCALES LINEARLY TO MULTIPLE CLIENTS

AI/ML, DEEP LEARNING DATA PIPELINE FOR AUTONOMOUS DRIVING

- REDUCED DATA EPOCH RUNTIME FROM 2 WEEKS TO 4 HOURS VIA MODERNISATION OF NETWORKING & STORAGE WITH WekaFS

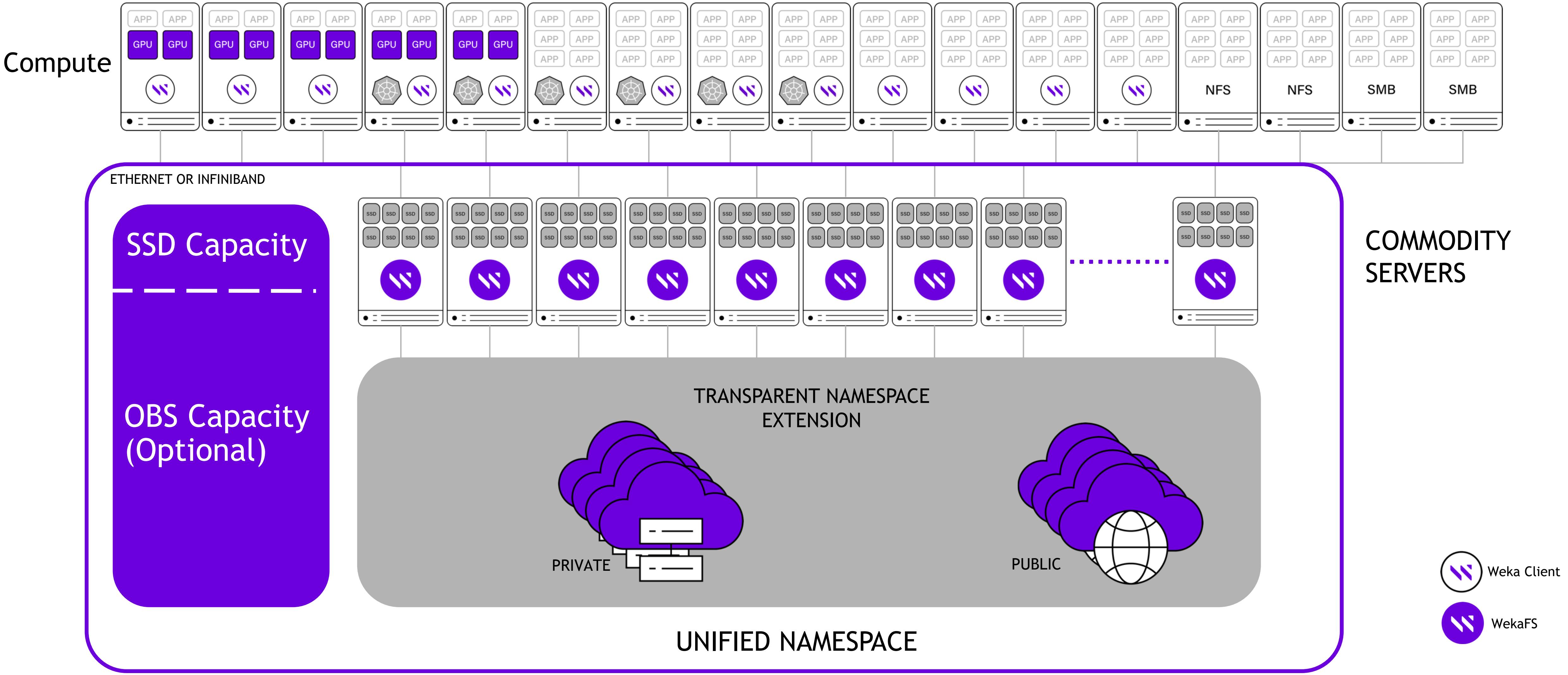


Benchmark	DAS (Optane SSD Server)	SAN (NVMe-oF)	NAS (All-Flash)	WekaFS (HPE NVMe Servers)
100T.YR1VWAB-12D-HO	15633	1886	4183	1028
100T.YR2VWAB-12D-HO	18114	1418	3294	892
100T.YR3VWAB-12D-HO	20730	1910	4773	1141
100T.YR4VWAB-12D-HO	24741	3317	7037	1550
100T.YR5VWAB-12D-HO	36888	22389	11376	4808
10T.YR2-MKTSNAP	176	355	6898	655
10T.YR3-MKTSNAP	176	358	7855	675
10T.YR4-MKTSNAP	149	375	8531	711
10T.YR5-MKTSNAP	155	393	8684	726
1T.2YRHIBID	645	374	1419	309
1T.3YRHIBID	1129	630	2737	480
1T.4YRHIBID	1957	1082	4881	804
1T.5YRHIBID	3234	1804	8589	1234
1T.OLDYRHIBID	61	46	129	48
1T.YR1VWAB-12D-HO	334	226	545	294
1T.YR2VWAB-12D-HO	394	268	632	355
1T.YR3VWAB-12D-HO	462	347	750	430
1T.YR4VWAB-12D-HO	553	517	928	547
1T.YR5VWAB-12D-HO	841	769	1298	732
50T.YR1VWAB-12D-HO	1089	1748	4302	2300
50T.YR2VWAB-12D-HO	1988	1774	4798	1971
50T.YR3VWAB-12D-HO	2865	2278	6253	2409
50T.YR4VWAB-12D-HO	4195	3118	8840	3077
50T.YR5VWAB-12D-HO	6731	4625	13597	4111
Average Result (lower is better)	**5968.33**	**2166.958**	**5097.041**	**1303.625**

FINANCE, TICK DATA ANALYTICS APPLICATION PERFORMANCE

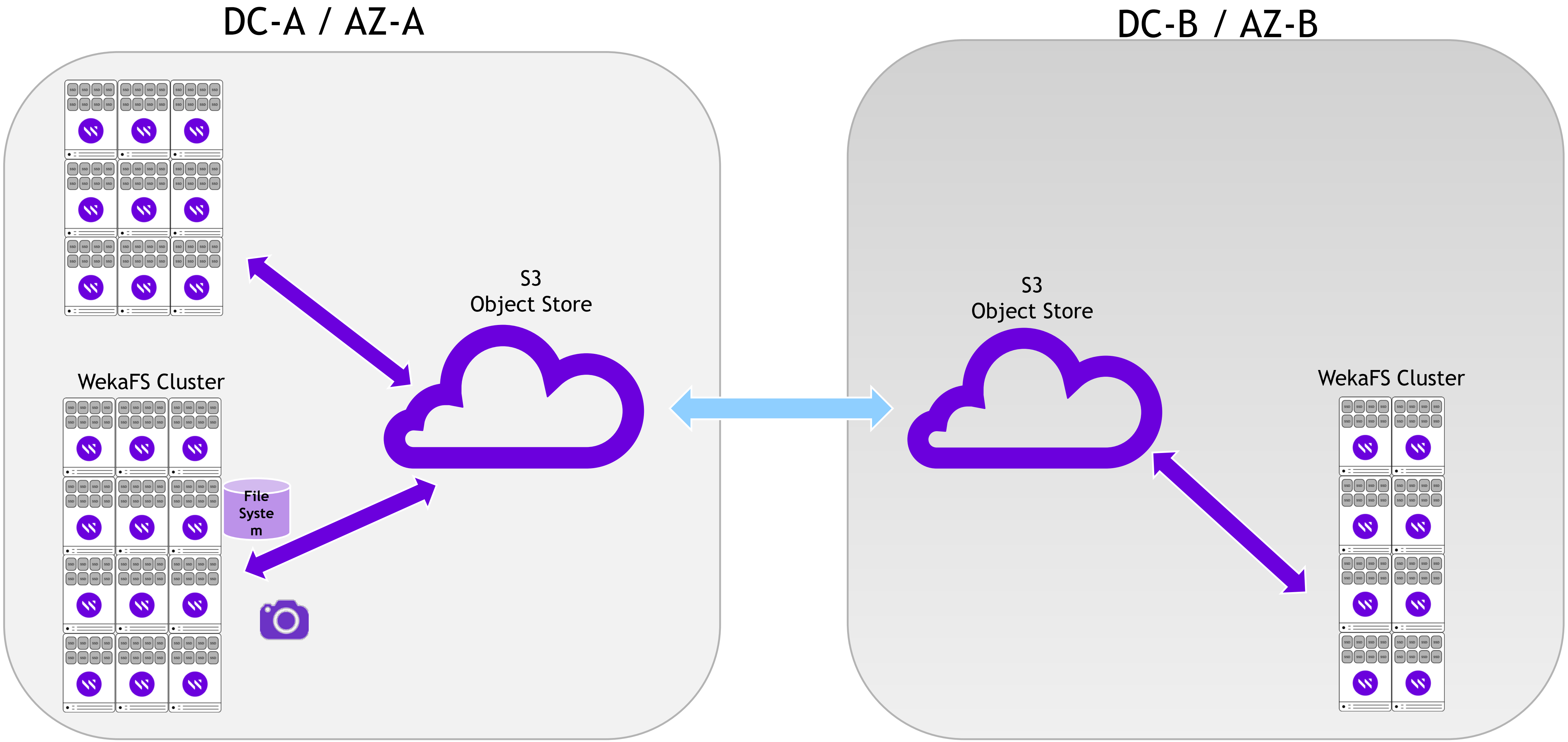
- 4X FASTER THAN ALL-FLASH NAS
- 4.5X FASTER THAN DAS WITH OPTANE
- 1.6X FASTER THAN ALL-FLASH ARRAY

THE WEKA FILE SYSTEM IN A PRODUCTION ENVIRONMENT



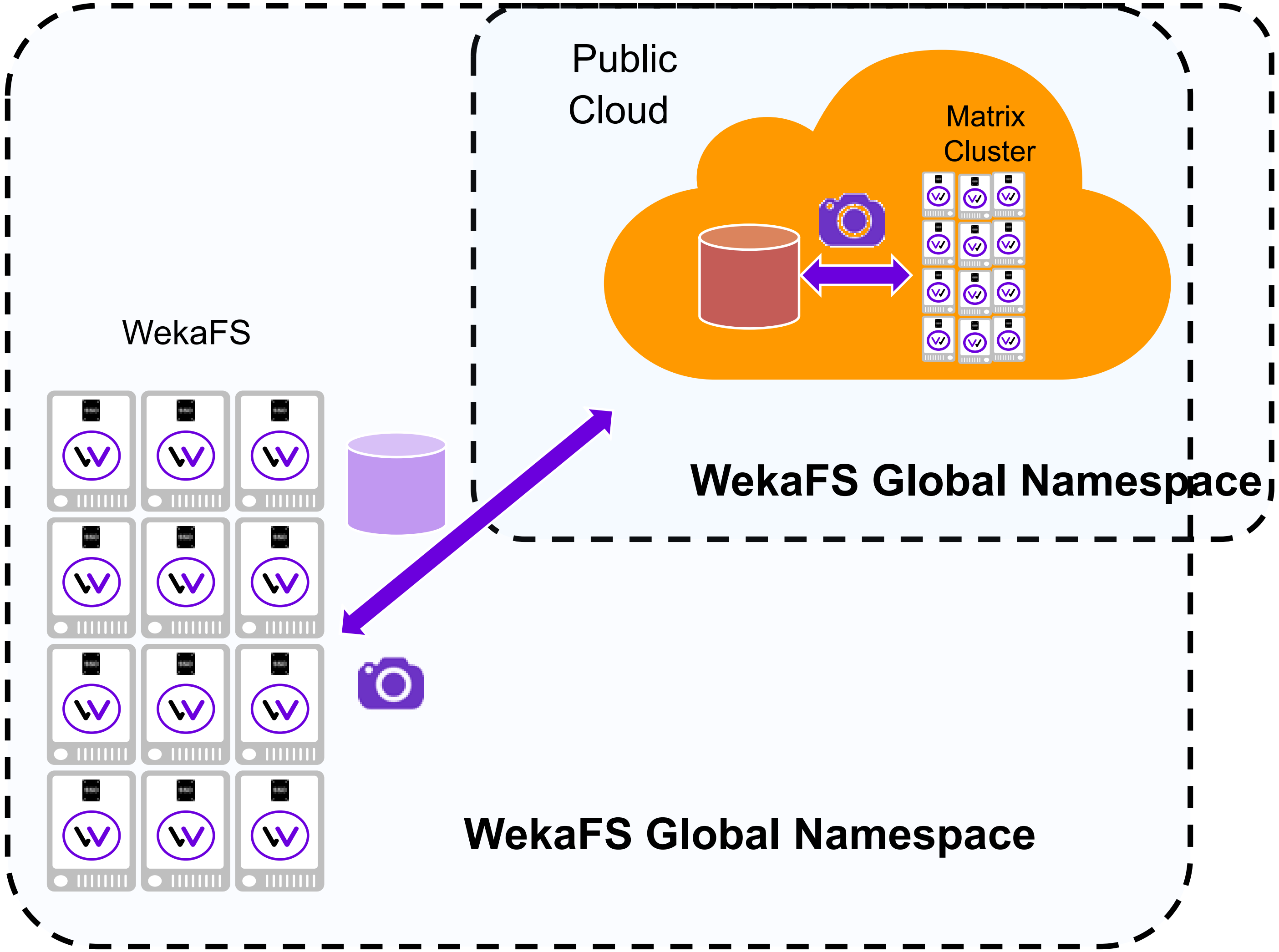
WEKA

Multi Datacenter replication (DR, Backup)



WEKA

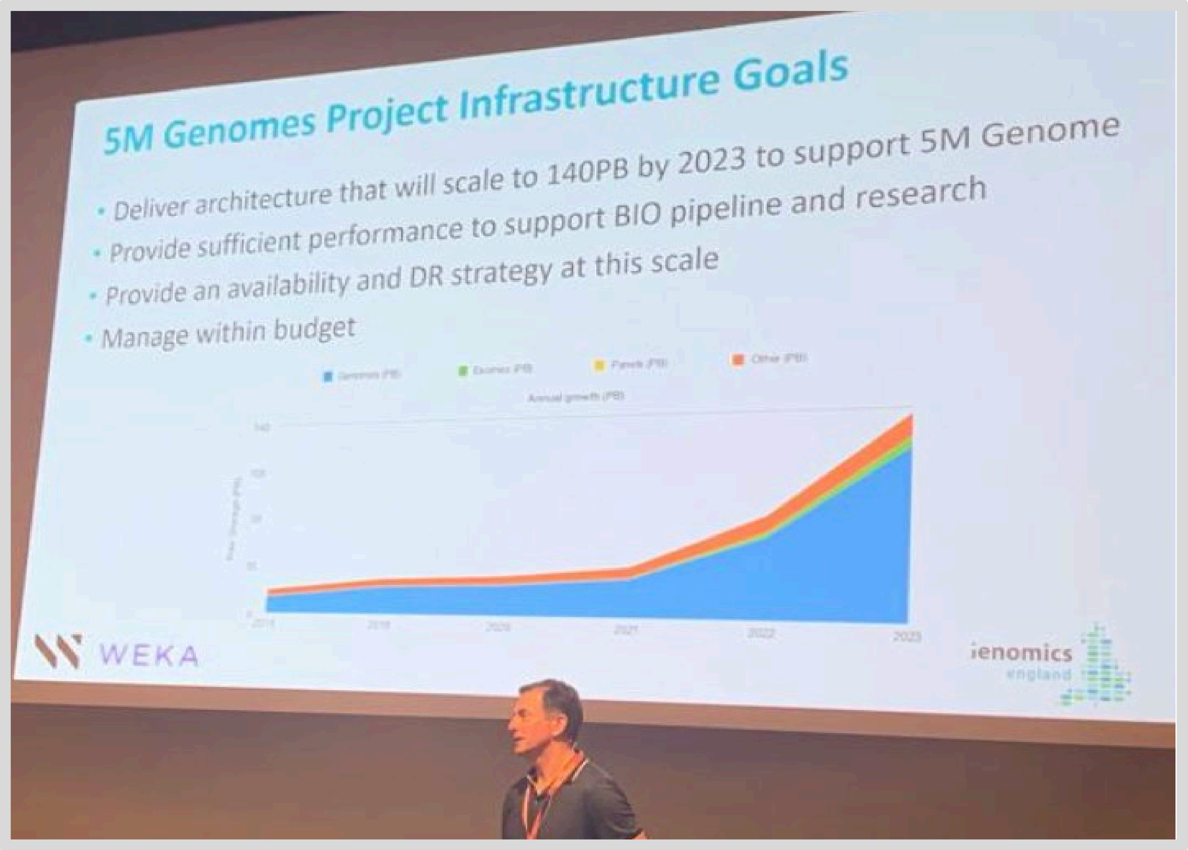
Snapshot-to-S3 for Infrastructure Elasticity & DR

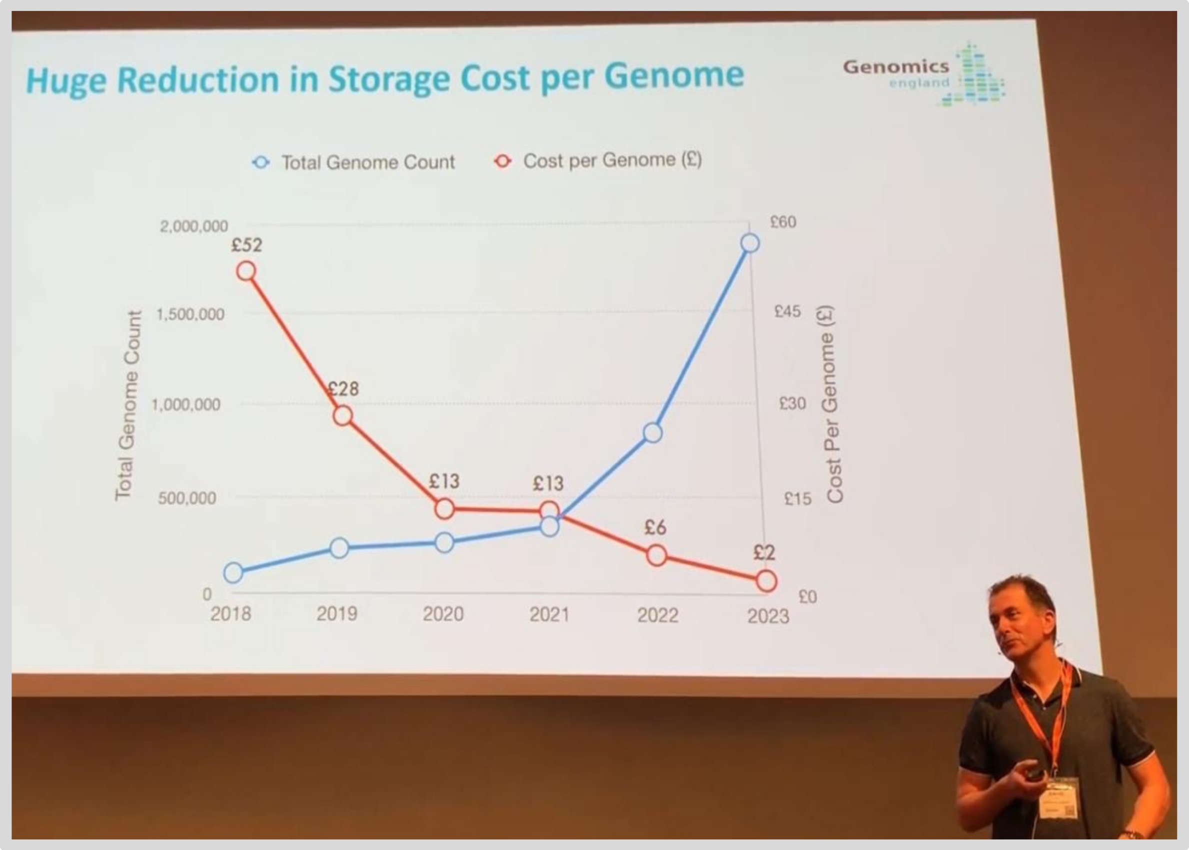


WEKA

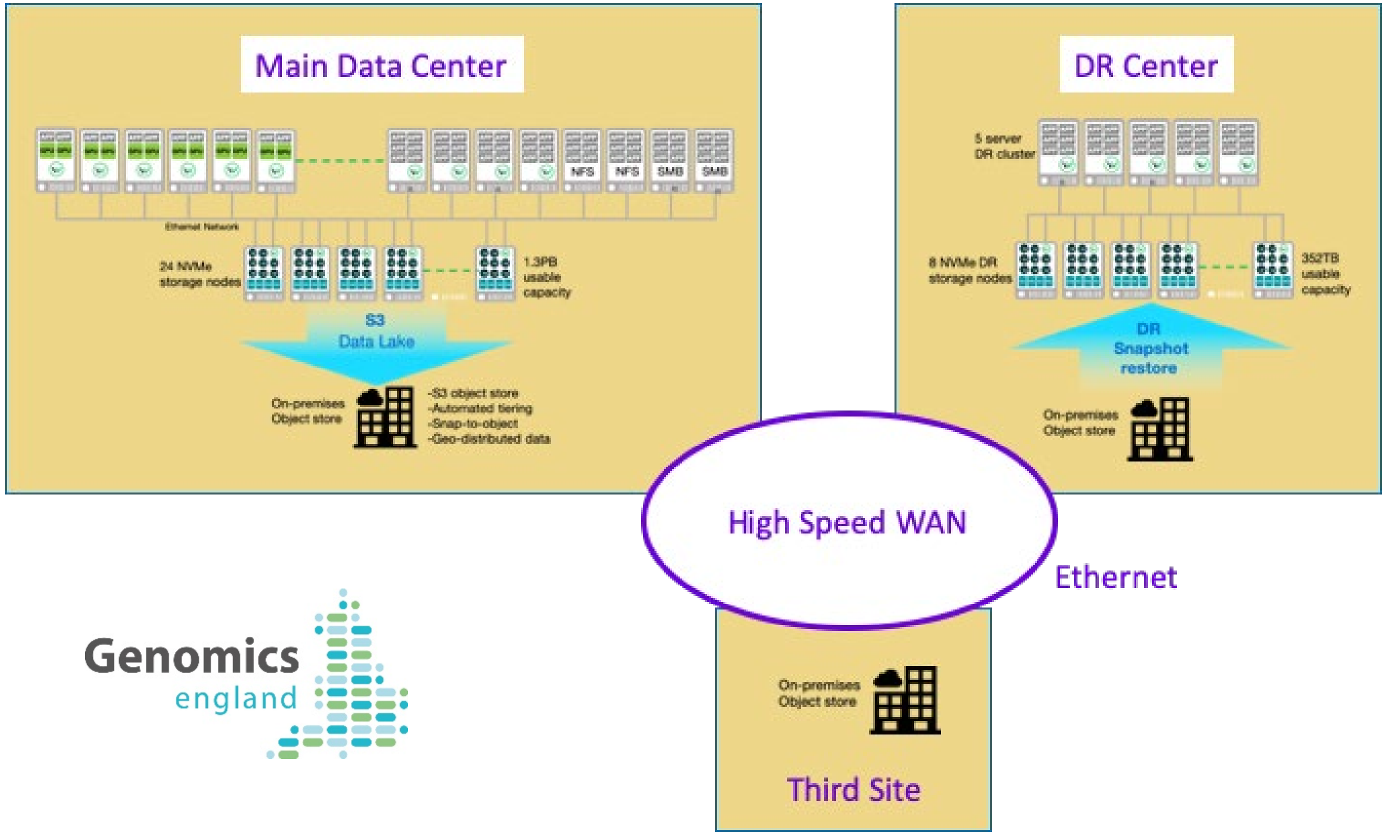
GENOMICS ENGLAND ANNOUNCED WEKA USE IN BIODATA WORLD CONFERENCE KEYNOTE IN BASEL

Presented by David Ardley, Director of infrastructure transformation at Genomics England





WekaFS in Production – Life Science



Genomics england

Weka Solution

- Two Tier Architecture
  - Hot 2.6PB NVMe on 40 nodes
  - Warm – 60PB OBJ geo-distributed
  - Small DR cluster in secondary data centre
- WekaIO integrates the two tiers to provide seamless data movement between the tiers
- Connectivity to the HPC via Mellanox Ethernet

WEKA

RESEARCH

Kioxia

- Matt Hallberg

  Sr. Product Marketing Manager

  www.kioxia.com

The Need for Feed

GPUs are always hungry for more data

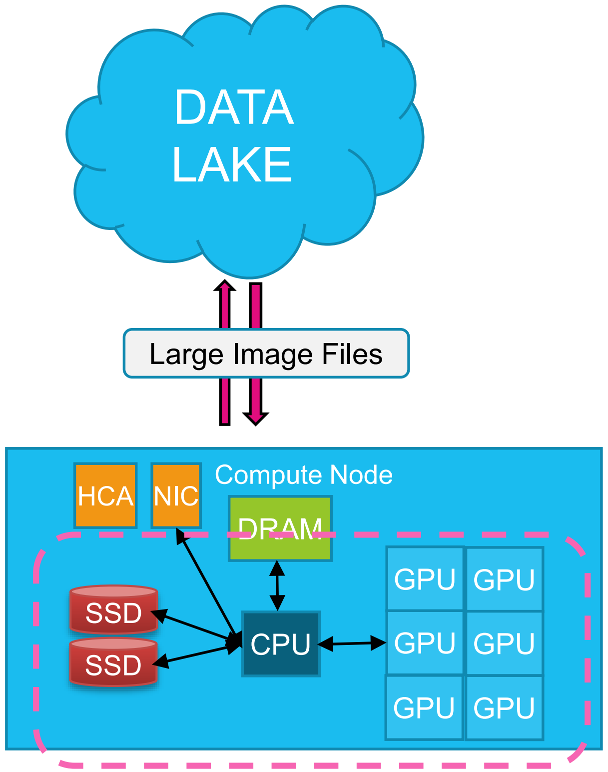
- Previous NVIDIA V100s were processing around 16GB/s and the latest A100 GPUs are multitudes faster!

Starved GPUs are a waste of resources and ultimately, money

- Starvation occurs while the GPU waits for DRAM and Storage to offload / reload / checkpoint training sets

Feed the need with fast local and remote storage, starring PCIe® 4.0 SSDs, NICs and NVMe-oF™ technology!

Why Local Storage Matters



- The training phase of machine learning is the most-resource intensive set of operations
  - Datasets are growing at a fast pace, MRIs can reach up to TBs, and training sets can be composed of thousands of images
  - Whether you are running on RAM or on local storage, the local storage needs to be able to handle reads and writes in blazing fashion and with little impact to overall latency
  - Moving the data in, moving the data out, and checkpointing all need to be completed quickly to minimize the idleness of the GPUs
- PCIe® 4.0 SSDs' noticeable benefits with file copying and other I/O tasks versus PCIe 3.0 SSDs
  - For sequential workloads
    - Up to 7000MB/s on Reads
    - Up to 4200MB/s on Writes
  - For random workloads
    - 1M+ IOPS on Random Reads
    - 70K+ on Random Writes
  - PCIe 4.0 SSDs are also able to take advantage of 3D NAND's higher densities, allowing for up to 30TB in a 2.5" SSD

PCIe is a registered trademark of PCI-SIG.

Storage Use Case: High Performance Nano-Scale Brain 3D Visualization

High performance SSDs enable interactive 3D and 8K visualization of brain images

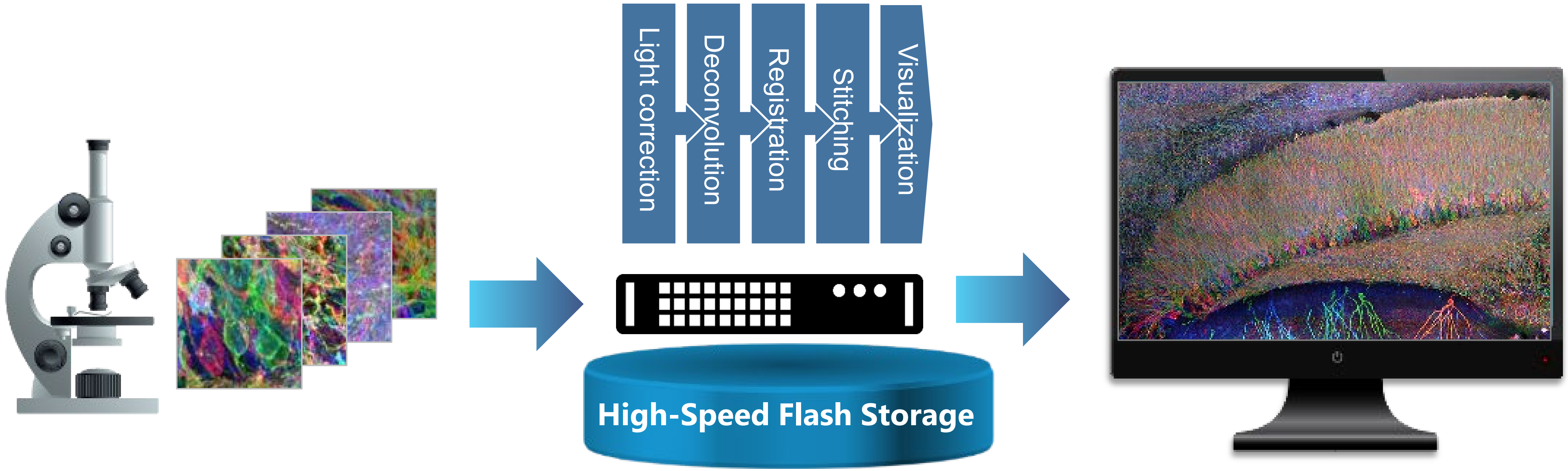


Image capture using Expansion Microscopy

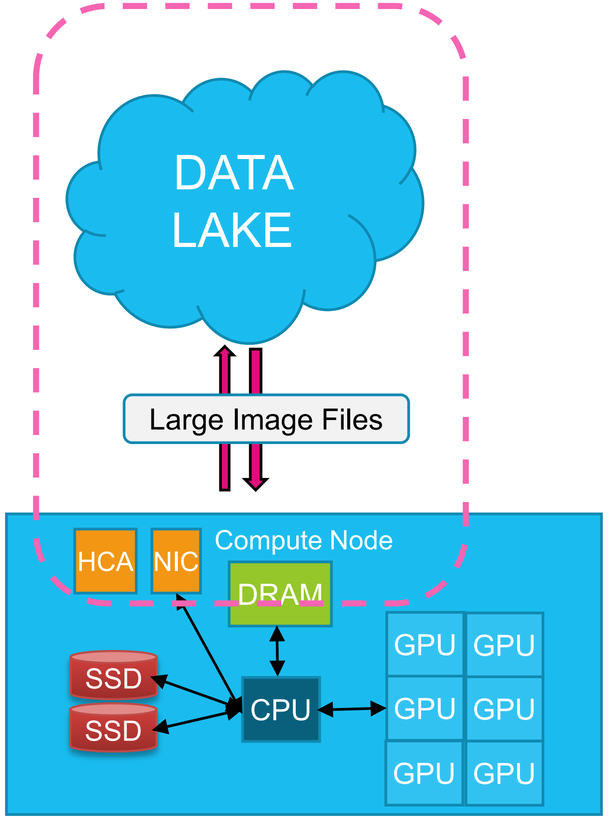
Flash-based high-speed data processing system

Real-time 3D exploration of whole brain images

*** Microscopy images courtesy of Shoh Asano, Edward Boyden, MIT Media Lab**

Additional Information: https://www.media.mit.edu/articles/mapping-the-brain-at-high-resolution/

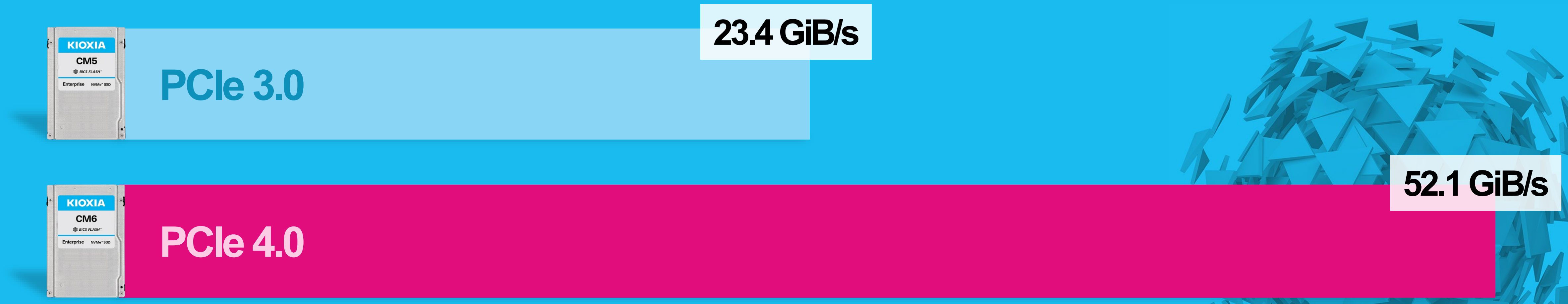
Why Remote Storage Matters



- Networking speeds, technologies, and topologies have greatly advanced over the past few years
  - 200GbE NICs
  - RoCEv2 / RMDA over Ethernet
  - NVMe® over Fabrics (NVMe-oF™) deployments
  - Etc.
- Data sets are comprised of thousands of files all of which need to be sent and received to and from the compute node to minimize downtime
- The storage behind the NIC(s) should be optimized for sequential performance to send data for processing and receive processed data. The faster the offload can occur, the faster the local storage can send the data to the GPUs
  - NVMe-oF deployments show their strength here by improving performance and reducing latency
  - Using NVMe SSDs for staging "warm" data vs cold data (cheaper storage) optimizes the spend on remote storage
  - GPUDirect® technology can also take advantage of the remote NVMe SSDs

NVMe is a registered trademark and NVMe-oF is a trademark of NVM Express, Inc. GPUDirect is a registered trademark of NVIDIA Corporation in the U.S. and/or other countries.

QCT / AMD / Broadcom / KIOXIA PCIe® 4.0 SSD Demo @ Microsoft Ignite 2019



Benchmark: Storage I/O for Virtualized App

- 100% sequential 129KiB reads
- More than double the performance over PCIe 3.0 configuration
- Each drive delivers over 7 GiB/s!

Microsoft Windows Server 2019, Azure Stack HCI 2 node cluster

- Build 177763.775 Data Center Edition
- Hyper-V and Storage Spaces Direct
- Two-way mirror volumes
- VM Fleet with diskspd 2.0.21a
- FIO

QCT D43K-1U Server(2)

- 2 x AMD Epyc 7742
- 512GiB DRAM
- 4x KIOXIA CM5 PCIe 3.0, 7.68T SSD
- 4x KIOXIA CM6 PCIe 4.0, 7.68T SSD
- Broadcom NetXtreme™-P2100G 200Gbps NIC

PCIe is a registered trademark of PCI-SIG. Microsoft Ignite, Azure, and Hyper-V are registered trademarks of Microsoft Corporation. NetXtreme is a trademark of Broadcom Corporation. All other company names, product names and service names may be trademarks of their respective companies.

There's more to come: Ethernet Bunch of Flash (EBOF) over NVMe-oF™

- There is a new type of storage appliance on the horizon: Ethernet Bunch of Flash (EBOF)
  - Utilizes the already proven performance gains and latency reductions of RoCEv2 / RDMA and NVMe-oF
  - Ethernet SSDs are attached to the network, offering up to 2 ports of 25Gbps Ethernet and RoCEv2 RDMA connections
  - No need for CPU or DRAM, allowing for direct access to storage
    - Avoids the "bounce buffer" of data needing to cycle through CPU and DRAM during offload to storage
  - Highly reduced system costs and cooling benefits
    - No CPU, DRAM, NIC, HBAs, etc.
    - Only cooling SSDs and components… 24 SSDs @ 18W = 432W!
  - EBOF could work with SDS for GPUDirect® / Magnum IO™ technology to allows direct GPU access to NIC(s)

- Current proof of concept EBOF system highlights
  - 2U Chassis with up to 24 drives
    - Can route for high availability (single drive going to 2 switches)
  - 600Gb/s storage throughput per 100Gb/s embedded network switch
  - High performance: 830K IOPS per drive, close to 20M IOPS per 24 bay system
  - Press Release: https://business.kioxia.com/en-us/news/2020/ssd-20200922-2.html

KIOXIA CM6 Series Enterprise NVMe SSDs

- Enterprise PCIe® 4.0, NVMe™ 1.4 SSDs
- Form factors: 2.5-inch, 15mm Z-height
- Proprietary KIOXIA architecture: controller, firmware and BiCS FLASH™ 96-layer 3D TLC memory
- SFF-TA-1001 conformant (U.3) works with Tri-mode controllers and backplanes
- Dual-port design for high availability applications
- 6th generation die failure recovery and double parity protection
- High performance with lower power consumption
- Power loss protection (PLP) and end-to-end data protection
- Suited for 24x7 enterprise workloads
- Data security options: SIE, SED, FIPS 140-2
- Six power mode settings
- Available now

			CM6 (Mixed-Use)					CM6 (Read-Intensive)					
Endurance		DWPD	3					1					
User Capacity*		GB	800	1600	3200	6400	12800	960	1920	3840	7680	15360	30720
Sequential Read	128KB(QD32)	MB/s	6900	6900	6900	6900	6900	6900	6900	6900	6900	6900	6850
Sequential Write	128KB(QD32)	MB/s	1400	2800	4200	4000	4000	1400	2800	4200	4000	4000	4000
Random Read	4KB(QD256)	KIOPS	800	1300	1400	1400	1400	800	1200	1400	1300	1400	900
Random Write	4KB(QD32)	KIOPS	100	215	350	325	330	50	100	170	170	170	70

* KIOXIA Corporation definition of capacity: 1 GB = 1,000,000,000 (10^9) bytes (see end of presentation for full capacity disclaimer).

Note: Specifications are subject to change

PCIe is a registered trademark of PCI-SIG. NVMe is a trademark of NVM Express, Inc.

KIOXIA CD6 Series Data Center NVMe SSDs

- Data Center PCIe® 4.0, NVMe™ 1.4 SSDs
- Form factors: 2.5-inch, 15mm Z-height
- Proprietary KIOXIA architecture: controller, firmware and BiCS FLASH™ 96-layer 3D TLC memory
- SFF-TA-1001 conformant (U.3) works with Tri-mode controllers and backplanes
- Single-port design, optimized for data center class workloads
- 6th generation die failure recovery and double parity protection
- Consistent performance and reliability in demanding 24x7 environments
- Designed for high density storage deployments
- Power loss protection (PLP) and end-to-end data correction
- Data security options: SIE, SED, FIPS 140-2
- Five power mode settings
- Available Now

			CD6 (Mixed-Use)					CD6 (Read-Intensive)				
Endurance		DWPD	3					1				
User Capacity*		GB	800	1600	3200	6400	12800	960	1920	3840	7680	15360
Sequential Read	128KB(QD32)	MB/s	5800	5800	6200	6200	5500	5800	5800	6200	6200	5500
Sequential Write	128KB(QD32)	MB/s	1300	1150	2350	4000	4000	1300	1150	2350	4000	4000
Random Read	4KB(QD256)	KIOPS	700	700	1000	1000	750	700	700	1000	1000	750
Random Write	4KB(QD32)	KIOPS	90	85	160	250	110	30	30	60	85	30

* KIOXIA Corporation definition of capacity: 1 GB = 1,000,000,000 (10^9) bytes (see end of presentation for full capacity disclaimer).

Note: Specifications are subject to change

Definition of capacity: KIOXIA defines a megabyte (MB) as 1,000,000 bytes, a gigabyte (GB) as 1,000,000,000 bytes and a terabyte (TB) as 1,000,000,000,000 bytes. A computer operating system, however, reports storage capacity using powers of 2 for the definition of 1GB = 2^30 = 1,073,741,824 bytes and therefore shows less storage capacity. Available storage capacity (including examples of various media files) will vary based on file size, formatting, settings, software and operating system, such as Microsoft Operating System and/or pre-installed software applications, or media content. Actual formatted capacity may vary.

Images are for illustration purposes only.

Datyra

- Keith Klarer
  Chief Executive Officer
  www.datyra.com

Managing Scale in Machine Learning

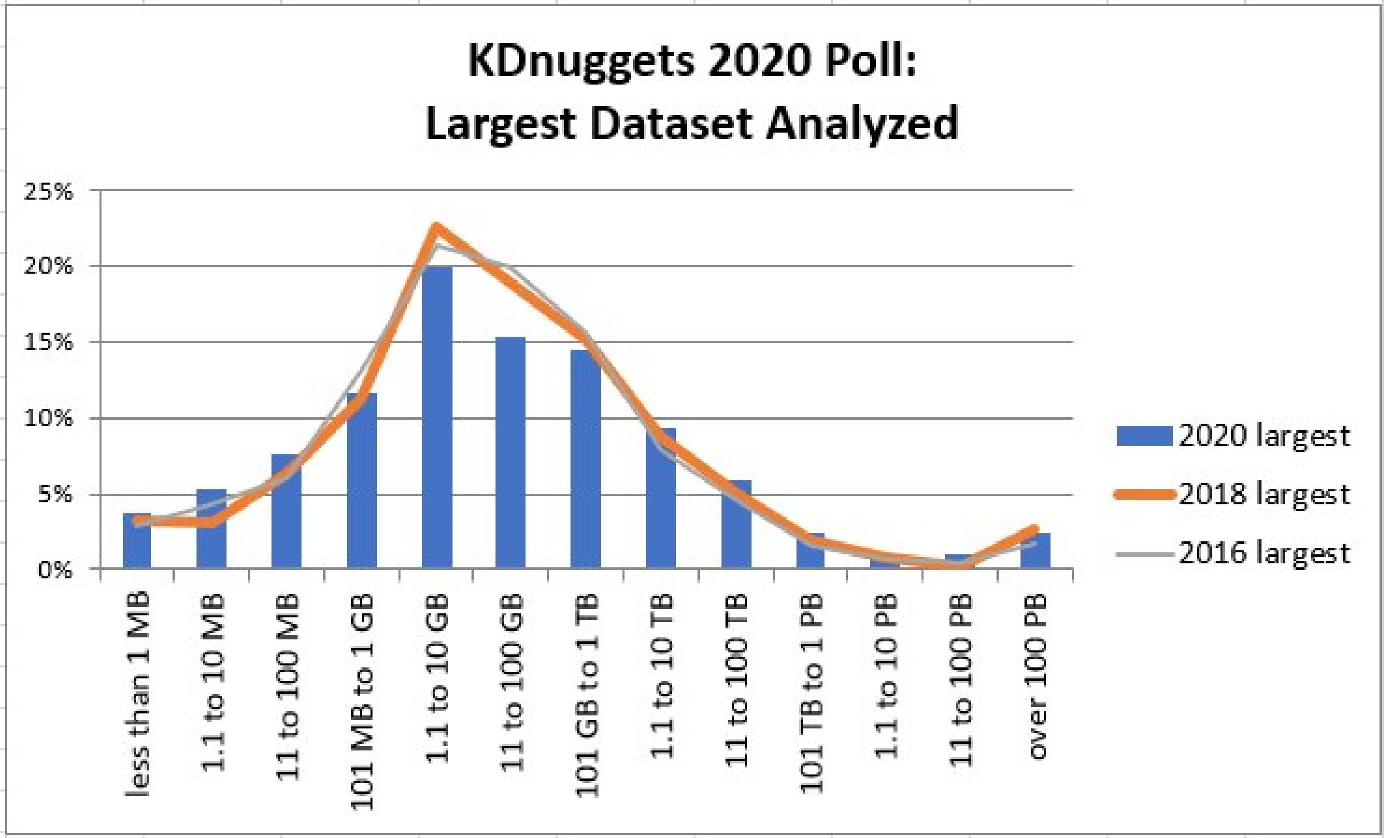
DATYRA

- There are many challenges when managing large scale data in machine learning projects
- The most apparent are related to the size and velocity of the data
- Less apparent are those related to cataloging, provenance, integrity and security
- Strict requirements surrounding medical data adds another layer of complexity

Managing Scale in Machine Learning

- What to be aware of when starting a project
- Planning your data organization
- An example infrastructure that can help when managing critical data at scale

Data Size Surveys



KDnuggets Poll
https://www.kdnuggets.com/2020/07/poll-largest-dataset-analyzed-results.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

Evaluate Your Dataset

- Many organizations have large information silos that must be integrated into the ML system
- Invest time to understand this information content and devise strategies to Extract, Transform and Load (ETL) the data
- Create an inventory list of the original and transformed data: counts, sizes, formats, ingestion schedules and data rates
- ETL development can dwarf actual ML development efforts
  - Strong planning is a important component of project success
  - Understand ingestion rates, error sources, cleaning effort

Data Organization Plan

- Describes how to manage the data generated in ML pipeline:
  - Data (structured and unstructured) from ETL
  - Models and data from ML training
    - Potentially 100’s of models can be active
  - Data from ML inference
  - Metadata from all the pipeline stages (ETL, MLT, MLT)

Data Organization Plan

- Decide on the technology used to store each data element:
  - Database (relational, graph, etc.), object store, file, etc.
- Understand regulatory compliance methods: GDPR guidelines, HIPA, corporate governance and country rules
  - Transparency, purpose limitation, data minimization
  - Data accuracy, integrity and confidentiality
- Identify the physical storage locations

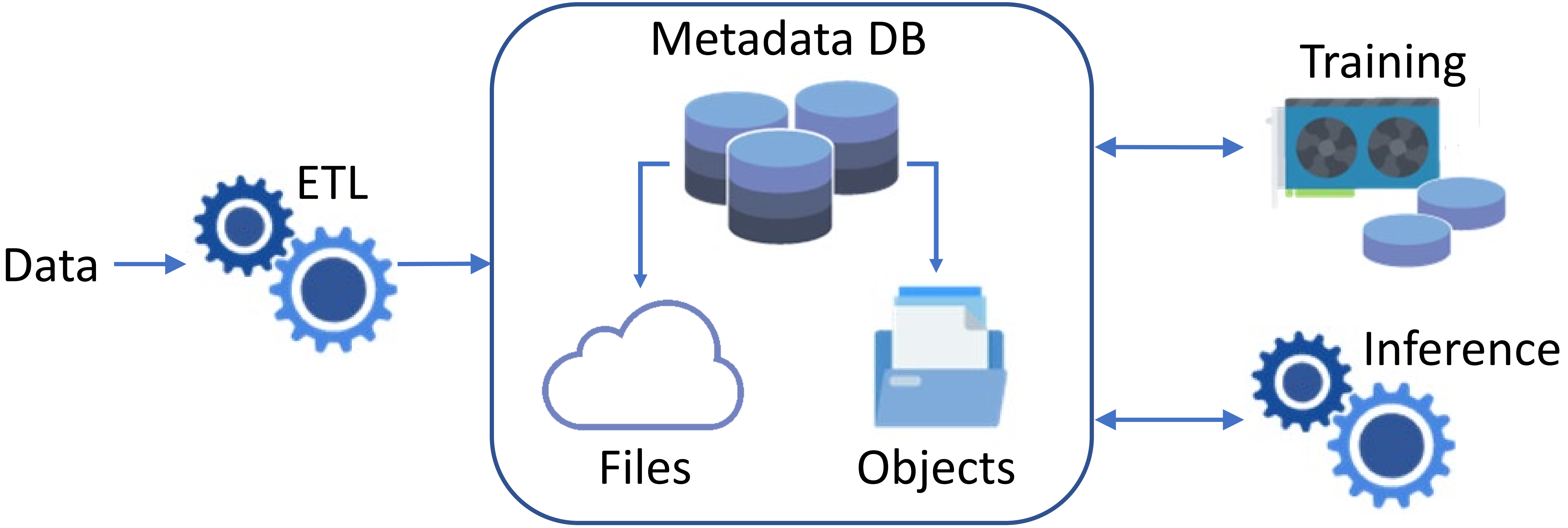
Metadata

- Metadata can be created at every stage of the pipeline:
  - ETL: data source, collection timestamp, transformation steps, integrity guarantees
  - MLT: data sets used to create the model, model parameters, version information
  - MLI: usage metrics, performance, outcomes

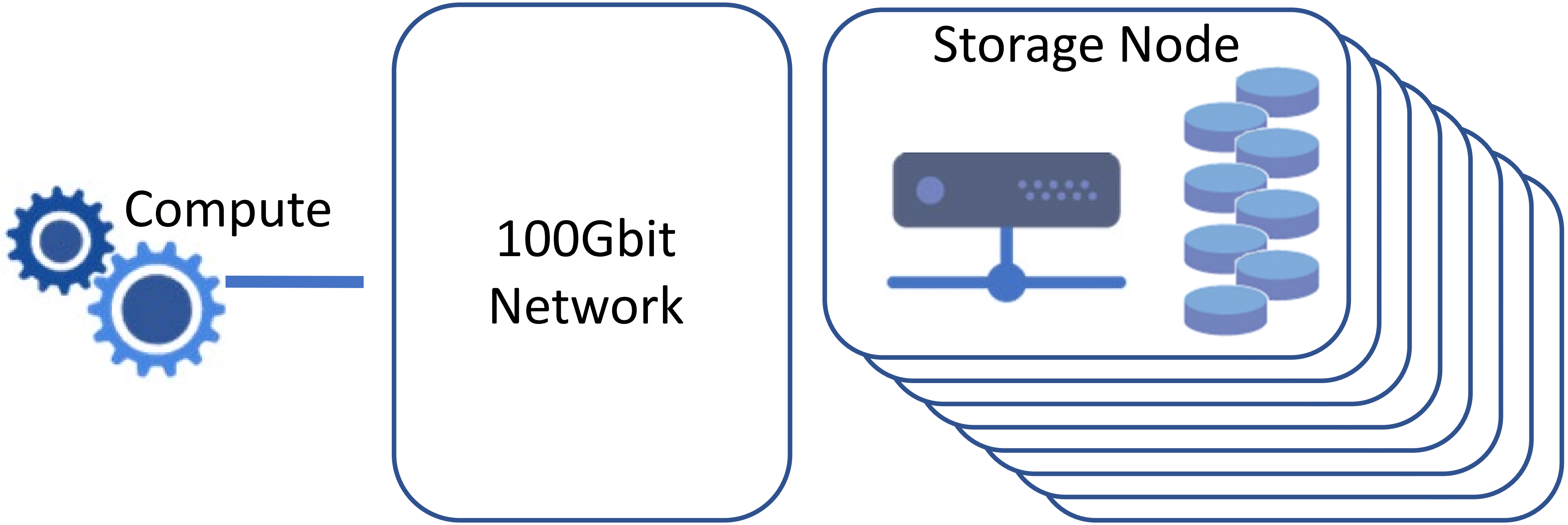
Metadata

- Insufficient or incorrect metadata:
  - Reduces the usability and potential value of the data
  - Hinders model creation/correctness/maintainability
  - Risks the accuracy and repeatability of the inference process
  - Puts in doubt the correctness of the ML pipeline results

Scaling Infrastructure



Performance Example: Minio Object Store



- 8 NVME SSDs in each storage node
- GET throughput per node: 5.7GB/s
- PUT throughput per node: 5.3GB/s
- 32 nodes: 183.2 GB/s (1.46 Tbit/s) and 171.3 GB/s (1.37 Tbit/s)

MinIO S3 Throughput Benchmark on NVMe SSD
December 2019

Thank You!

- For more information https://datyra.com/publications/
- Or contact me at mailto:k@datyra.com

Panel Questions and Audience Surveys

Panel Question # 1

- What are the biggest bottlenecks to building and deploying large-scale AI/ML solutions for healthcare?
  - NVIDIA
  - Weka
  - Kioxia
  - Datyra

Audience Survey Question #1

- How large are your organization's AI/ML training data sets? (check one):
  - Greater than 5 PB: 0%
  - Between 1PB and 5PB: 7%
  - Between 250TB and 1PB: 13%
  - Between 50TB and 250TB: 13%
  - Less than 50TB: 67%

G2M RESEARCH

Panel Question #2

- What are some best practices to efficiently move, store, and manage the large amounts of data required for the training of AI/ML-based medical imaging solutions?
  - Weka
  - Kioxia
  - Datyra
  - NVIDIA

G2M RESEARCH

Audience Survey Question #2

- What are your greatest concerns when building very large AI/ML training data sets? (check one):
  - The amount of time it will take to run the training data through the model: 39%
  - The cost of the hardware required to run the training model: 33%
  - Managing the various training and verification datasets: 6%
  - Managing and archiving the results of training runs: 11%
  - Other issues: 0%
  - No opinion: 11%

G2M RESEARCH

Panel Question # 3

- How do new technologies such as GPUDirect®, NVMe™, and NVMe-oF™ help to accelerate both the deployment and performance of AI/ML solutions for healthcare?
  - Kioxia
  - Datyra
  - NVIDIA
  - Weka

G2M RESEARCH

Audience Q&A

G2M RESEARCH

Thank You For Attending

G2M

RESEARCH