G2M RESEARCH

Storage Architectures, AI, and Biotech

Multi-Vendor Webinar

Tuesday March 23, 2021

Webinar Agenda

How Did Storage Architectures Perform for Biotech AI Modeling & What Can We Learn From This?

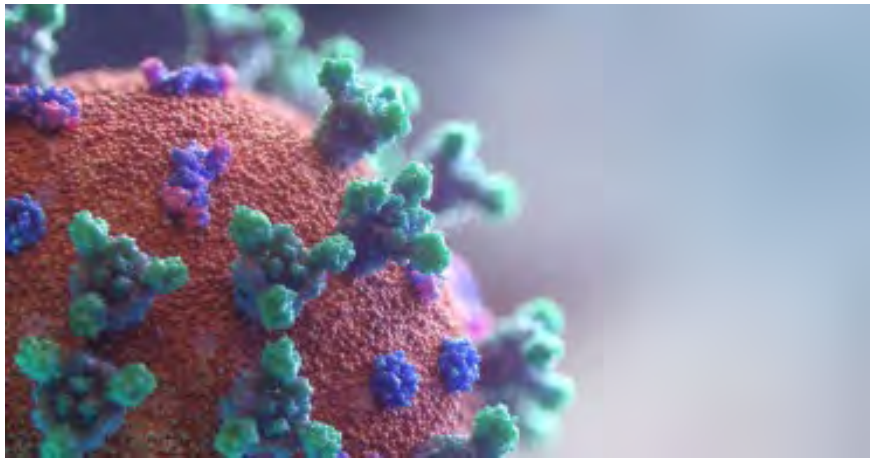
9:00-9:05	Ground Rules and Webinar Topic Introduction (G2M Research)
9:06-9:37	Sponsoring Vendor presentations on topic (8 minute each)
9:38-9:44	Panel Discussion Question #1
9:45-9:46	Audience Survey #1
9:46-9:52	Panel Discussion Question #2
9:53-9:54	Audience Survey #2
9:54-10:00	Panel Discussion Question #3
10:01-10:08	Audience Q&A (8 minutes)
10:09-10:10	Wrap-Up

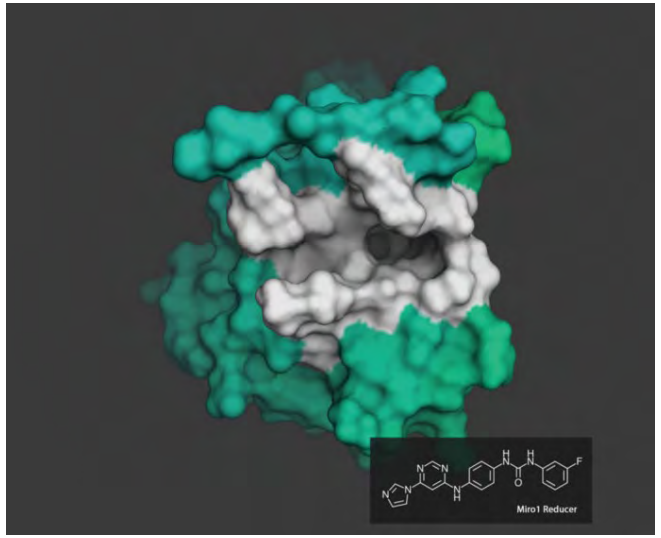
G2M Research Introduction and Ground Rules

Mike Heumann
(Managing Partner, G2M Research)

Artificial Intelligence and Biotech

- Artificial intelligence (AI) has demonstrated its value in a number of biotech areas
  - COVID screening, gene analysis
  - Understanding protein binding
  - Disease identification and diagnostics
- AI has enabled biotech companies to speed up drug analysis, design and screening
  - Screening potential drugs against thousands of molecules
  - Assessing impact of multi-drug treatments for complex diseases
  - Development of biomarkers





Storage Architectures for Biotech AI

- Most storage vendors are now optimizing their architectures for AI workloads
- Both cloud and on-prem solutions are now available for AI workloads
- These systems must provide large, scalable storage with high performance
- These storage architectures must also be able to provide data management to store training data sets and training results

Panelists

NetApp

Esteban Rubens
Principal, Healthcare AI Practice
www.netapp.com

NGD systems

Scott Shadley
VP of Marketing
www.ngdsystems.com

WEKA

Greg Mazzu
Sr. Systems Engineer
www.weka.io

PANASAS

Adam Marko
Director, Life Sciences
www.panasas.com

Andrew Bartko
Executive Director,
CMI/UCSD

NetApp

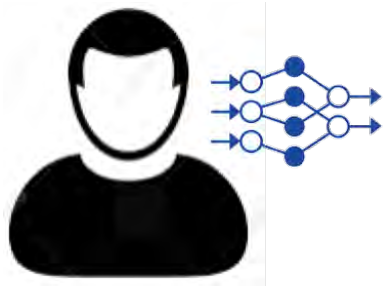
Esteban Rubens

Principal, Healthcare AI Practice

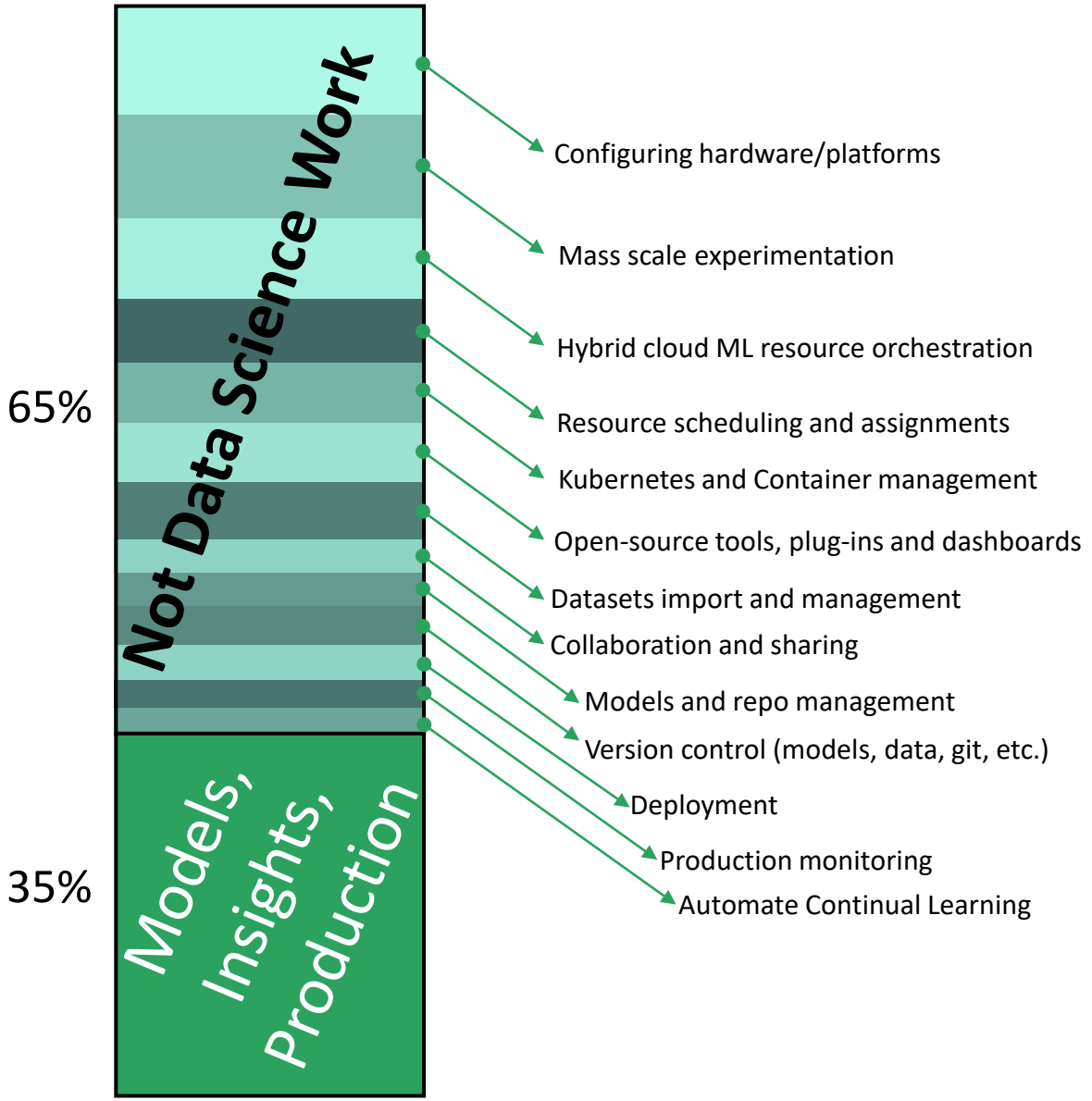
www.netapp.com

Industry pain point

Data Scientists are doing very little Data Science



Data Scientist
Work Distribution



50%
OPEX WASTE

2X
SLOWER
DEVELOPMENT TIME

30%
LOST REVENUE

Bring a cloud consumption model to data science

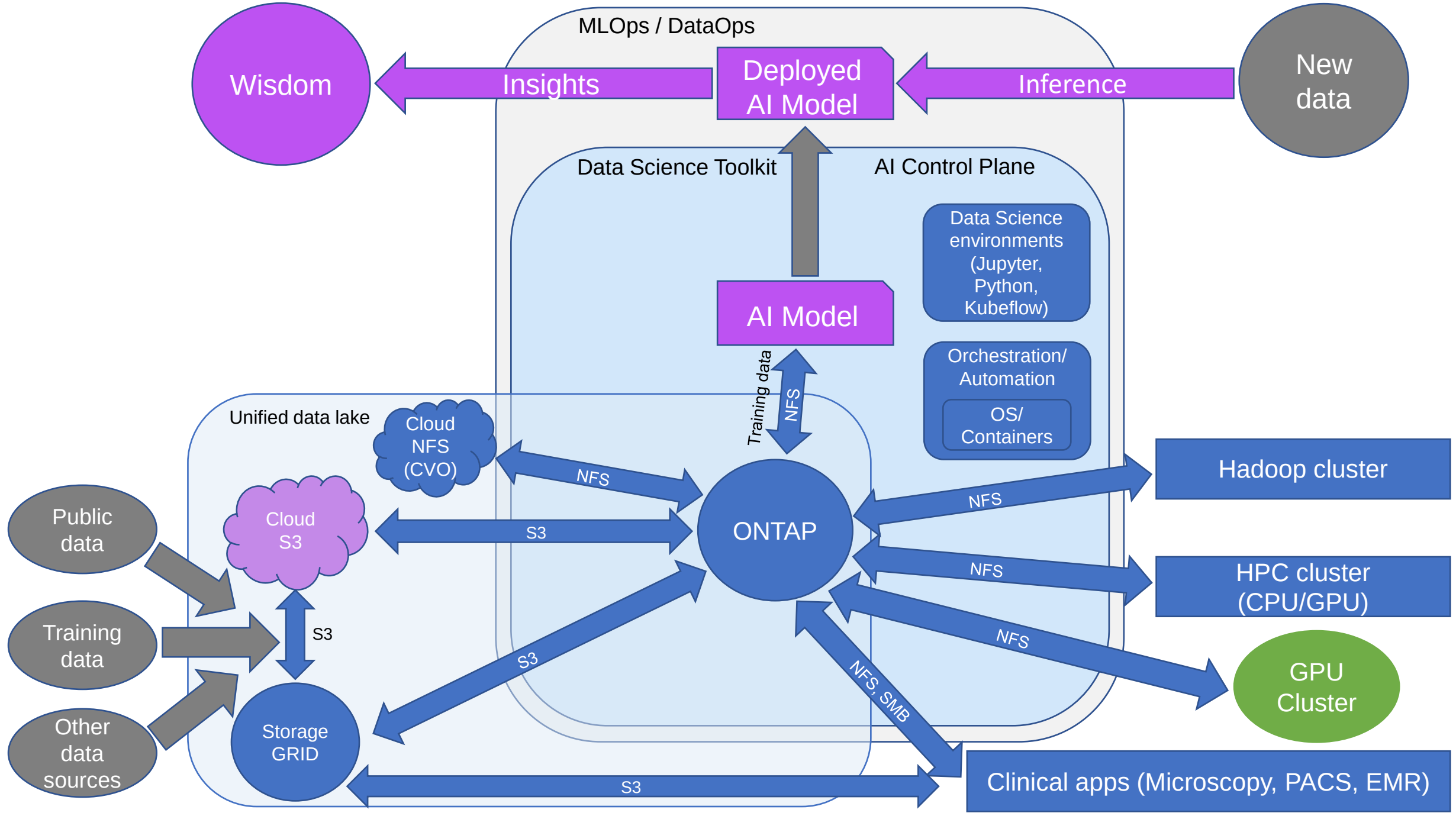
Data Science as a Service benefits

- No need to reinvent the wheel: a turnkey solution for data science
- Minimize the time data scientists and engineers spend wrangling data
- Make tools available to data scientists while reducing the burden on IT
- Maximize resource utilization
- Avoid shadow AI
- Improve training data availability with caching

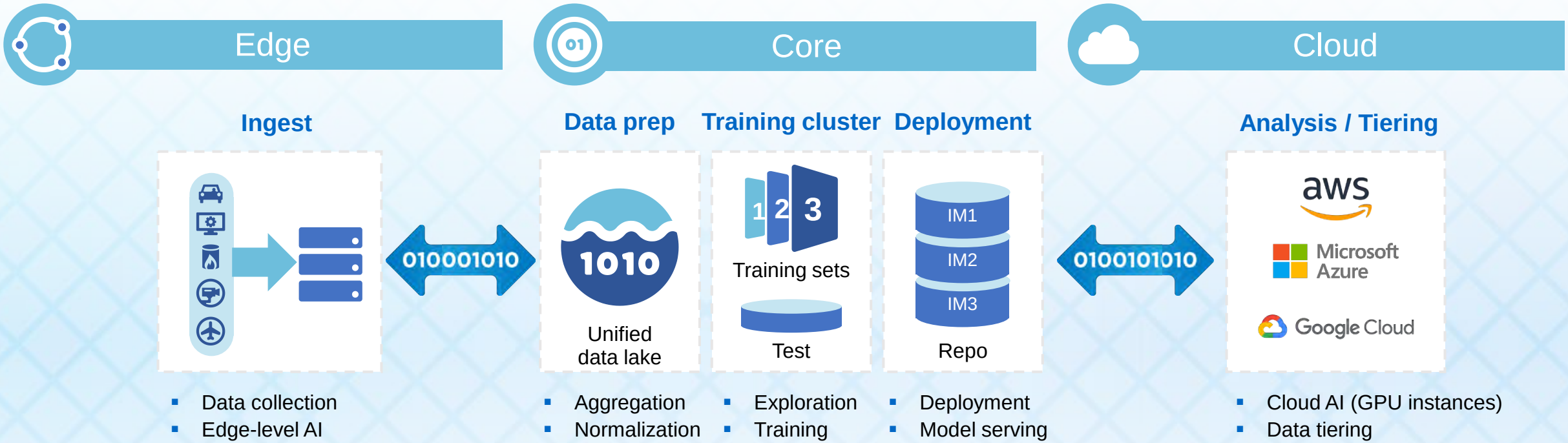
Data Science as a Service competitive differentiators

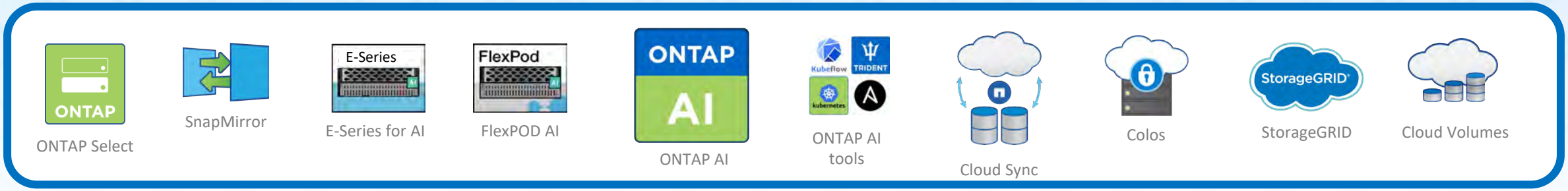
- Tools for data scientists built by data scientists
  - NetApp AI Control Plane: a full-stack AI data and experiment management solution
  - NetApp Data Science Toolkit: a Python program that makes it simple for data scientists and data engineers to perform advanced data management tasks
- Simplify results reproducibility with the Machine Learning Version Control framework
- Data Fabric: seamlessly move data
  - To and from clouds
  - Between different clouds
  - Between ONTAP systems (edge, core, and cloud)
  - ONTAP on AFF for high-performance model training paired with an S3 data lake
- Cloud cost optimization

Data Fabric: The right data in the right place at the right time, seamlessly


MLOps / DataOps
Wisdom
Insights
Deployed AI Model
Inference
New data
Data Science Toolkit
AI Control Plane
Data Science environments (Jupyter, Python, Kubeflow)
AI Model
Orchestration/ Automation
OS/ Containers
Training data
NFS
Unified data lake
Cloud NFS (CVO)
NFS
Hadoop cluster
Public data
Cloud S3
S3
ONTAP
NFS
HPC cluster (CPU/GPU)
Training data
S3
NFS
S3
NFS, SMB
GPU Cluster
Other data sources
Storage GRID
S3
Clinical apps (Microscopy, PACS, EMR)

NetApp's portfolio spans the data pipeline





PANASAS®

Andrew Bartko
Executive Director, CMI/UCSD

Adam Marko
Director of Life Sciences Panasas

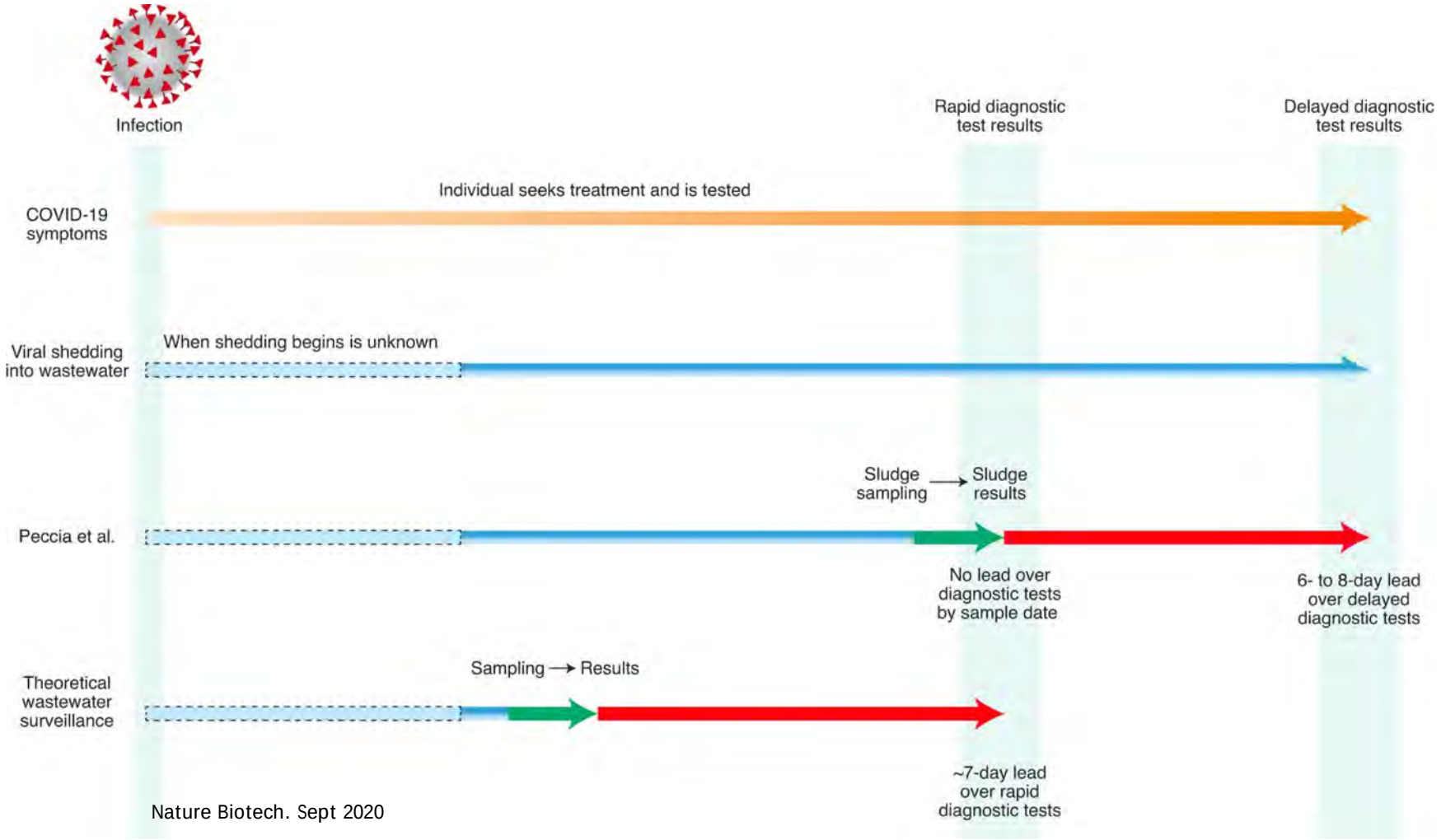
www.panasas.com

UC San Diego | CENTER FOR microbiome INNOVATION

Combating COVID-19 : Reducing SARS-CoV-2 transmission with early detection through wastewater monitoring

Dr. Andrew Bartko
Executive Director of the Center for Microbiome Innovation (CMI)
Professor of Practice – Bioengineering
UC San Diego

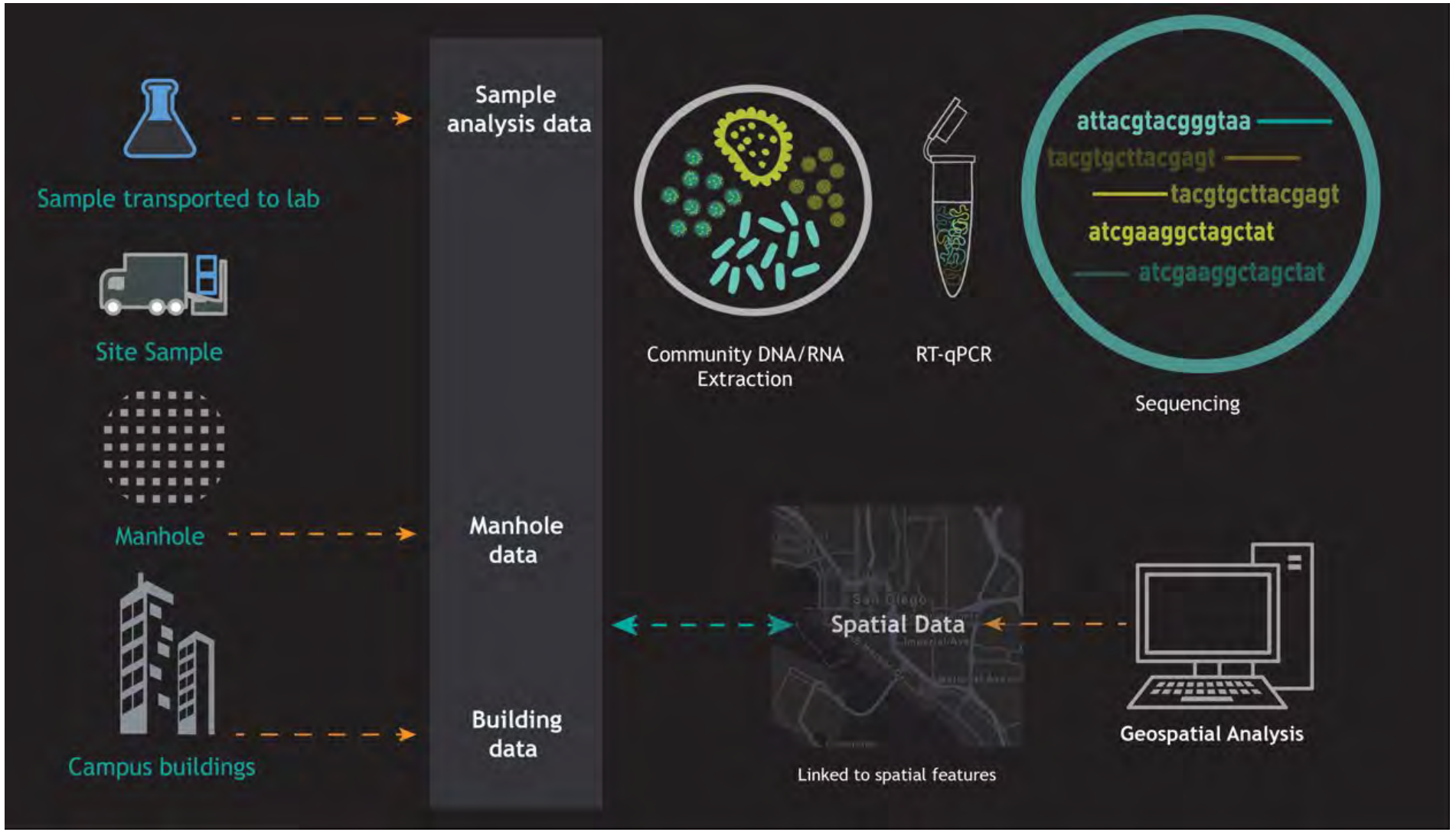
Observed and theoretical time lags between infection and detection of increasing SARS-CoV-2 transmission in wastewater and the health system



Nature Biotech. Sept 2020

UC San Diego

CENTER FOR microbiome INNOVATION

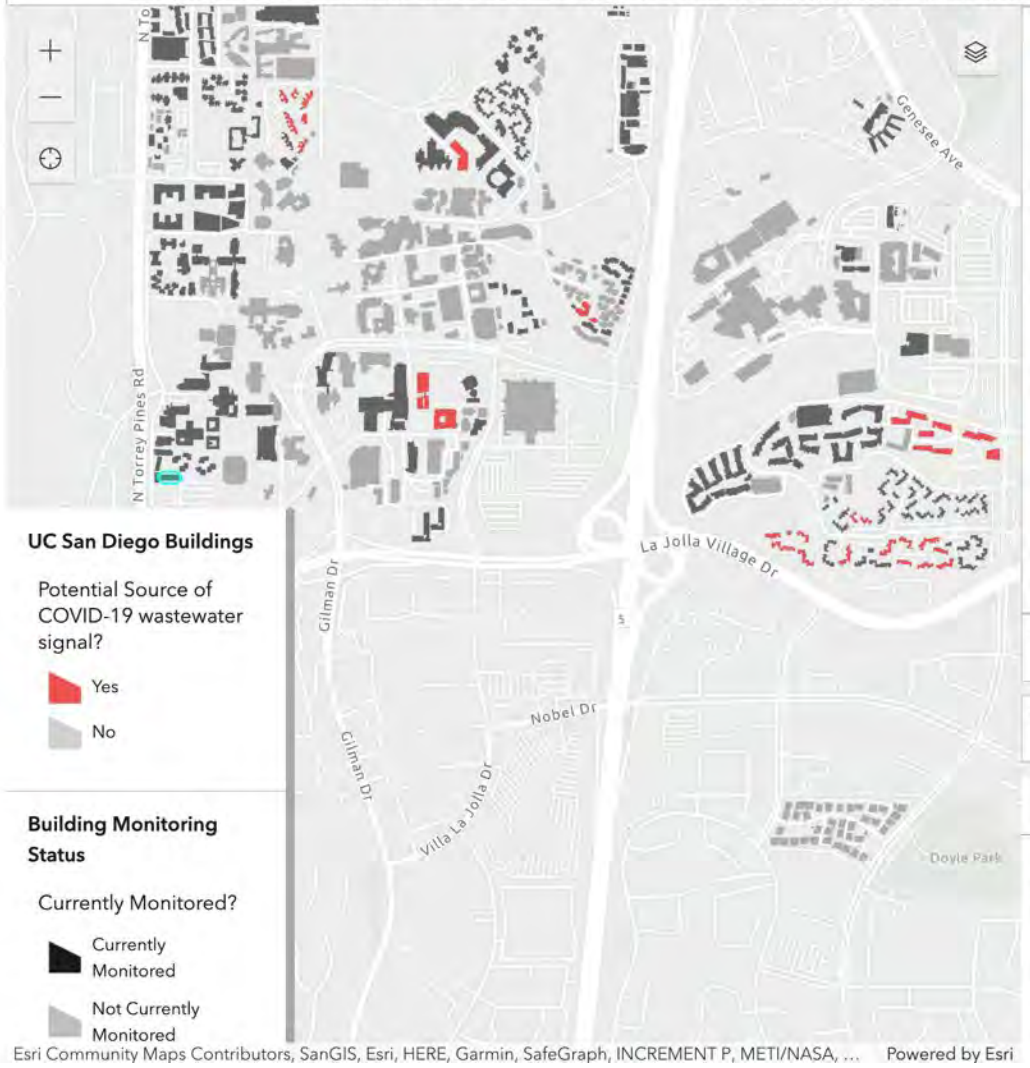

Sample transported to lab
Sample analysis data
Site Sample
Manhole
Manhole data
Campus buildings
Building data
Community DNA/RNA Extraction
RT-qPCR
attacgtacgggtaa
tacgtgcttacgagt
tacgtgcttacgagt
atcgaaggctagctat
atcgaaggctagctat
Sequencing
Spatial Data
Linked to spatial features
Geospatial Analysis

UC San Diego | Return to Learn

Wastewater Monitoring Dashboard

Wastewater collection time period:
11AM PST February 12, 2021 -
11AM PST February 13, 2021

Click here to find further information on our wastewater monitoring program and how to interpret these data on our Return to Learn FAQ page.



UC San Diego Buildings

Search

Name	PossibleSource	Mo
Revelle Apartments - Keeli...	No	Yes
Revelle Residence Hall - M...	No	Yes
Revelle Residence Hall - G...	No	Yes
Revelle Residence Hall - A...	No	Yes
Revelle Apartments - Keeli...	No	Yes
Revelle Residence Hall - D...	No	Yes
Revelle Residence Hall - B...	No	Yes
Revelle Residence Hall - C...	No	Yes
Revelle Apartments - Keeli...	No	Yes

Filter for possible sources of COVID-19 wastewater signal

Filter for currently monitored buildings

< 1 of 602 >

Buildings: Revelle Apartments - Keeling 3

Name	Revelle Apartments - Keeling 3
Currently Monitored?	Yes
Potential Source of COVID-19 wastewater signal?	No

UC San Diego

CENTER FOR microbiome INNOVATION

Acknowledgements

Knight Lab

Expedited COVID IdenTification Environment (EXCITE Lab) at

UCSD

UCSD RTL Team

https://returntolearn.ucsd.edu/about/program-leadership/

Facilities Management

Point Loma WWTP

Center for Microbiome Innovation

Panasas

UC San Diego

CENTER FOR microbiome INNOVATION

PANASAS

PanFS® on ActiveStor® Ultra for Life Sciences HPC Storage

G2M 2021

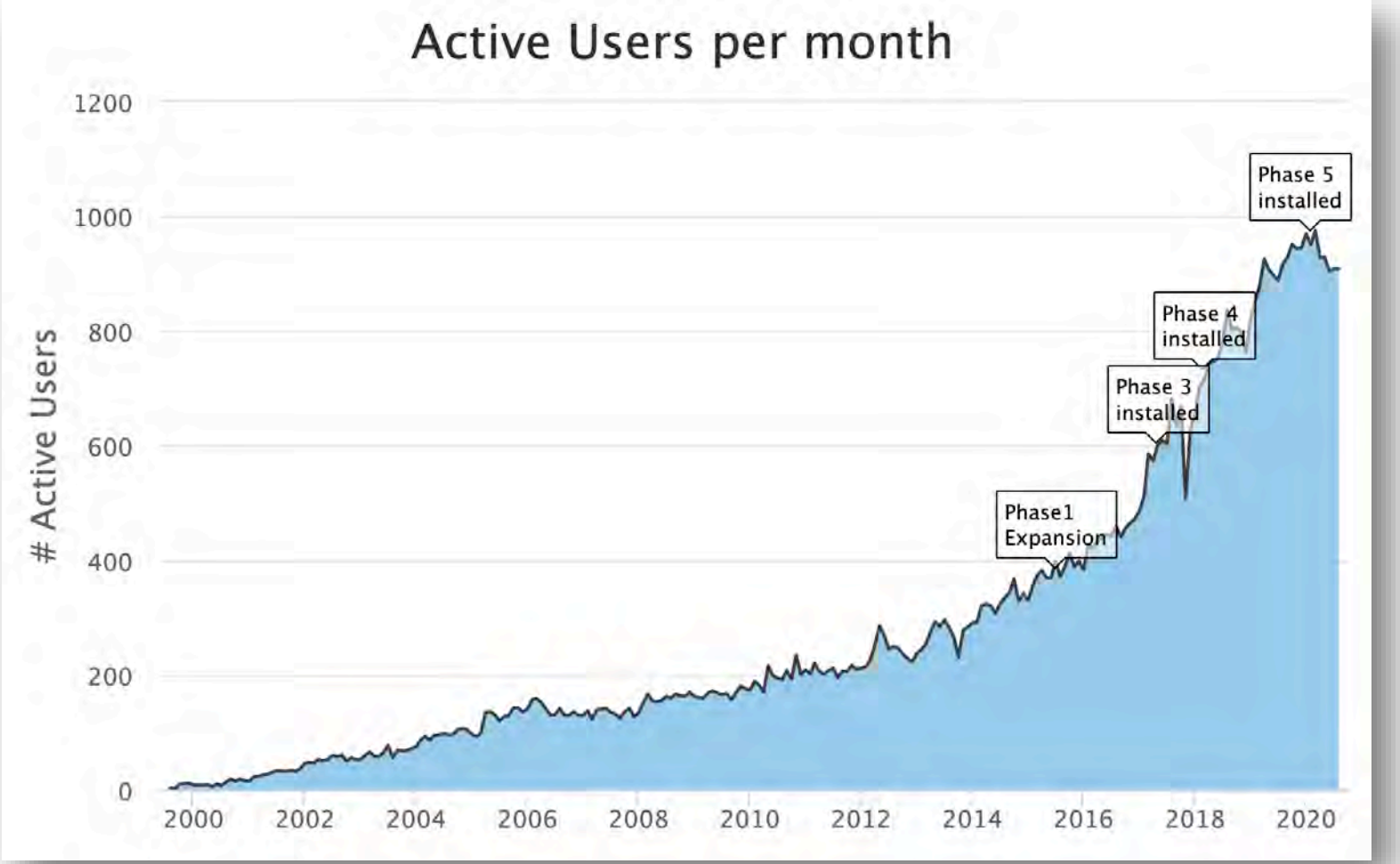
Adam Marko, Director of Life Science Solutions
amarko@panasas.com

Researchers needing HPC

Source: NIH Biowulf Cluster Public Data

HPC users are growing at a faster rate than headcount

- *HPC is needed in more research areas than ever before*

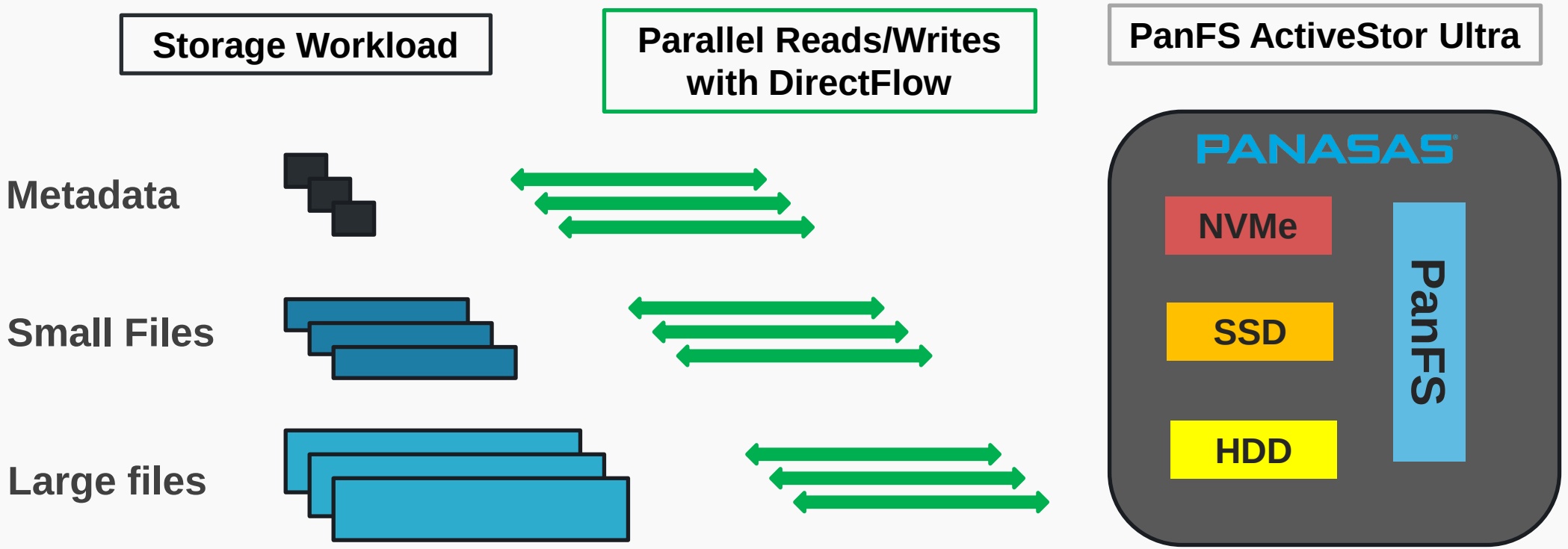


What is Panasas?

PANASAS®

- High Performance, Mixed Storage Media, File System Appliance
- NVMe, SSDs, and HDDs optimized without user config
- Scalable performance and data protection
- Data is auto balanced to ideal media transparently
- Users get seamless total performance from all storage media simultaneously

What makes Panasas Unique?



Dynamic Data Acceleration Enables Simultaneous Storage Access

Data Types are Mapped to Ideal Storage Media

Thank You!

Adam Marko, Director of Life Science Solutions
amarko@panasas.com
info@panasas.com

APPENDIX

What is Panasas?

PANASAS®

- A High Performance, Scalable, Parallel File System
  - Designed to meet dynamic workflows
  - Data protection, reliability, and performance increase as system grows
- In business for 20 years serving the HPC community
- Provide short- and long-term storage for HPC customers globally
- Greatly improved price/performance with Ultra release

Research Cost of Ownership (RCO)

~~TCO/ROI~~

The Effect of Infrastructure on Research Productivity

Researchers

- Publications and research outcomes
- Timetables and collaborations
- Poor morale and frustration

IT

- Annual goals
- Daily responsibilities
- Impact on other projects

Leadership

- Cost of reduced staff productivity
- Effect on research goals
- Reputation

Using RCO

- Ask your Researchers, IT, and Leadership questions
- Don’t make decisions based on side-by-side cost comparison alone

Panasas for AI and ML

AI without all flash

Storage System (Half Rack)	V100 GPUs	Capacity	Cost
Panasas ActiveStor® Ultra	42	1.5PB raw	$
Dell Isilon F800	52	768TB raw	$$$$$

- **PanFS supports 100s of GPUs per rack for demanding AI/ML applications**
  - Configured for AlexNet with TensorFlow using ImageNet dataset
- **Compared to Isilon F800 All-Flash, PanFS provides:**
  - ~ 2x the capacity
  - 1/5th the cost
  - Similar number of GPUs supported

https://www.delltechnologies.com/resources/en-us/asset/white-papers/products/storage/h17361_wp_deep_learning_and_dell_emc_isilon.pdf

Greg Mazzu
Senior Systems Engineer
www.weka.io

Welcome to Weka

Our Mission

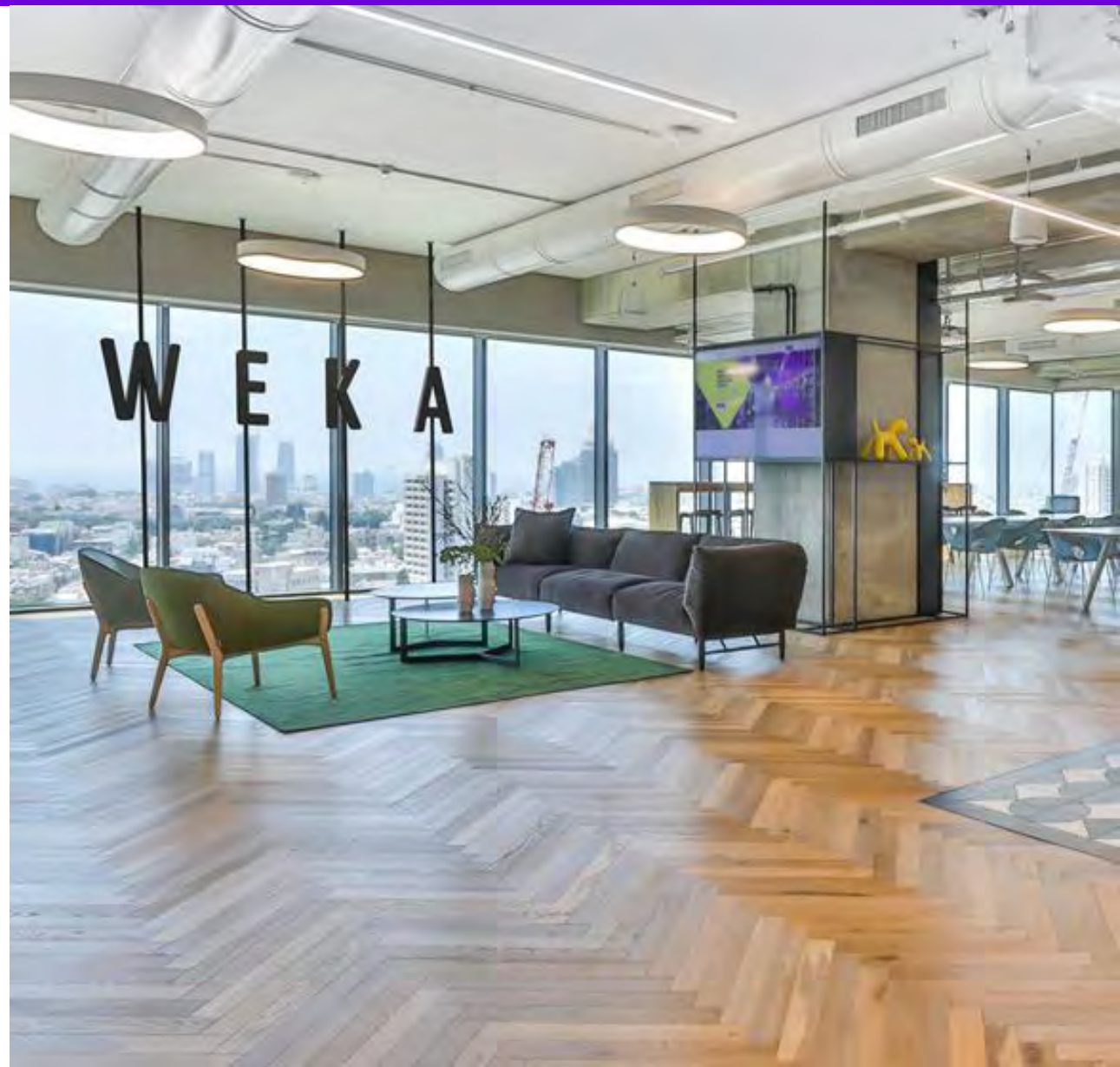
Make storage a utility by delivering simplicity, speed, scale, and better economics

8 of the Fortune 50 are customers

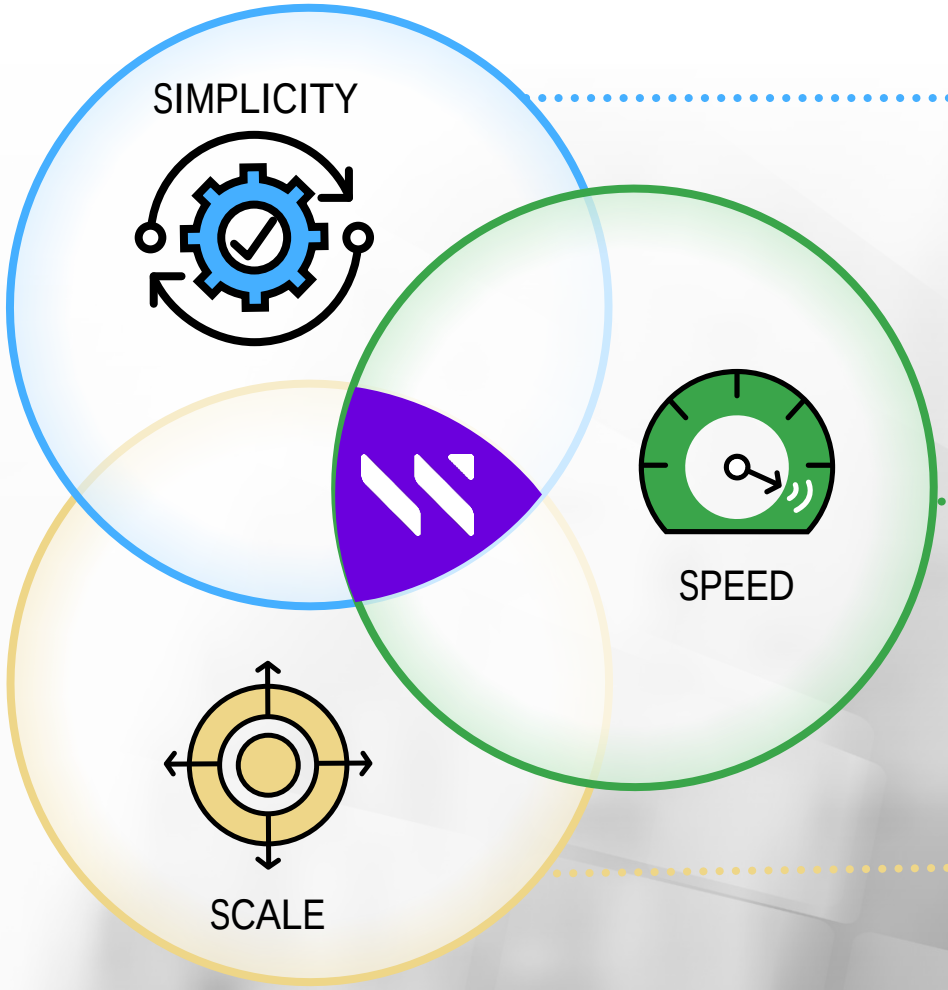
Backed By Industry Leaders

cisco · Hewlett Packard Enterprise · NVIDIA · Mellanox Technologies · Qualcomm · Micron · Western Digital · SEAGATE

WEKA



Limitless Data Platform for Health & Life Sciences



SIMPLICITY

- Unifies all your data
- Single pane of management
- Any platform- cloud or on-premises

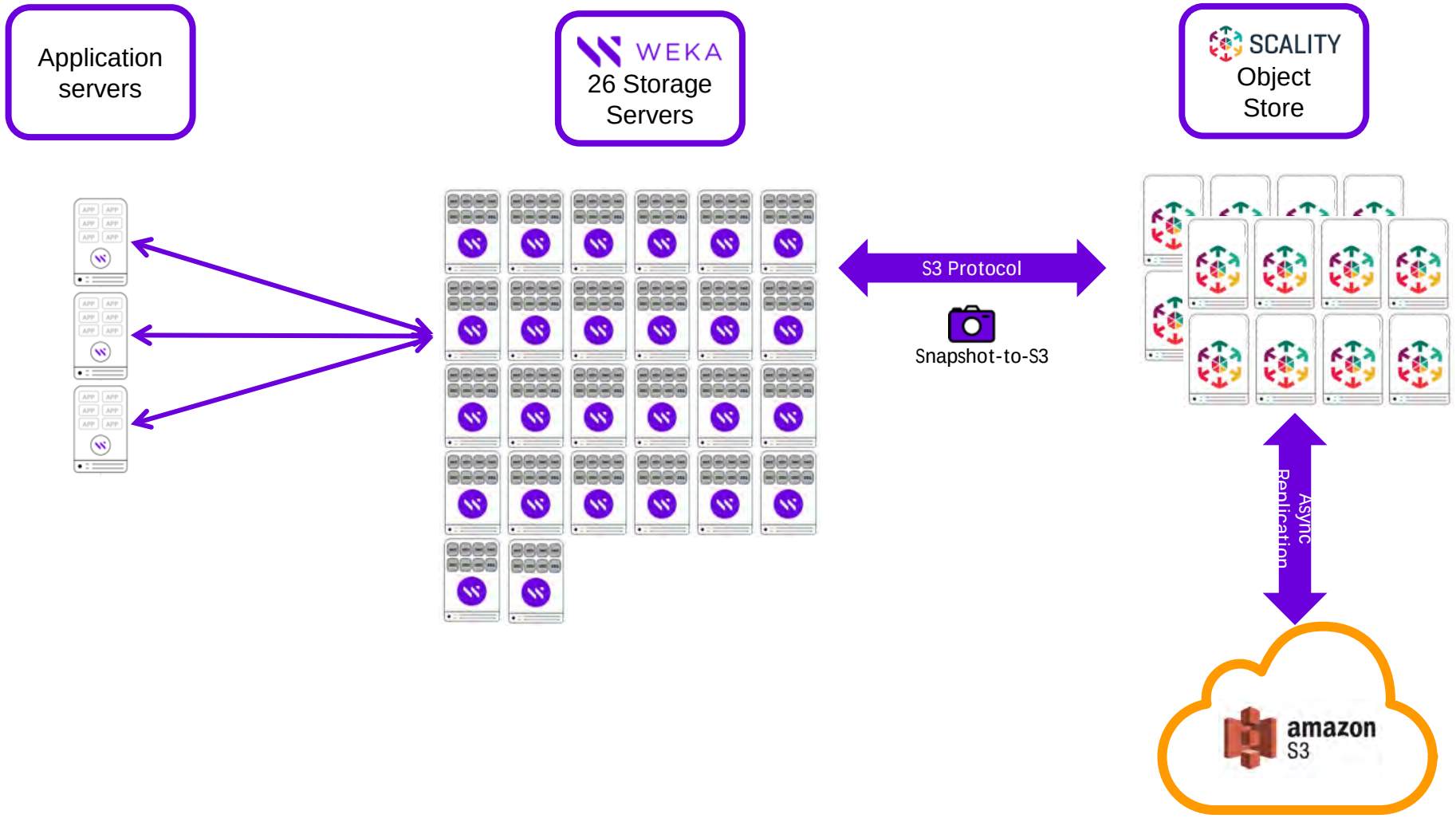
SPEED

- Industry Leading Performance
- Resiliency without performance impact
- Amazing latency across diverse workloads

SCALE

- From Terabytes to Exabytes
- Data mobility to cloud
- Flash & disk for best economics

Boston Pharmaceutical Solution



One solution for many applications

Scientists never wait for technology

Integrated backup

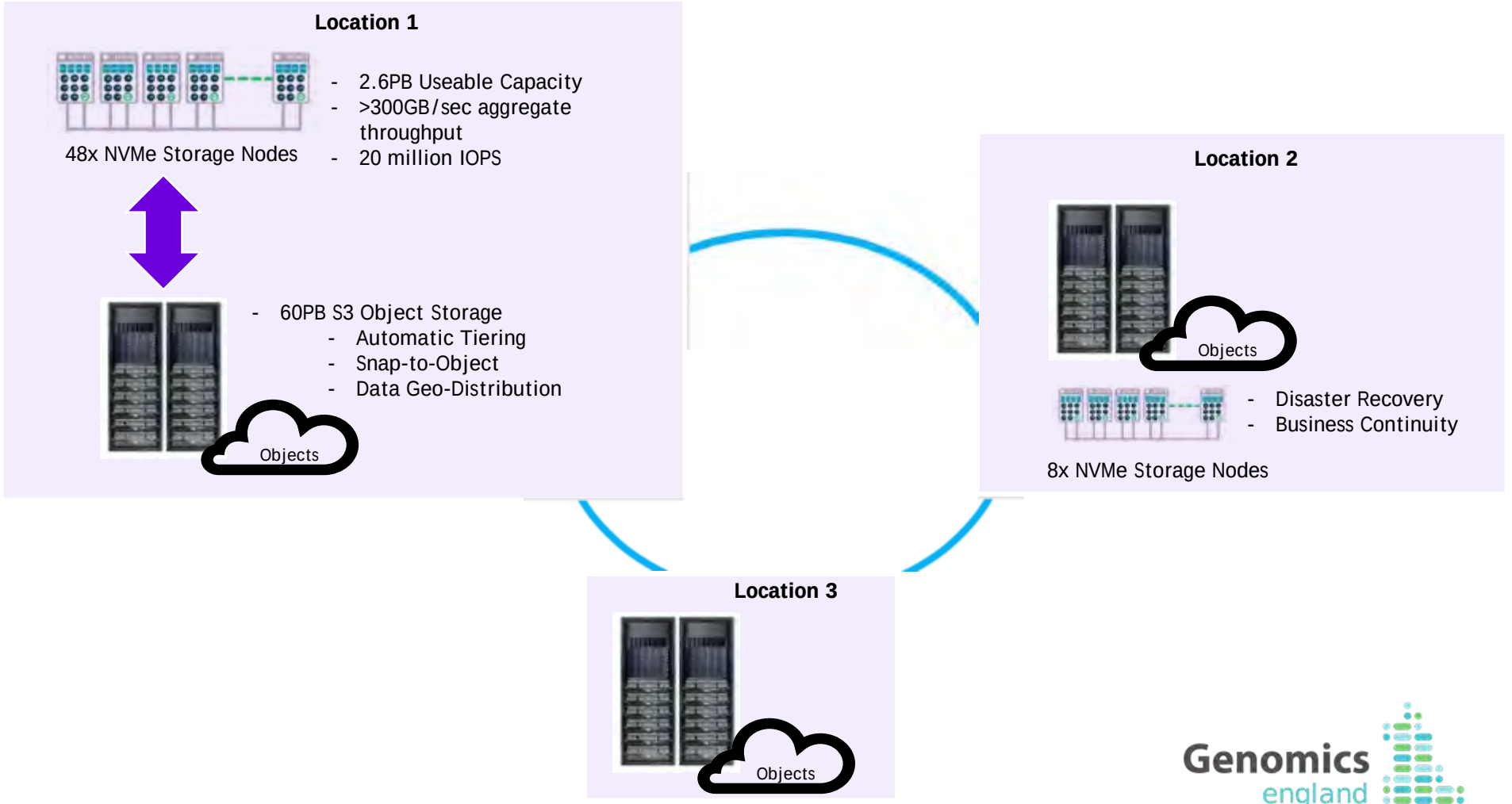
AWS is now DR and BC

Drastic improvement in availability and performance

Technology is an enabler

Weka Deployment at Scale - Genomics England

- Simple single shared namespace
- Dramatic reduction in cost – 26x
- Weka will scale to 160PB within project lifetime
- Over 2,500 Researchers access data for drug discovery & COVID research



WEKA

Genomics england

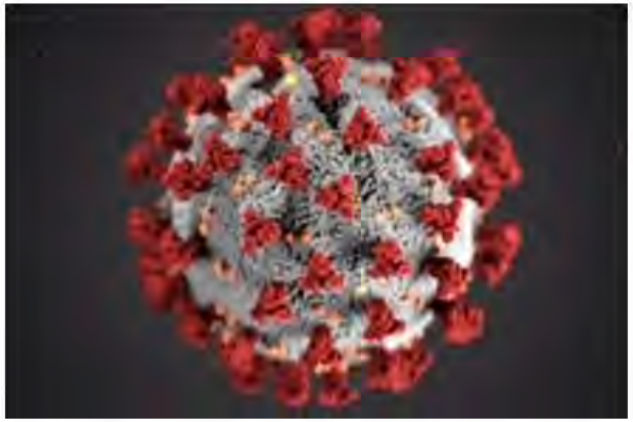
Weka on AWS for NVIDIA Clara Parabricks

AI / DEEP LEARNING

Analyzing Genome Sequence Data on AWS with WekaFS and NVIDIA Clara Parabricks Pipelines

By Robert Clark and Bob Bakh | October 2, 2020 Tags: AWS, Clara, Computational Chemistry, genomics, Healthcare, mitigate, Parabricks, Weka 2 Comments

Whole genome sequencing has become an important and foundational part of genomic research, enabling researchers to identify genetic signatures associated with diseases, differentiate sequencing errors from biological signals, and better characterize the genomes of various organisms. With the ongoing COVID-19 pandemic threatening the globe, characterizing, and understanding genomes is now more crucial than ever. Commercially available, next-generation sequencing platforms allow researchers to decode an entire human genome in less than a day. This helps in understanding the susceptibility with infection by SARS-CoV-2; can be used as the basis for vaccine creation; and can be used for therapy selection for an individual based upon their unique genetic signatures, along with many other use cases.



Traditionally, sequencing a whole human genome takes multiple days coupled to a heavy compute power using CPU resources. The Genome Analysis Toolkit (GATK), developed by the Broad Institute, is a multi-purpose software suite for analyzing DNA- and RNA-based sequence data with the primary goal to identify genetic variants. It is generally considered the industry standard toolset. The GATK suite contains several tools, including the Base Quality Score Recalibration (BQSR), Burrows Wheeler Aligner Maximal Effect Match (BWA MEM) for aligning and calibrating genomes, and HaplotypeCaller, which efficiently identifies variants in sequences. These tools are all run as sub-stages in the germline pipeline that identifies germline variants and is one of the algorithms used in this post.

NVIDIA Clara Parabrick Pipelines replicate the functionality of GATK while harnessing the power of GPUs to provide the fastest method to analyze genomes. The software, leveraging the NVIDIA CUDA libraries to accelerate key algorithms, dramatically reduces the time required to analyze a sequence compared to the CPU-only GATK tools and also includes the DeepVariant algorithm developed by researchers at Google. Combined with the power of Amazon Web Services (AWS) and NVIDIA storage partner WekaIO (Weka), the high-performance sequence analysis of NVIDIA Clara Parabricks integrates with the convenience of the cloud to offer a robust, fast, and simple solution to enable faster genomic research over existing state-of-the-art CPU-based solutions. It offers over 33x improvements in performance with over 99.9% concordance of results.

WEKA

Scott Shadley
VP of Marketing
www.ngdsystems.com

Computational Storage

NVMe SSDs with Compute on Drive

Driving Compute and Storage in any Environment

The Market Needs a New Way to Look at Storage.

Pain Points

Physical Space

Available Power

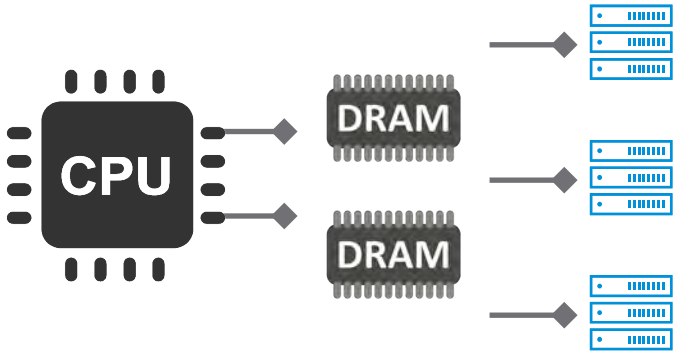
Scaling Mismatch

Bottleneck Shuffle

Traditional storage architectures are in **trouble**.

Current Compute Systems are limited to a few CPU count per system

The Processing Unit (CPU/GPU/TPU) all require scale for BioTech



Finding ways to process data more efficiently is needed

This support needs to come **without wholesale change**

The Market Solution Using Computational Storage.

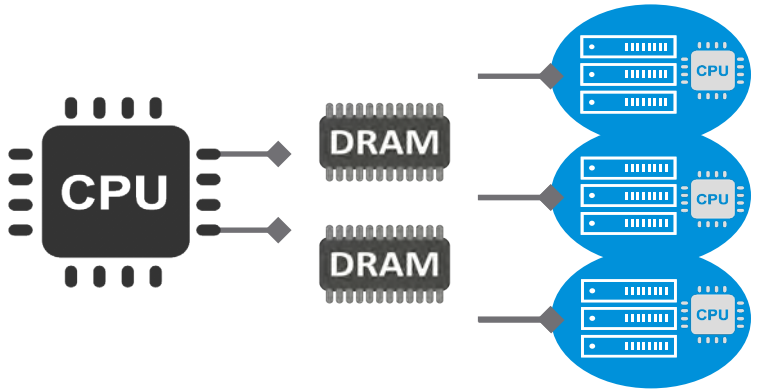
Value Add

- ✓ Distributed Processing
- ✓ Faster Results
- ✓ Lower Power
- ✓ Smaller Footprint

Scaling compute resources within storage provides accelerated results

Computational Storage resources 'offload' work from the limited CPU count

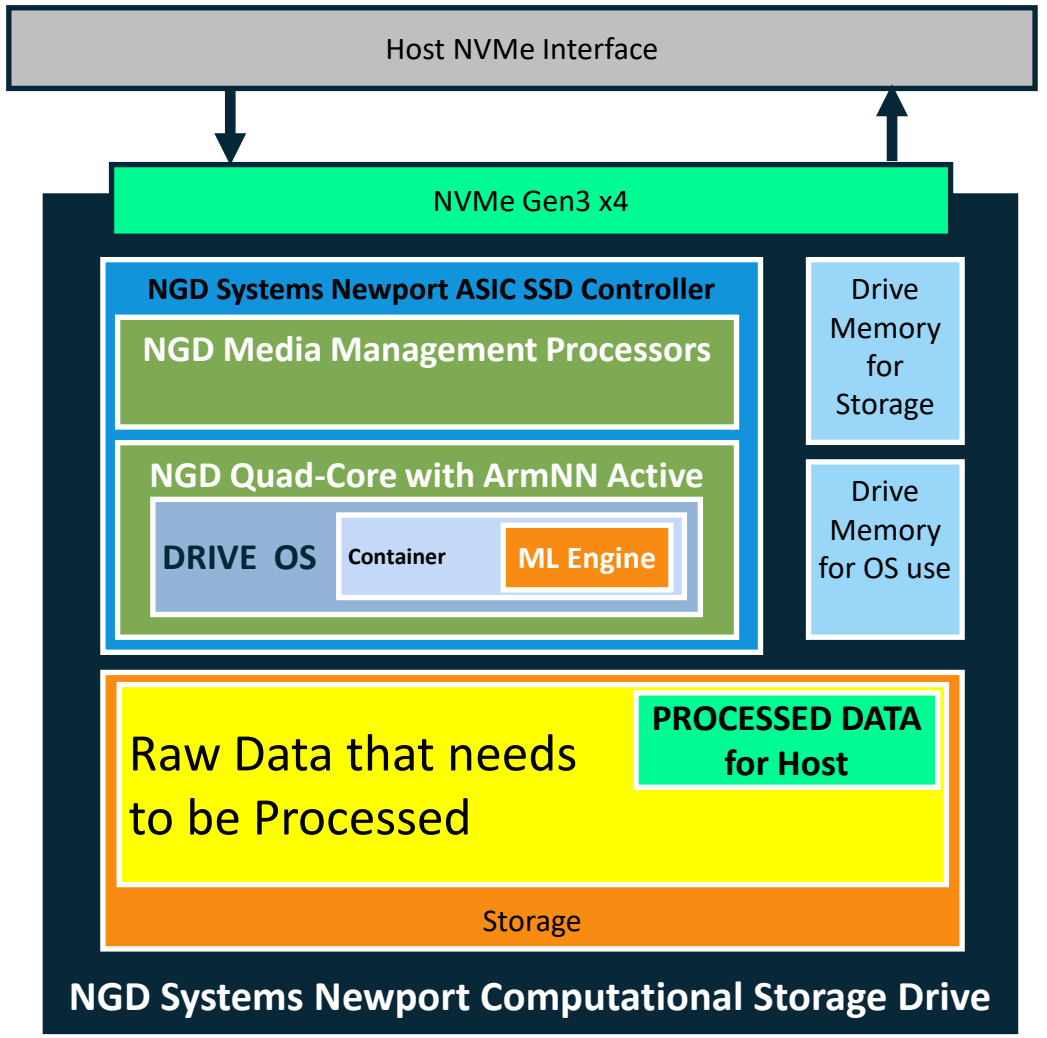
Seamless architectures create new 'servers' for more effective analytics



This **'Server in a Server'** Architecture provides value across many use cases

Additional CPU resources for the cost of Storage without added Resources

ARM NN Deployment for Parallel & Distributed Processing



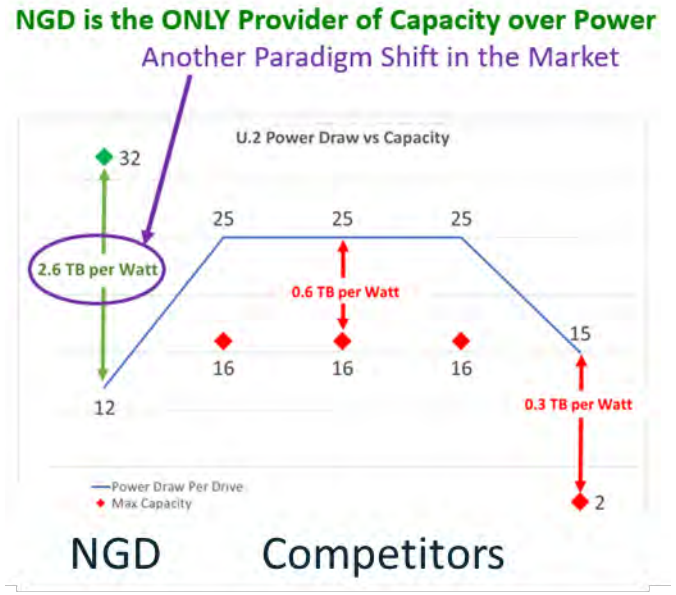
- More than **100k data points per drive** are processed without sending the data to host. Saving Time, Network and Power.
- **The output of the application is written directly to a MongoDB database implemented inside In-situ engine, and user can access the output data using MongoDB APIs**.

VIDEO DEMO WITH Accelerated response

The dataset is acquired from: http://cocodataset.org

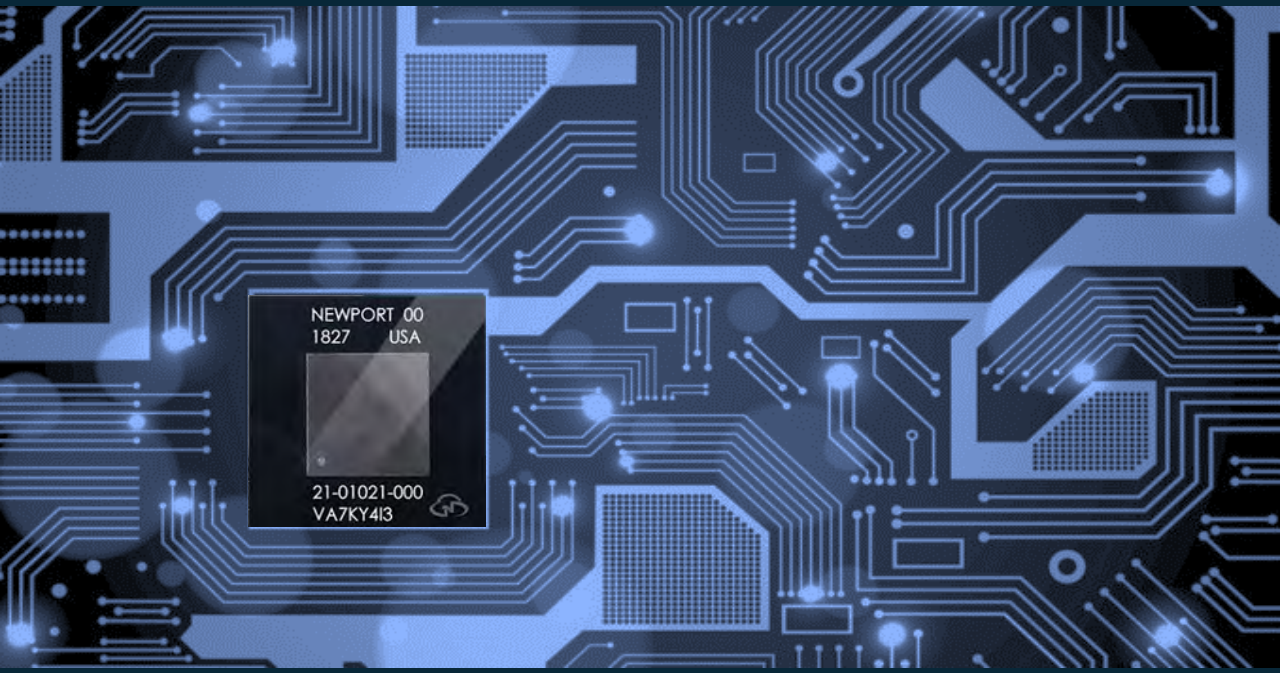
NGD NVMe SSD Products at a Glance.

- Large breadth of **SSD** solutions and capacity options
- Leading **TB/W** Energy Efficiency
- Industry's only **16-Channel** 14nm SSD SoC
- Industry's Largest capacity NVMe SSDs
- Quad-Core **Computational Storage CPUs**



Form Factor	Availability	Raw Capacity TLC (TB)	MAX Power (W)
M.2 22110	NOW	up to 8	8
U.2 15mm	NOW	up to 32	12
EDSFF E1.S	NOW	up to 12	12

Computational Storage, Some Real World Results.



- **AI Inference Offload**
  - Similarity search, Tracking
- **Distributed Machine Learning**
  - Identification, Training, Tagging
- **Data Search, Compare**
  - Finding the Needle, Making sure it is!
- **Database Acceleration**
  - Data is stored somewhere, drive it efficiently

Protein Sequencing – BLAST® Accelerated

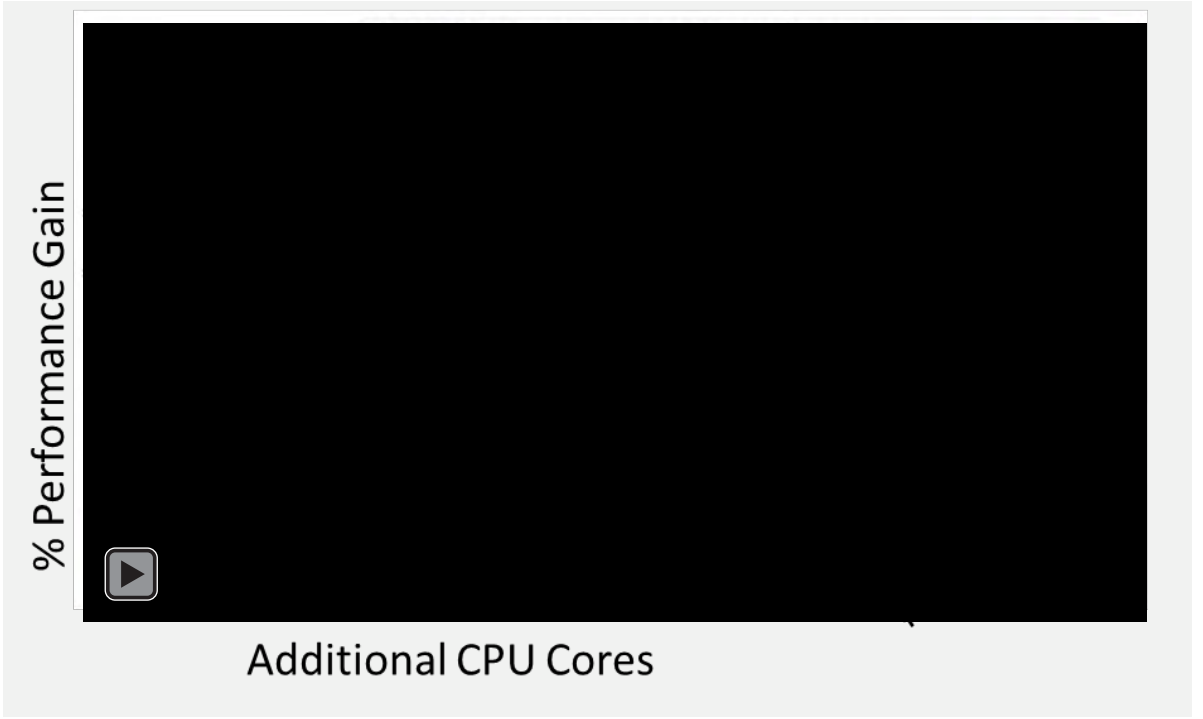
- DNA and Protein alignment Database Management

NCBI BLAST

The Basic Local Alignment Search Tool (BLAST) finds regions of similarity between sequences.

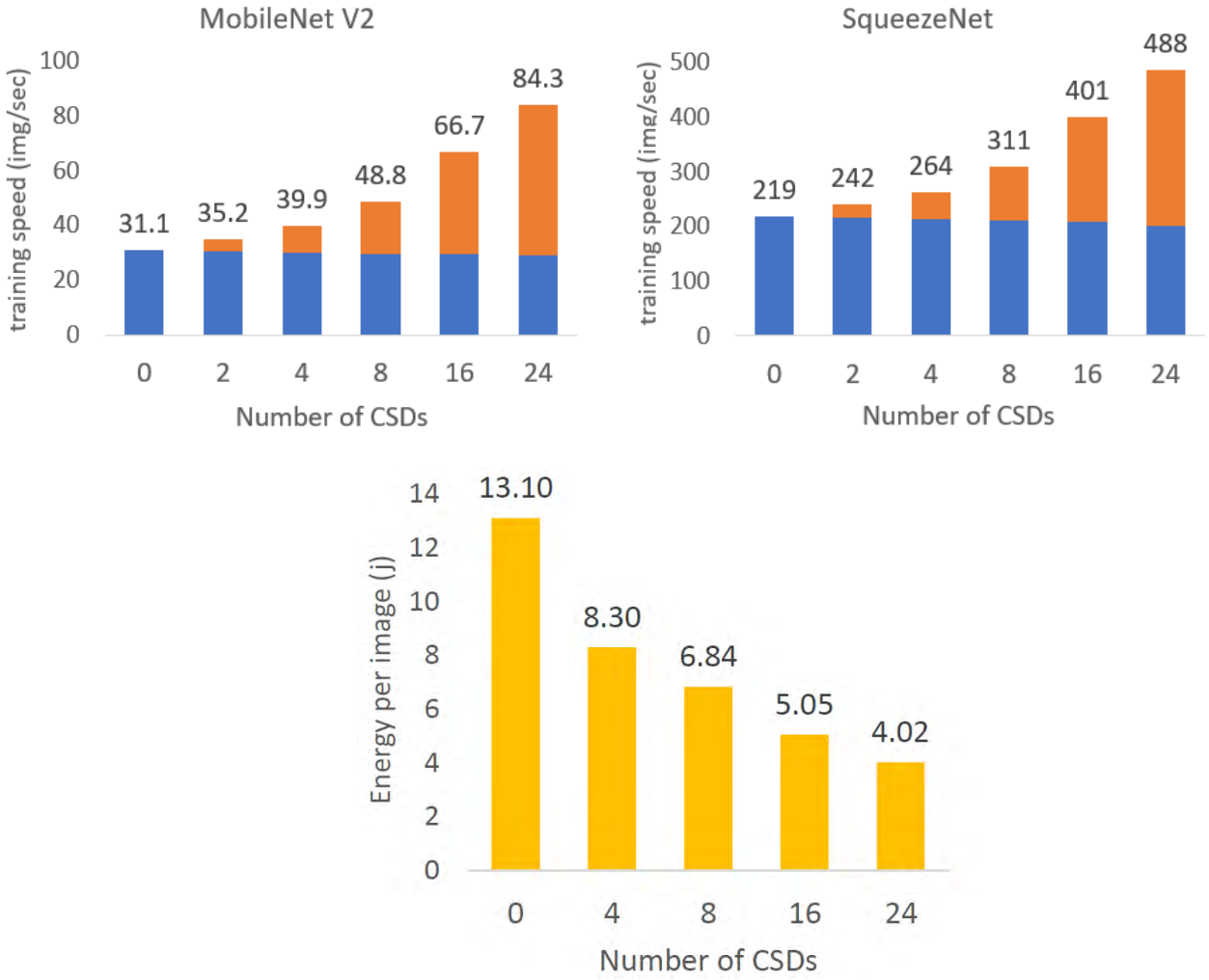
- Not balanced dataset
  - Computation time varies a lot across files
  - Different number of sequences per file

- By combining with Computational Storage SSDs and using the 4 cores per drive, you gain **up to 100% more** performance at **no cost** in CPU or Memory



Accelerated ML, Faster, More Efficient

- Four neural networks Evaluated
  - **MobilenetV2**
  - NASNet
  - **SqueezeNet**
  - InceptionV3Quad-core
- Training data stored on CSDs
- Tested with 24 CSDs
  - 32TB capacity each
  - Quad-core CPUs on Drive
  - 4x NEON SIMD engines
- Using an AIC 2U-FB201-LX server
  - Intel® Xeon® Silver 4108 CPU
  - 32GB DRAM



Panel Questions and Audience Surveys

Panel Question # 1

- How does storage for biotech AI applications differ from other AI applications and/or other HPC applications?
  - NetApp
  - Panasas
  - Weka
  - NGD Systems

Audience Survey Question #1

- To what extent is your organization utilizing AI and ML in your business processes (check one):
  - We use AI and ML widely in our business/organization: 20%
  - We have several problem domains which utilize AI and ML: 20%
  - We have a few problem domains in which we utilize AI and ML: 0%
  - We are exploring the use of AI and ML for new workloads, but have not yet implemented any AI/ML solutions: 30%
  - We have no plans to utilize AI and ML workloads in our organization in the near future: 30%
  - Don't know:

Panel Question #2

- We have all heard about the importance of managing training data sets, even as they evolve over time. What technologies are key to integrating and managing different data types?
  - Panasas
  - Weka
  - NGD Systems
  - NetApp

Audience Survey Question #2

- When looking at implementing AI and ML storage solutions, who do you look to for guidance on hardware and software choices (check all that apply):
  - My current IT server vendor(s): 21%
  - My current IT storage vendor(s): 43%
  - My current IT integrators: 0%
  - AI/ML application/framework vendors: 43%
  - AI/ML application/framework integrators: 14%
  - Other: 21%

Panel Question #3

- All of your companies provide on-premise storage solutions. How does the cloud fit in to storage architectures and solutions for AI applications?
  - Panasas
  - Weka
  - NGD Systems
  - NetApp

Audience Q&A


Thank You For Attending!