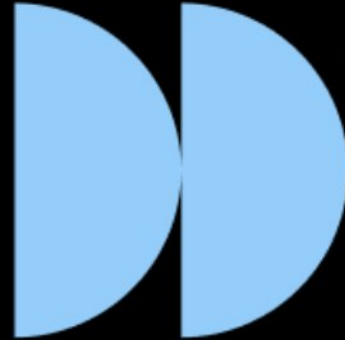


Welcome to the IBM TechXchange Community

Connect and engage to get answers, discuss best practices, and continually learn more about IBM solutions.



Chris Maestas
IBM CTO, Data and AI Storage
Solutions
Chief Troublemaking Officer
UNM Alum!

Accessing Data Anywhere and
Everywhere
with
IBM Storage Scale – a Global
Platform for Storage (GPFS)
Services

The world is not flat...nor is data



Round Earth Clues: How Science Proves that our Home is a
Globe | University of Nevada, Las Vegas

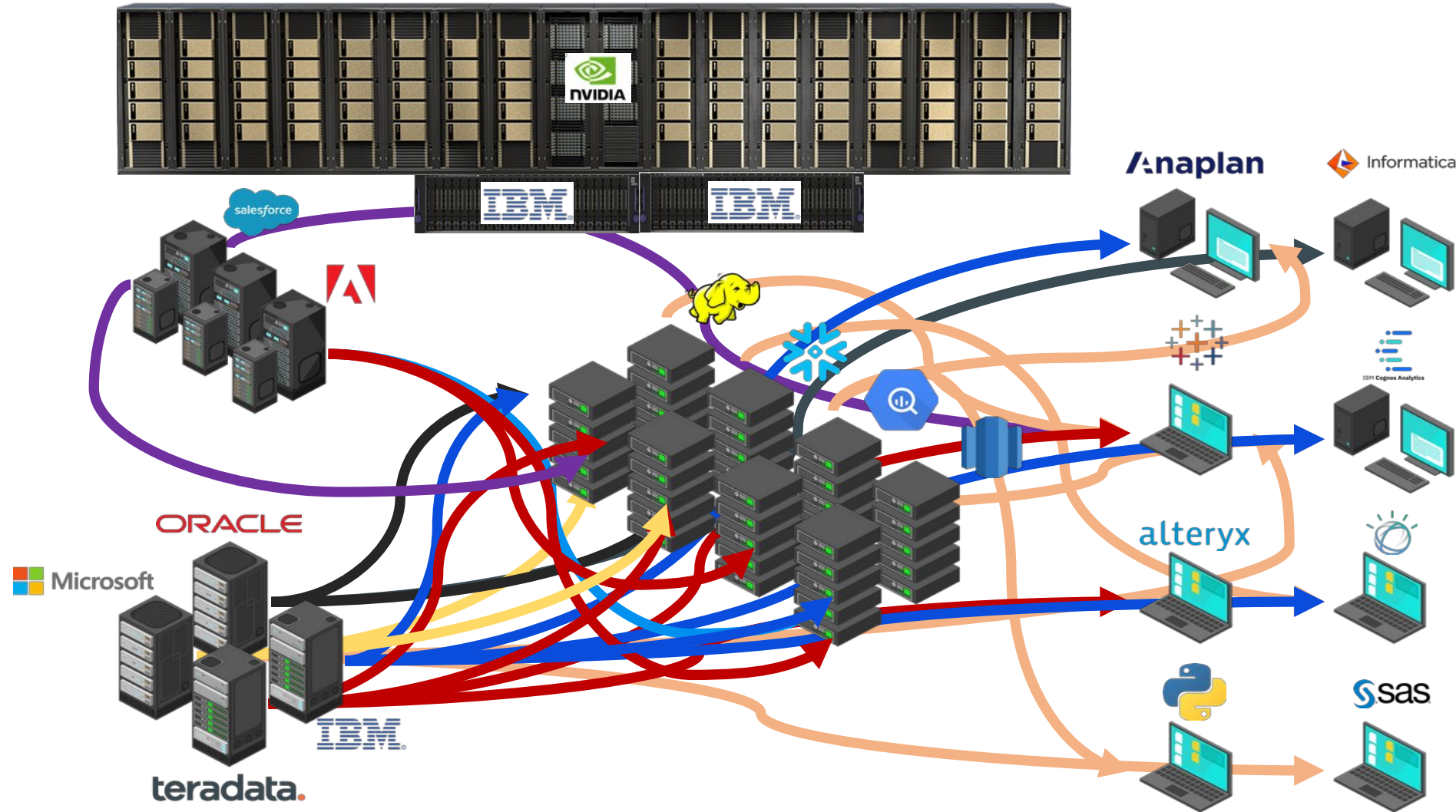
- Up to 80% of data is unstructured, growing 55-65% per year, edgedelta.com
- Data has a life cycle
- This requires intelligent software, policy management, tiering, and more...

STORAGE for AI MATTERS

The fastest servers in the world are the world's
slowest servers if they are waiting for data
ALL GPUs & CPUs WAIT AT THE SAME SPEED

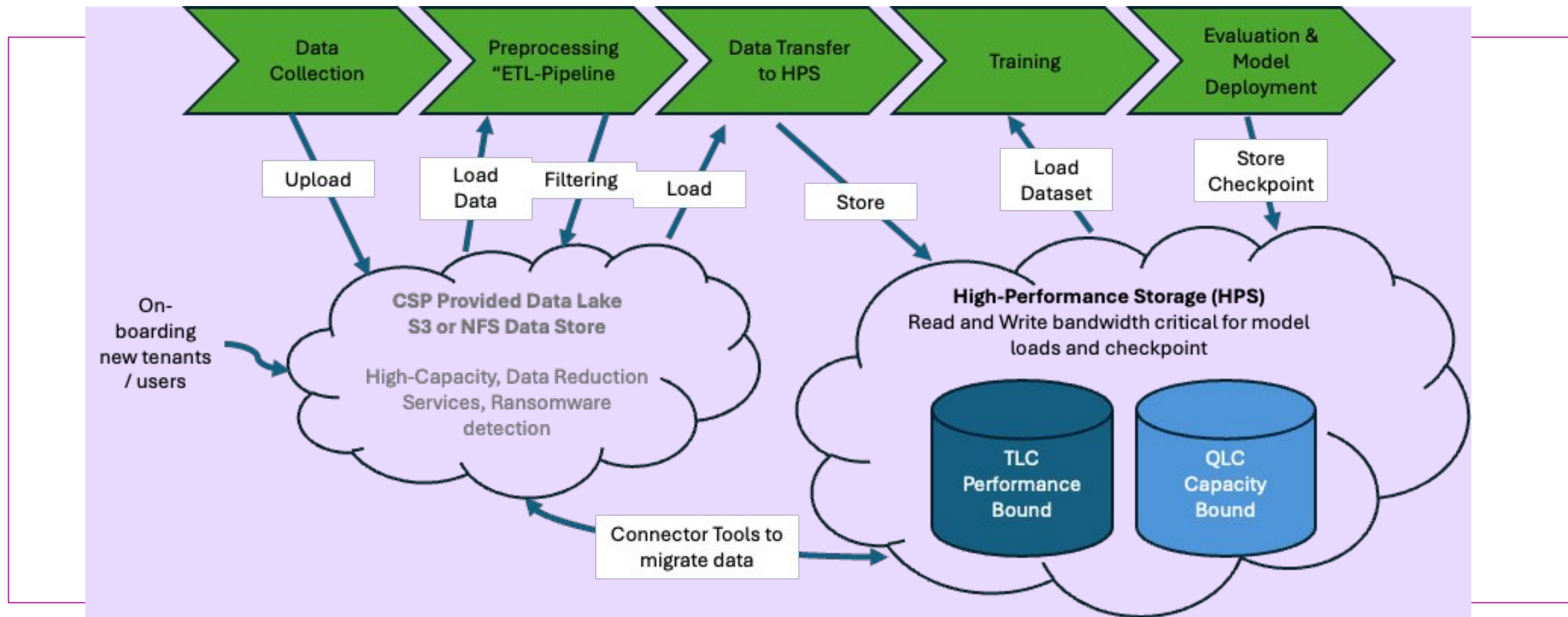


Feeding HPC and AI data comes from everywhere and IBM has a roadmap for data access, not another storage silo



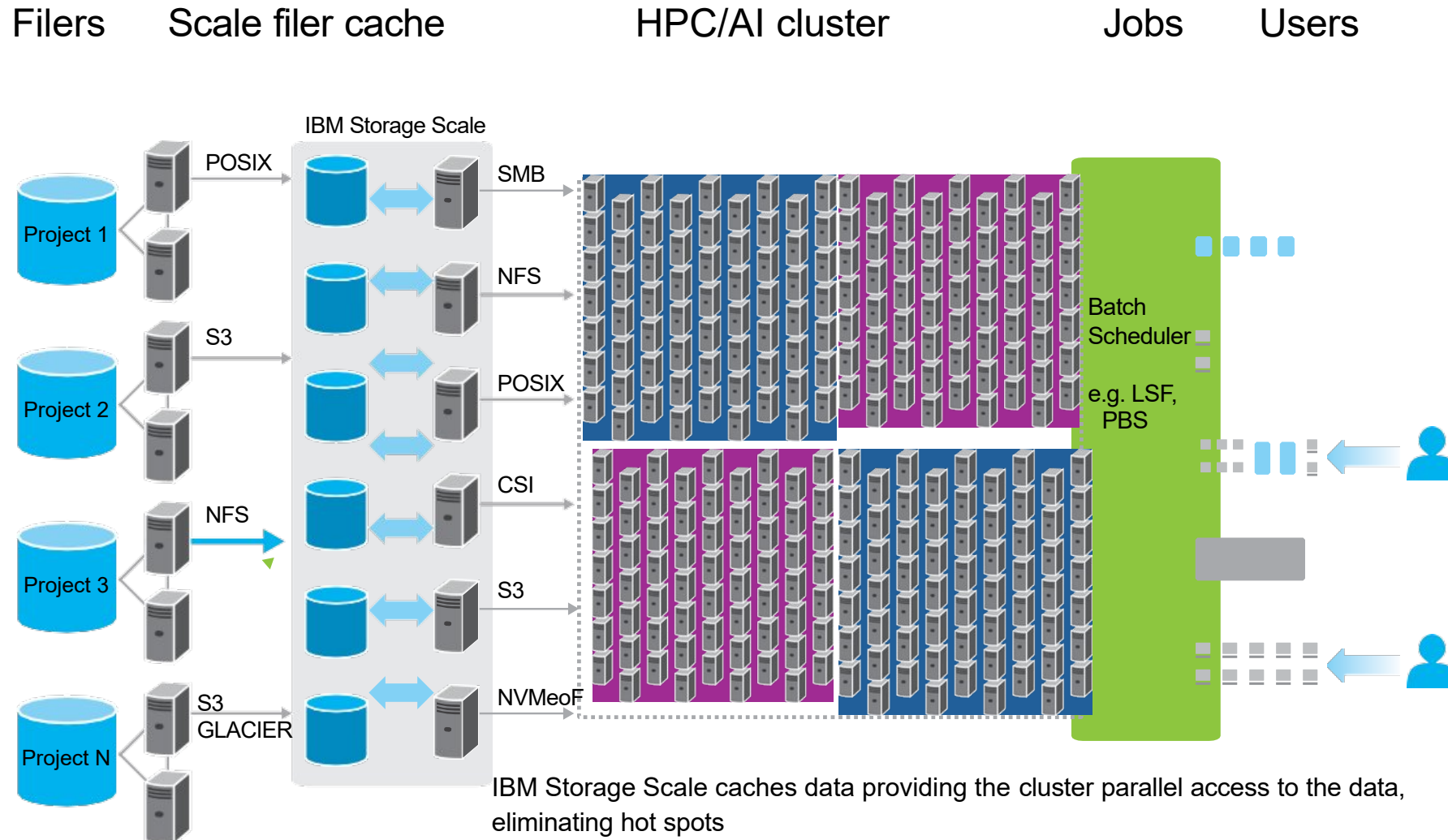
High-Performance Storage

Offered by Certified Storage Partners



Secure, Multi-tenancy, Highly-Available, and Resilient

HPC and AI with IBM Storage Scale



Identifying HPC and AI Related Opportunities

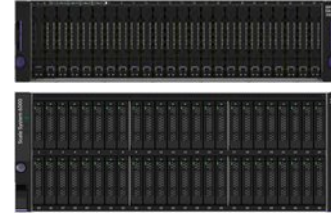
AI, Accelerated
Computing &
HPC



IBM watsonx.ai
RHEL AI
Open Source



IBM Storage Scale and Storage Scale System



- HPC compute, GPU, DPU
- AI Training
- Feed NVIDIA DGX servers or SuperPODs
- Extreme price perf overall, per U, and per Watt
- OpenShift and Non OpenShift workloads
- One of only three approved for NVIDIA SuperPODs



AI, Hybrid Cloud, &
App Modernization



watsonx

IBM CloudPaks
ISVs on OCP
In-house SW Dev
VMware alternative



IBM Fusion & Fusion HCI



- Fastest way to deploy **watsonx**
- AI Acceleration tier as a turnkey HCI appliance
- AI Inference
- Watsonx RA with Ceph for x.Lakehouse
- OpenShift applications only

CLOUDERA

Hadoop Data Lake
Modernization



S3 Data Lakehouse

watsonx.data

Cloud Repatriation
Open Source Tools

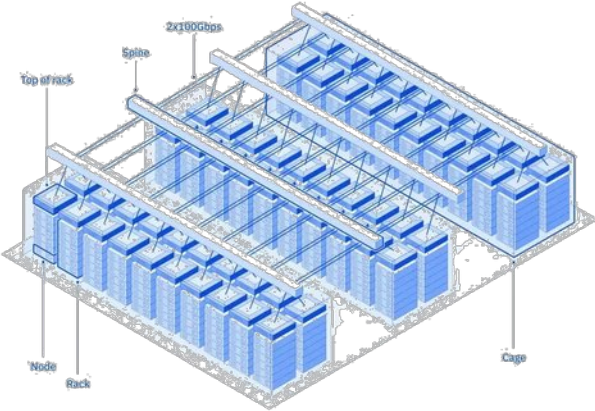
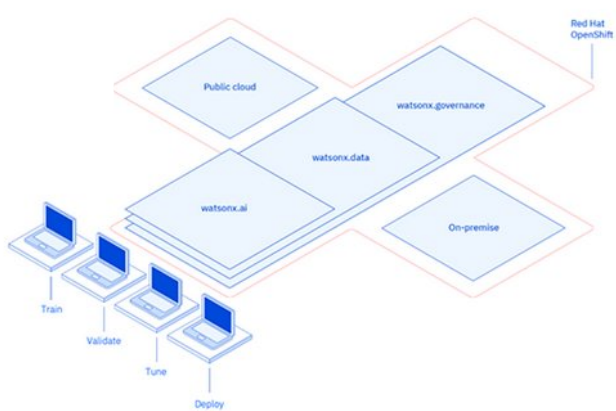


IBM Storage Ceph & Storage Ready Nodes



- Building blocks for an S3 data lakehouse
- Modular and flexible; 1-stop support from IBM
- De facto data lakehouse for **watsonx**

Storage requirements for HPC and AI



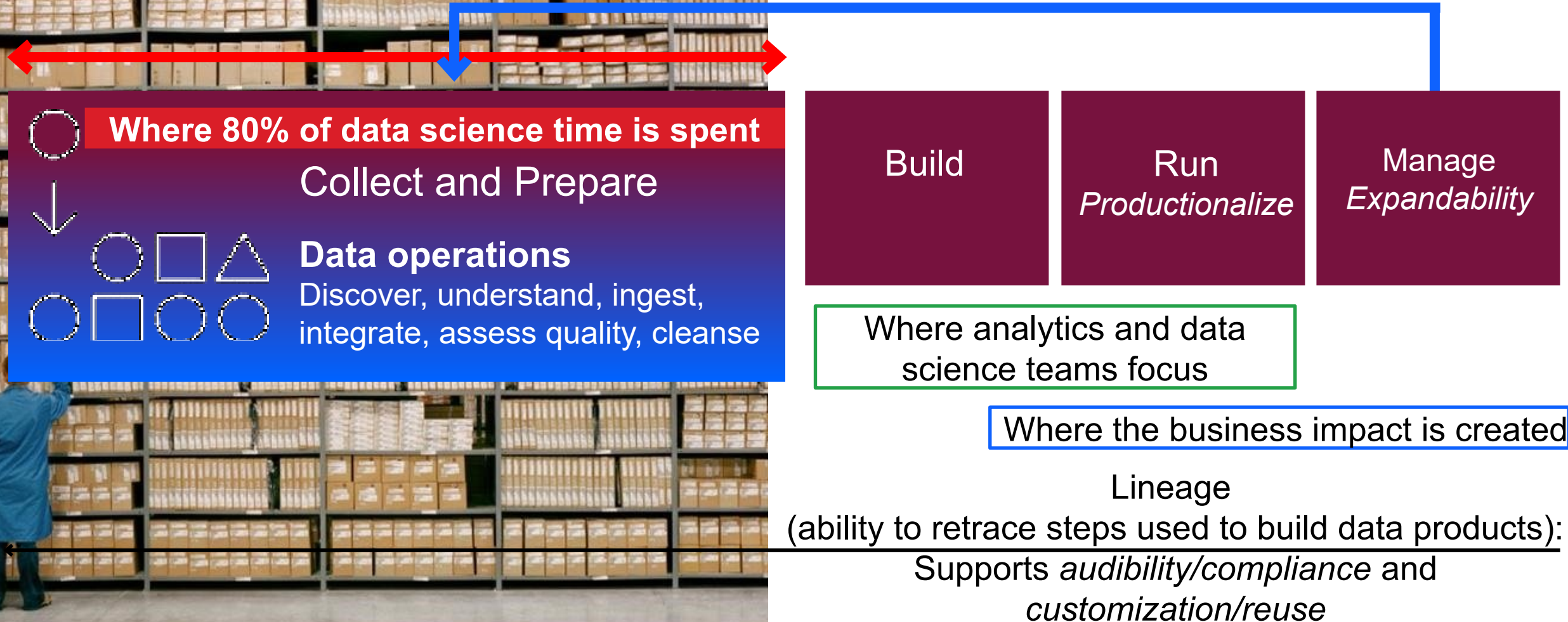
AI Tuning/Inferencing/Postprocessing

Storage Acceleration	Efficient GPU support	Rapid deployment
	Storage Abstraction	HA/DR/Backup
	Metadata catalog integration	Simplified Day-2 operations

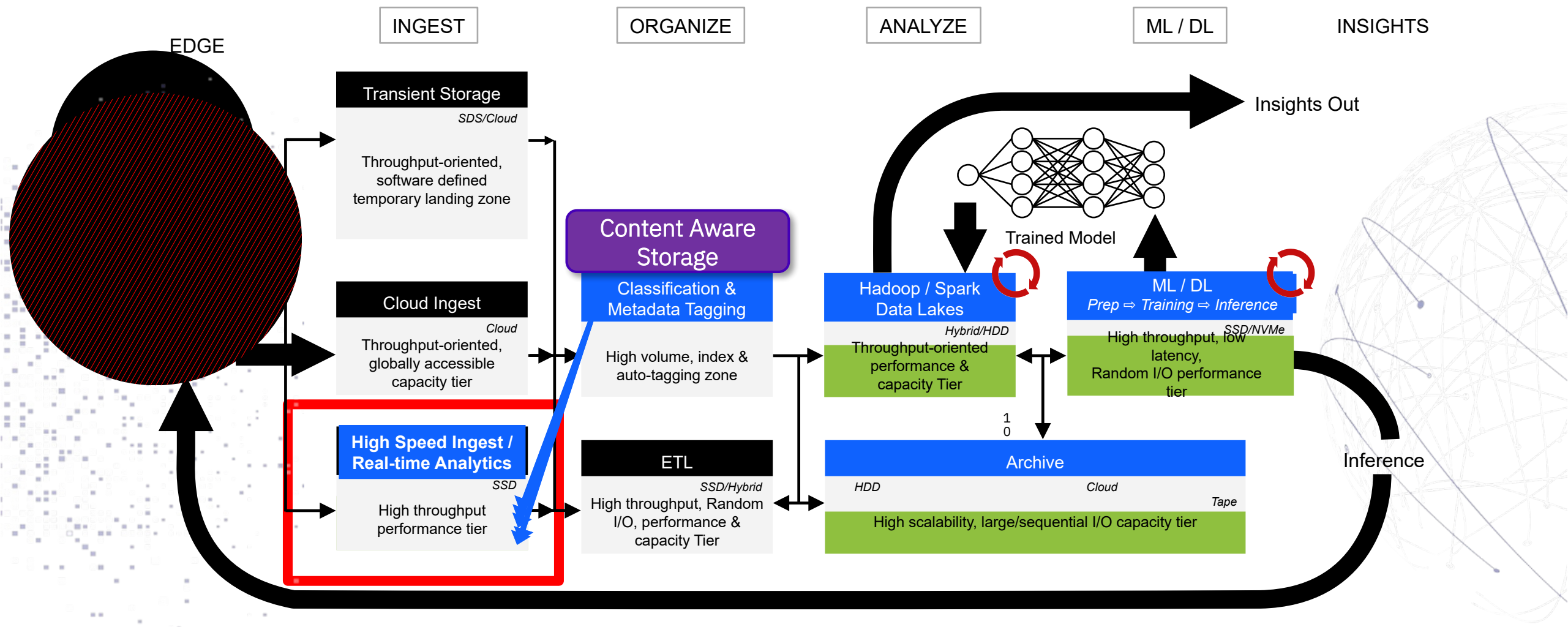
HPC and AI Training

Maximum Performance	Efficient GPU support
	High bandwidth
	Low latency
Scalability	Scalable performance
	Linear scaling of performance and capacity
	High density

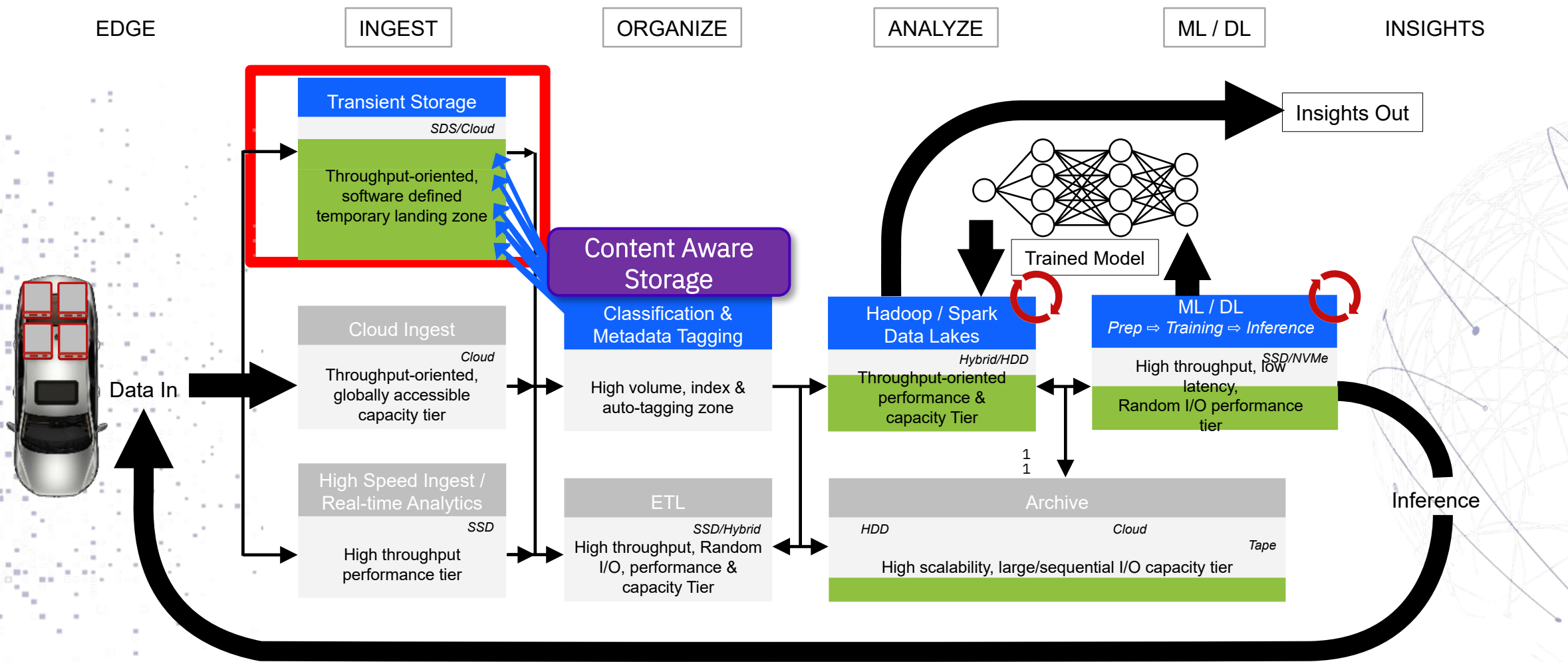
Making your data ready for high value analytics is widely time-consuming today



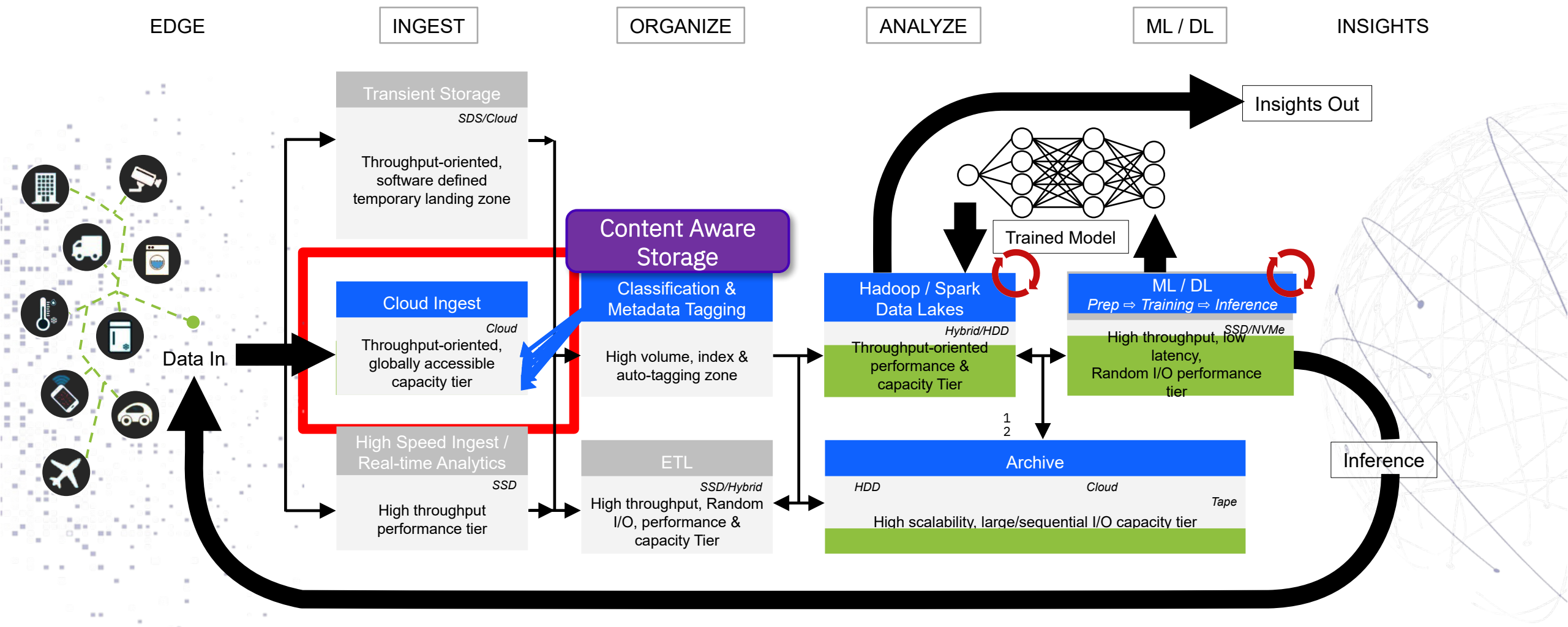
AI and Analytics Data Flow - High Speed Ingest (Local Windows Boxes)



AI and Analytics Data Flow – Transient Ingest (Temporary)



AI and Analytics Data – Cloud Ingest (Internet of Things (IoT))



Is there another case?

Discovering your Archive

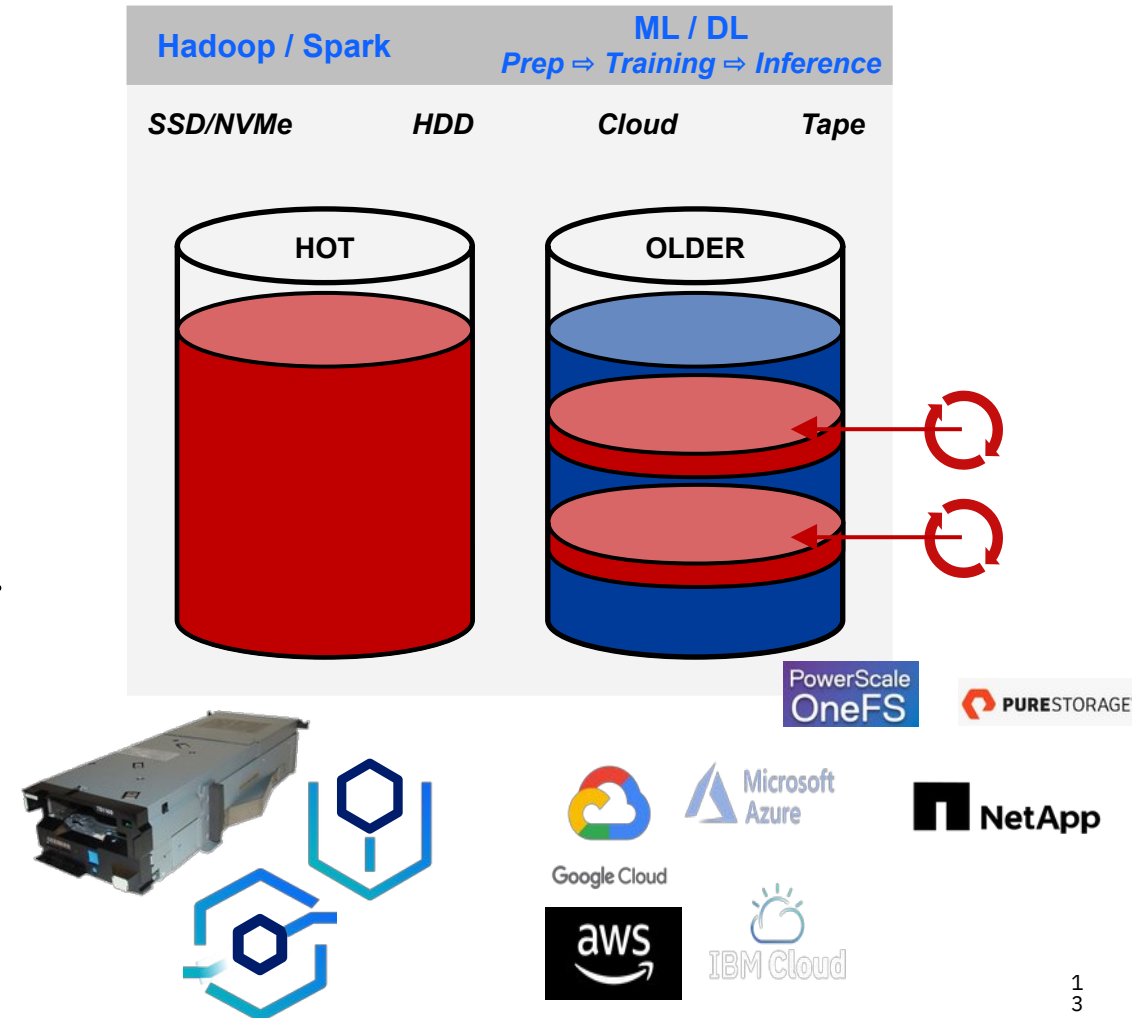
We see:

Customers across all verticals are creating large PB to EB data stores.

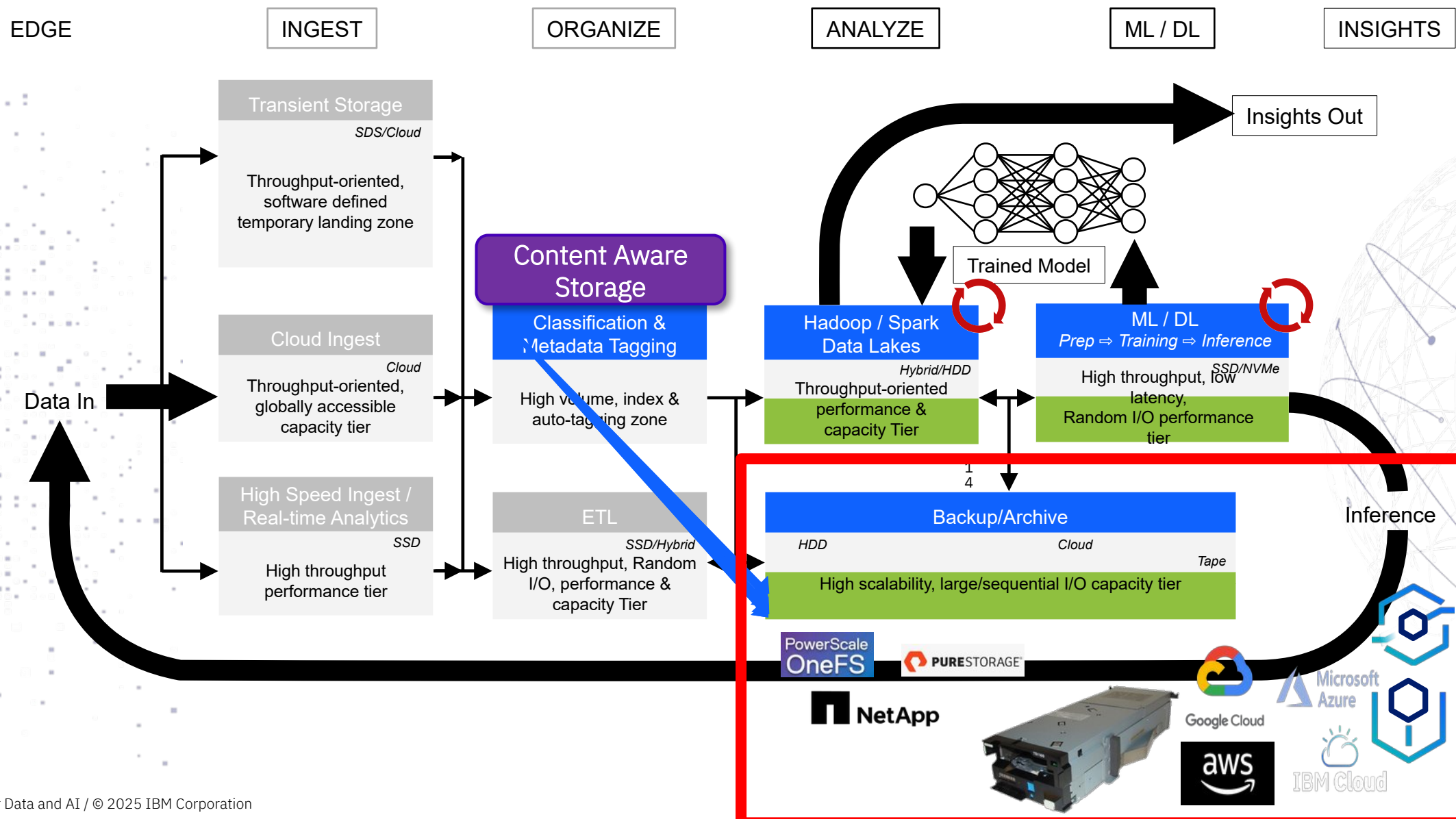
Other just does not mean cloud or tape,
it could be existing storage

Vast majority of data is relatively (c)old,
but still required for periodic trend analysis.

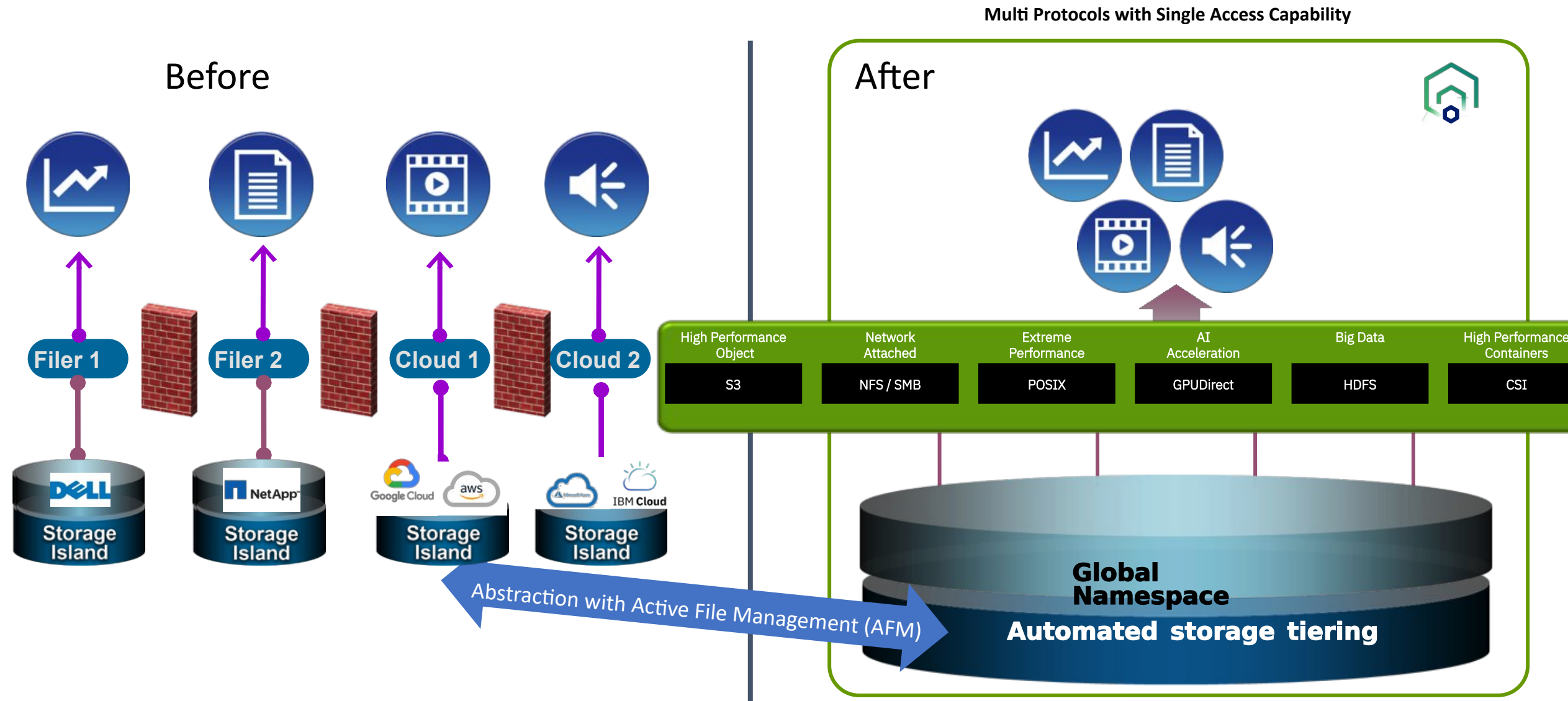
AI / Analytics require high performance,
low latency storage to keep expensive
CPU / GPU / TPU / FPGA busy.



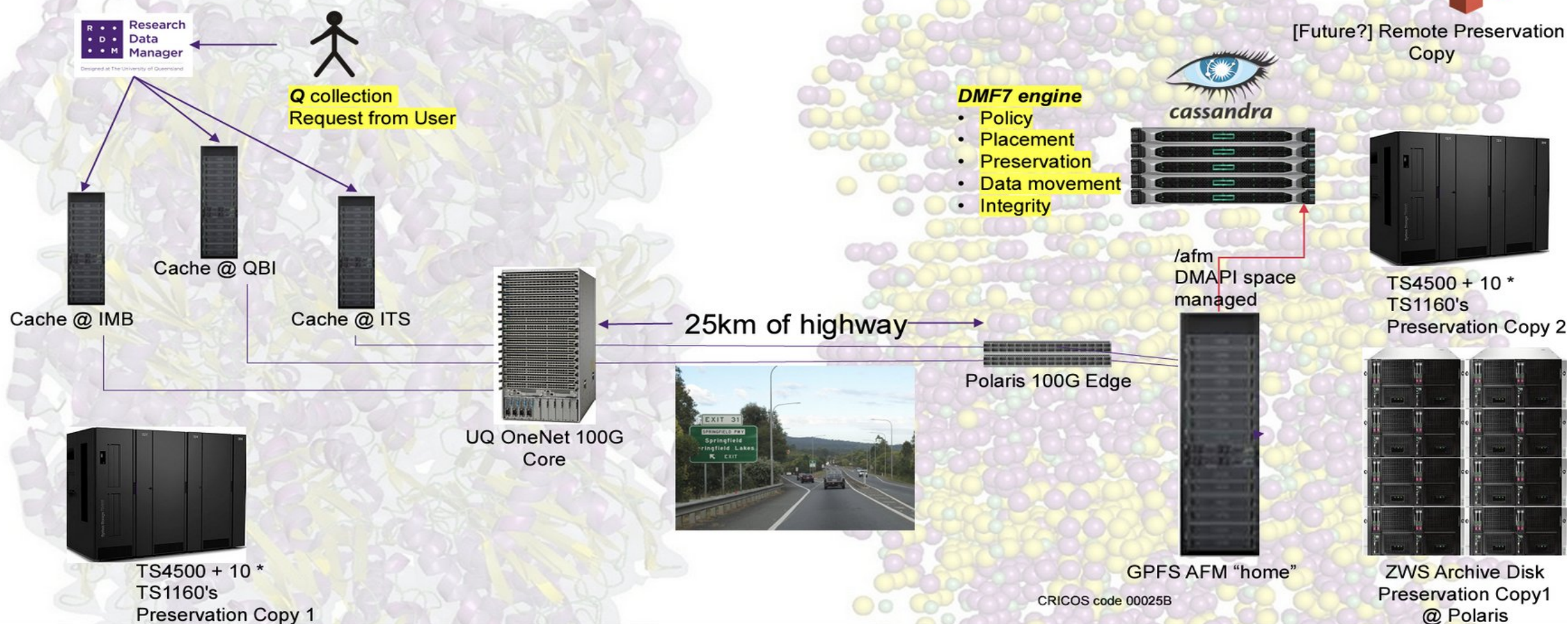
AI Data Assistant for your existing data!



Abstract your Storage with Unified Access

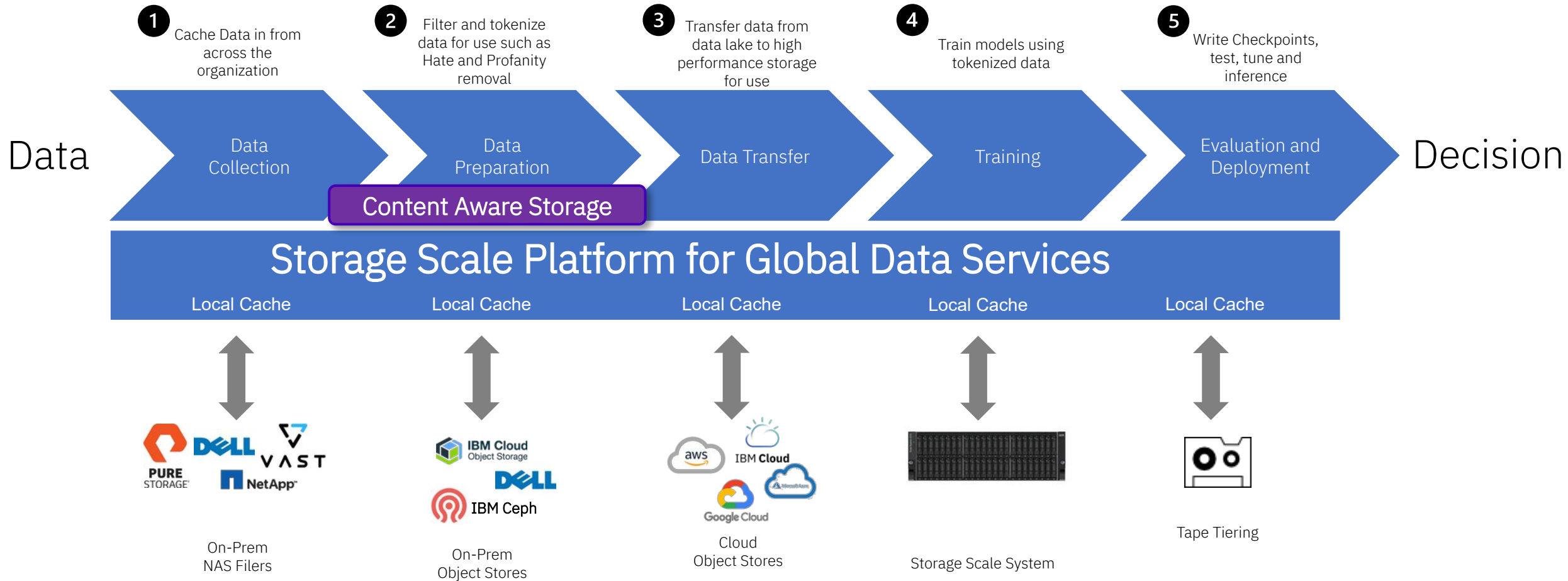


A more holistic picture of the new architecture shape...



Storage for Data and AI

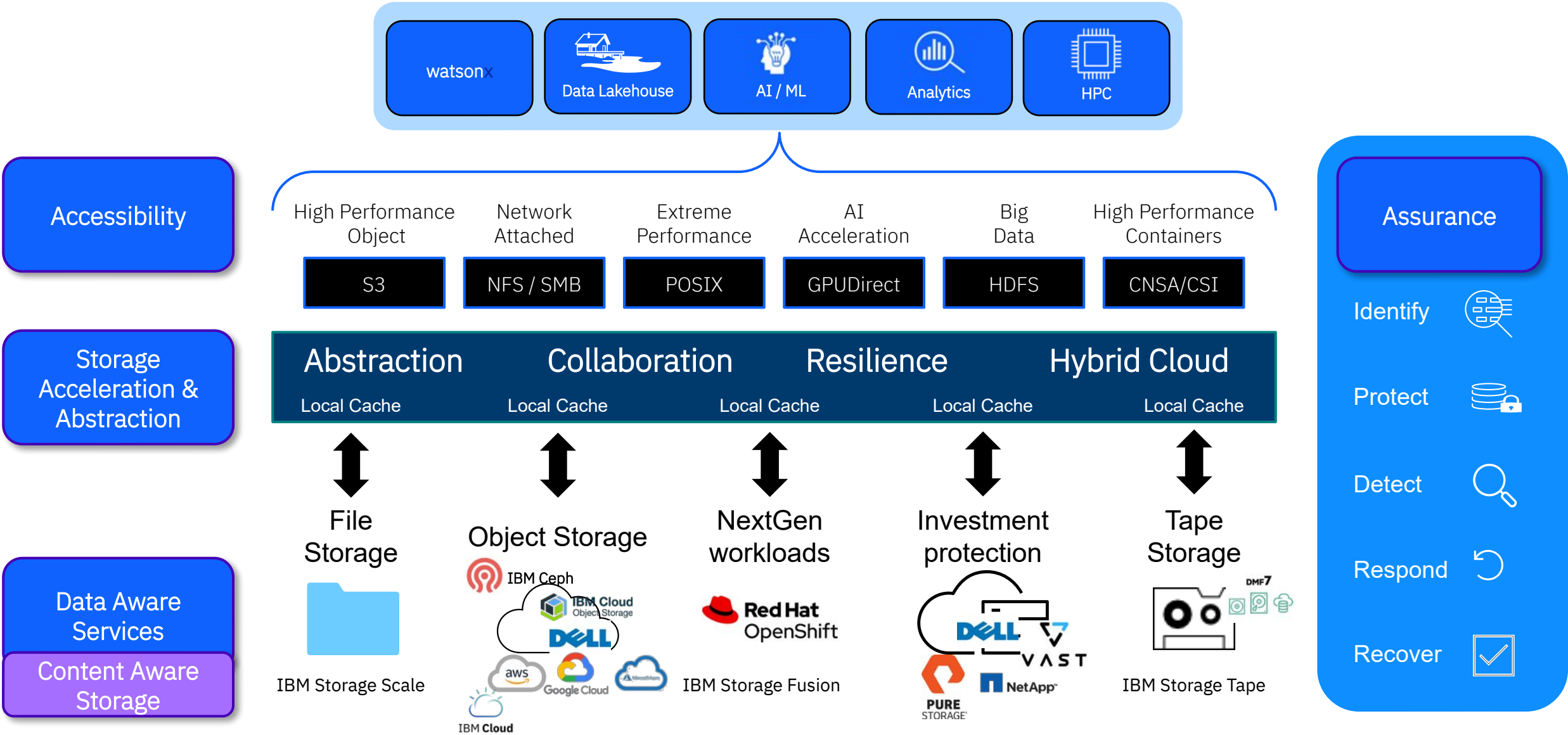
Phases of AI – Scale Optimized



Cache the latest necessary Data - No full Copies!

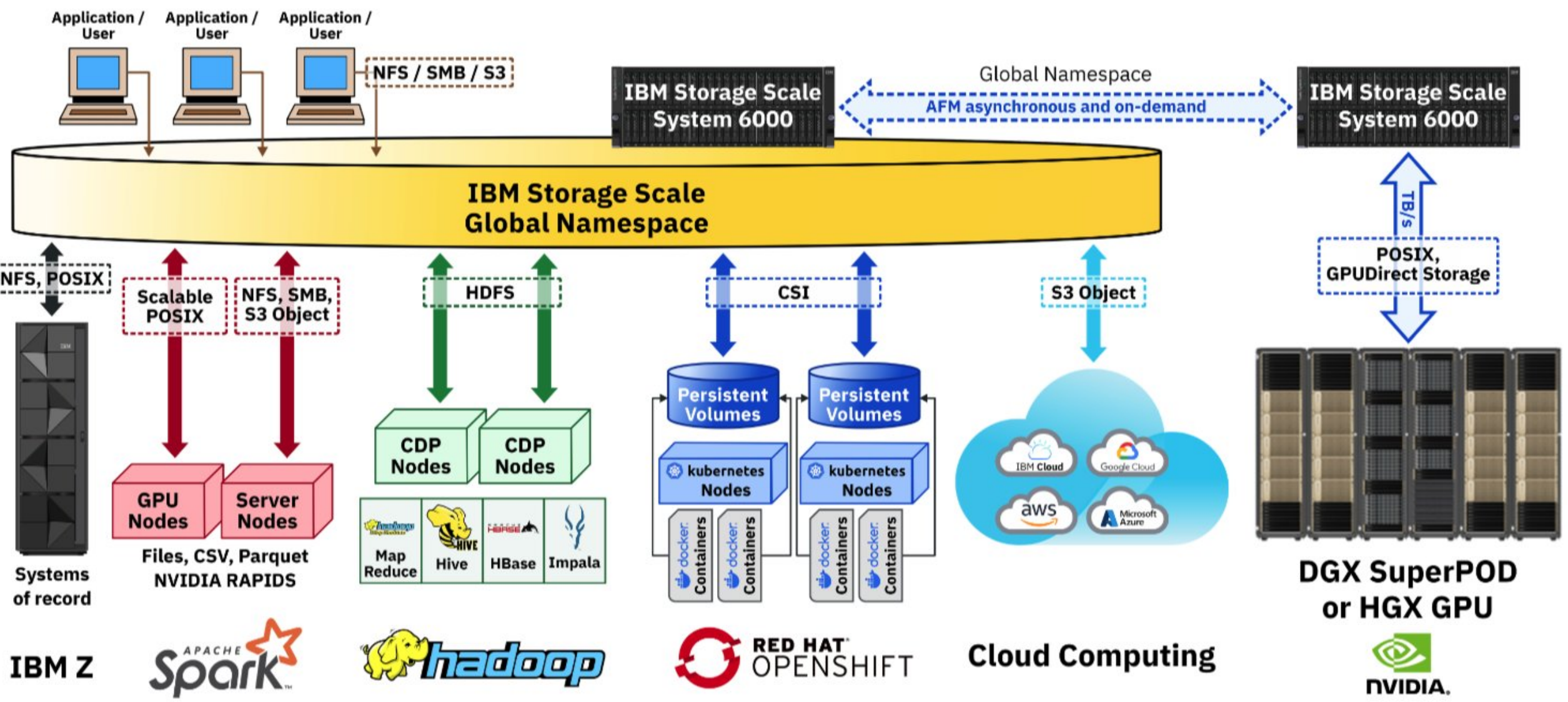
IBM Storage Scale - a global platform for storage (gpfs) services of unstructured data

1) Accessibility, 2) Storage Acceleration and Abstraction 3) Data Aware Services, and 4) Assurance



Example IBM Storage Scale global data platform

Performance and scalability optimized



MN5: IO Partition

ESS model	#ESS	Drive Capacity	Total # drives	Raw capacity	Net capacity	Read perf	Write perf
ESS 3500 Capacity model	50	NL-SAS 18TB	20400	367PB	248 PB (8+3P)	1.6TB/s (IOR 100%read)	1.2TB/s (IOR 100%read)
ESS 3500 Performance model	13	NVMe 15.36TB	312	4.79PB	2.81PB (8+2P)	600GB/s 1Mio iops 4KB	600GB/s 500K iops 4KB

<https://www.spectrumscaleug.org/wp-content/uploads/2023/08/SSUG23UK-Barcelona-Supercomputing-Center-Site-Update.pdf>



Total net storage capacity: 650 PB

Element	Element	Size
IBM TS4500	2	
Tape Enterprise	20100	400 PB
Drives	64	



IBM's Storage Scale System works with NVIDIA solutions!

<https://www.nvidia.com/en-us/data-center/dgx-superpod/>



NVIDIA
Enterprise
Reference
Architecture (ERA)



NVIDIA
DGX BasePOD



NVIDIA
DGX SuperPOD



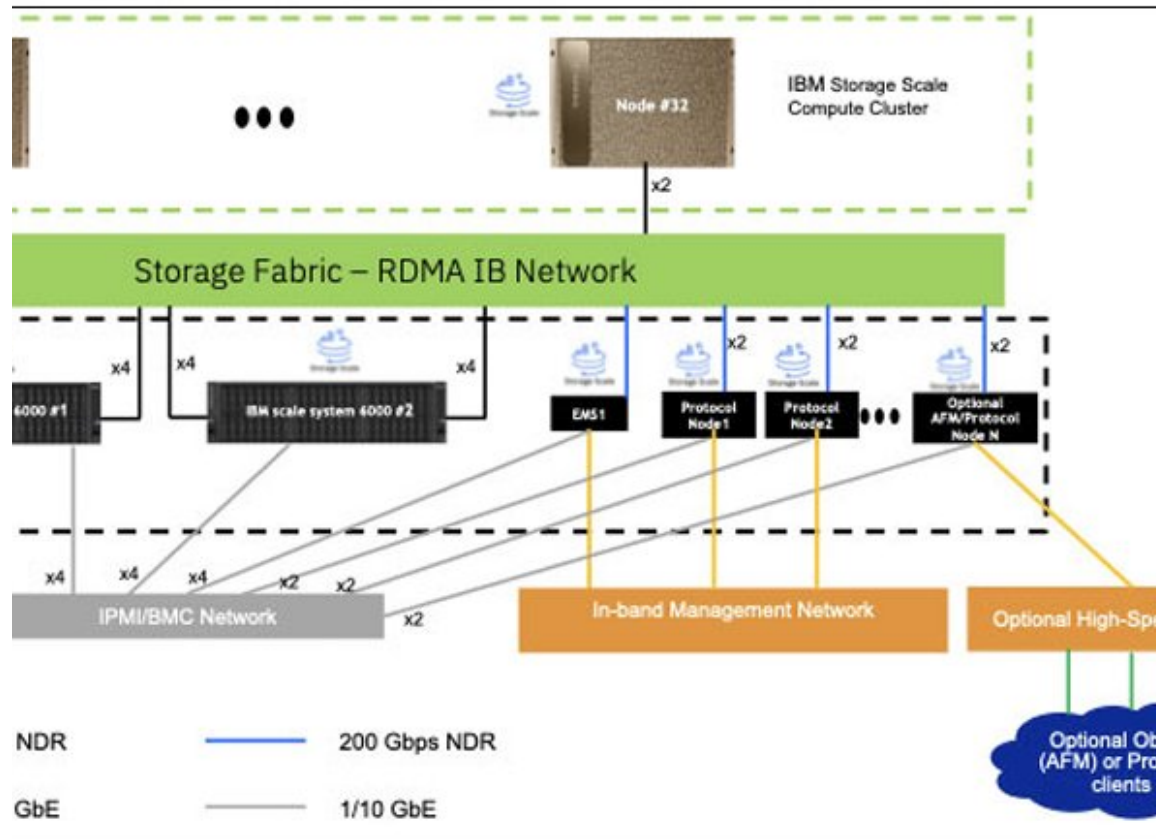
NVIDIA
Cloud Partner
(NCP)

And wait there's more! NVIDIA certified system vendor based on HGX

<https://www.nvidia.com/en-us/data-center/products/certified-systems/>

IBM and NVIDIA Redbook

<https://redbooks.ibm.com/abstracts/redp5746.html>



network diagram with optional AFM connectivity

IBM Storage Scale System 6000 with NVIDIA DGX SuperPOD Deployment Guide

Chris Maestas
Ana Gabriela Iturbe Desantis
Phillip Gerard
Kiran Ghag
Nikhil Khandewal
Matthew Kios
John Lewars
Jesus Daniel Munoz Lopez
Roger E. Sanders
Sanjay Sudam
Lindsay Todd
Joanna Wong



Julich Lab Jupiter Exascale AI: IBM Storage, NVIDIA GPU and ARM

JUPITER + IBM

IBM

A new class of
supercomputers
for AI-driven
scientific
breakthroughs

Extreme-scale computing for
AI powered by the NVIDIA
Grace Hopper™ and IBM
Storage Scale System

Hosted at the Forschungszentrum Jülich facility in Germany, JUPITER, the world's most powerful AI supercomputer, is being built in collaboration with NVIDIA, ParTec, Eviden and SiPearl to accelerate the creation of foundational AI models in climate and weather research, material science, drug discovery, industrial engineering and quantum computing.



6000 compute nodes to achieve the compute performance of 1 ExaFLOP/s, featuring NVIDIA Grace-Hopper processors. To keep the data flowing, a 21 Petabyte Flash Module (ExaFLASH) is provided based on the IBM *Storage Scale* software and a corresponding storage appliance based on IBM ESS 3500 building blocks.

More information @ <https://www.fz-juelich.de/en/ias/jsc/jupiter/tech>

Scale System models are built for speed and capacity



IBM Storage Scale System 6000

4U48 Enclosure
24 or 48 NVMe drives



ISS Performance
Machine type 5149-F48

Up to 310+ GB/s
IOR 100% read, InfiniBand



Tech Days

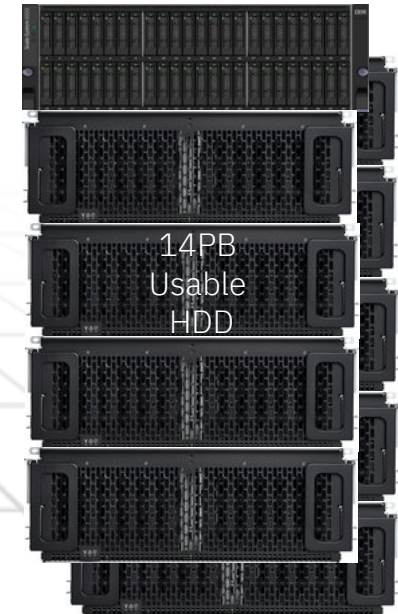
IBM Storage Scale System 6000 Hybrid (NVMe + HDD)



Up to 9x 4U91 JBOD
ISS 6000 Hx
Machine type 5149-F48

Up to 280 GB/s – NVMe
IOR 100% read InfiniBand
Up to 90 GB/s – HDD
IOR 100% read InfiniBand

IBM Storage Scale System 6000 Capacity (HDD-Only)

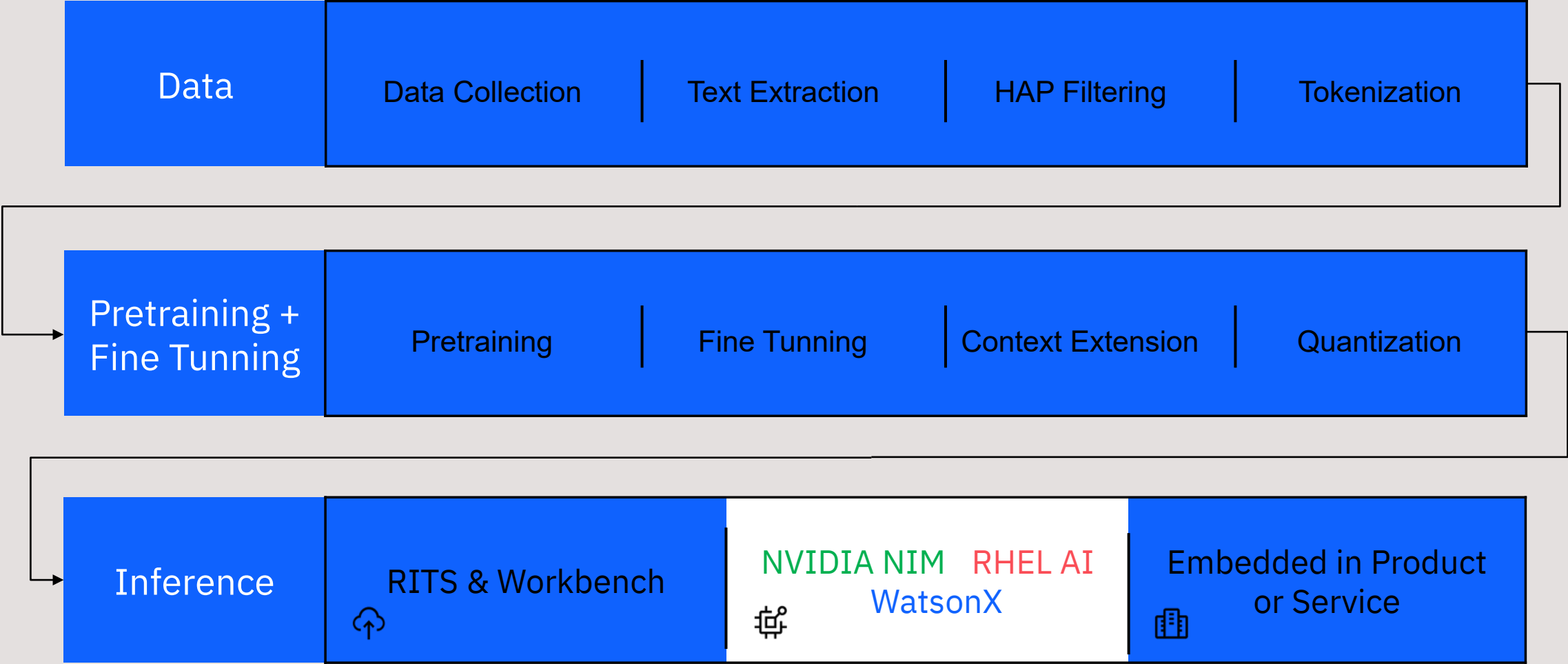


Up to 9x 4U91 JBOD

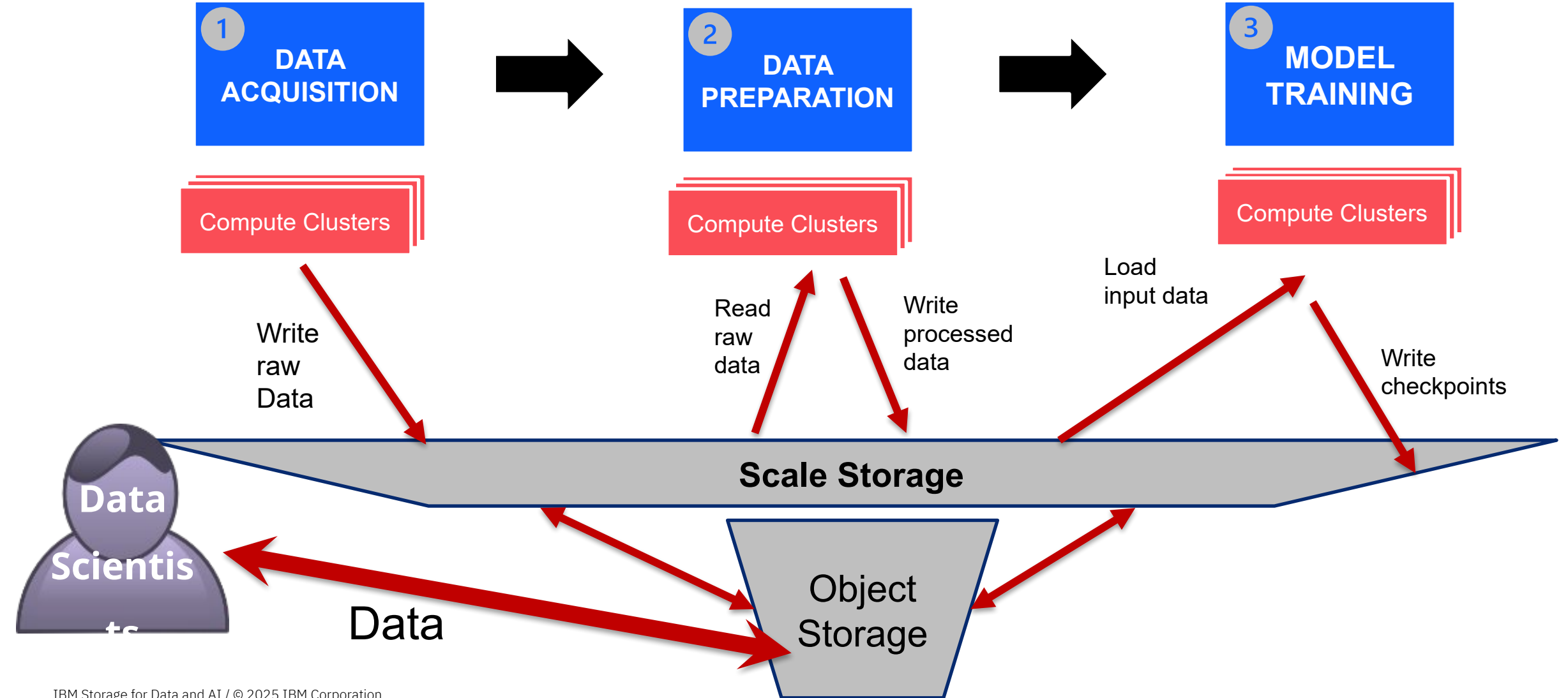
ISS 6000 Cx
Machine types
5141-FN2

Up to 90 GB/s
IOR 100% read C9 model, InfiniBand

AI Workflows @ IBM

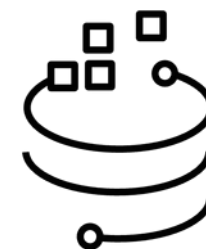
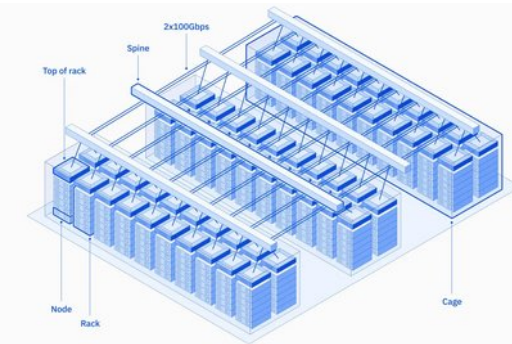


Data Access in HPC and AI workflows @ IBM



Vela: AI Supercomputer in IBM Cloud

- IBM Research needs to train, fine-tune, and experiment w/ ML models
 - Large language models, foundational models – billions of parameters
- Let's build AI Supercomputer in IBM Cloud
- Virtual Machines (VSIs) as a building block (not baremetal)
- New type of servers added to IBM Cloud
 - 200 nodes, 8× NVIDIA A100 80GB
- Red Hat OpenShift (OCP) for resources and training job management
 - MCAD Scheduler for training jobs
https://pytorch.org/torchx/latest/schedulers/kubernetes_mcad.html
(not <https://pytorch.org/torchx/latest/schedulers/lsf.html>)
- May 2024: IBM Open sources IBM's Granite
 - <https://research.ibm.com/blog/granite-code-models-open-source>



Storage Scale



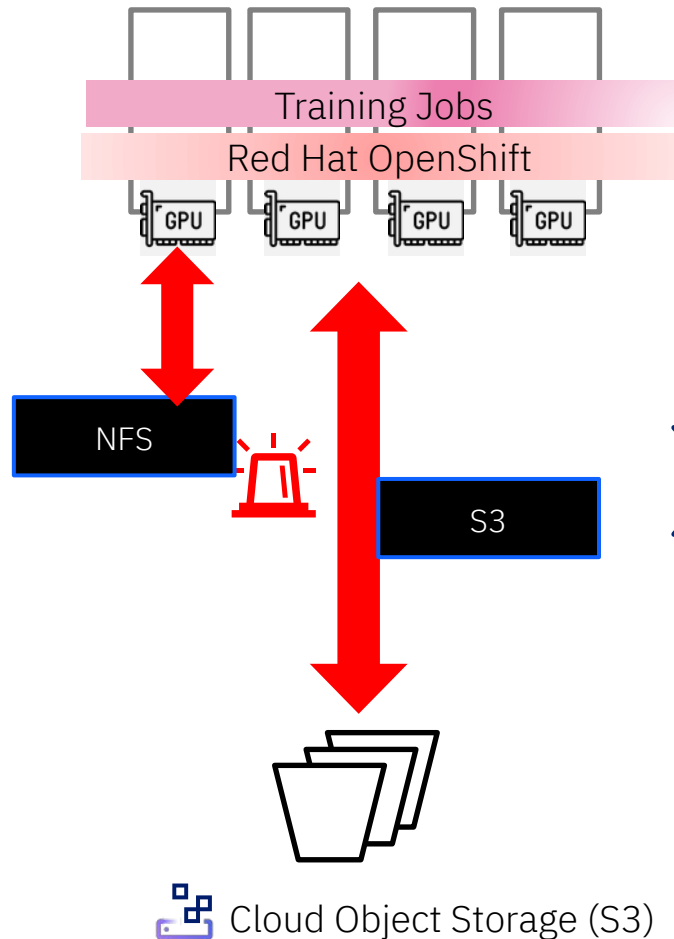
Open Shift



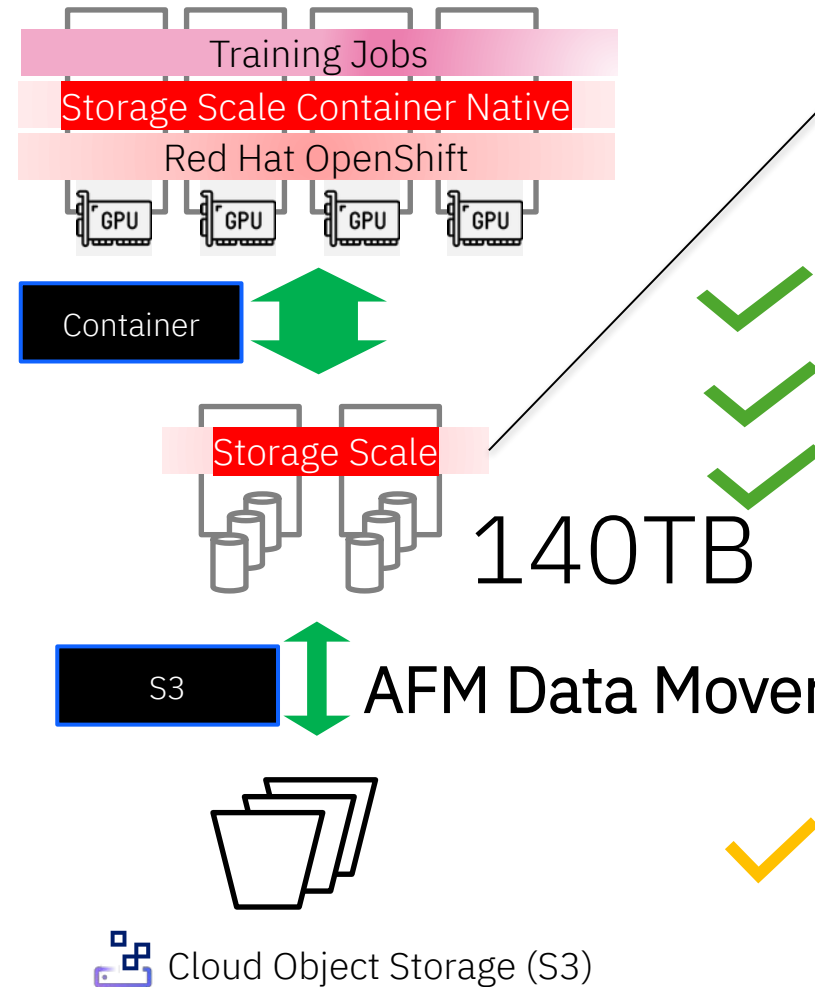
IBM Cloud

Multi-user Elastic Cache w/ Storage Scale

BEFORE



NOW



Large, persistent,
and high-performance
data cache dedicated
to the AI cluster

- 1. Fast data load from cache
- 2. Fast checkpoints
- 3. No backend overload

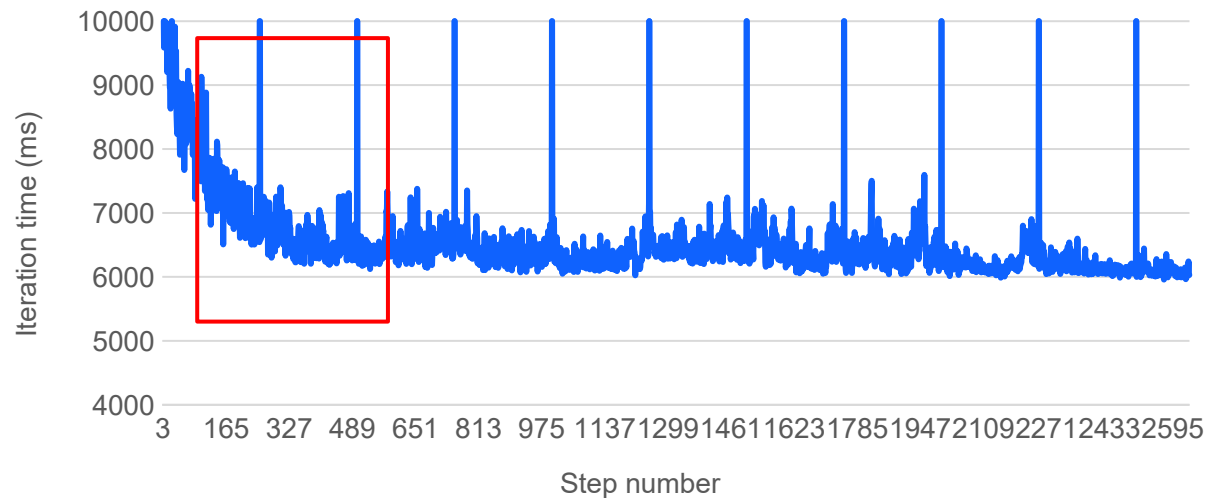
AFM Data Mover - Critical Differentiator

- 4. Automated data
movement

Results

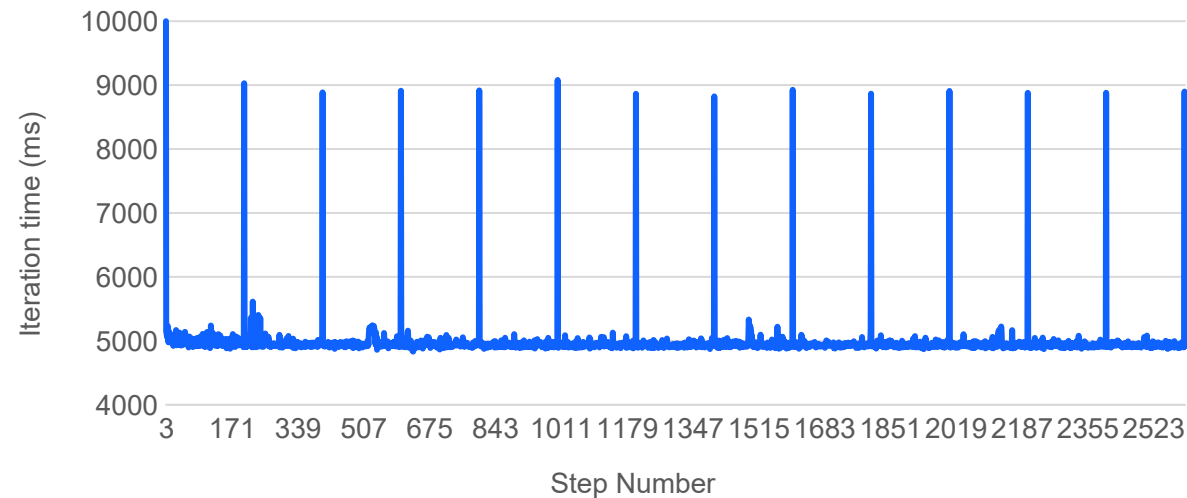
- ✓ Checkpoint times are **3x shorter** compared to COS. E.g., 50B takes ~50 seconds vs 3 minute
- ✓ Input read times are less variable

Granite-13B-4k iteration times with NFS



Step times: 9 – 6 seconds, up to 50% variation

Granite-8B-4K-iteration times with Spectrum Scale File system

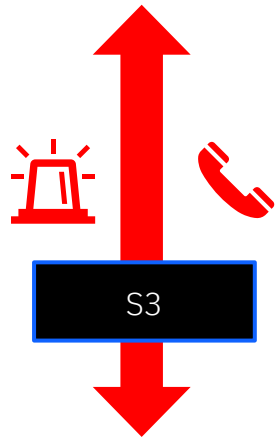
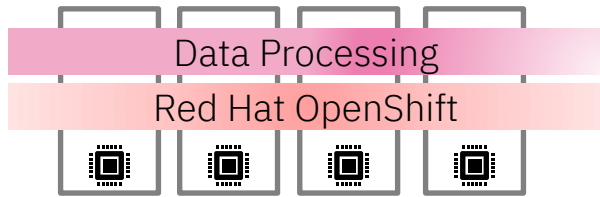


Step times: 4.8 to 5.2 seconds, less than 10% variation

- ✓ Job restart no longer takes a long time to achieve steady state
- ✓ Step time variation reduces from 50% to under 10%

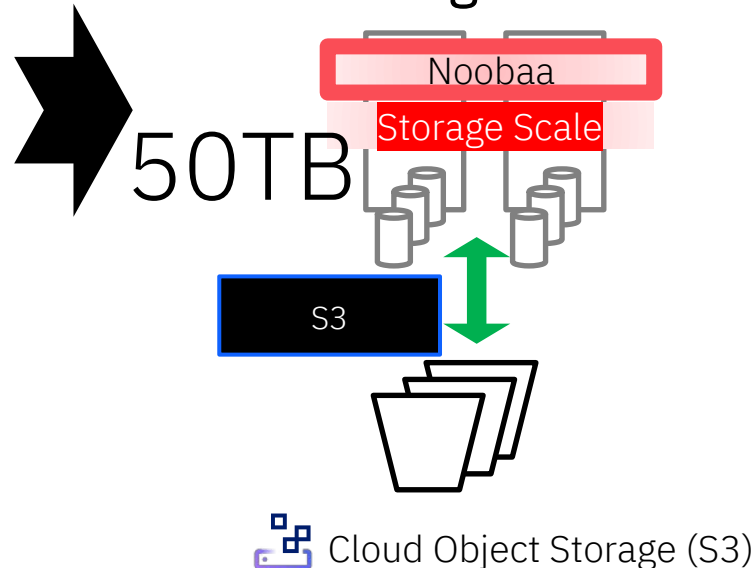
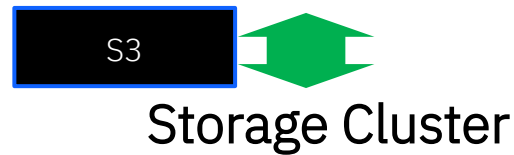
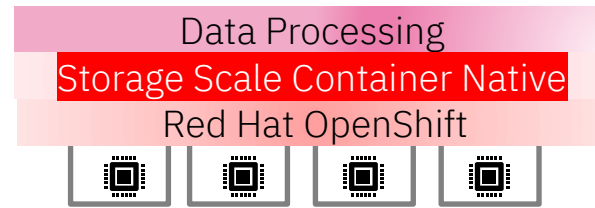
Storage for Data Preparation

Compute Cluster (ROKS)



 Cloud Object Storage (S3)

Compute Cluster (ROKS)



 Cloud Object Storage (S3)

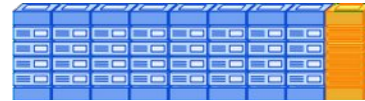
- Embarrassingly parallel workload for which one can deploy larger compute clusters to accelerate data preparation
- **Problem.** When using over 70 compute nodes, data preparation applications overload COS backend.
Limits the speed of data preparation.
- Deploying cache allows to flatten I/O burst, not overload COS, and run data preparation at larger scale – 200 nodes (target).
 - Lots of data reuse
- Data factory applications require object storage interface (S3), not file system. Scale provides object storage (S3) interface through Noobaa technology
 - No need to deploy Scale Cloud Native ☺
- Status
 1. Functional PoC completed
 2. Performance PoC in progress: Larger scale, reduced peak load on COS, higher throughput

IBM Blue Vela- HGX “SuperPOD” Storage Fabric (IBM Cloud/ IBM Research/NVIDIA/Dell) working together to delivery an AI solution

DELLTechnologies



Storage Scale 6000



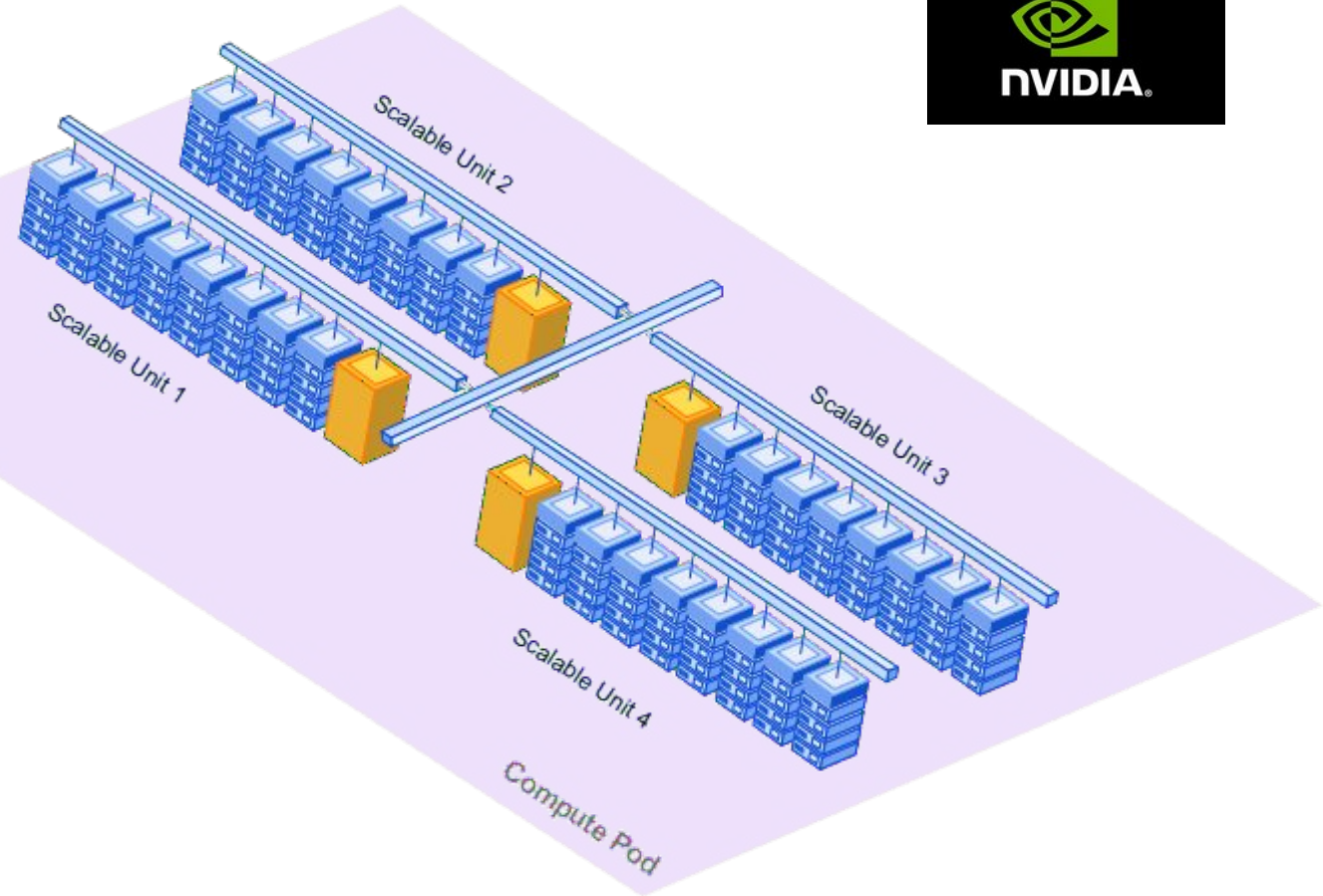
Scalable Unit

- 32 Compute Nodes
- 256 H100 GPUs



Compute Pod

- 4 Scalable Units
- 128 Compute Nodes
- 1024 H100 GPUs
- 82 TB of GPU Ram
- 12,288 Physical Cores
- 256 TB of RAM
- 3481 TB NVME Local Storage



IBM Storage Scale System 6000 Now a Certified NVIDIA Cloud Partner



<https://community.ibm.com/community/user/storage/blogs/mike-kieran/2025/01/10/ibm-storage-scale-system-6000-now-a-certified-nvidia>

IBM Storage Scale System 6000 is now a certified NVIDIA Cloud Partner (NCP) for HGX H100/H200/B200 systems. As a certified high performance storage partner for NCP, IBM Storage Scale System 6000 has demonstrated that it can deliver scalable high-performance IO to the most demanding AI training and inferencing workloads deployed on NVIDIA HGX GPUs in the cloud.



“The supercomputer will leverage **IBM Storage Scale System 6000 flash** technology to deliver high-performance storage for AI, data analytics, and other demanding workloads.

As part of this agreement, CoreWeave customers can access the IBM Storage platform within CoreWeave’s dedicated environments and AI cloud platform.”

CoreWeave Partners with IBM to Deliver New AI Supercomputer for IBM Granite Models



NEWS PROVIDED BY
CoreWeave →
Jan 15, 2025, 08:00 ET

<https://www.prnewswire.com/news-releases/coreweave-partners-with-ibm-to-deliver-new-ai-supercomputer-for-ibm-granite-models-302351465.html>

- One of the first deployments of NVIDIA GB200 NVL72 at supercomputing scale
- Supercomputer will leverage IBM Storage Scale System to power AI research and development



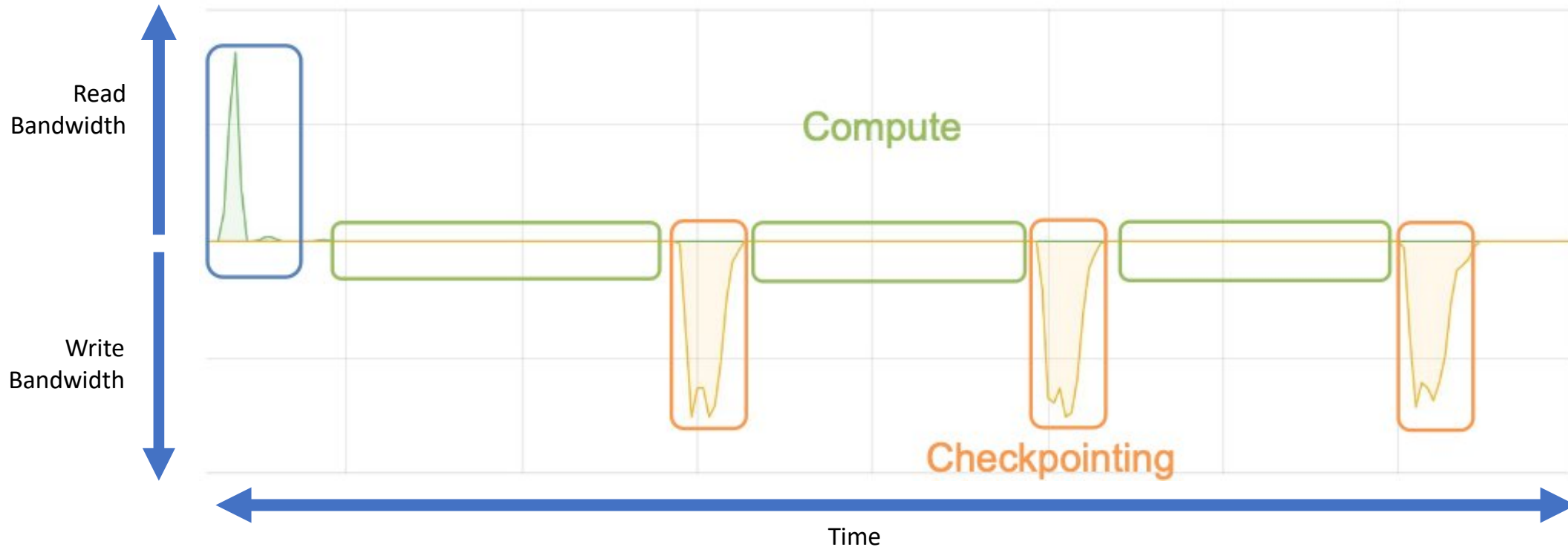
LLM I/O Patterns

3 Phases

Init read phase – Model Load

Compute Phase – Iterative GPU Phase

Checkpoint write phase – every N compute iterations (buffered I/O)



LLM Data Set Size

Checkpoint time reduced to 1% an hour => 36 seconds

Model Specific Example for Synchronous

Checkpoint

Tensor Model Parallel Size determines how many GPUs participate in the checkpoint.

For example, If Tensor Parallel size is set to 8, 1 out of 8 GPUs will participate in the checkpoint

~14 bytes per Parameter

Example

175B Parameter Model: ~2.4TB Data set size

512B Parameter Model: ~7.2TB Data set size

1T Parameter Model: ~14TB Data set Size

3 x IBM Storage Scale 6000 load in 36 seconds

Total Number of GPUs is 4000 GPUs

Only 512 GPUs will participate in the checkpoint
($4000_GPUs / 8_Tensor_Parallel_Size$)

175B: ~4.8GB data set / GPU

512B: ~15GB data set / GPU

1T: ~28GB data set / GPU

Model Load

Using the same Tensor Parallel Size of 8 from the checkpoint

8x the data set size will need to be loaded across all GPUs

Example

175B Parameter Model: ~19TB Data set size

512B Parameter Model: ~58TB Data set size

1T Parameter Model: ~110TB Data set Size

3 x IBM Storage Scale 6000 load in < 2 min

Total Number of GPUs is 4000 GPUs

Data set per GPU is the same, but now all GPUs participate in the Model Load

175B: ~4.8GB data set / GPU

512B: ~15GB data set / GPU

✓^T **READ: 310+ GB/s**
✓ **WRITE: 155+ GB/s**



Checkpointing Large Models

Parameters	Nodes (8 GPUs/node)	GPIUs	Total Size of Checkpoint	Write Bandwidth	Per Thread Write	Per Node Write	Overall Write Performance	Number of Storage Scale Systems
175B	128	1,024	2.45 TB	68.06 GB/s	0.53 GB/s	4.25 GB/s	~155 GB/s	1 Scale System 6000
530B	256	2,048	7.42 TB	206.11 GB/s	0.74 GB/s	5.89 GB/s	~300 GB/s	2 Scale System 6000
1T	512	4,096	14 TB	388.89 GB/s	0.76 GB/s	6.08 GB/s	~450 GB/s	3 Scale System 6000
4T	2048	16,384	56 TB	1,555.7 GB/s	0.76 GB/s	6.08 GB/s	~1650 GB/s	11 Scale System 6000
8T	4096	32,768	112 TB	3,111.1 GB/s	0.76 GB/s	6.08 GB/s	~3150 GB/s	21 Scale System 6000

- Number of bytes per parameter: **14**
- Checkpoint write time percentage of total training time: **1%**
- Checkpoint interval: **1 hour**

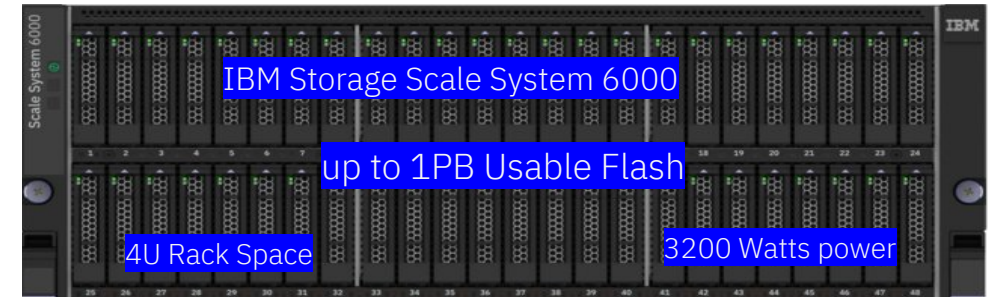
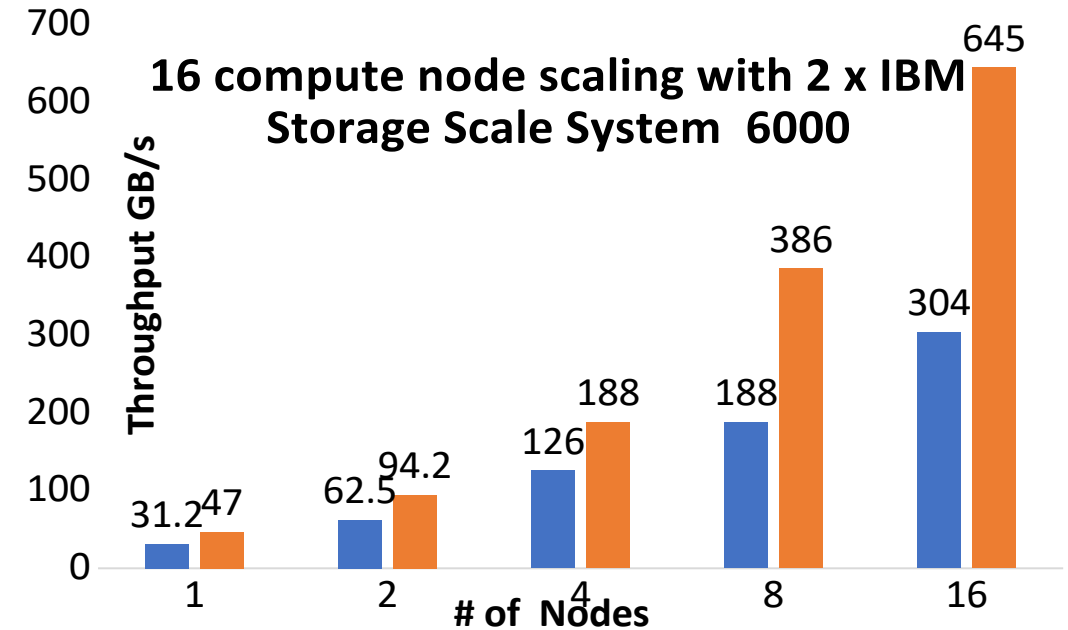
NVIDIA DGX SuperPOD solution with IBM Storage Scale System 6000

NVIDIA SuperPOD Storage Guidelines <https://docs.nvidia.com/dgx-superpod/reference-architecture-scalable-infrastructure-b200/latest/storage-architecture.html>

Performance Characteristics	NVIDIA Storage Guidance		IBM Storage Sizing Guidance
	Standard (GBps)	Enhanced (GBps)	
Single SU aggregate system read	40	125	1x IBM Storage Scale system 6000 Performance model Up to 320 GB/s read & 150 GB/s write
Single SU aggregate system write	20	62	
4 SU aggregate system read	160	500	2x IBM Storage Scale system 6000 Performance model Up to 640 GB/s read & 300 GB/s write
4 SU aggregate system write	80	250	

- High performance, resilient, parallel filesystem storage recommended for multi-threaded read and write operations across multiple nodes
- Peak performance for writes and read are needed for creating and reading checkpoint files
- Concurrent filesystem access minimizes step time

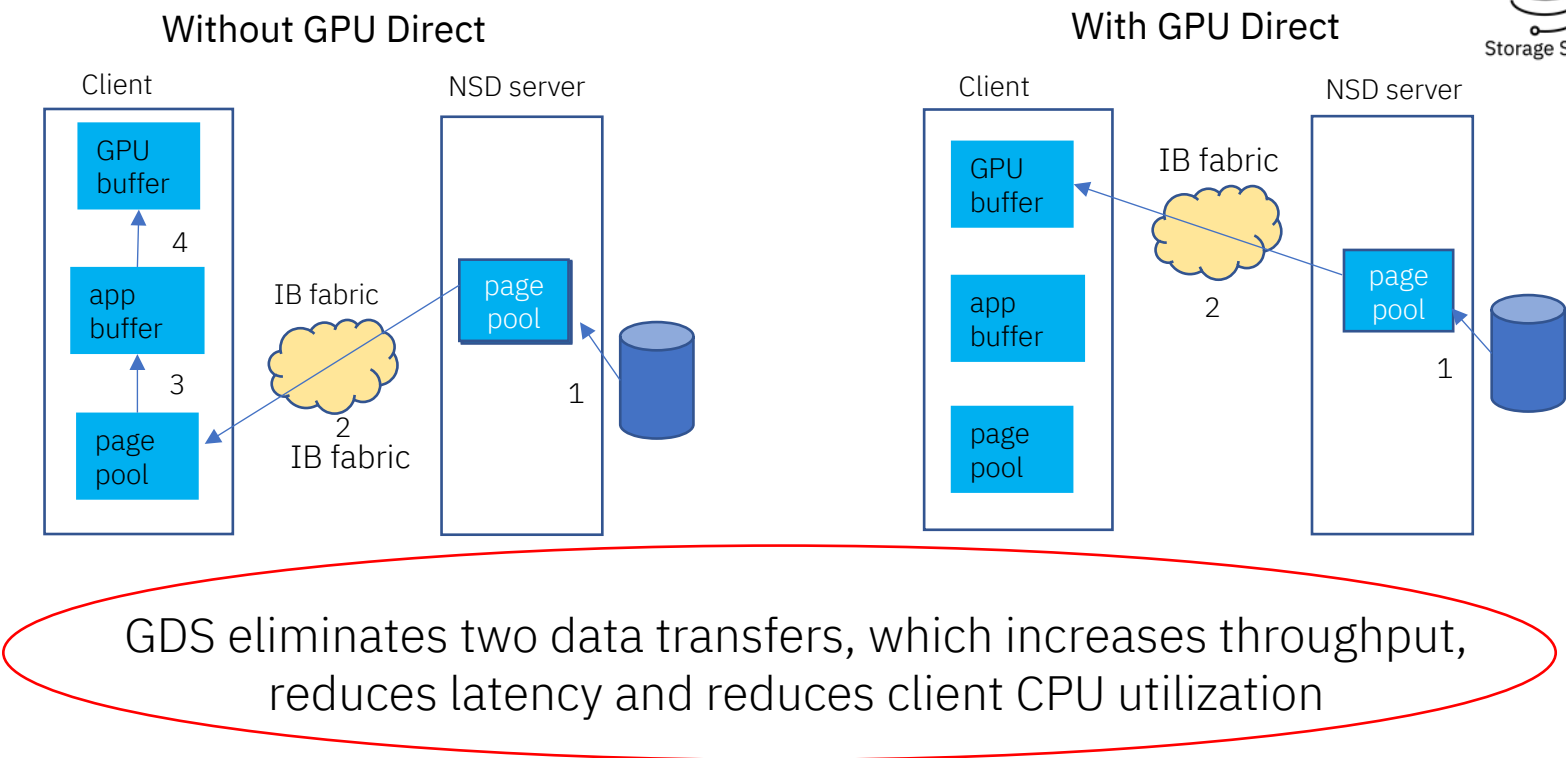
IBM Storage Scale system 6000 delivers “Enhanced” in class performance



IBM Storage Scale

GPU Direct Storage

GPUDirect Storage (GDS) for Storage Scale enables an RDMA (remote direct memory access) path between GPU memory and storage.



- NVIDIA Magnum IO
- Family of I/O Optimizations for GPU accelerated data centers.
- GPUDirect RDMA: Access peer node's memory without copying through host memory
- GPUDirect Storage: Transfer data between GPU and storage without involving CPU and CPU memory

- NVIDIA CUDA Toolkit
- GDS API is in the CUDA toolkit
- A development environment for GPU accelerated applications
- Libraries, compilers, debuggers, optimizers and tools
- Leading GPU compute platform since 2006

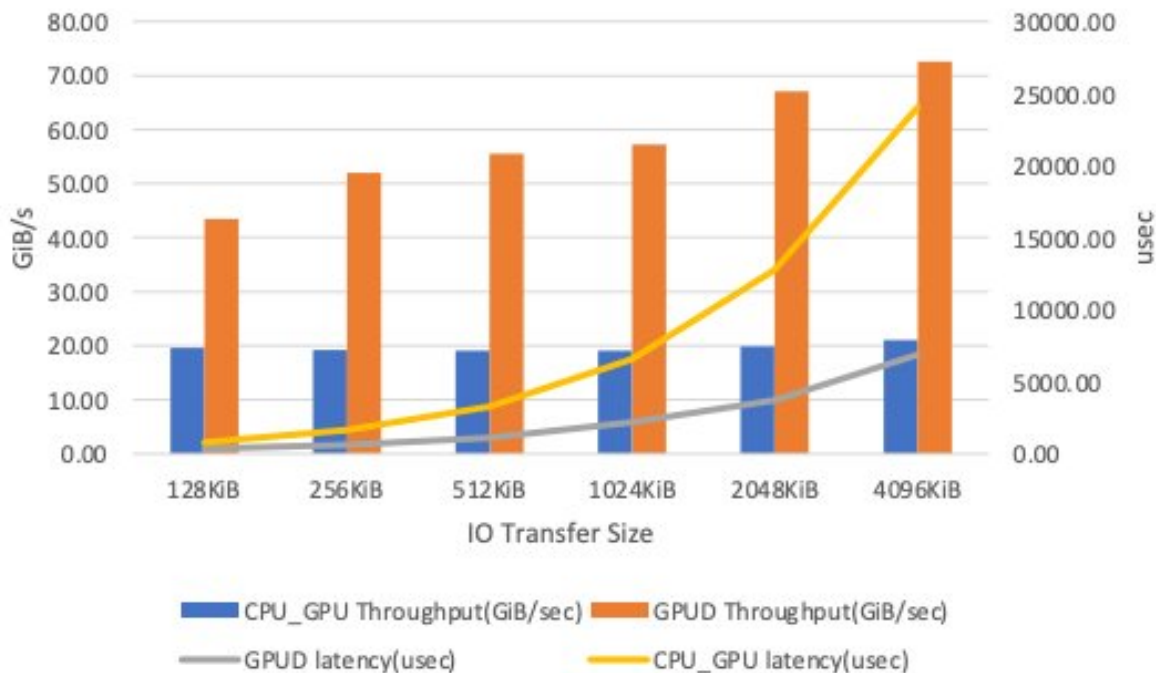
- GDS for Applications
- Invoked using the CUDA Toolkit (cuFile) API
- GDS APIs must be explicitly called by the applications
- Storage must be GDS enabled. If not, GDS APIs use regular data movement.

- Why it matters
- AI, HPC, ML and analytics are data hungry and require a very high data throughput.
- GPUs can be starved by slow I/O due to multiple data transfers on the client.

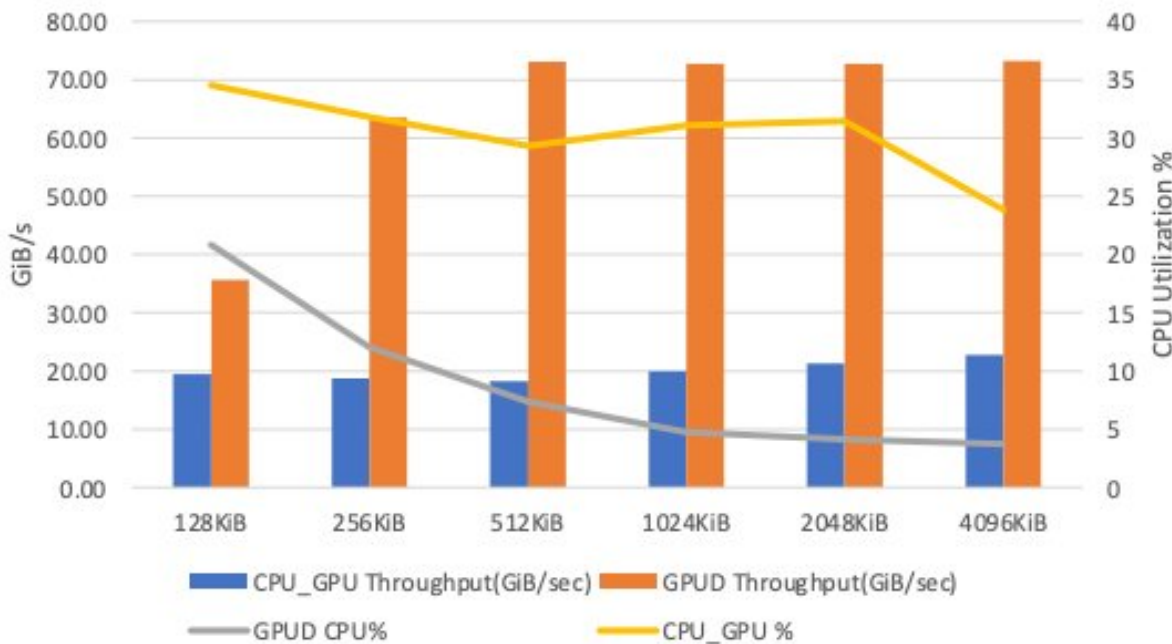
IBM Storage Scale

GPU Direct Storage Comparison

GPU Direct IO comparison with latencies



GPU Direct IO comparison with CPU Utilization



Note : 16 Threads per GPU; Total 128 threads for read

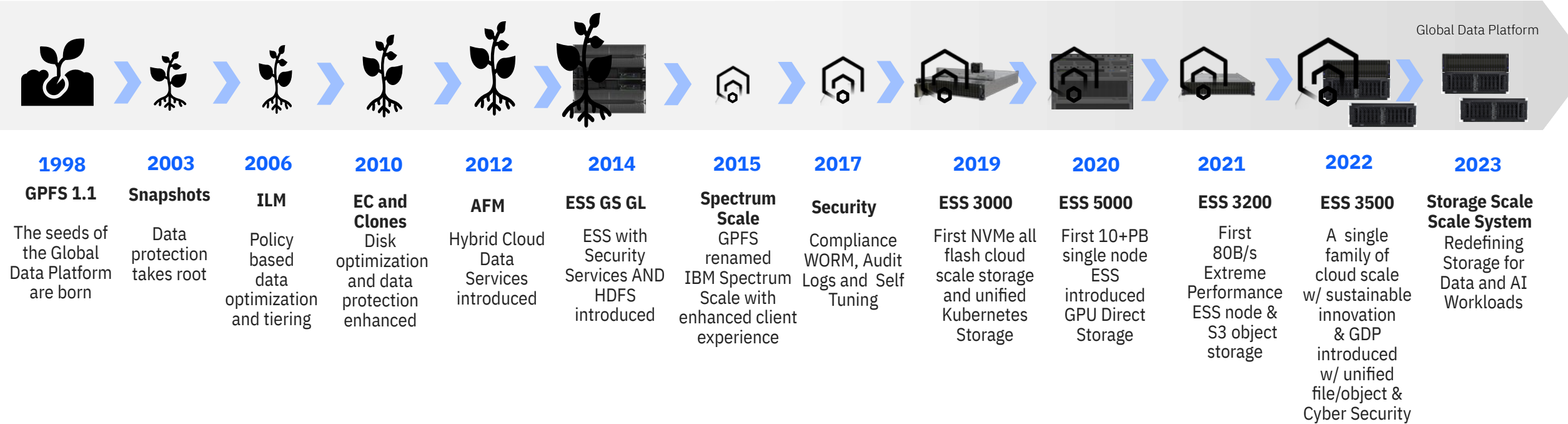
Up to 3.5x Higher Bandwidth; 50% reduction in latencies

IBM Storage Scale has Evolved for Over 25+ Years

Unified Data and Parallel Storage Access that Redefines a Global Platform For Services

Spanning a Millennium of Storage Innovation

8192 Disks 7GB/s 1 Billion Files Scanned in 3 hrs. Over 100,000 Systems Worlds Fastest Supercomputer Data Over 6,000 Customers 80GB/s 720TB 2u 91B/s 2u



IBM Storage Scale Developer Edition

<https://www.ibm.com/products/storage-scale>

IBM Storage Scale

Accelerate AI and unlock value from your data



17 Reviews - G2 Crowd

Try the free developer edition →

Schedule a free demo →



Scale User Group

The Scale (GPFS) User Group is free to join and open to all using, interested in using or integrating IBM Storage Scale.

The format of the group is as a web community with events held during the year, hosted by our members or by IBM.

See our web page for upcoming events and presentations of past events. Join our conversation via mail and Slack.

www.storagescale.org

IBM Storage Scale Developer Edition Labs

Resources

★★★★★ (1)

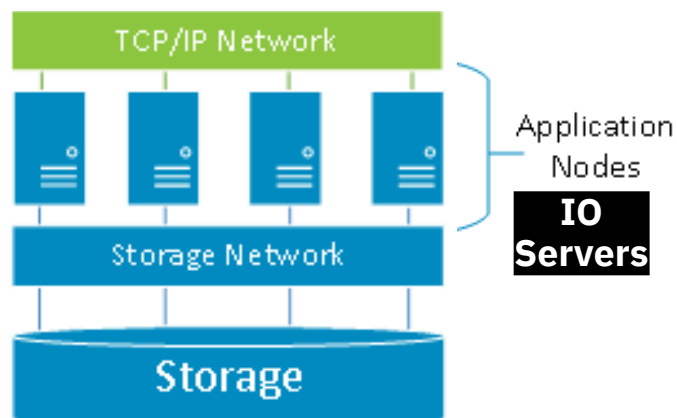
Rate this resource

Edit

Aug 25, 2024 Ibmcloud 2: us-east, us-south, ca-tor, eu-gb, eu-de, jp-tok, jp-osa, eu-es IBM Storage Scale Developer Edition - Installation Experience IBM Storage Scale Developer Edition - Installation Lab	Aug 25, 2024 Ibmcloud 2: us-south, us-east, ca-tor, eu-de, eu-gb, jp-tok, jp-osa, eu-es IBM Storage Scale Developer Edition Experience IBM Storage Scale Developer Edition Installed on a 5 node system consisting of a GUI, 2 clients and 2 storage servers.	Aug 25, 2024 Ibmcloud 2: us-south, us-east, ca-tor, eu-de, eu-gb, jp-tok, jp-osa, eu-es IBM Storage Scale Developer Edition Lab - Cyber Security Experience with IBM QRadar IBM Storage Scale Developer Edition Installed on a 5 node system consisting of a GUI, 2 clients and 2 storage servers along with IBM QRadar. Visibility IBMers, Business Partners	Aug 25, 2024 Ibmcloud 2: us-south, us-east, ca-tor, eu-gb, eu-de, eu-es, jp-tok, jp-osa IBM Storage Scale High Availability Experiences Setup clusters for: 1. Erasure Coding 2. Active File Management Disaster Recovery 3. an RPO =0 Active/Active Stretch Cluster 4. a multi-cluster remote mount with AFM-POSIX or NSD remote mount Visibility IBMers, Business Partners	ew amples 15 ata
		Reserve	Reserve Platform powered by Storage Scale	

Scale Deployment model comparison

Storage Area Network (SAN) (NVMeoF, Fiber Channel, iSCSI)



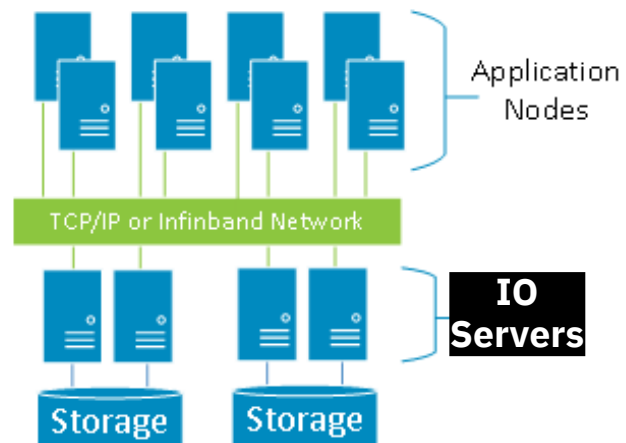
PureScale

Sailfish

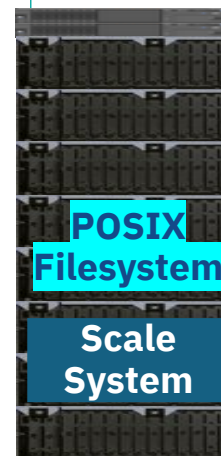
DS8k SDS
Scale

Unify and parallelize storage silos

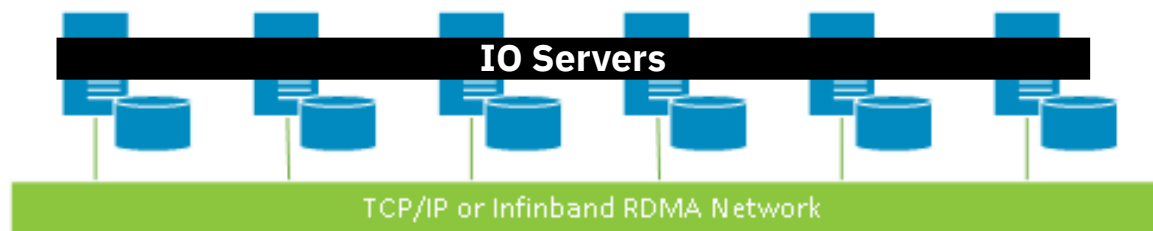
Twin tailed storage with erasure coding



Modular High-Performance Scaling



Shared Nothing Cluster (SNC) Model (Storage Rich Servers (replication, erasure code))



IBM Storage Ceph

IBM Cloud
Object Storage
(COS)

Hadoop

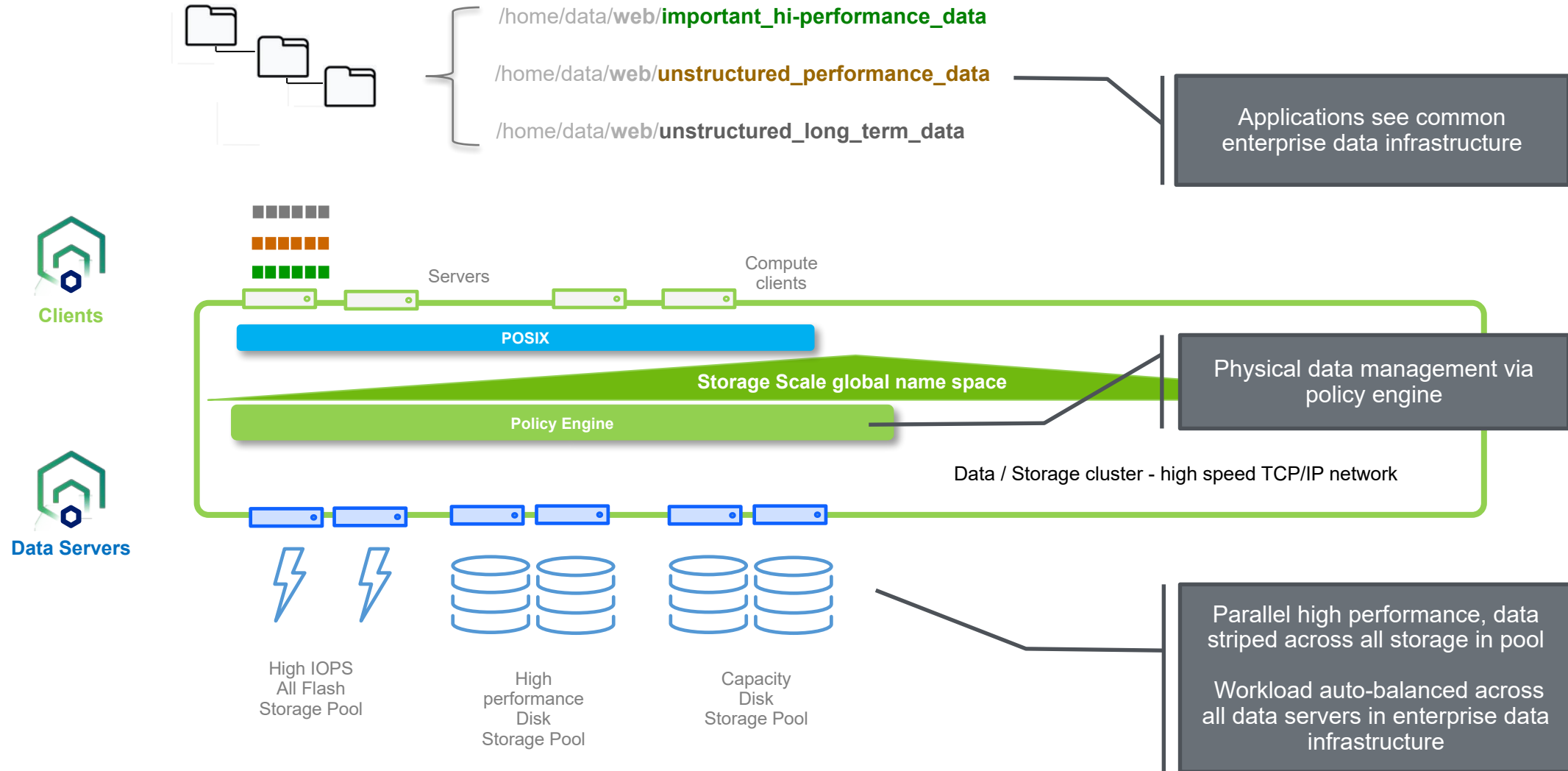
Scale (FPO)

Scale Erasure
Code Edition (ECE)

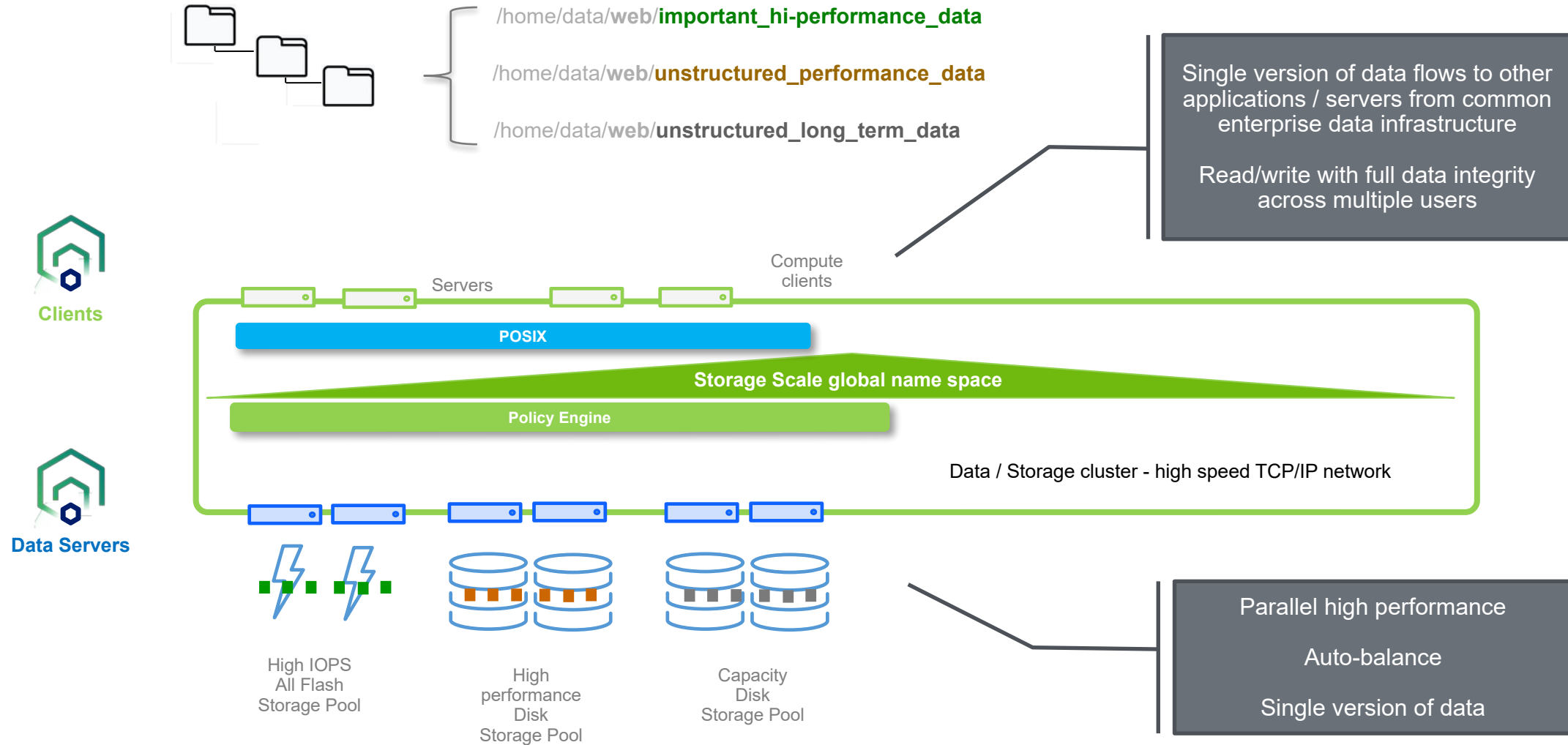
Fusion HCI

Span storage rich servers for converged architecture or HDFS deployment

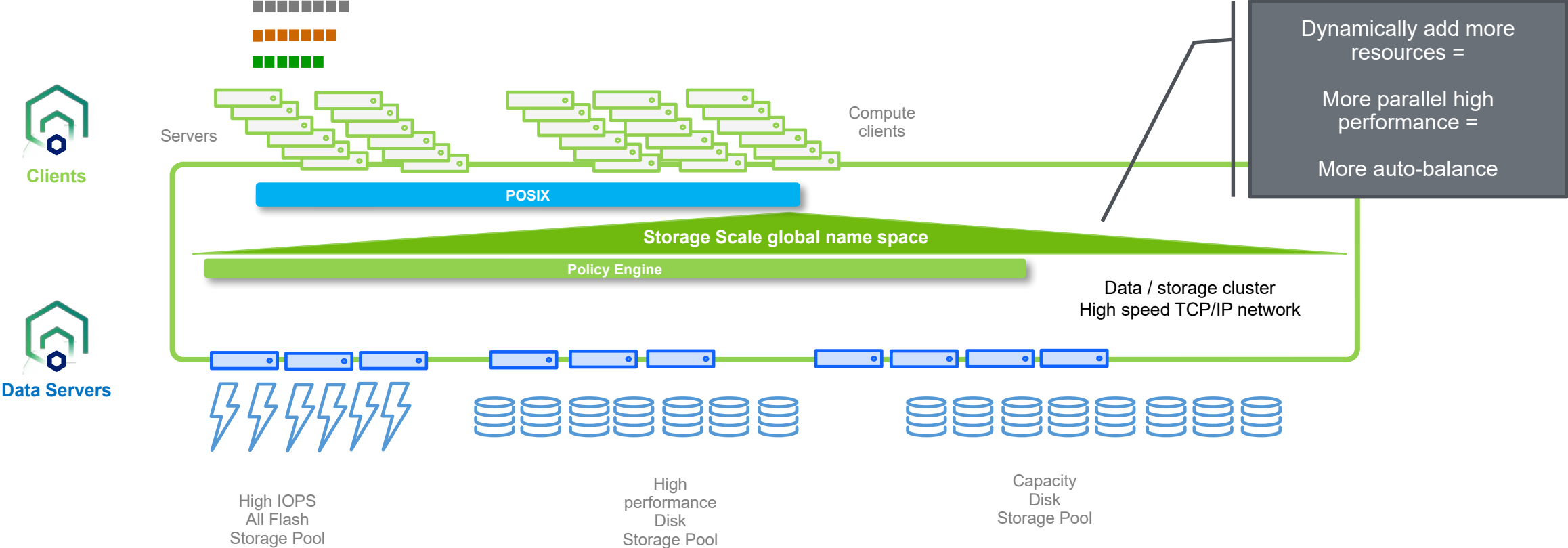
Storage Scale foundations 1 – place data where you need it!



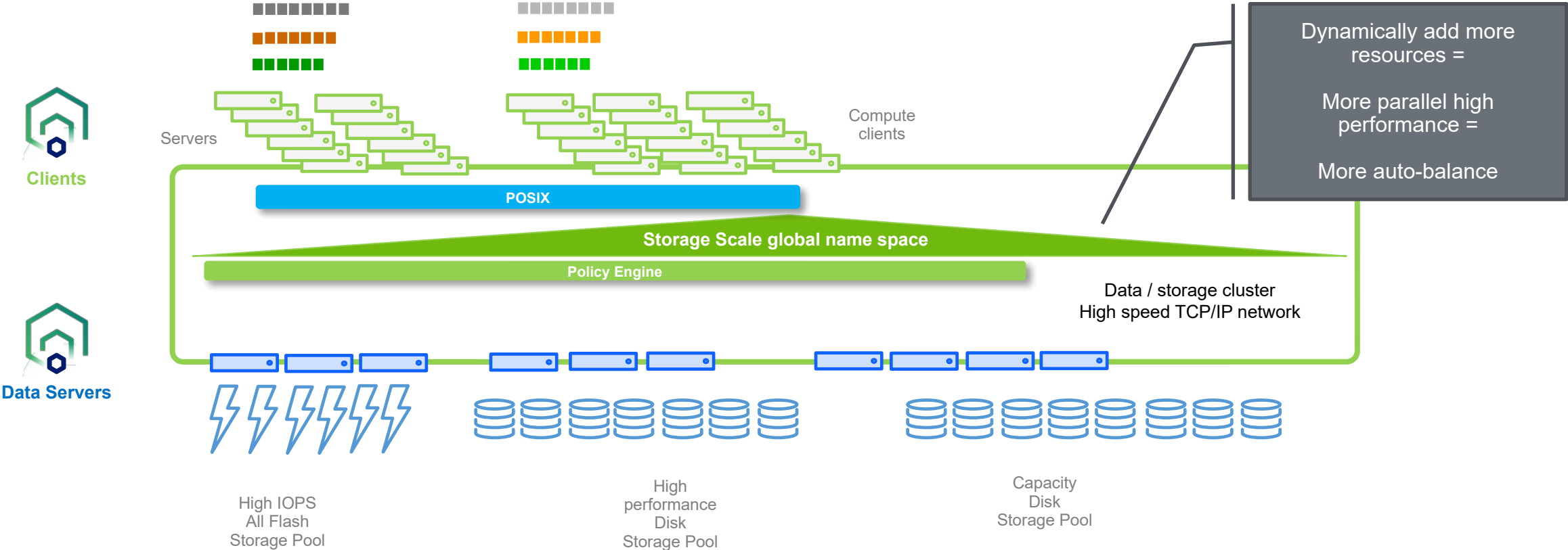
Storage Scale foundations 1.1 – just read it when you need it!



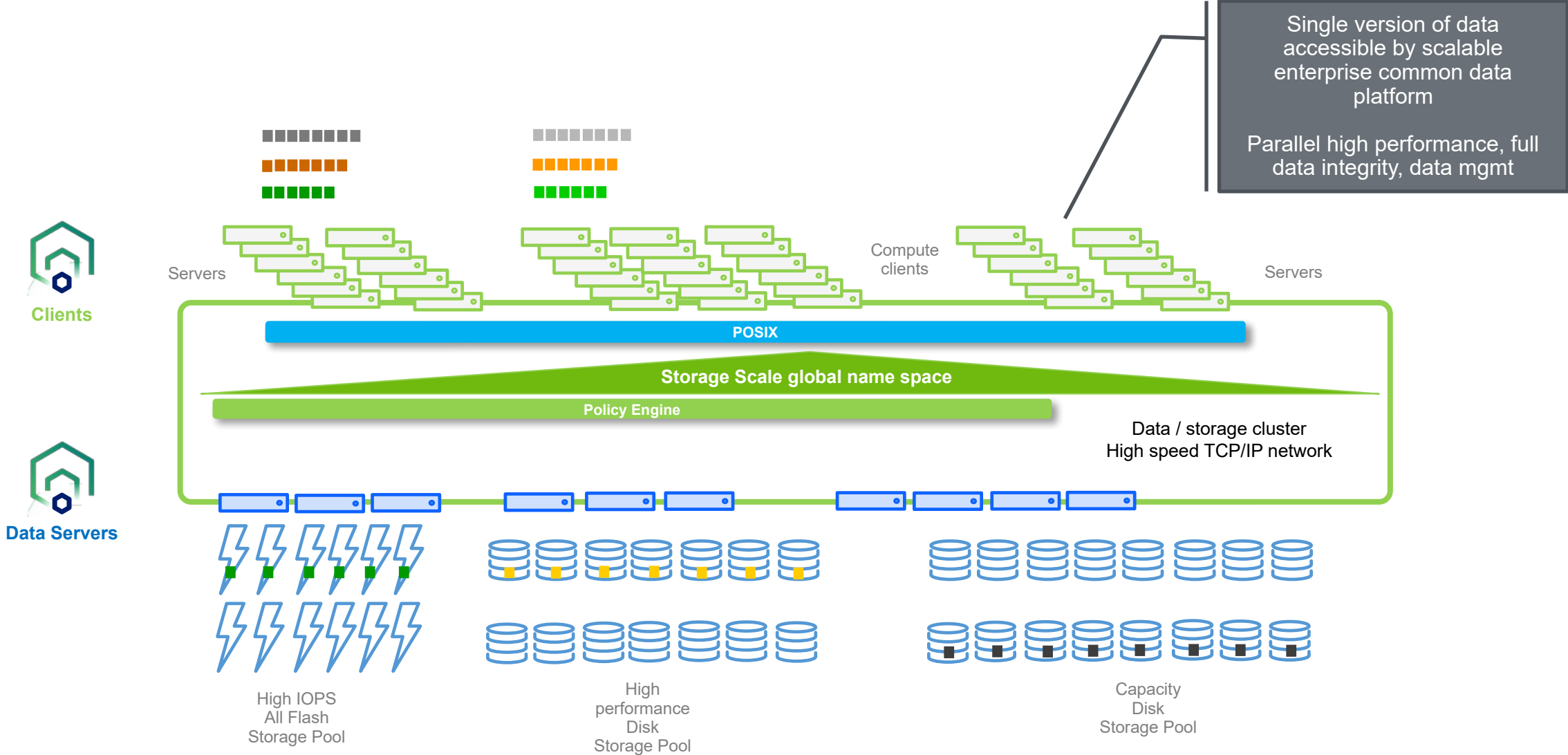
Storage Scale foundations 2 - scalability



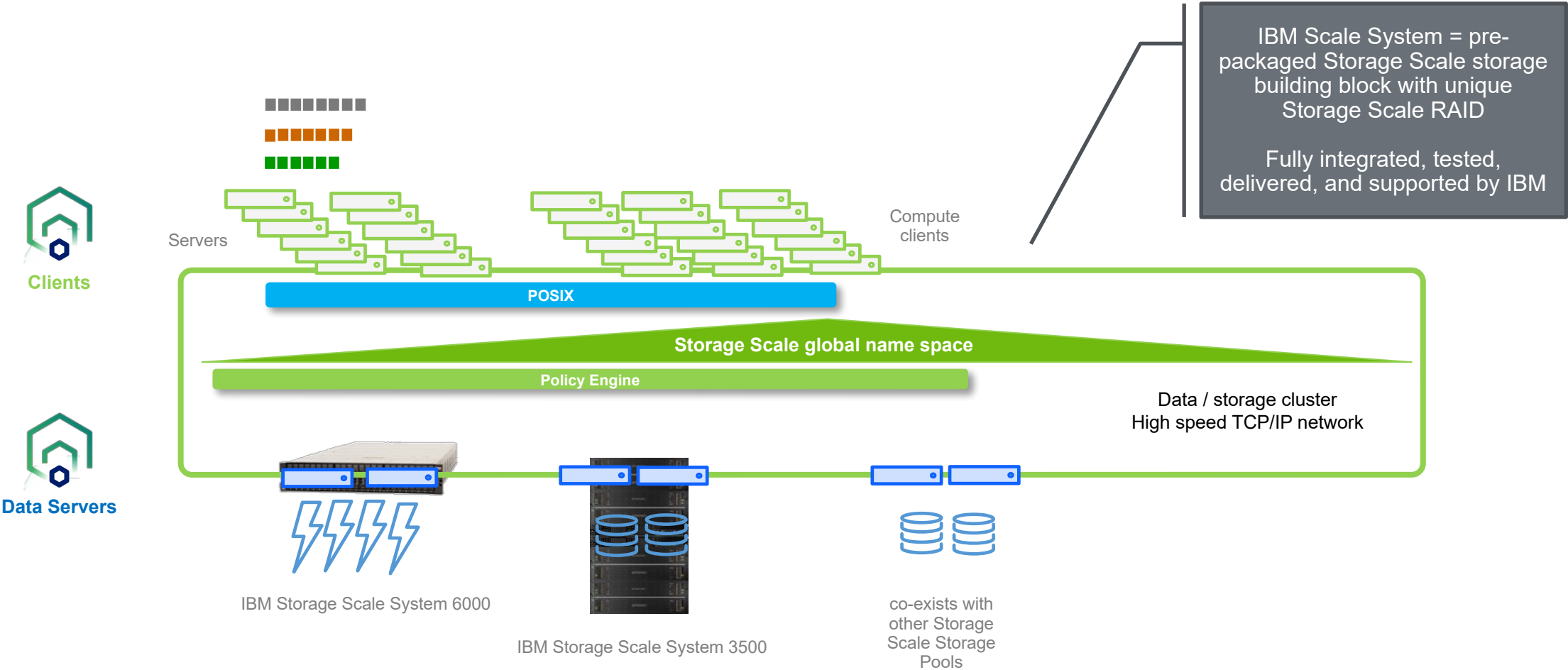
Storage Scale foundations 2.1 – more scalability



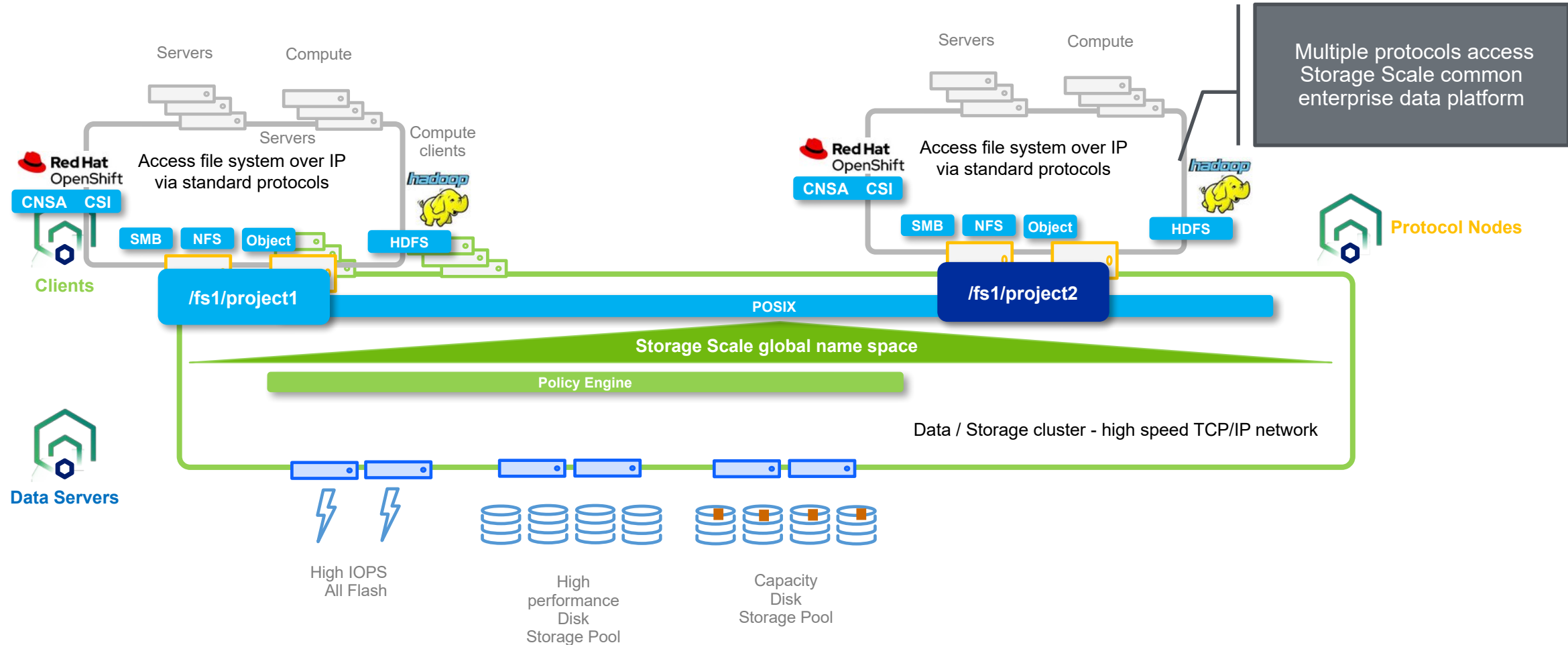
Storage Scale foundations 2.2 – even more scalability



Storage Scale foundations 2 – Scale System 3500/6000

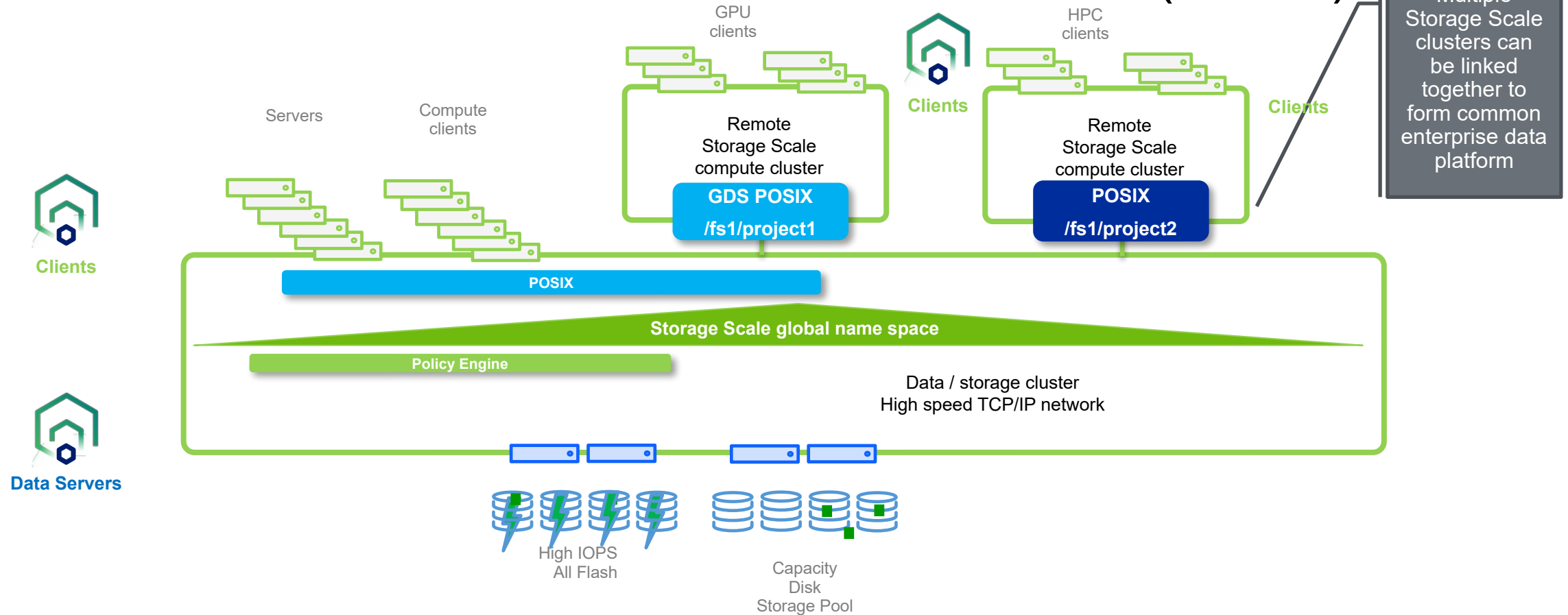


Storage Scale foundations 3 – access by more protocols



Storage Scale foundations 3.1 – access by multiple clusters\

“remote mount” Remote Fileset Access Control (RFAC)

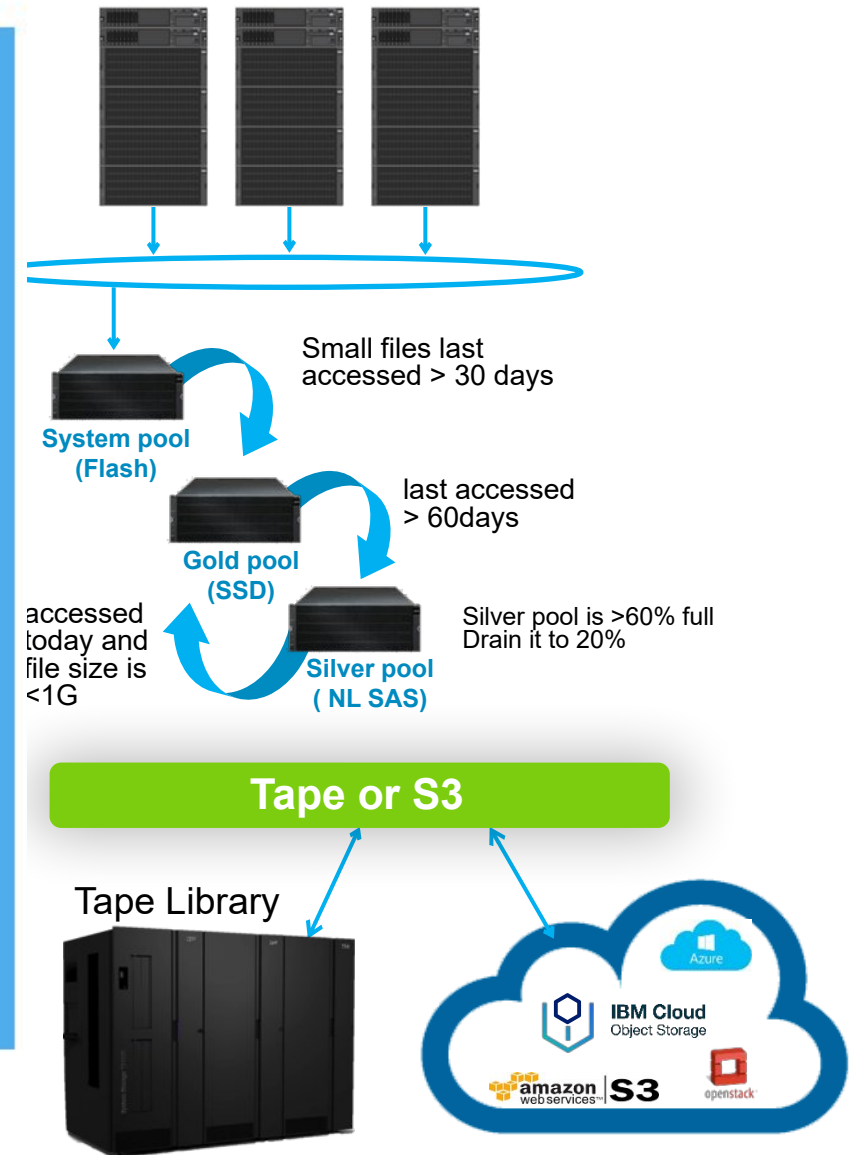
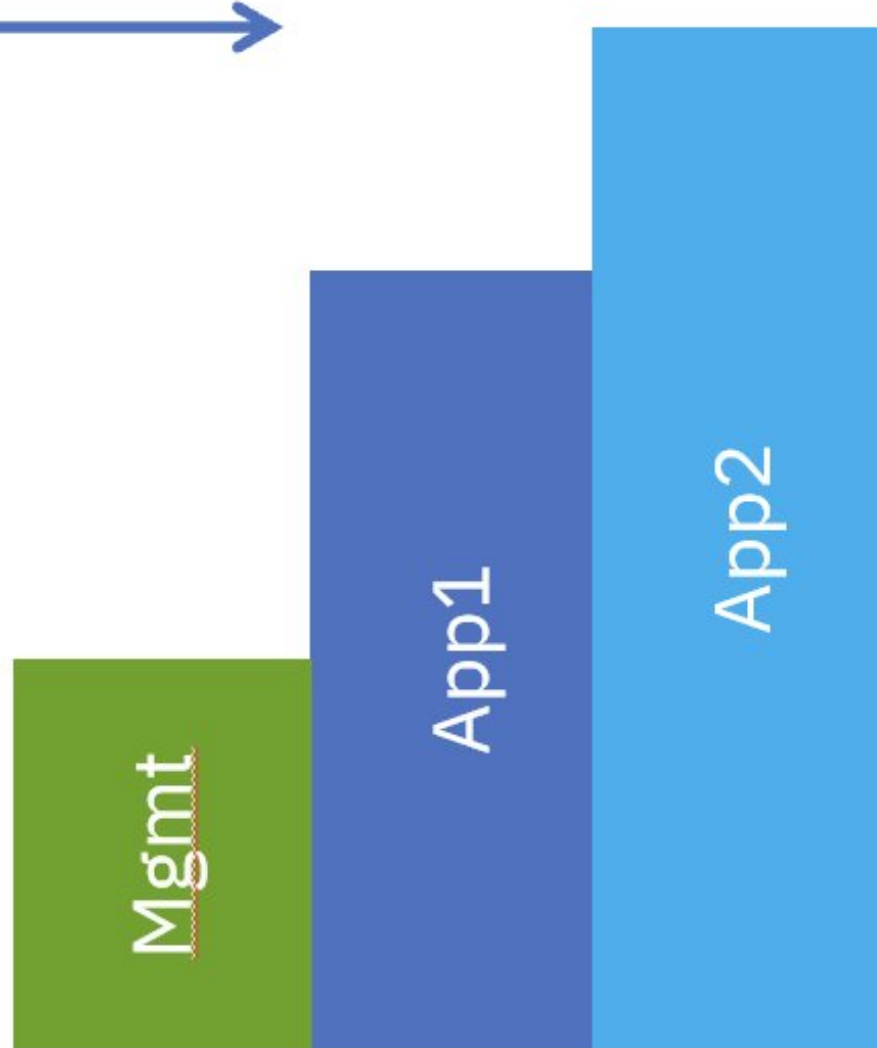


Data aware cost optimization

Total IO Capability
B/W or IOPS



Management
Capped at 20%
(example)



A Global Data Platform in action

EDGE

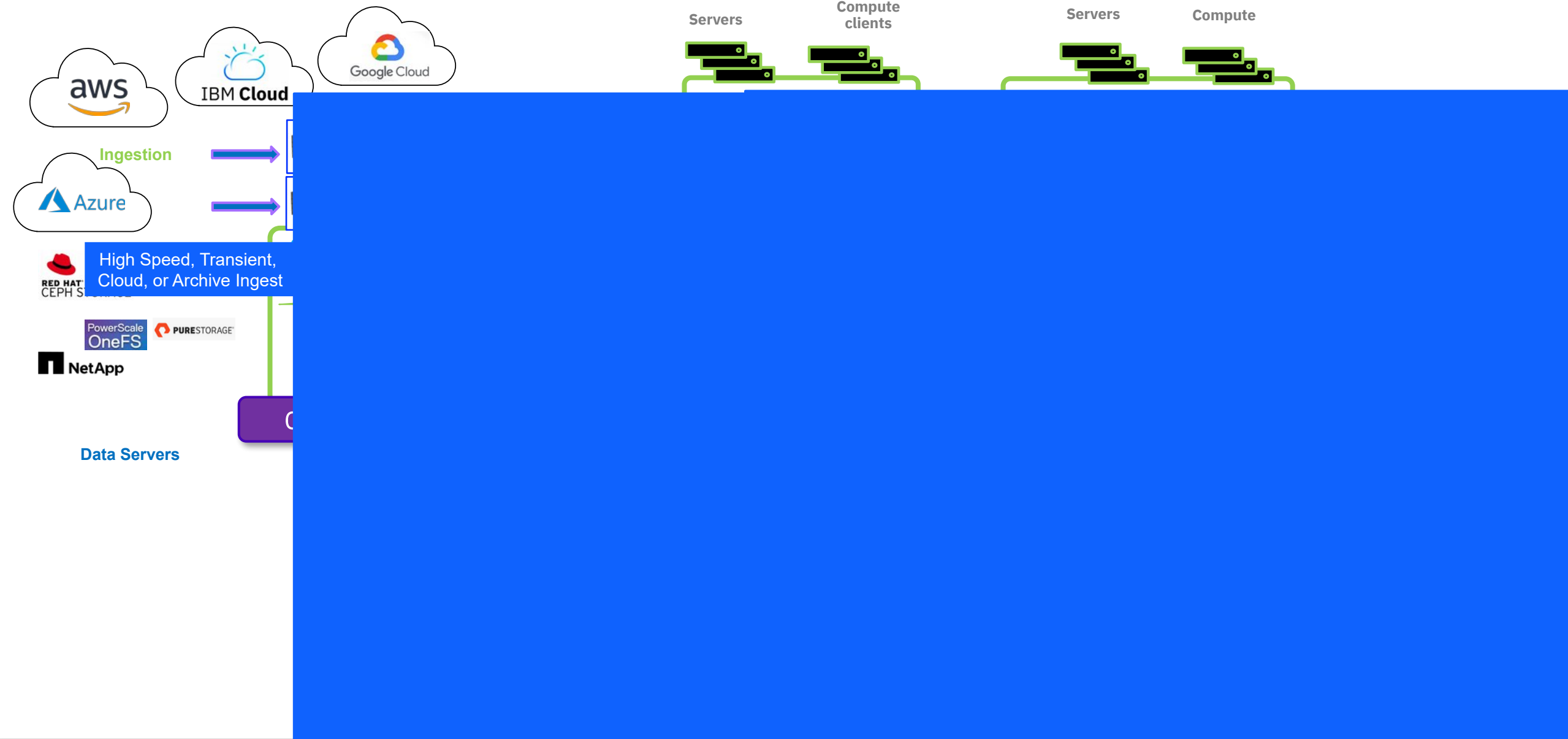
INGEST

ORGANIZE

ANALYZE

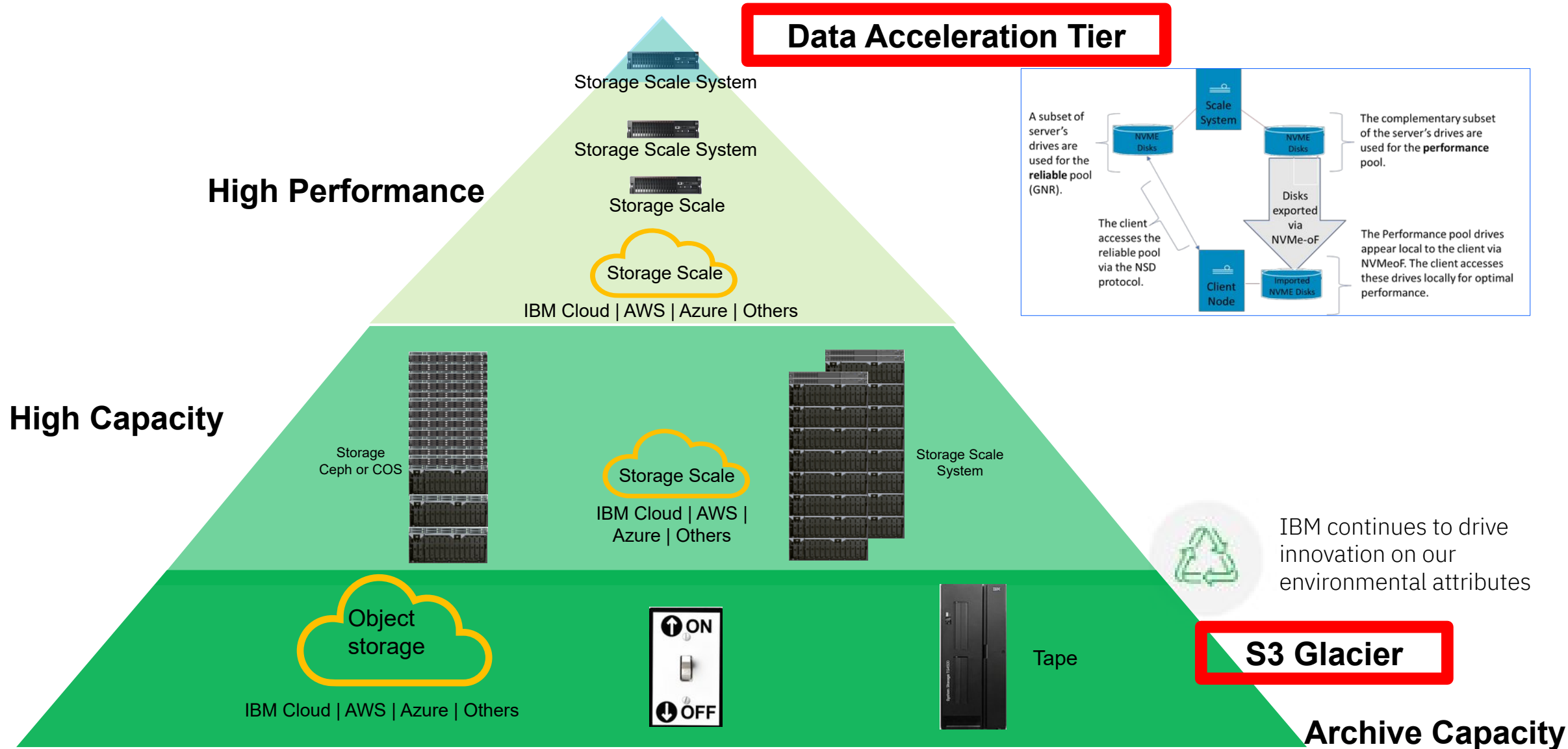
ML / DL

INSIGHTS



Tiering models to move the data to and from the compute

Lowers Energy Consumption and Costs with Data Lifecycle Management



Self-managed S3 Deep Archive



- **Dense** – 27PB of uncompressed data per library, delivering unparalleled data center economics
- **Deep Archive Configuration** – IBM's Diamondback uses the S3 API for deep archive with glacier storage classes
- **Ease of deployment** – As a fully isolated unit, a Diamondback can be shipped pre-populated with media and deployed in under 30 minutes
- **Operationally light** – With IBM's Expert Care offering included with every Diamondback, no specific tape expertise is needed to use a Diamondback for day-to-day tasks
- **Secure** - each Diamondback is an air gapped vault for your data, with WORM protection and built-in encryption available

