SNIA CSTI | CLOUD STORAGE TECHNOLOGIES

High Performance Storage at Exascale

Live Webcast

February 2, 2022

10:00 am PT / 1:00 pm ET

Today's Presenters

Alex McDonald
Independent Consultant
Chair, SNIA Cloud Technologies Initiative

Dr. Torben Kling Petersen
Principal Engineer
HPC Storage BU
HPE

Michael Hennecke
Principal Engineer
HPC Storage
Intel

Glyn Bowden
Chief Architect, AI & Data Science Practice
HPE

SNIA. CSTI | CLOUD STORAGE TECHNOLOGIES

SNIA-at-a-Glance

180
industry leading organizations

2,500
active contributing members

50,000
IT end users & storage pros worldwide

Learn more: **snia.org/technical** @SNIA

SNIA CSTI | CLOUD STORAGE TECHNOLOGIES

What We Do

Educate vendors and users on cloud storage, data services and orchestration

Support & promote business models and architectures: OpenStack, Software Defined Storage, Kubernetes, Object Storage

Understand Hyperscaler requirements
Incorporate them into standards and programs

Collaborate with other industry associations

SNIA Legal Notice

The material contained in this presentation is copyrighted by the SNIA unless otherwise noted.

Member companies and individual members may use this material in presentations and literature under the following conditions:

- Any slide or slides used must be reproduced in their entirety without modification
- The SNIA must be acknowledged as the source of any material used in the body of any document containing material from these presentations.

This presentation is a project of the SNIA.

Neither the author nor the presenter is an attorney and nothing in this presentation is intended to be, or should be construed as legal advice or an opinion of counsel. If you need legal advice or a legal opinion please contact your attorney.

The information presented herein represents the author's personal opinion and current understanding of the relevant issues involved. The author, the presenter, and the SNIA do not assume any responsibility or liability for damages arising out of any reliance on or use of this information.

NO WARRANTIES, EXPRESS OR IMPLIED. USE AT YOUR OWN RISK.

Agenda

- HPC’s evolution and real-world examples
- Technology evolution of HPC Storage
- HPC Data Storage and Data Mobility

HPC Evolution and Real World Examples

Torben Kling Petersen

“Not Your Grandfathers HPC ....*”

- Throughput is no longer the criteria by which HPC storage are designed
- Data generation is replaced by data ingestion ....
- New workloads require different metrics
  - IOPS for random access of small data
  - “find” on steroids
  - extensive metadata requirements on stat, read, write AND delete
  - many new file formats requiring careful handling ....
  - deleting files no longer a trivial process
  - storage tiering requires massive data migration capabilities
  - location, location, location .....

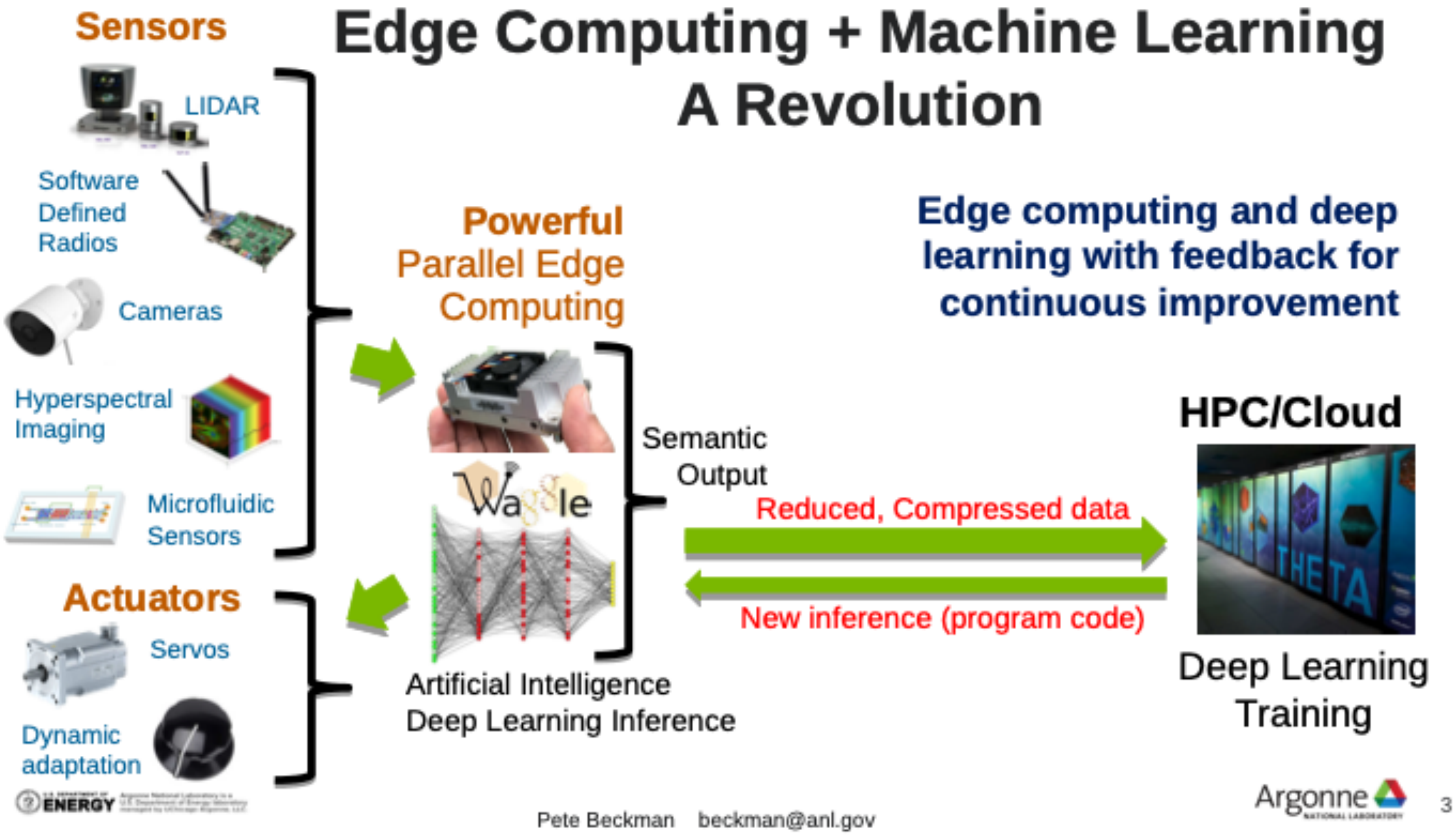
* Adapted from Rick Stevens (ANL)

Example Case: The Waggle Project

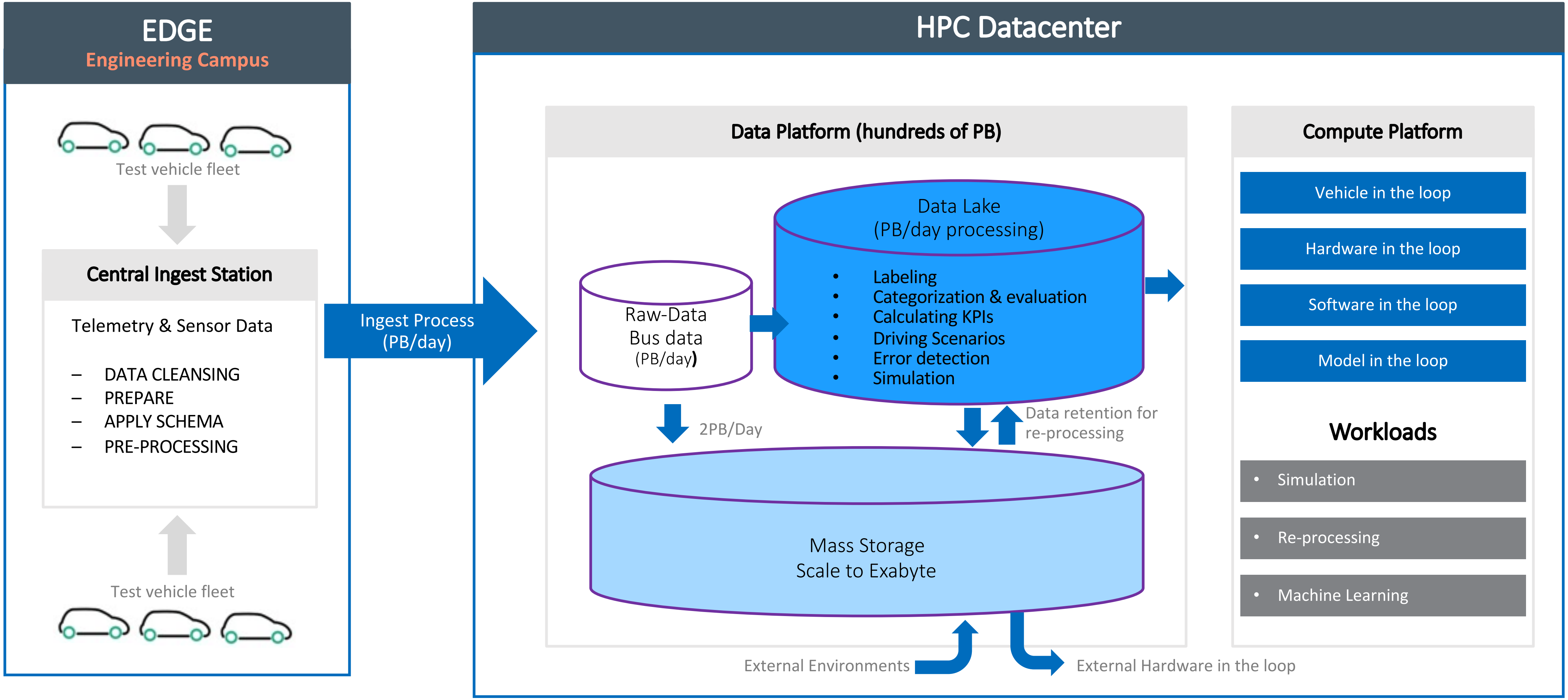
500 sensors

PBs of raw data

Compression, pre-processing, data pushed up, Deep learning, inference ...

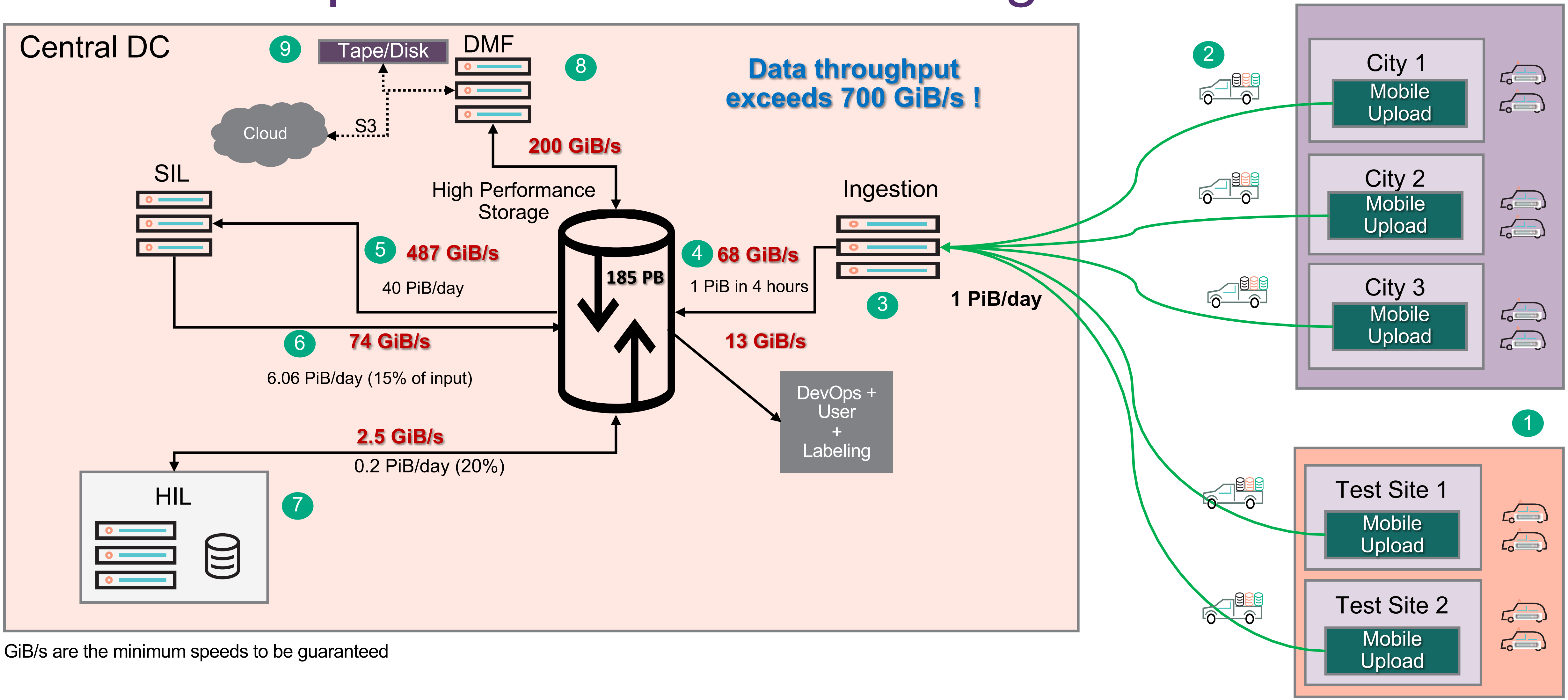


Example Case: Autonomic Driving



SNIA. CSTI | CLOUD STORAGE TECHNOLOGIES

HAD Development: The Data Challenge

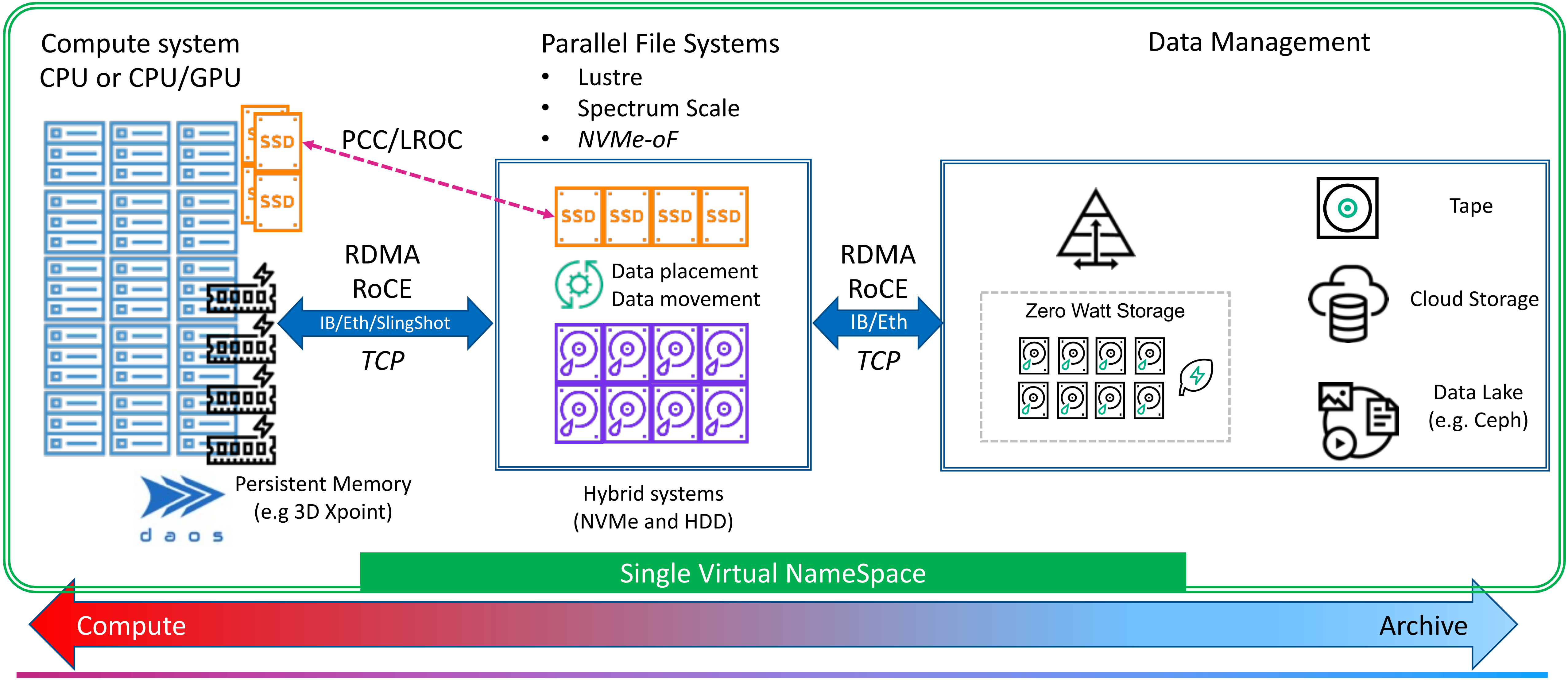


GiB/s are the minimum speeds to be guaranteed

Enabling Technologies

- PCIe Gen 4 -> Gen5
- Networks 200G -> 400G
- Dedicated protocols such as GPU Direct Storage
- VLM interaction with storage systems
- .....

The ”NEW” World – On Prem or Cloud Based



High Performance Data IO

- RDMA Based non-blocking fabrics
- Flexible file system IO
  - Client based caches (LROC or PCC)
  - Ephemeral file systems (e.g. NVMe-oF)
  - Persistent Memory
- Data aware application frameworks
  - PGAS (non-POSIX IO)
  - DAOS
  - Cortx
  - DeltaFS

Data Services Requirements

- Data movement NVMe <-> HDDs
  - Policy based data migration based on capacity and age
  - Manual data migration
  - File purging policies
  - WLM directives
  - Data integrity and Security
- Rapid search facility
  - External to file system -> low impact
  - Workload Manager Integration
  - Query function for advanced searching
  - HSM aware

Data Management

- Data movement - Primary FS to:
  - hot archive
  - object store
  - tape
  - cloud
- Policy based data migration based on
  - Age
  - Size
  - Type
  - Project
  - Classification
  - Usage history etc
- Manage multiple front ends
- Horizontal data movement
- Maintain full namespace mirror
- HSM **and** Incremental Backups
- Tiers gated by:
  - Cost per PB
  - Capacity growth
  - Retention requirements
  - Access performance

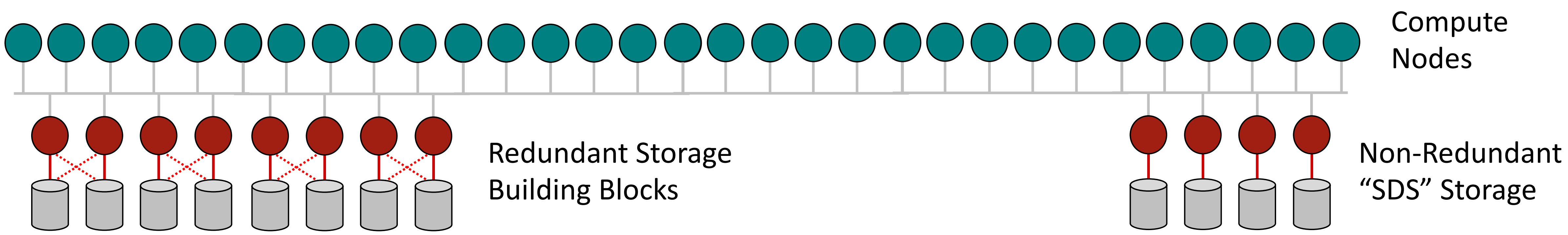
Technology Evolution of HPC Storage

Michael Hennecke

SNIA. CSTI | CLOUD STORAGE TECHNOLOGIES

What Characterizes HPC Storage?

- Massive **Scale-Out** (on an HPC Fabric)

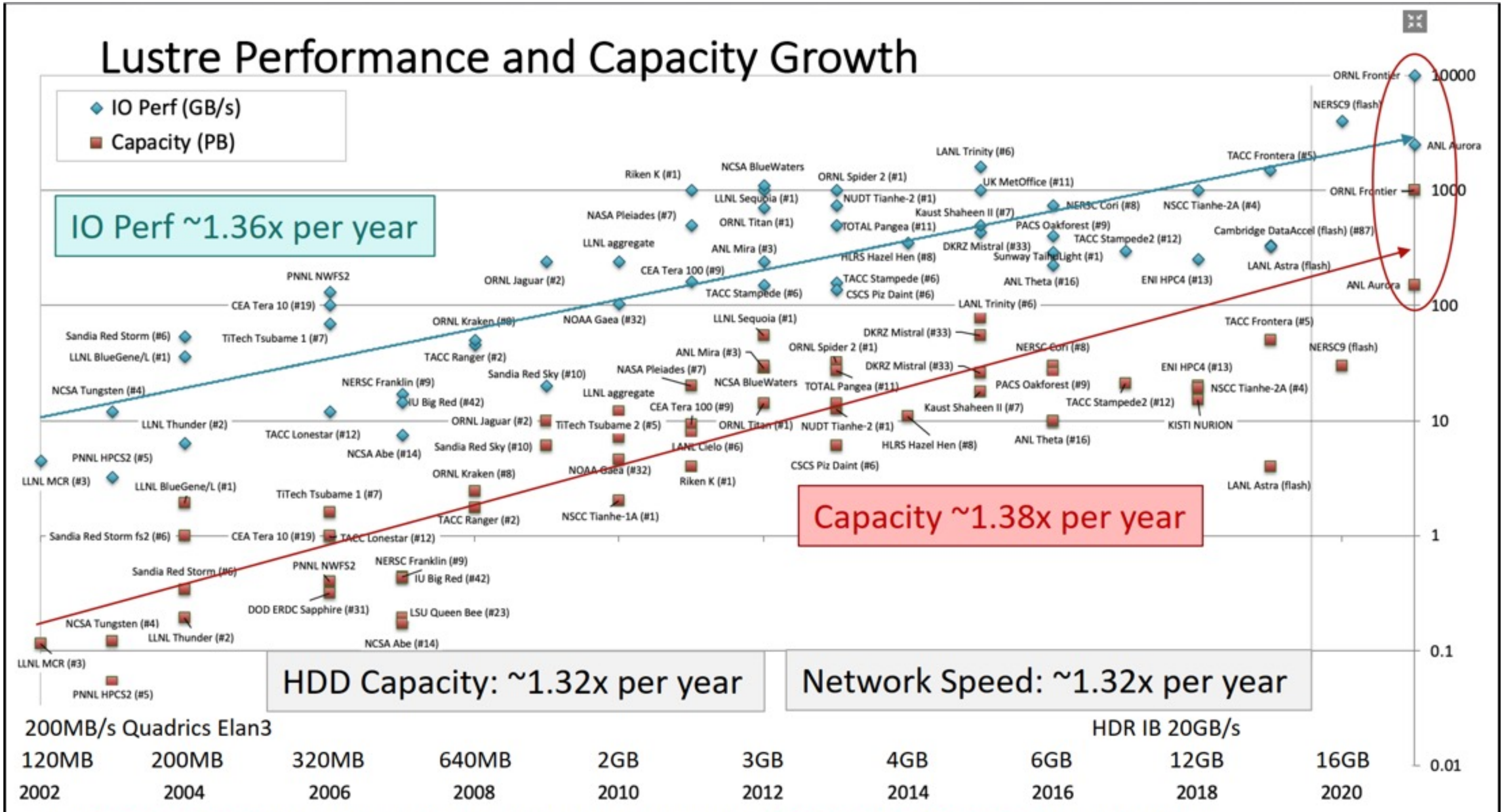


- Global **Parallel Filesystems** (not just block-layer storage access)

Spectrum Scale (GPFS)

lustre

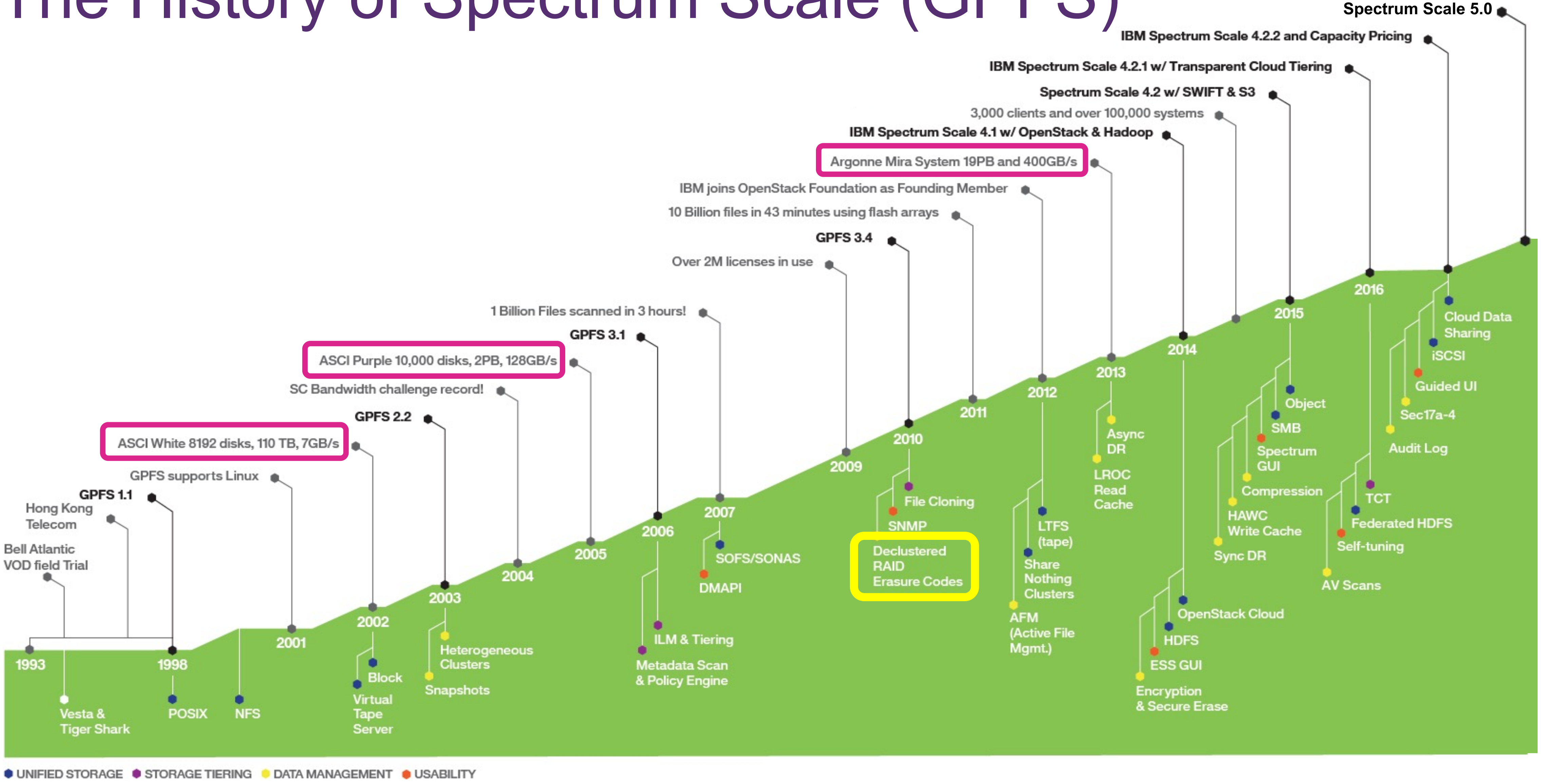
- More and more **storage tiers** (to integrate e.g. HDDs and SSDs)
  - Plus "node-local" storage – often as "stop-gap" solutions for parallel FS limitations



Source: *Rock Hard Lustre*, Nathan Rutman, Cray (with updates for recent years); *Disk Drive Prices (1955-2019)*, John C. McCallum

Source: https://hps.vi4io.org/_media/events/2019/hpc-iodc-lustre_next_20_years-dilger.pdf

The History of Spectrum Scale (GPFS)



A Look at the Top500.org (Nov/2021)

Rank	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	**Supercomputer Fugaku** - Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D, Fujitsu RIKEN Center for Computational Science Japan	7,630,848	442,010.0	537,212.0	29,899
2	**Summit** - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM DOE/SC/Oak Ridge National Laboratory United States	2,414,592	148,600.0	200,794.9	10,096
3	**Sierra** - IBM Power System AC922, IBM POWER9 22C 3.1GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM / NVIDIA / Mellanox DOE/NNSA/LLNL United States	1,572,480	94,640.0	125,712.0	7,438
4	**Sunway TaihuLight** - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway, NRCPC National Supercomputing Center in Wuxi China	10,649,600	93,014.6	125,435.9	15,371
5	**Perlmutter** - HPE Cray EX235n, AMD EPYC 7763 64C 2.45GHz, NVIDIA A100 SXM4 40 GB, Slingshot-10, HPE DOE/SC/LBNL/NERSC United States	761,856	70,870.0	93,750.0	2,589

432 racks; 158k nodes

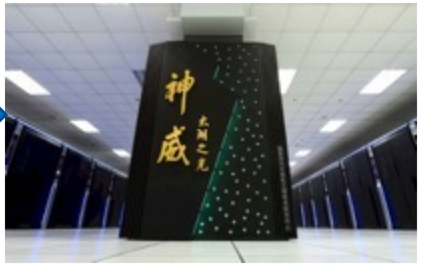
Phase1: 12 racks; 1.5k nodes

FEFS (Fujitsu Exabyte File System) – Lustre based
[+ LLIO – SSD in 1:16 nodes]

250 PB disk ; 2.5TB/s
Spectrum Scale RAID
[+ 7.3 PB node-local NVMe]

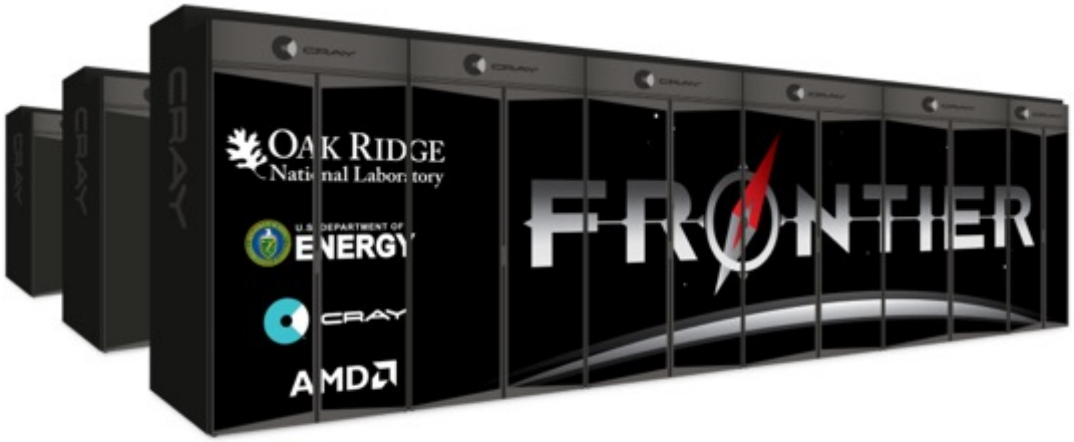
120 PB disk ; 1.0TB/s
Spectrum Scale RAID
[+ node-local NVMe]

???



35 PB all-flash ; 5TB/s
Lustre scratch
[+ 60 PB NERSC CFS – Spectrum Scale]

What's Coming … Future Exascale Systems in the USA



>1.5 ExaFlops (FP64) with >100 racks
(Node: 1x AMD EPYC + 4x AMD Radeon Instinct)
Slingshot fabric; multiple NICs/node

>2 ExaFlops (FP64) with >9000 nodes
(Node: 2x Intel SPR + 6x Intel Xe "Ponte Vecchio" GPUs)
Slingshot fabric; 8 fabric endpoints/node

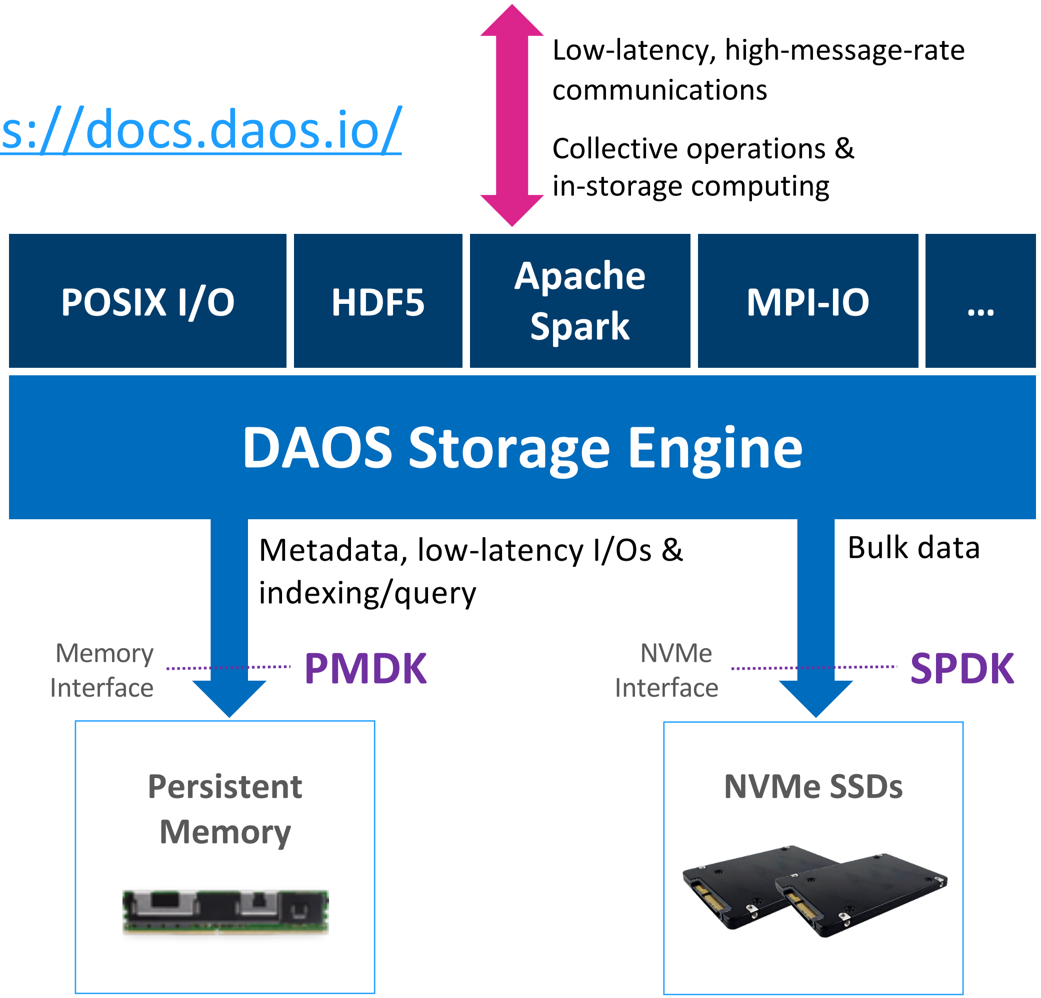
Multi-Tier 679 PB Lustre (+ ZFS Filesystem)
Disk: 679 PB on 47700 disks ; ~5 TB/s
Flash: 11 PB on 5400 NVMe ; ~10 TB/s read
[+ in-system storage layer 75 TB/s read, 35 TB/s write]

Primary storage: DAOS (NVMe + PMem)
>230 PB ; >25 TB/s

What is DAOS?

Distributed Asynchronous Object Storage → https://docs.daos.io/

- Scalable storage based on SCM (Persistent Memory) and NVMe SSDs
- Delivers exceptionally high bandwidth and IOPS, and low latency, on commodity servers
  - Innovative architecture overcomes industry bottlenecks in block-based IO and POSIX
  - Latency in tens of microseconds
- Can be utilized either as a standalone file system, or as a performance tier integrated with other storage systems



More IOPS and bandwidth per dollar. Supports many data models beyond POSIX.

SNIA. CSTI | CLOUD STORAGE TECHNOLOGIES

PMDK Libraries

http://pmem.io
https://github.com/pmem/pmdk

High Level Interfaces

- C++ Persistent Containers
- LLPL (Java) Low-Level Persistent Library
- **pmemkv**

Language support

- C++
- C
- Java
- Python

Transaction Support

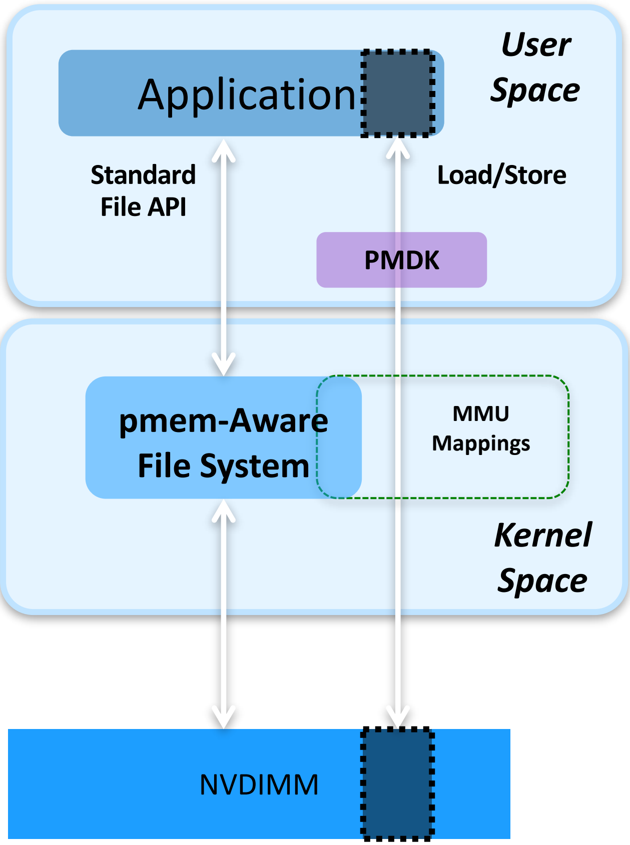
- **Interface to create a persistent memory resident log file** — **libpmemlog**
- **Interface for persistent memory allocation, transactions and general facilities** — **libpmemobj**
- **Interface to create arrays of pmem-resident blocks, of same size, atomically updated** — **libpmemblk**

Support for **volatile** memory usage

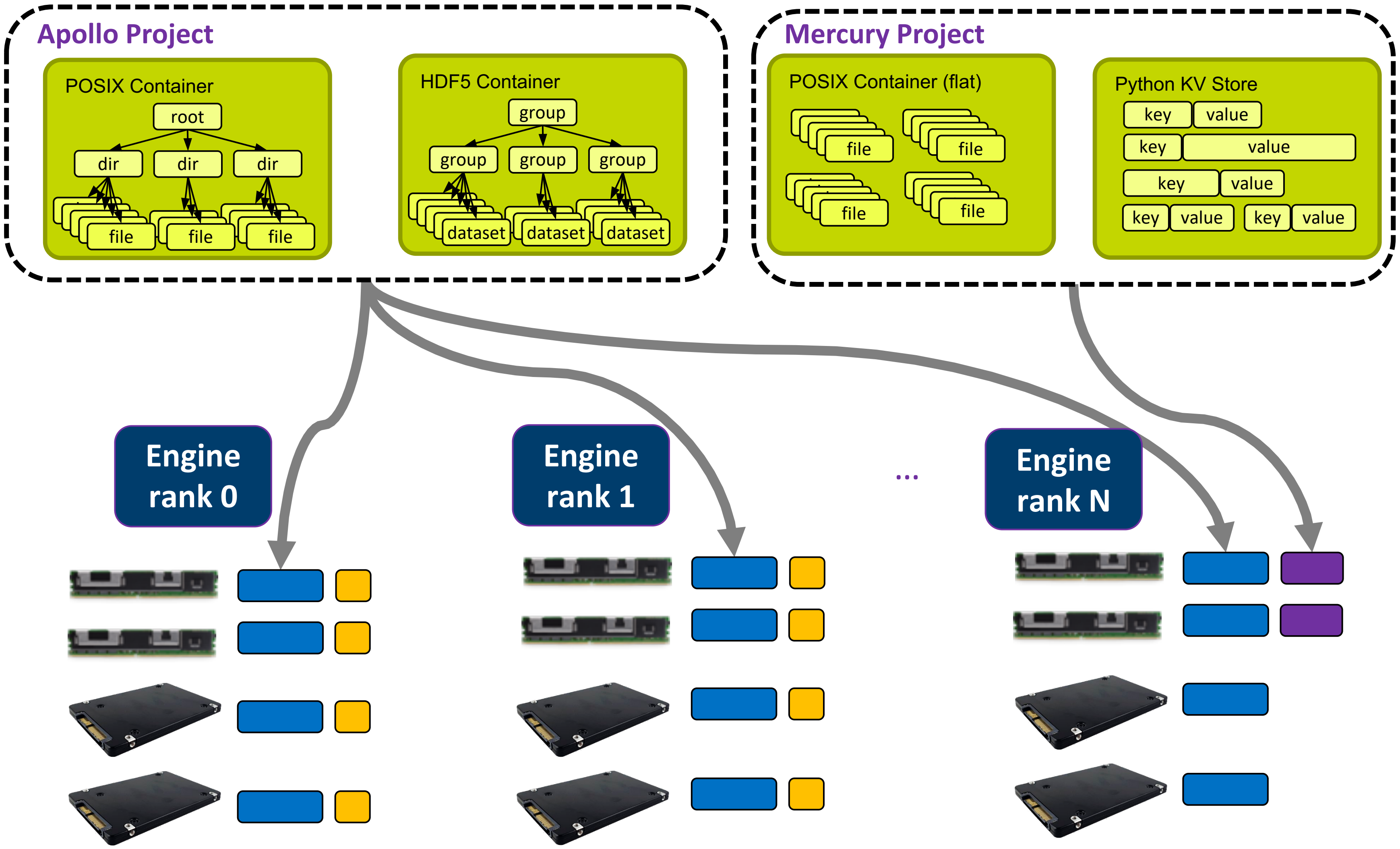
- **memkind**
- **vmemcache**

Low-level support

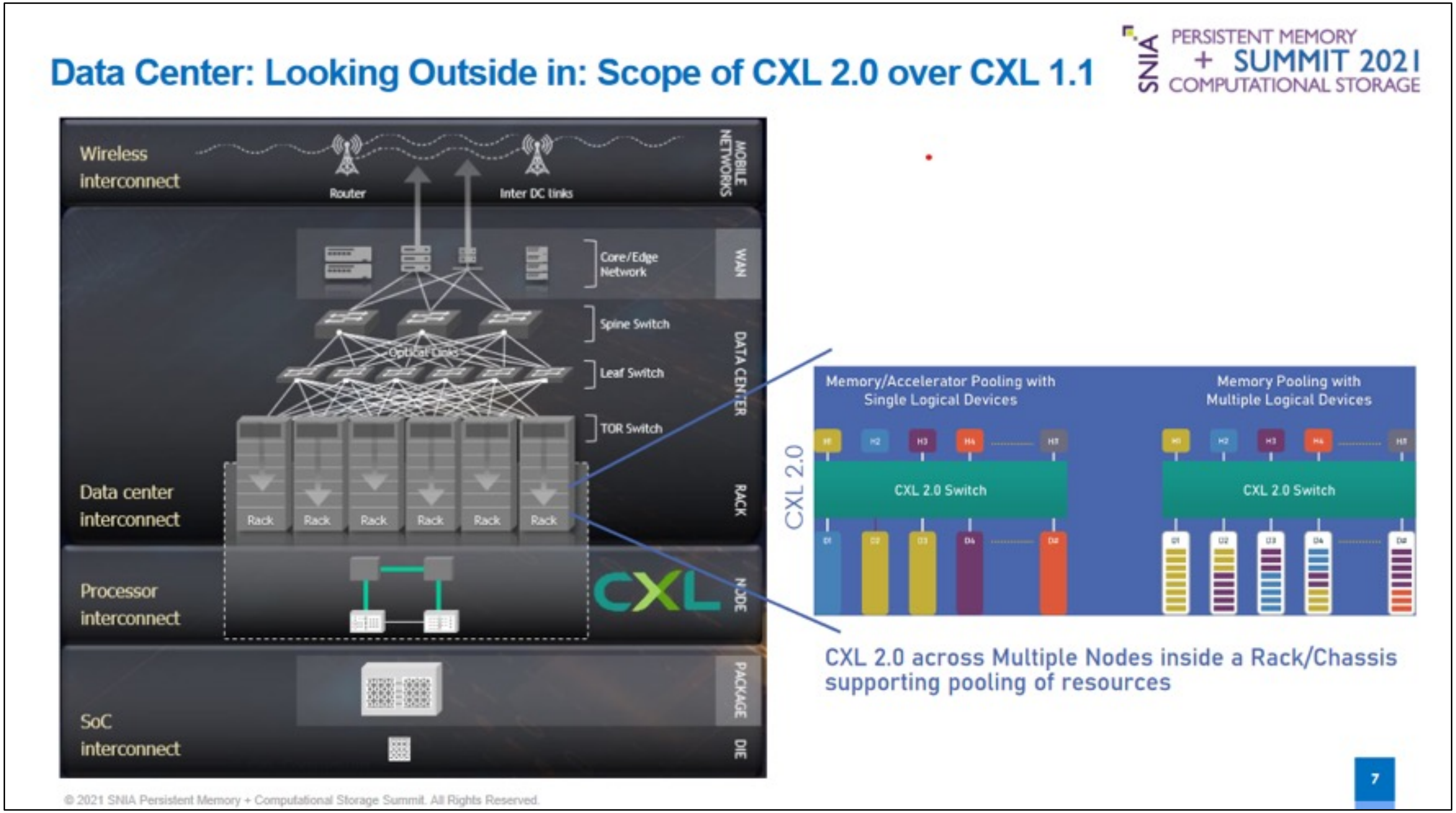
- **Low level support for local persistent memory** — **libpmem**
- **Low level support for remote access to persistent memory** — **librpmem**



DAOS Data Model: Storage Pools and Containers



On the Horizon: CXL 2.0 introduces CXL.mem



SNIA. CSTI | CLOUD STORAGE TECHNOLOGIES

HPC Data Storage and Data Mobility

Glyn Bowden

Current HPC Storage Ecosystem

Operating Environment

- The operating system of the node
- Supporting binaries and libraries for the nodes hardware
- Telemetry and operational tooling for cluster management

Scratch

- A high-performance disk space
- Typically a POSIX compatible filesystem
- Used to right temporary data during computation

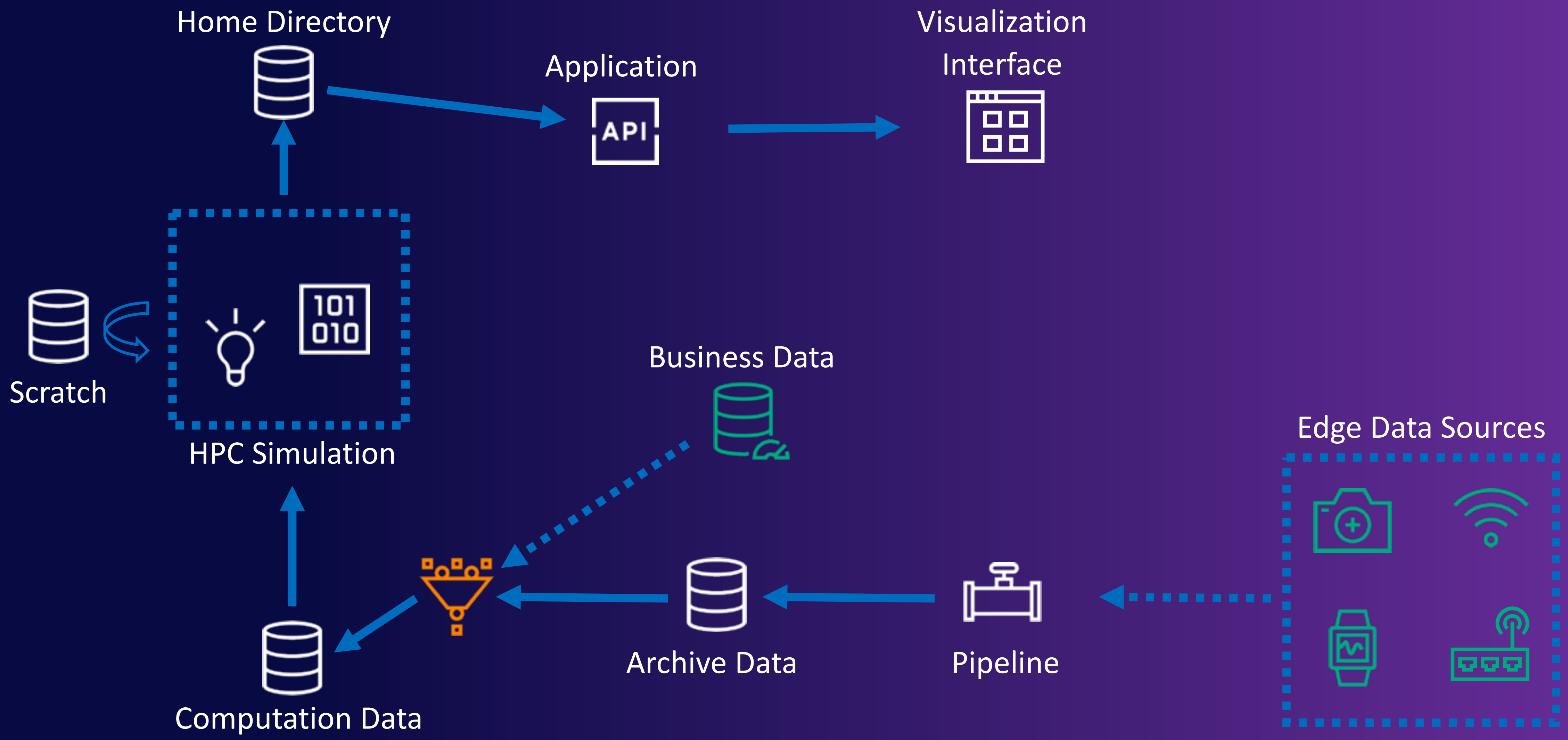
Data for Computation

- The large data sets that are to be processed
- Fundamental to the jobs
- Need to be accessible on all nodes
- Can be a subset of a larger dataset
- Can be a combination of multiple data sets

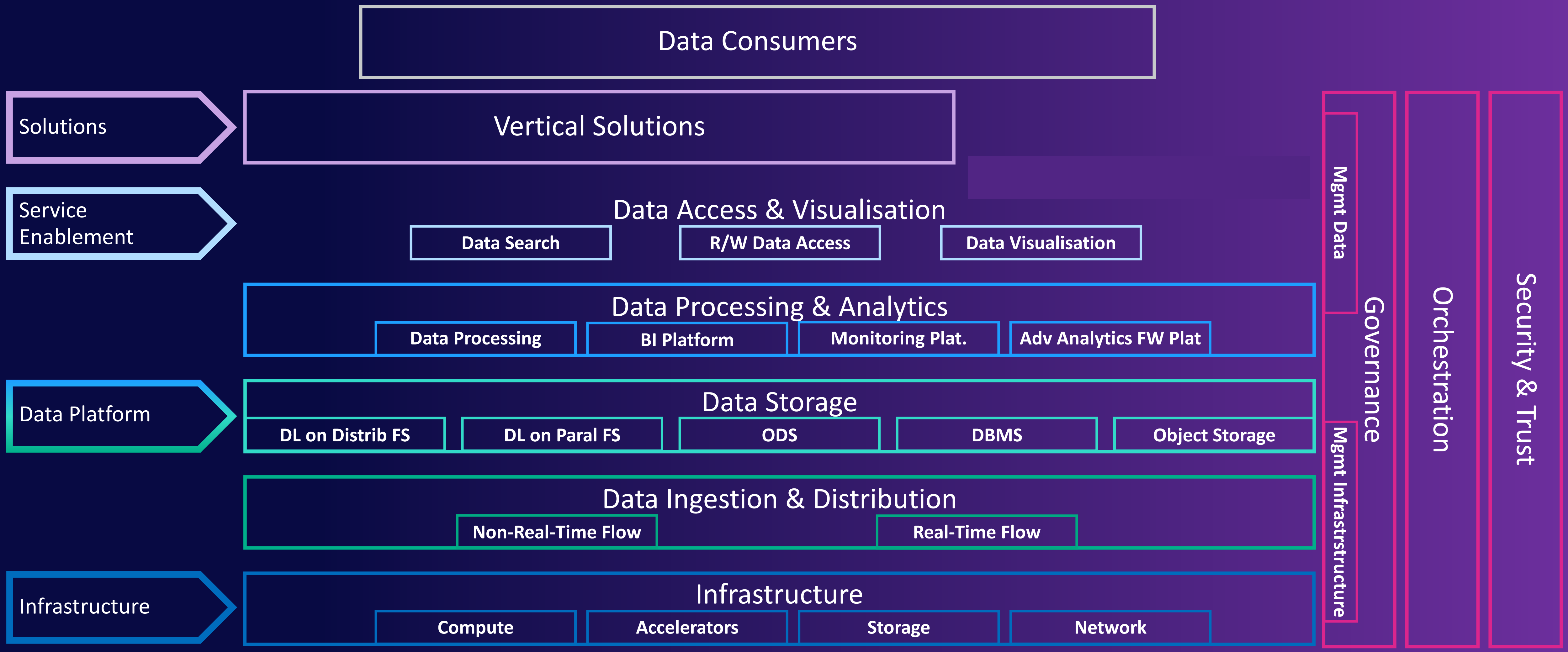
Users Home Directory

- Contains user specific data
- Program binaries
- Users source code / scripts
- User experiment data and results

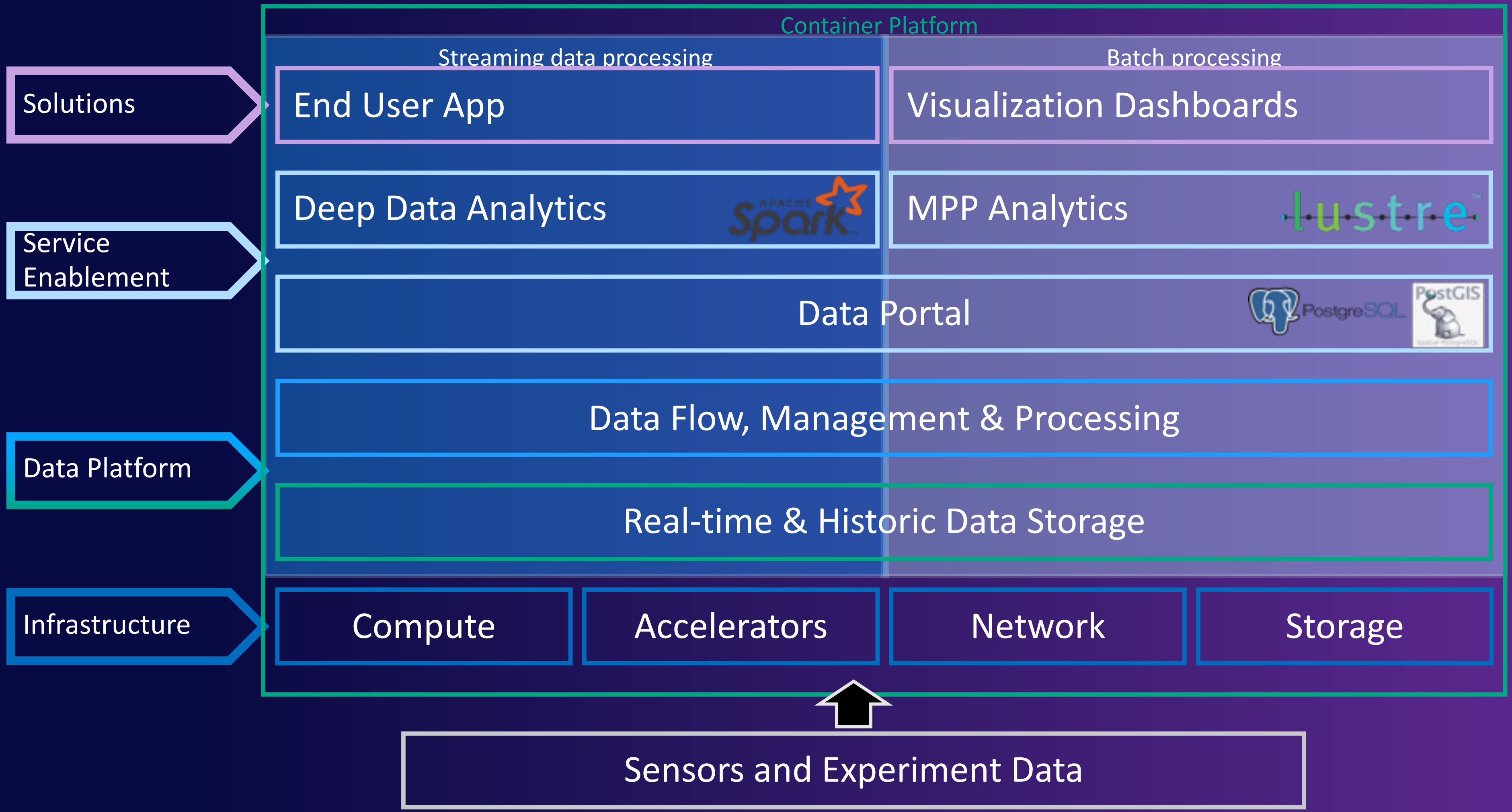
Data Mobility in HPC



Data Platform Functional Architecture



Example Data Platform Solution



Summary

- Moving more toward a convergence of AI and HPC workloads
- HPC needs to learn from Enterprise regarding managing multiple, disparate data stores
- Ingest has always been a problem, now it could potentially be the literal "Blocker"
- Exascale is going to require both data AND compute mobility as jobs span edge, data center and cloud

Thanks for Viewing this Webcast

- Please rate this presentation and provide us with feedback
- This webcast and a copy of the slides will be available at the SNIA Educational Library https://www.snia.org/educational-library
- A Q&A from this webcast will be posted to the SNIA Cloud blog: www.sniacloud.com/
- Follow us on Twitter @SNIACloud

