Hands on exercises: https://github.com/radix-io/hands-on

Survey of Storage Systems

ATPESC 2021

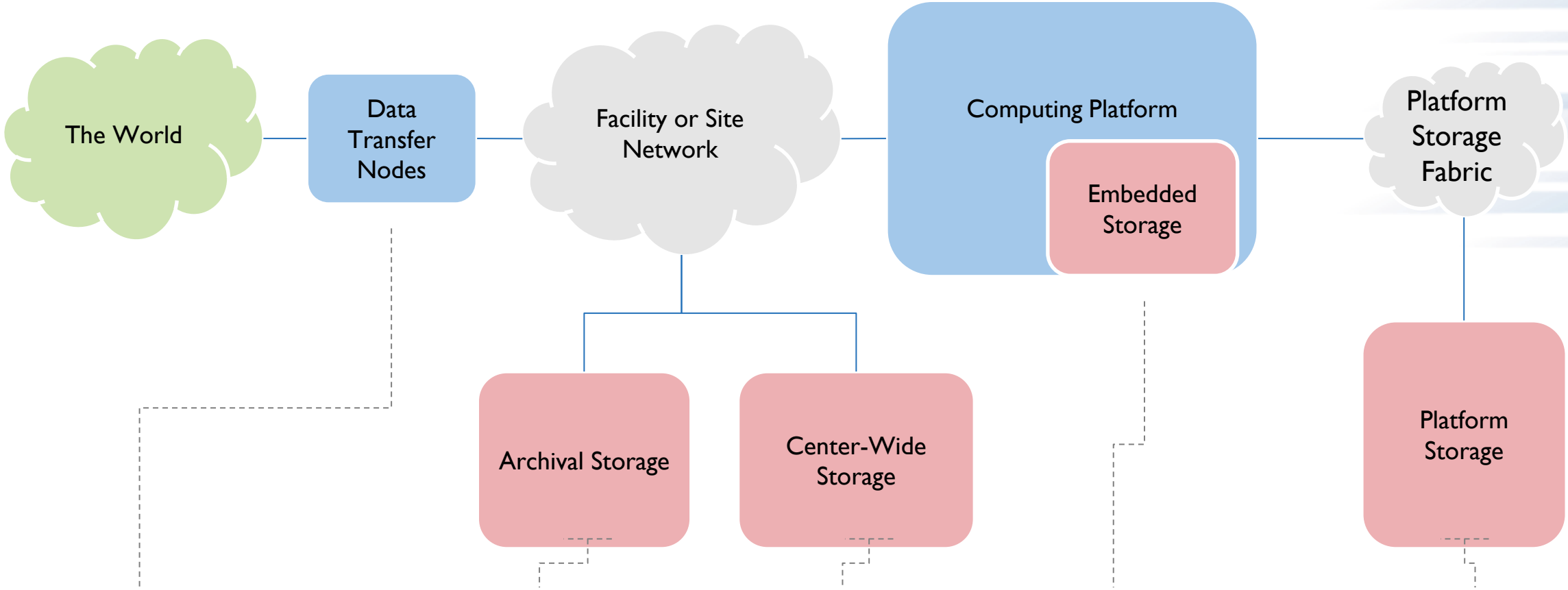
Phil Carns
Mathematics and Computer Science Division
Argonne National Laboratory

Thank you to Rob Ross, Kevin Harms, Glenn Lockwood, Sarp Oral, and Jim Ahrens for the background material used throughout this presentation.

August 6, 2021

Hands on exercises: https://github.com/radix-io/hands-on

A generalized view of storage at a DOE compute facility



Data transfer nodes facilitate bulk data movement in/out of the facility.

Archival storage holds infrequently accessed data for extended time periods.

Center-wide storage allows sharing of data between systems.

Embedded storage is high BW, low latency, and located on node or in the platform fabric.

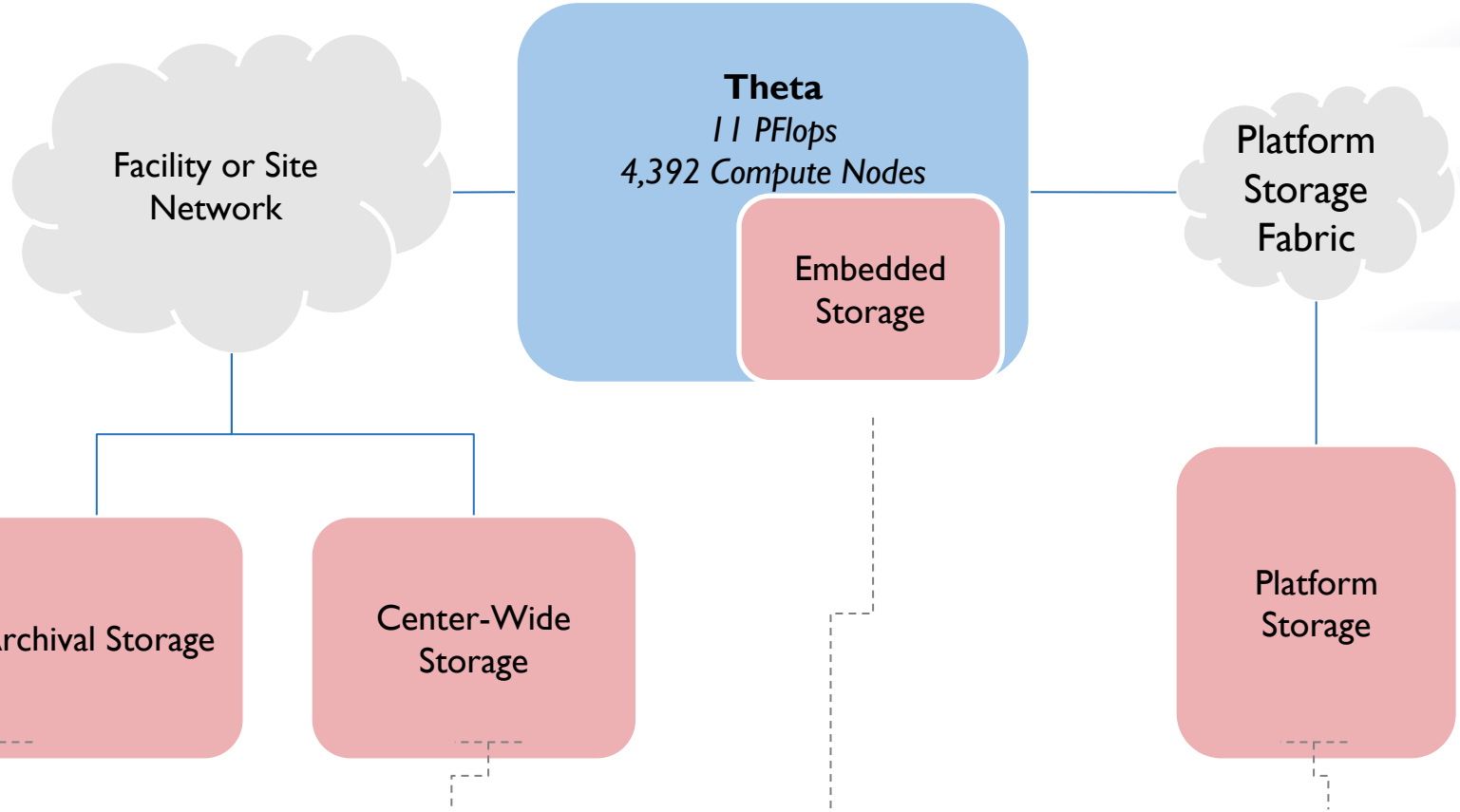
Platform storage provides high capacity storage for data between runs.

Argonne NATIONAL LABORATORY

ECP EXASCALE COMPUTING PROJECT

Hands on exercises: https://github.com/radix-io/hands-on

Example in practice: ALCF Theta



HPSS archival store
*305 PB capacity **
9.4 PB cache
90 GB/sec disk cache

Grand and Eagle PFSes
200 PB aggregate
650 GB/sec (each FS)
Lustre

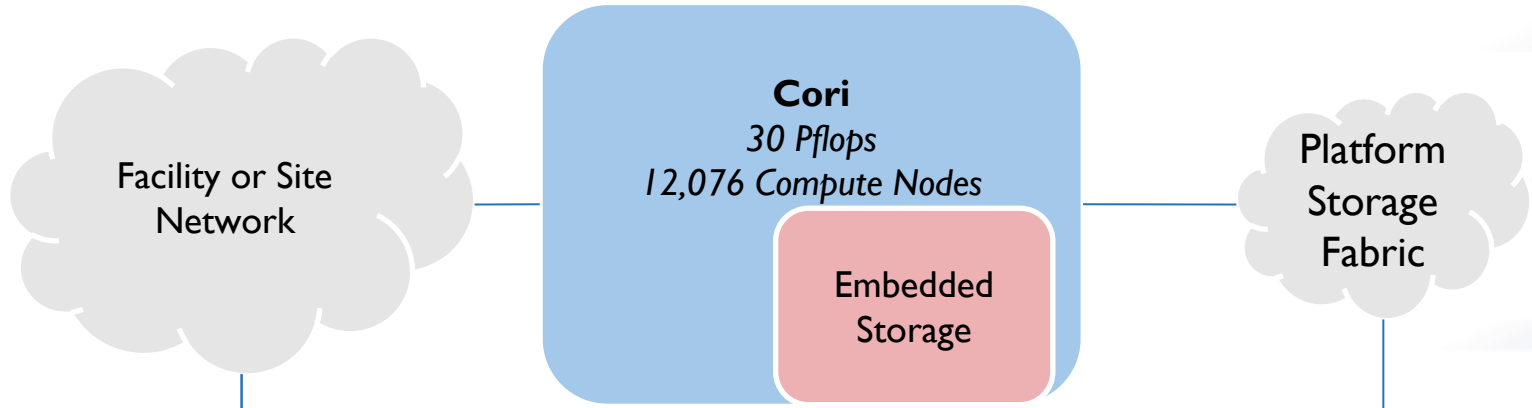
Node-local SSD
549 TB capacity (aggregate)
6.6 TB/sec write, 9.2 TB/sec read
ext3 (no global FS)

theta-fs0
10 PB capacity, HDD based
240 GB/sec
Lustre

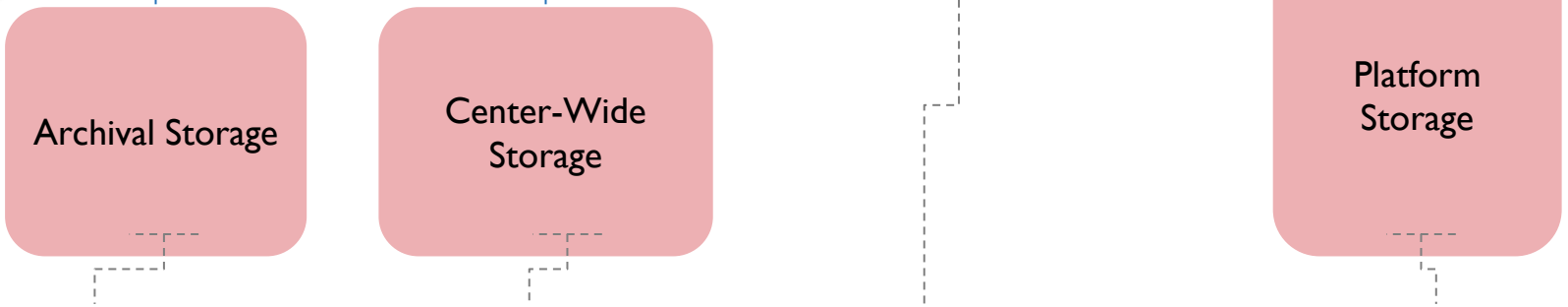
Argonne NATIONAL LABORATORY
ECP EXASCALE COMPUTING PROJECT

Hands on exercises: https://github.com/radix-io/hands-on

Example in practice: NERSC Cori (NERSC-8)



Uses burst buffer (partionable shared scratch space) rather than local devices for embedded storage.



Archival storage holds infrequently accessed data for extended time periods.

Center-wide storage allows sharing of data between systems.

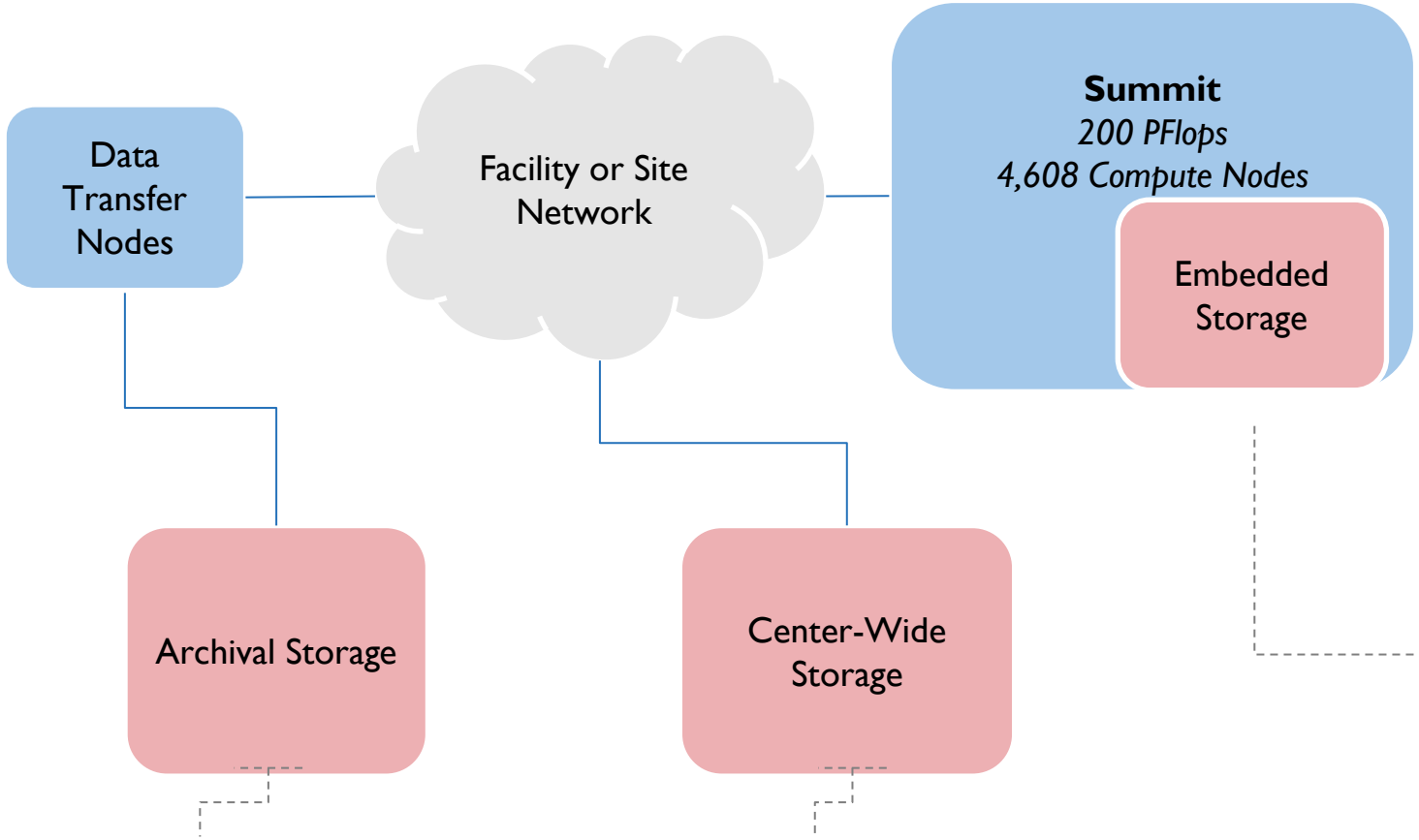
Burst Buffer
1.8 PB capacity, SSD
1.7 TB/s read/write
DataWarp

Cori Scratch
30 PB capacity
700 GB/s read/write
Lustre

Argonne NATIONAL LABORATORY

ECP EXASCALE COMPUTING PROJECT

Example in practice: OLCF Summit



No platform storage: OLCF instead invested in significant center-wide storage improvements to support Summit.

Archival storage holds infrequently accessed data for extended time periods.

Alpine parallel file system
250 PB capacity
2.2-2.5 TB/sec write (540MB/sec/node)
IBM SpectumScale (GPFS) POSIX FS

Node-local NVMe SSDs
7.3 PB capacity (aggregate)
9.7 TB/sec write, 27TB/sec read
XFS (no global FS, Spectral I/O copy agent)

Argonne NATIONAL LABORATORY
ECP EXASCALE COMPUTING PROJECT

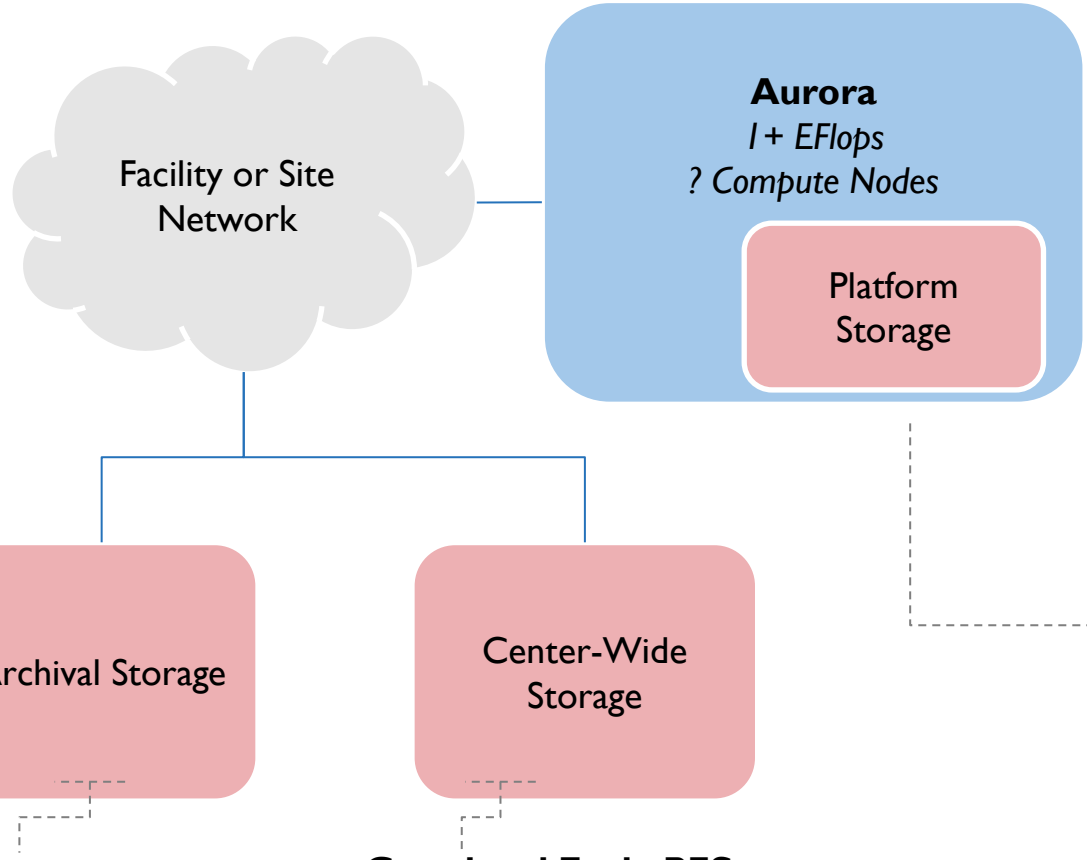
Current Platforms

Hands on exercises: https://github.com/radix-io/hands-on

	ALCF Theta				NERSC Cori				OLCF Summit			
	Archive	**Center**	**Platform**	**Embedded**	**Archive**	**Center**	**Platform**	**Embedded**	**Archive**	**Center**	**Platform**	**Embedded**
SW	HPSS	Lustre	Lustre	ext3	HPSS	GPFS	Lustre	DataWarp	HPSS	GPFS	*N/A*	XFS
HW	LTO8 Tape/ HDD	SSD/HDD	HDD	SSD	3592 Tape/ HDD	HDD	HDD	SSD	3592 D Tape/ HDD	HDD		SSD
Capacity	305 PB	200 PB	10 PB	549 TB	230 PB	128 PB	30 PB	1.8 PB	130 PB	250 PB		7.3 PB
BW	90 GB/s (cache)	650 GB/s	240 GB/s	6.6-9.2 TB/s	100 GB/s (cache)	100 GB/s	700 GB/s	1.7 TB/s	120 GB/sec (cache)	2.2-2.5 TB/s		9.7-27 TB/s
Usability	Medium	High (POSIX)	High (POSIX)	Low (per-node)	Medium (hsi, htar, ftp, globus)	High (POSIX)	High (POSIX)	Medium (POSIXy)	Medium (his, htar, globus)	High (POSIX)		Low (per-node)

Note: No effort has been taken to try to uniformly measure BW or IOPS; consider these as estimates.

Projected deployment: ALCF Aurora



HPSS archival store
*305 PB capacity **
9.4 PB cache
90 GB/sec disk cache

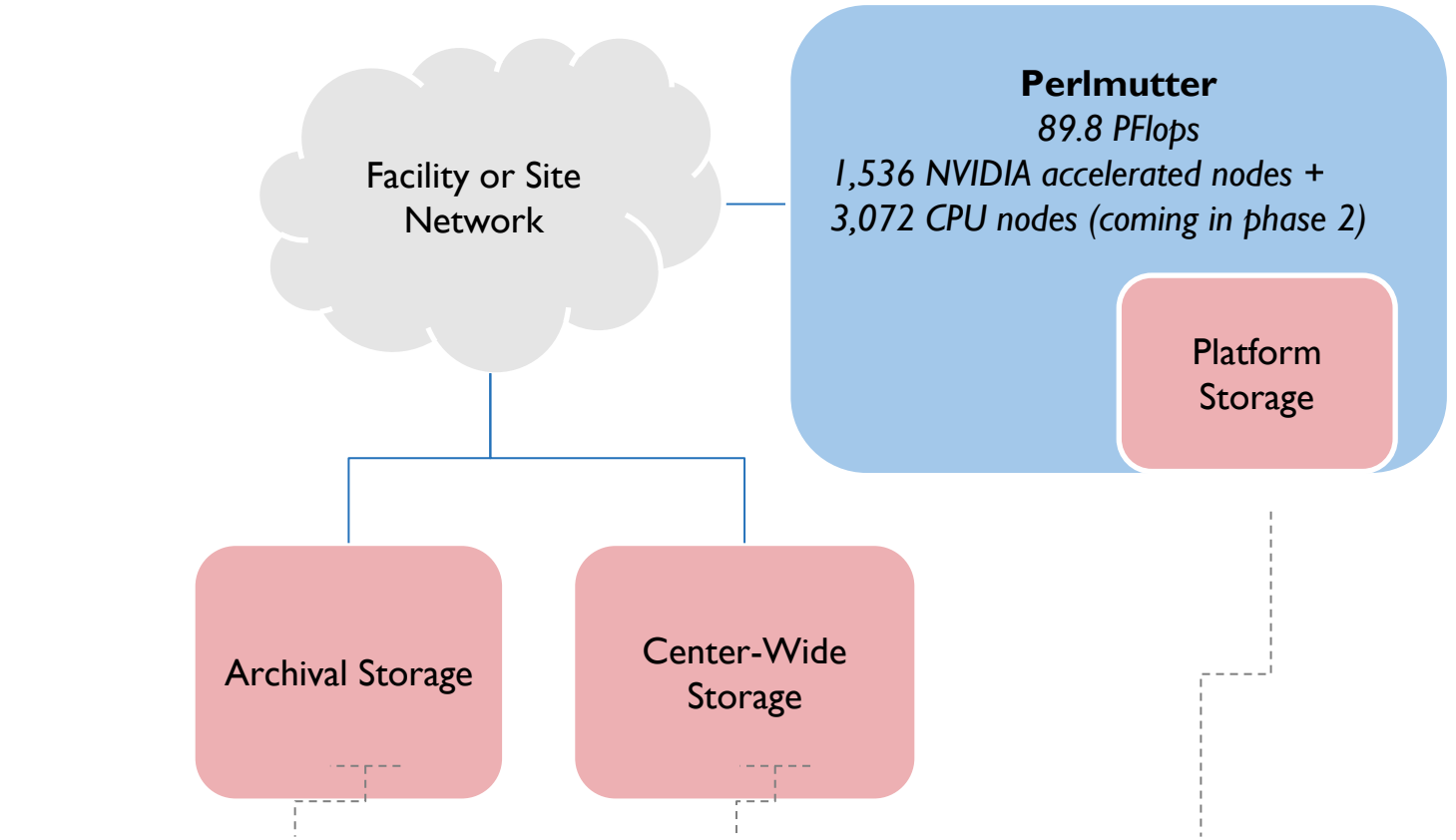
Grand and Eagle PFSes
200 PB aggregate
650 GB/sec (each FS)
Lustre

DAOS
220 PB capacity, PM/NVME based
25+ TB/sec
DAOS (non-POSIX)

Platform storage shifts from a POSIX FS to DAOS. DAOS includes a POSIX compatibility layer but is likely to perform better when treated as an object or key/value store.

Argonne NATIONAL LABORATORY
ECP EXASCALE COMPUTING PROJECT

Projected deployment: NERSC Perlmutter (NERSC-9)



Archival storage holds infrequently accessed data for extended time periods.

Center-wide storage allows sharing of data between systems.

Perlmutter Scratch
35 PB capacity, NVME SSD
5 TB/s read/write
Lustre

Platform storage is an all-flash parallel file system rather than a burst buffer as on Cori.

Upcoming Platforms

Hands on exercises: https://github.com/radix-io/hands-on

	ALCF Aurora				NERSC Perlmutter				OLCF Frontier			
	Archive	**Center**	**Platform**	**Embedded**	**Archive**	**Center**	**Platform**	**Embedded**	**Archive**	**Center**	**Platform**	**Embedded**
SW	HPSS	Lustre	DAOS	*N/A*	HPSS	GPFS	Lustre	*N/A*	*TBD*	Lustre	*N/A*	XFS
HW	LTO8 Tape/ HDD	SSD/HDD	PM/ NVME		3592 Tape/ HDD	HDD	NVME			SSD/HDD		*TBD*
Capacity	305 PB	200 PB	220 PB		230 PB	200 PB	35 PB			700 PB		*TBD*
BW	90 GB/s (cache)	650 GB/s	25+ TB/s		100 GB/s (disk)	500 GB/s	5 TB/s			10 TB/s		*TBD*
Usability	Medium	High (POSIX)	? (non-POSIX)		Medium (hsi, ftp, globus)	High (POSIX)	High (POSIX)			High (POSIX)		Low (per-node POSIX)

Note: These systems haven't been deployed, so configurations may change.

Hands on exercises: https://github.com/radix-io/hands-on

Existing DOE HPC I/O software is being updated for use on projected platforms

I/O Libraries
- ADIOS
- HDF5 (ExaIO)
- Parallel NetCDF
- MPI-IO

Checkpoint/Restart
- SCR
- VeloC
- UnifyFS

Understanding
- Darshan

Compression
- SZ
- ZFP

Performance optimization in progress. POSIX API usage will continue with internal APIs to support new back-ends.

These teams have already been working with local storage and hierarchies for some time – adapting to new layers is relatively easy.

Darshan team is developing new modules to capture information from emerging storage platforms.

Compression algorithms are largely orthogonal to storage hardware technology; ongoing work is focused on making it more accessible.

Hands on exercises: https://github.com/radix-io/hands-on

Thank you!