Community Storage
using Lustre and Globus Sharing

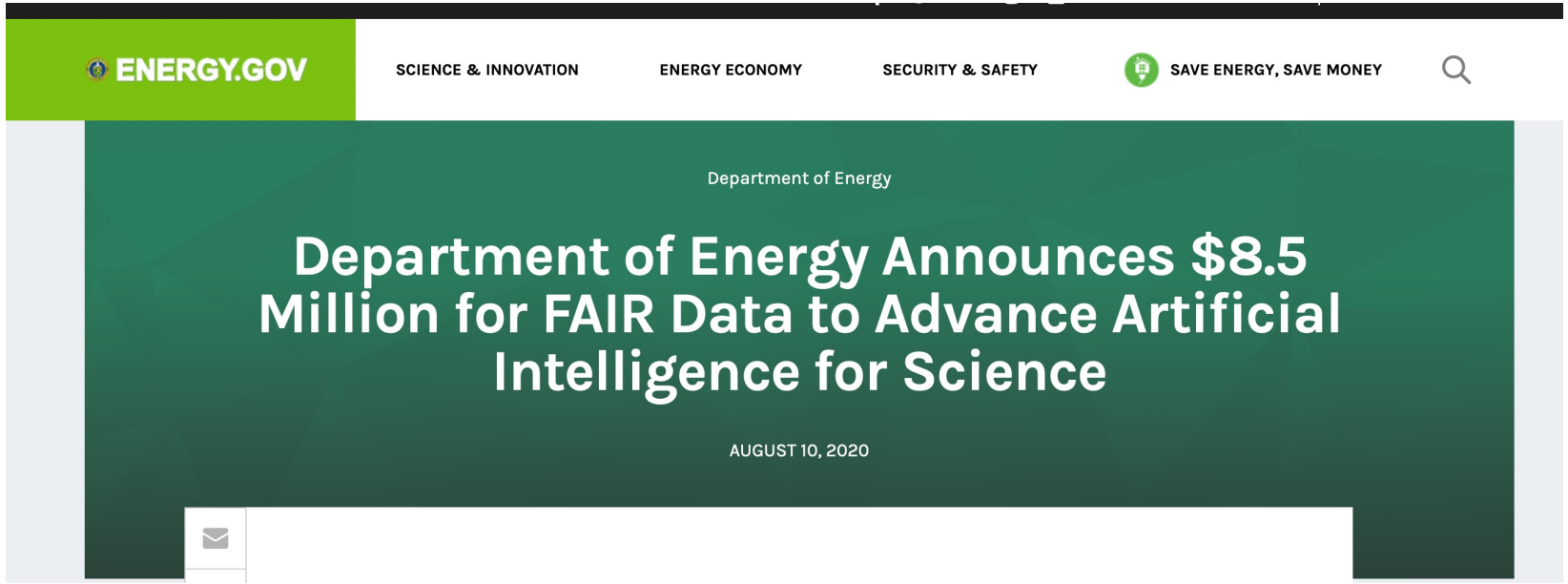
Alex Kulyavtsev, **alexku@anl.gov,** **Argonne National Laboratory**
Andrew Cherry, acherry@alcf.anl.gov, Argonne National Laboratory
Kevin Harms, harms@alcf.anl.gov, Argonne National Laboratory
Gordon McPheeters, gmcpheeters@anl.gov, Argonne National Laboratory

Argonne Leadership Computing Facility (ALCF)

- Provides HPC resources via a competitive allocation process (INCITE & ALCC) where major projects are peer reviewed and time is awarded
  - Small allocation grants for project startups and investigations
- Results in a user base that consists of both internal (Argonne) users and external (anywhere else in the world) users
- Projects have collaborators across government, academia, and industry
- What do ALCF users do if they want to share this data?
  - Previously, any person the Principal Investigator (PI) wanted to share data with had to have
    - An ALCF account
    - Part of the PIs project (which implies they can runs jobs, charge hours, create data, read all data, etc.)
    - This type of access was generally only given to project members
- Globus Sharing
  - Enables PI to provide access to specific data that resides in Eagle
  - Does not require ALCF account or project access
  - PI defines access rights and privileges to the shared data
  - Data can be transferred to other systems via Globus transfer mechanisms

FAIR

- US Department of Energy (DOE) promotes any DOE funded research apply FAIR principles to data
  - https://www.energy.gov/articles/department-energy-announces-85-million-fair-data-advance-artificial-intelligence-science
- FAIR – Findable, Accessible, Interoperable and Reusable
- Funded researches should make generated data ”FAIR”
- ALCF strives to provide mechanism to enable researchers to provide “FAIR” data
  - Globus Sharing is a step in that process



Argonne NATIONAL LABORATORY

ALCF Systems

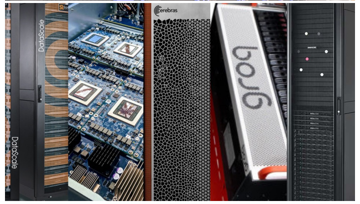
ALCF Systems

- **Polaris** (CPU+GPU)
  - —Top500: Rmax 25.82 PFlop/s, Rpeak 34.16 PFlop/s
  - —560 nodes: 1x AMD EPYC Milan 7543P + 4x NVIDIA A100

- **Theta** (CPU)
  - —Top500: Rmax 6.92 PFlop/s, Rpeak 11.66 PFlop/s
  - —4392 nodes: 1x Intel Xeon Phi 7230 (KNL)

- **ThetaGPU** (CPU+GPU) part of Theta
  - —GPU-accelerated computing pathfinder, Rpeak 3.9 PFlop/s
  - —24 nodes: 2x AMD EPYC Rome 7742 + 8x NVIDIA A100

- **Cooley** (CPU+GPU)
  - —Visualization + Data Analysis, Rpeak 0.3 PFlop/s
  - —126 nodes: 2x Intel Haswell E5-2620 + 1x NVIDIA Tesla K80



- **AI Testbed** (various AI accelerators)
  - —Available for Allocation Requests (DD): Cerebras CS-2, SambaNova DataScale
  - —Access Forthcoming: Graphcore MK-1, Groq, Habana Gaudi

Argonne NATIONAL LABORATORY

Polaris

https://www.alcf.anl.gov/polaris

- ALCF's latest computational resource
  - Provides on-ramp to Aurora
- Generally available ALCF resource
  - INCITE, ALCC, DD

14	**Polaris** - Apollo 6500, AMD EPYC 7532 32C 2.4GHz, NVIDIA A100 SXM4 40 GB, Slingshot-10, HPE DOE/SC/Argonne National Laboratory United States	256,592	25.81	34.16

Top500 June 2022

Aurora

Argonne's upcoming exascale supercomputer will leverage several technological innovations to support machine learning and data science workloads alongside traditional modeling and simulation runs.

PEAK PERFORMANCE
≥2 Exaflop DP

Intel® Xe ARCHITECTURE-BASED GPU
Data Center GPU Max Series

INTEL® XEON® SCALABLE PROCESSOR
Intel Xeon CPU Max Series

PLATFORM
HPE Cray EX

Compute Node
2 Intel® Xeon® CPU Max Series processors
6 Intel® Data Center GPU Max Series GPUs
Unified Memory Architecture
8 fabric endpoints; RAMBO

GPU Architecture
Intel® Data Center GPU Max Series
Tile-based chiplets, HBM stack,
Foveros 3D integration, 7nm

CPU-GPU Interconnect
CPU-GPU: PCIe
GPU-GPU: Xe Link

System Interconnect
HPE Slingshot
Dragonfly topology with adaptive routing

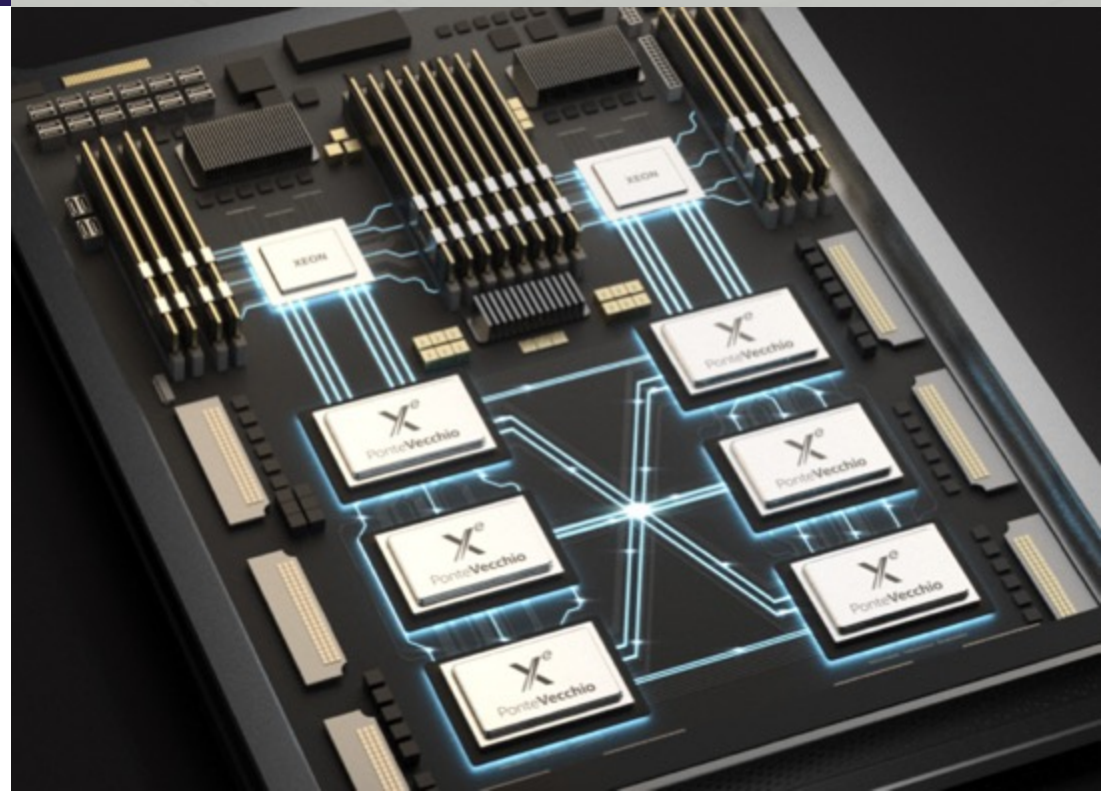
Network Switch
25.6 Tb/s per switch,
from 64 – 200 Gbs ports
(25 GB/s per direction)

High-Performance Storage
≥230 PB, ≥25 TB/s (DAOS)

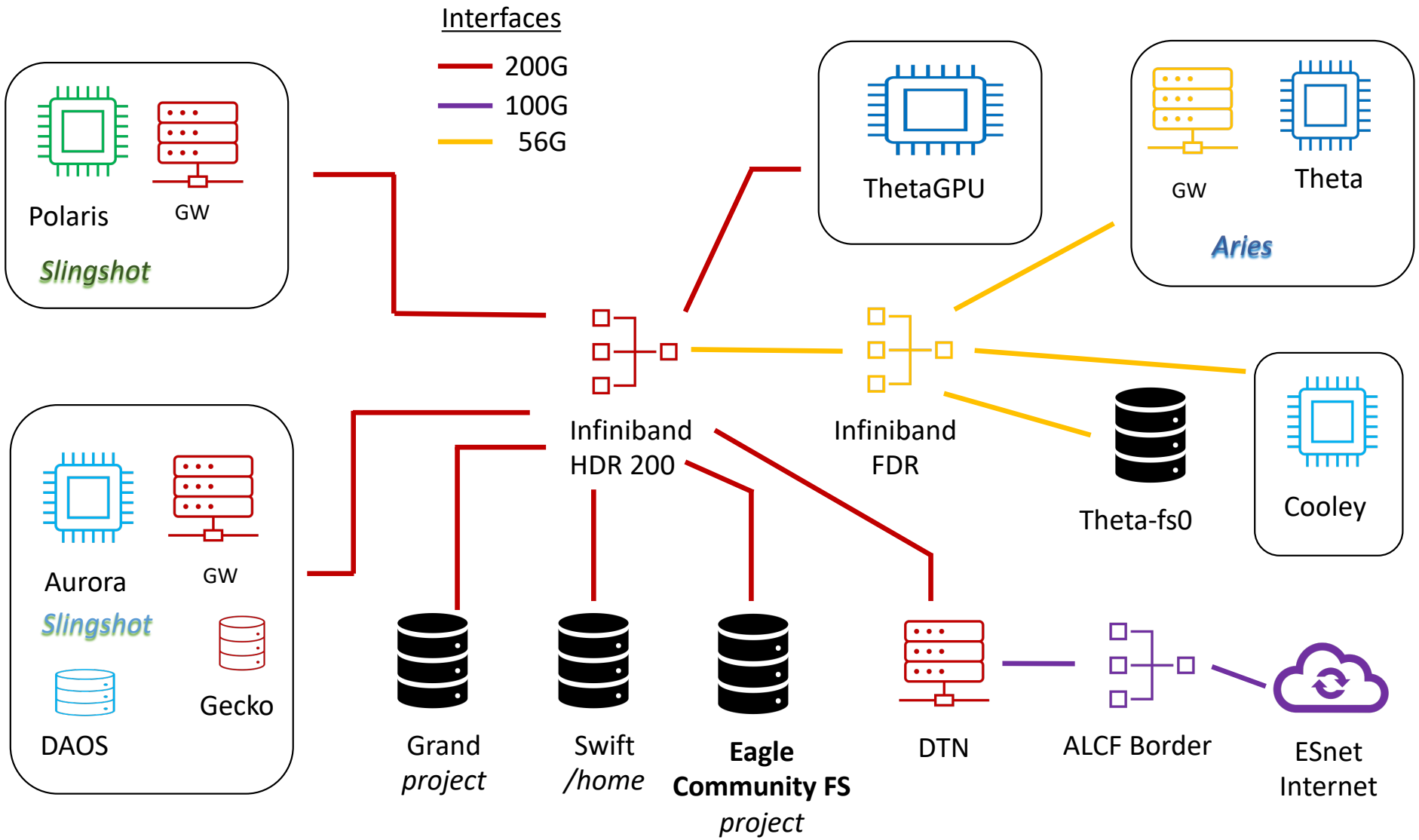
Programming Models
Intel oneAPI, MPI, OpenMP, C/C++,
Fortran, SYCL/DPC++

Node Performance
>130 TF

System Size
>10,000 compute nodes



System Architecture



Eagle - Hardware

- HPE ClusterStor E1000
- 100 PB of useable capacity
- 650 GB/s sustained large block I/O

- Total 8480 HDD + 284 NVMe drives
- Composed into
  - — 10 racks
  - — 10 Metadata Management Units (2 MDS+4 MDT)
  - — 20 Scalable Storage Units
    - Each has 2 OSS controlling 8 HDD enclosures
    - Journal and WIB SSD
- 20 MDS + 40 OSS
- 160 OST (two per 4u106 HDD enclosure)
  - — 53x16TB HDD/OST (5*(8d+2p)+3)
  - — HPE GridRAID (RAID6)
- 40 MDT (four per enclosure)
  - — 24 x 3.84 TB SSD (22d+2s)
  - — RAID10
- Eight racks of Ten shown ->

Figure 2. Rack Layout Diagram – Campaign Storage System (1 of 2)

Eagle - Lustre

- HPE Neo 4.5-011
- Lustre: cray-2.12.6.1
- ldiskfs based OST
- HDR Infiniband (200 Gb/sec)

- project quotas
- DNE

- Planning upgrade to Neo-6.x
  - —lustre 2.15

IOR results from Polaris Acceptance Testing

Polaris / Grand E1000

Eagle has same configuration

	OST	GW	size	CN	ppn	Write	Read
FPP	156	32	4GiB	512	8	366 GB/s	270 GB/s
Shared	158	32	12GiB	512	9	254 GB/s	371 GB/s

32 gateways: HDR 200 Gbit/s to SS10 100 Gbit/s

BW limited by GW count 32 used for test (out of 50)
and SS 10 (100 Gbit/s)
some OST was in rebuild

Argonne NATIONAL LABORATORY

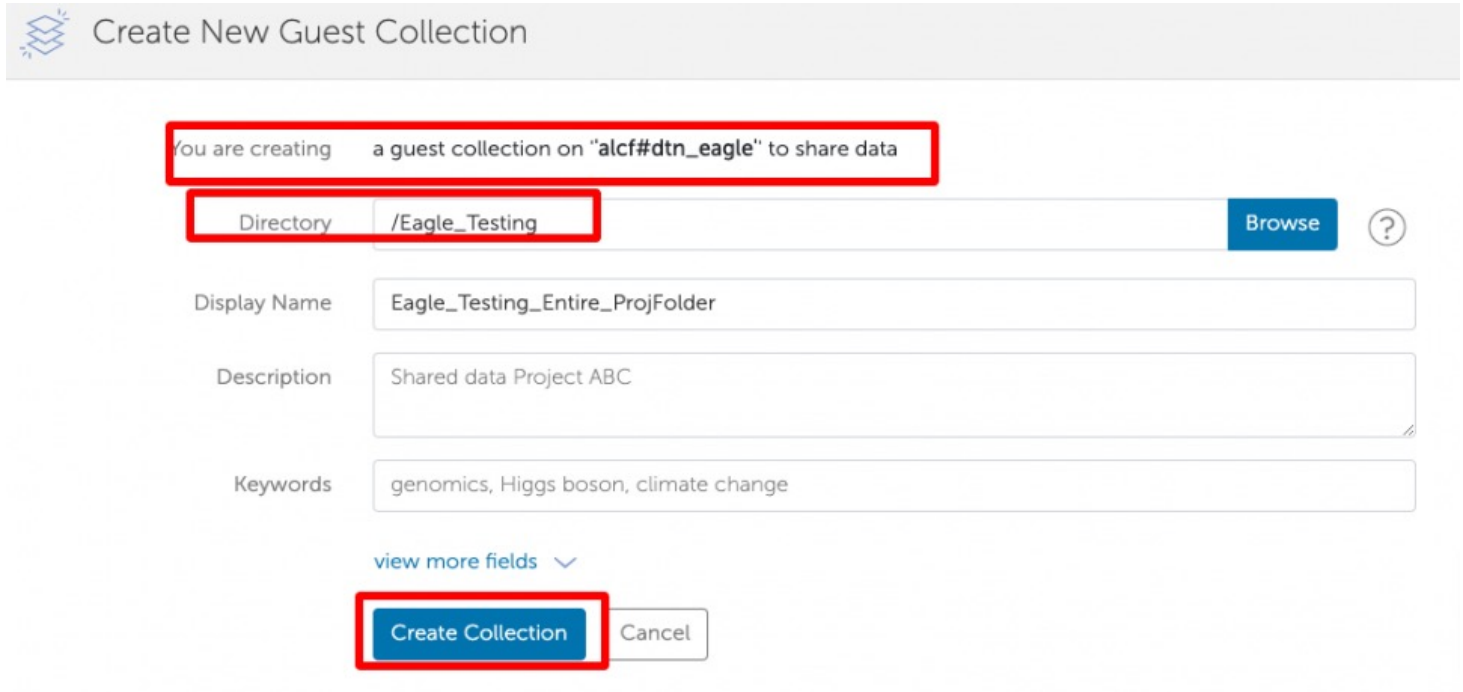
Globus

https://docs.alcf.anl.gov/data-management/acdc/eagle-data-sharing/

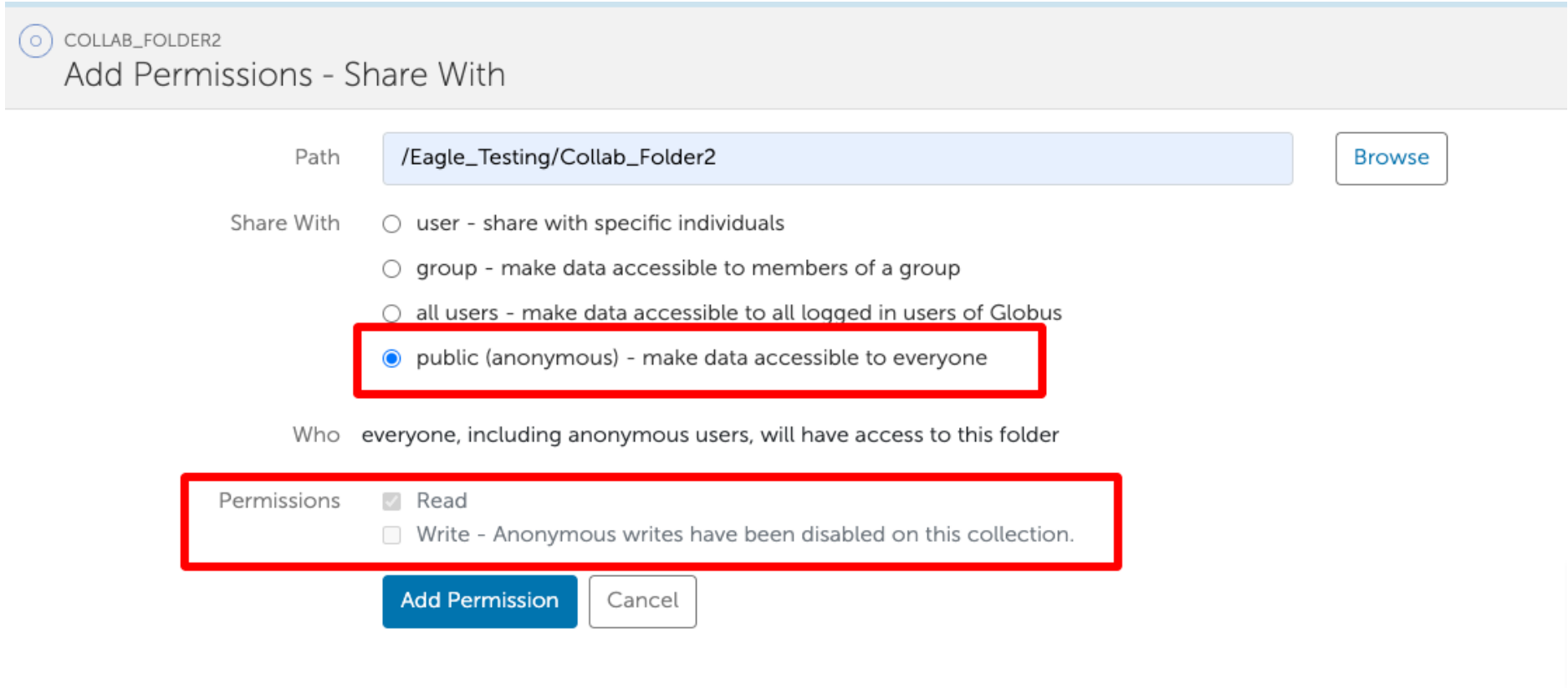
Globus Sharing

- Globus
  - Globus is a service that provides research data management, including managed transfer and sharing. It makes it easy to move, sync, and share large amounts of data. Globus will manage file transfers, monitor performance, retry failures, recover from faults automatically when possible, and report the status of your data transfer.
  - Utility that enables efficient transfer between major LCF data centers as well as support to local clients, like a desktop or laptop
- Globus Sharing
  - Provides the ability to share data with other users even if they do not have an account on the system where the data resides
  - Data on ALCF systems can easily be shared with collaborators who are at ALCF or elsewhere. The PI has full control over which files a collaborator can access, and whether they have read-only or read-write permissions.
  - The PI can also choose to share their data to all authenticated Globus users from any institution that uses Globus, or can even allow anonymous public access on a read-only basis (anonymous write is disallowed).
  - Transfers from one Globus collection to another is performed using the Globus File Manager web application at https://app.globus.org/
  - Direct web browser access to individual shared files using https is also possible by using the "Get Link" function to publish or share a URL.

Sharing in Pictures



Sharing Permissions



Data Transfer Node (DTN) configuration

- Four data transfer nodes provide Globus access to all ALCF Lustre filesystems (sharing is only permitted on Eagle)
- DTN specifications
  - —64 processor cores per node
  - —External network connectivity via 2 bonded 100Gbit Ethernet interfaces
  - —200GBit IB (4X HDR) for internal DTN-to-storage network
- Lustre client : HPE cray-2.12.B58
- TCP tuning per ESnet guidelines: https://fasterdata.es.net/host-tuning/linux/
  - —TCP buffer sizes set to maximum (2GB)
  - —CPU governor set to 'performance'
  - —Disable irqbalance and manually configure IRQ affinity for network interfaces (only necessary on multi-socket systems)
- Globus endpoint is configured to allow a maximum of 96 files in flight simultaneously per transfer job, with up to 16 parallel network streams per file. In general use, preferred concurrency is 64 files simultaneously with 4 streams per file.

Globus Setup

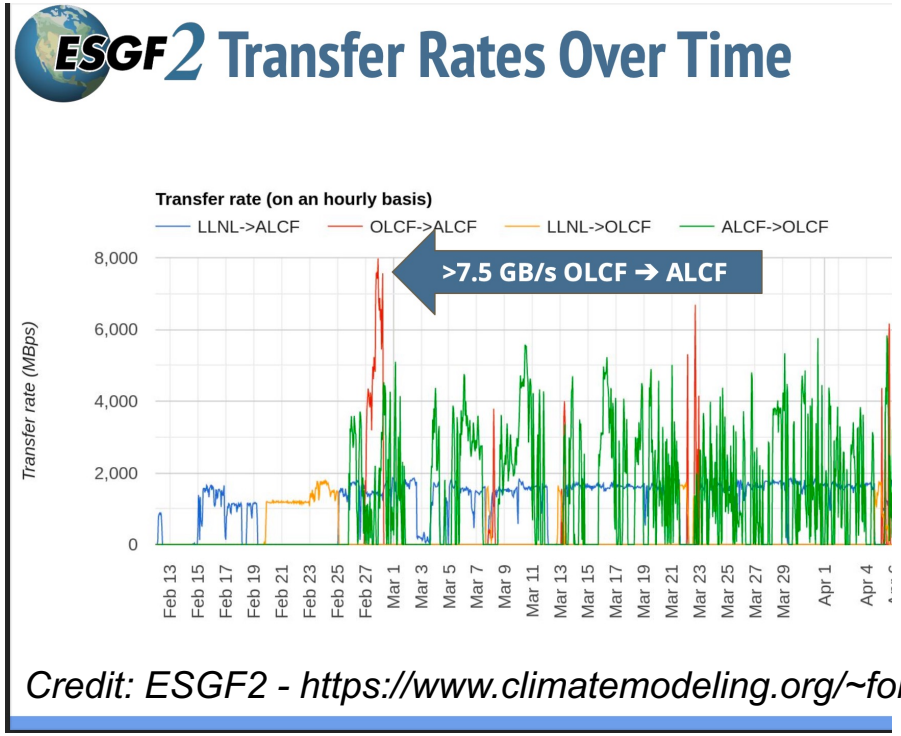
- Globus
  - Web-based application service operated by globus.org
  - Sites maintain “endpoints” which consist of data transfer nodes (DTN) hosting relevant file systems.
  - An endpoint can host multiple "collections", which are essentially views of specific filesystem locations. "Mapped collections" are maintained by ALCF admins (one per filesystem) and provide authenticated access for ALCF users. On Eagle, users may also create their own "guest collections" which provide shared access to non-ALCF users.
  - ALCF endpoints run Globus Connect Sever v5.4
- All ALCF project members can access their project directory via the **alcf#dtn_eagle** mapped collection
- Only the PI for a project is allowed to share, and they are only permitted to share from their own project space.
- Guest collections are tied to the PI's account and rely on that account remaining active (e.g. they are limited to the lifetime of the project)
- Direct https access to collections is enabled (allows direct download of files using a web browser)
- UDT available as an alternate (non-default) transport, can improve performance for high-delay network paths
- ALCF had Globus add several new features to support the Eagle use case:
  - Fine grained sharing policies to ensure PIs can only share from their own project directories - extra layer of security on top of file permissions
  - The option to administratively prohibit write access for anonymous sharing

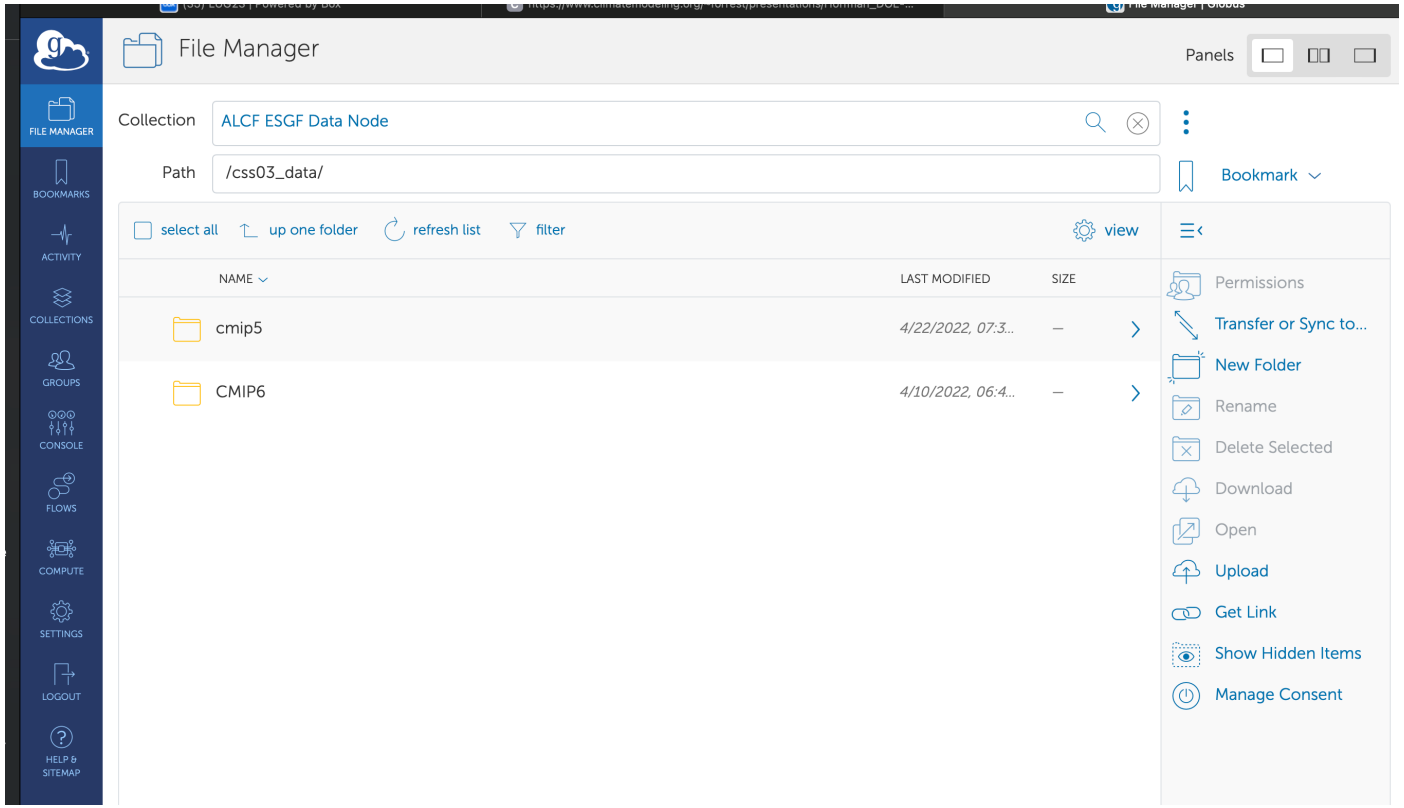
Considerations

- Eagle is mounted on compute nodes, which facilitates I/O being written directly to shared directories
- Provides easy method to share data with no need to copy or have duplicate data
  - —No need to keep data synchronized
  - —Saves time and space
- Consequently, all project members can read/write data into directories based on common unix group used by project
  - —PI could share a folder with some data
  - —Project member could write new data into this directory and it would become shared

Use Case: Earth System Grid Federation (ESGF) 2

- ESGF2 creating replication of existing dataset consisting of ~7.5 PB of data
  - —Provides access to Earth System Model simulation and analysis data
- Data transferred from LLNL to ALCF @ 1.5 GB/s between mid-February to May 4, 2022
  - —17,347,671 directories
  - —28,907,532 files
  - —4-6 GB/s ALCF -> OLCF

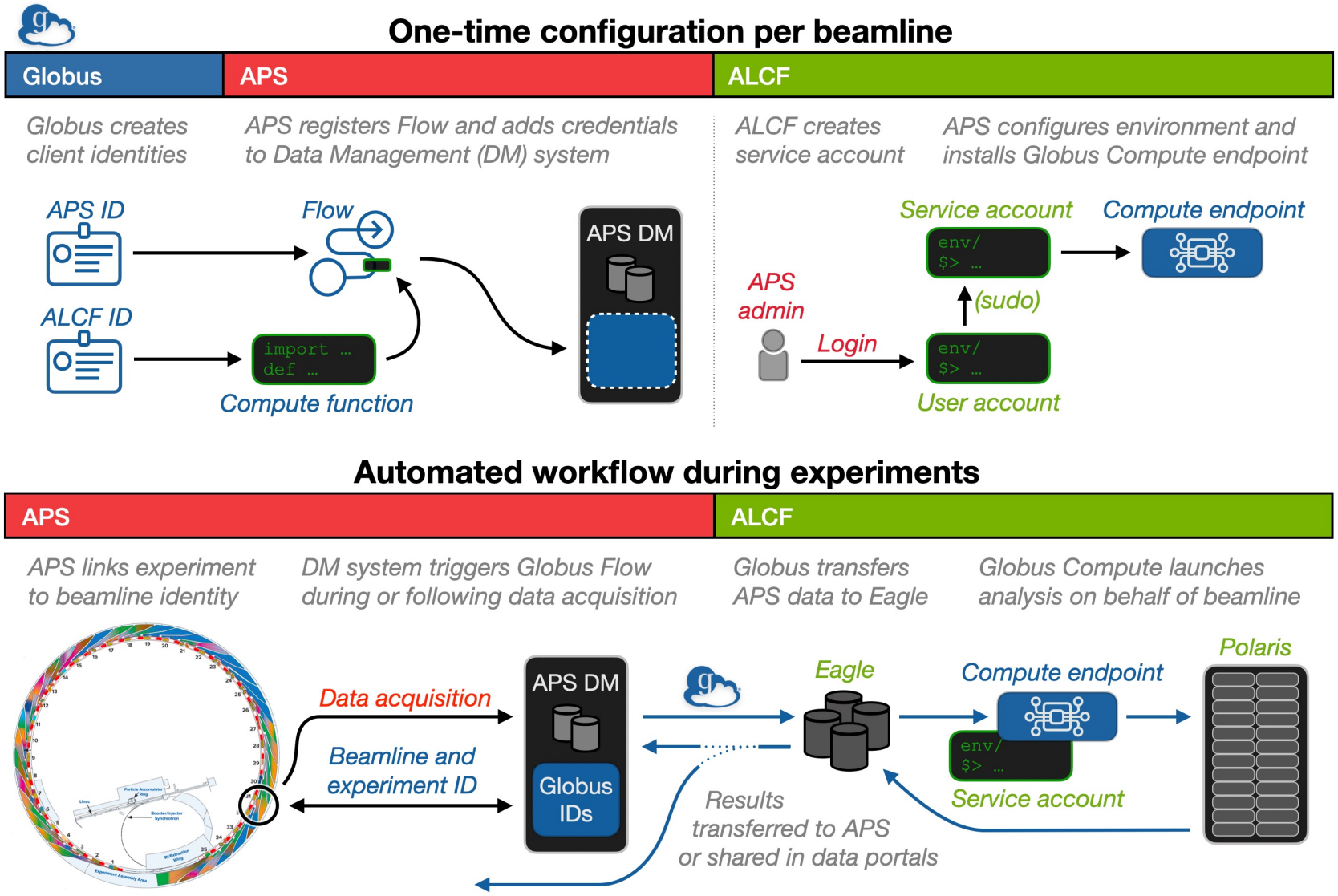




Credit: ESGF2 - https://www.climatemodeling.org/~forrest/presentations/Hoffman_DOE-Data-Days_20220602.pdf

Use Case: Automated APS Workflow With On-Demand Computing

- APS Data Management system
  - —Trigger workflows following data acquisition
- Globus Flows
  - —Automate workflows securely without human intervention
- Globus Compute
  - —Run jobs at ALCF on behalf of beamline
- On-demand computing
  - —Preemption to execute APS jobs
- Globus Share
  - Results shared from Eagle filesystem



Conclusion

- Provided a high-performance Lustre based parallel file system
  - Center-wide
  - Support Checkpoint I/O, Analysis I/O, input data, code, …
  - POSIX compliance
- Enable scientists to share data produced with collaborators or to the public
  - Minimal overhead
  - No requirements for ALCF accounts or project membership
  - Efficient data transfer provided via Globus and Globus endpoints (ALCF DTNs service Eagle)
- Enable scientists to implement FAIR principles for data generated from DOE funded science

Acknowledgements

This research used resources of the Argonne Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC02-06CH11357.