DAOS @ ARGONNE

DAOS USER GROUP 2018

KEVIN HARMS

ARGONNE LEADERSHIP COMPUTING FACILITY

Argonne NATIONAL LABORATORY

COLLABORATION

- Intel and Argonne collaborating on technologies used in DAOS
  - Mercury
    - RPC framework geared toward HPC
    - https://mercury-hpc.github.io
    - Support for multiple transports (OPA, Infiniband, Aries)
    - Used in DAOS CART implementation
  - Argobots
    - User-level threading
    - http://www.argobots.org
  - Gossip Protocol [1]
    - SWIM
    - Weakly consistent
    - Fault detection and dissemination
  - Object Placement [3]

- Argonne work done under the Mochi project [2]
  - *See next slide*

Mochi

http://www.mcs.anl.gov/research/projects/mochi/

Vision

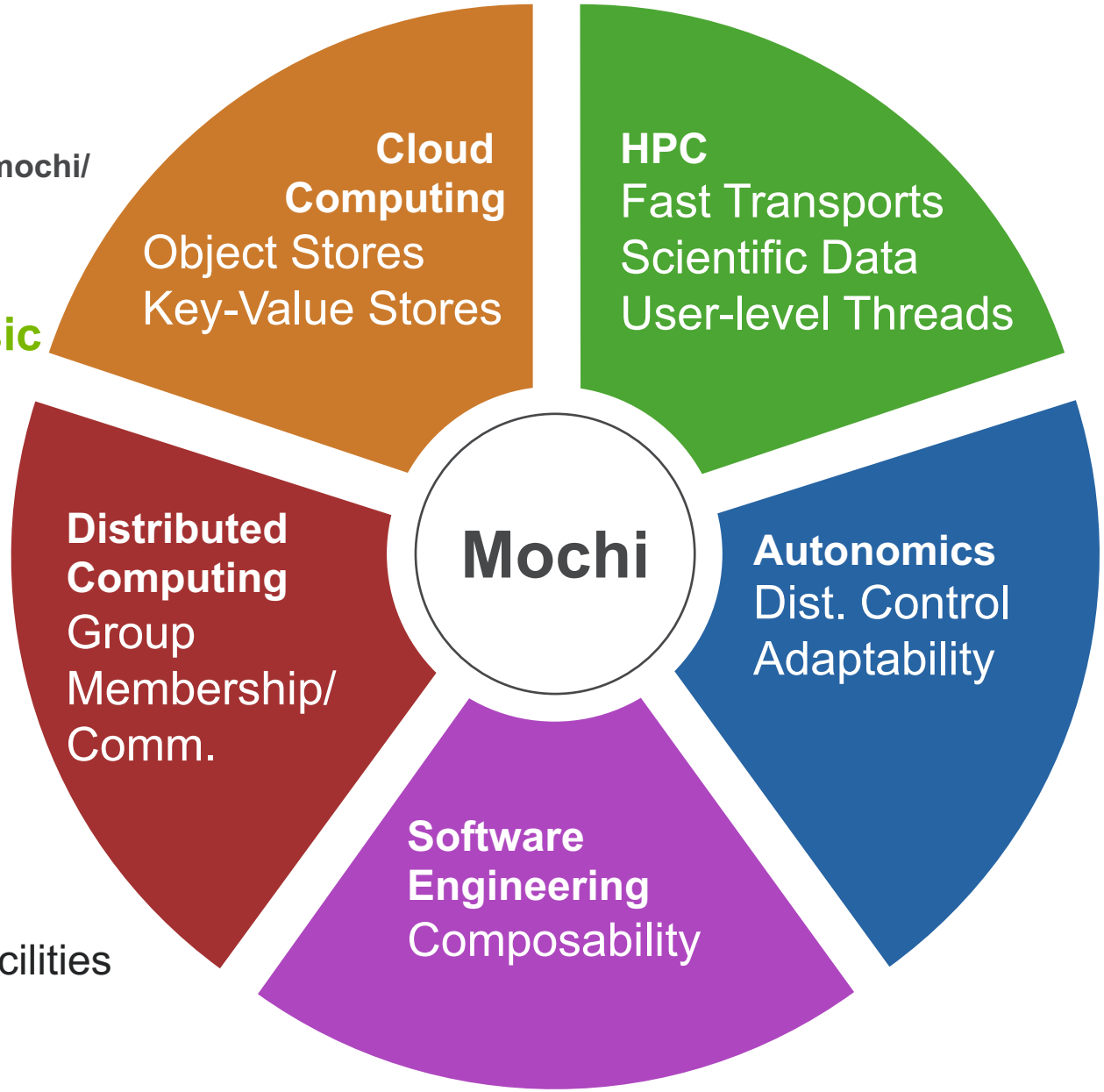
Specialized data services composed from basic building blocks

Approach

- Lightweight, user-space components
- Easy to build, adapt, and deploy
- Modern hardware support

Impact

- Better, more capable services for DOE science and facilities
- Significant code reuse
- Ecosystem for service development



Credit: Rob Ross, ANL

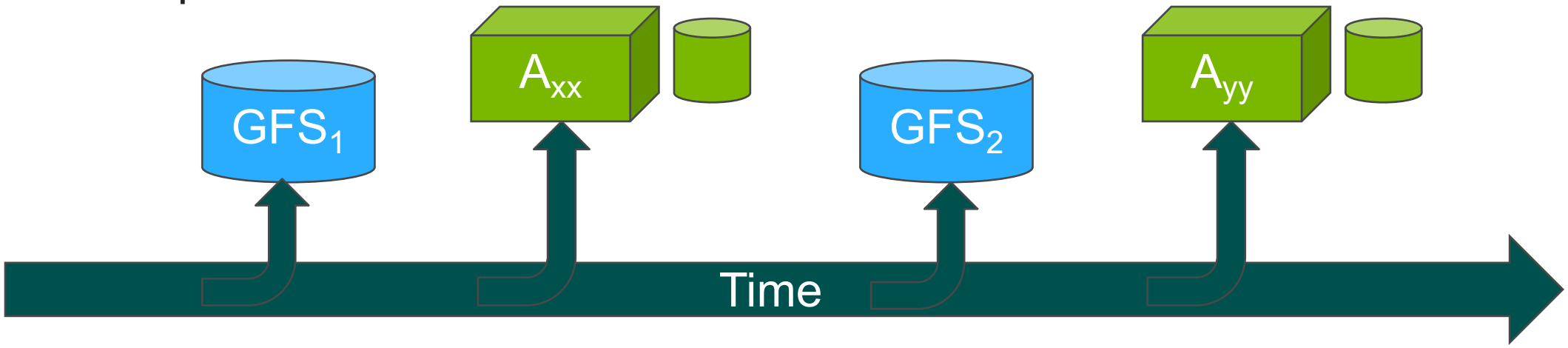
Carnegie Mellon University · HDF · Los Alamos National Laboratory EST. 1943 · Argonne National Laboratory

EVALUATION

- JLSE system – CY2019
  - 18 storage nodes
    - Allows testing 14+2 parity with 2 nodes for failover/rebuild
    - Cascade Lake / 3D XPoint NVDIMMs / 3D NAND NVMe / OPA
  - Evaluating
    - System management/administration functions
    - Failure and Recovery scenarios
    - Application / Middleware correctness
    - Performance of components
- Theta
  - Containerized (singularity) DAOS software stack
  - Utilizing local NVMe and DDR for storage
  - Evaluate scalability

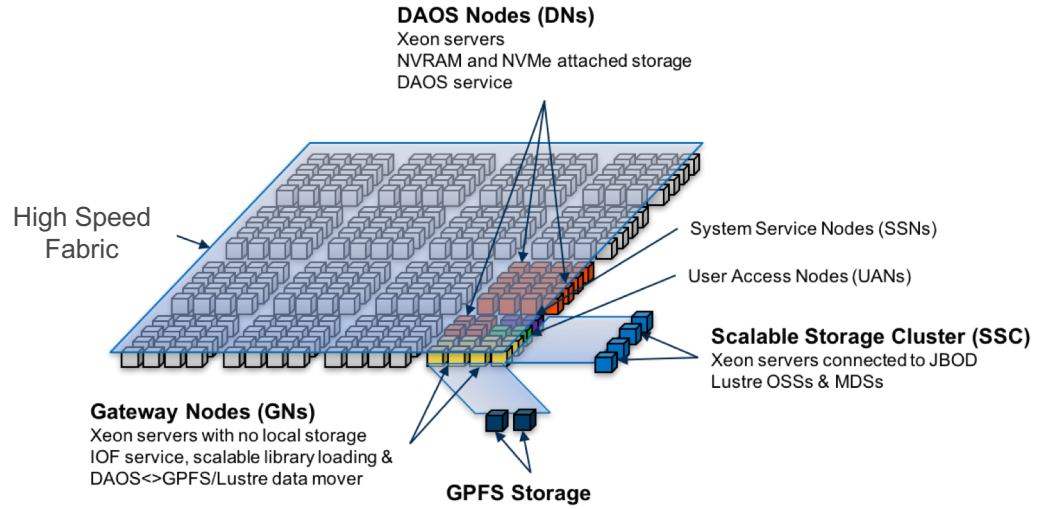
STORAGE PLAN

- Campaign/project storage
  - Capacity storage for extended project data life time
  - Strong support for legacy interfaces
  - Out-of-band from system acquisitions
- Dedicated storage with systems
  - Performance focused
  - Adequate capacity for month(s)



AURORA

- DAOS will be a component of the Aurora system
  - Exact details yet to come
- Desire to exploit DAOS performance capabilities using NVM technologies
  - High IOPs
  - High BW
- Traditional PFS augments DAOS by providing traditional project (campaign) storage
  - Legacy compatibility
  - Bulk capacity



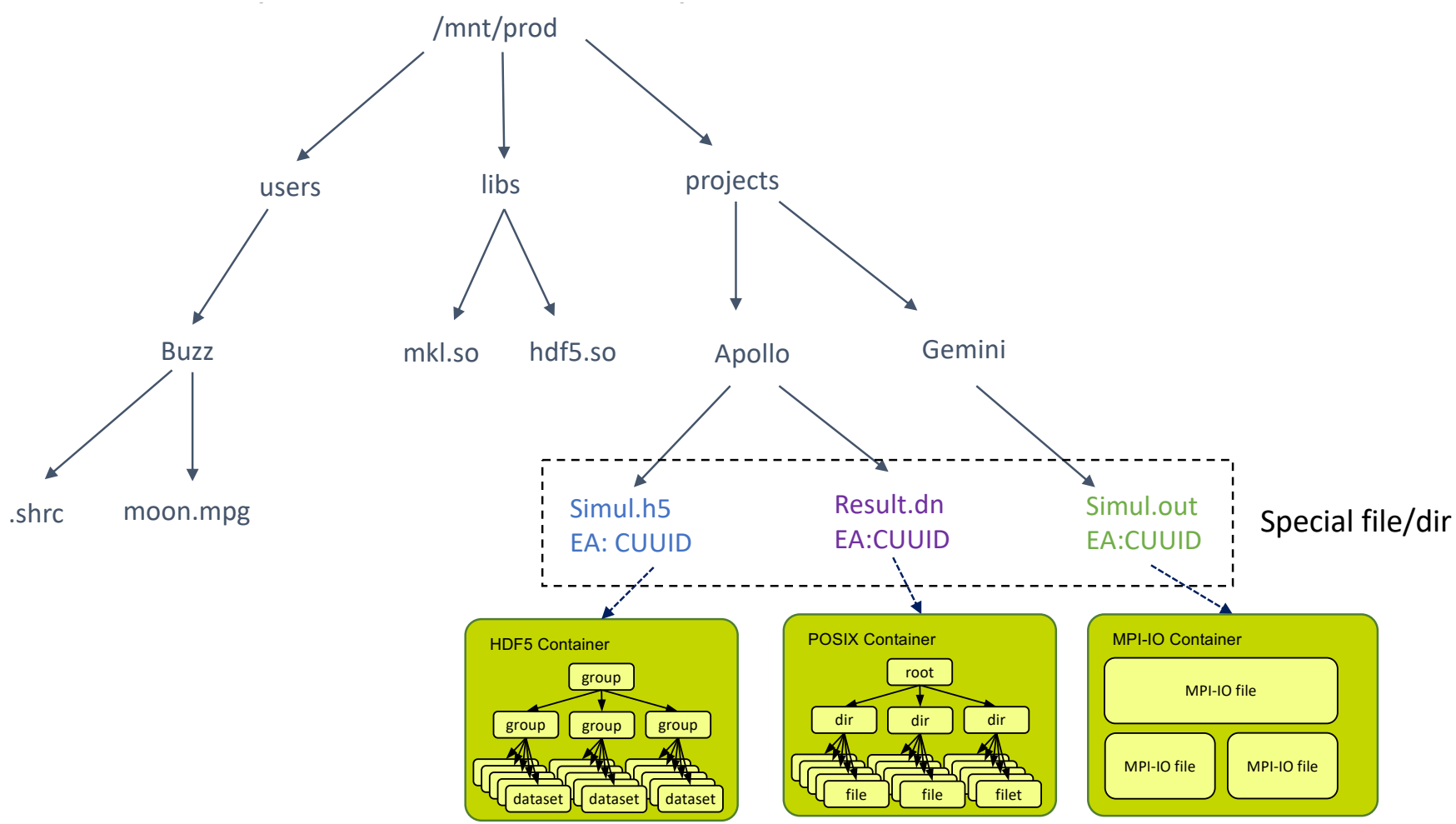
LUSTRE AND DAOS INTEGRATION

- Single unified name space
  - Make it easy for users to work with both Lustre and DAOS
  - Transparently recognize DAOS objects and
    - Interpret the contents of a container if possible
    - Pass off operations to DAOS client

- Presented at LAD by Johann
  - https://www.eofs.eu/_media/events/lad18/15_johann_lombardi_intel_cross_tier_unified_namespace_v3.pdf

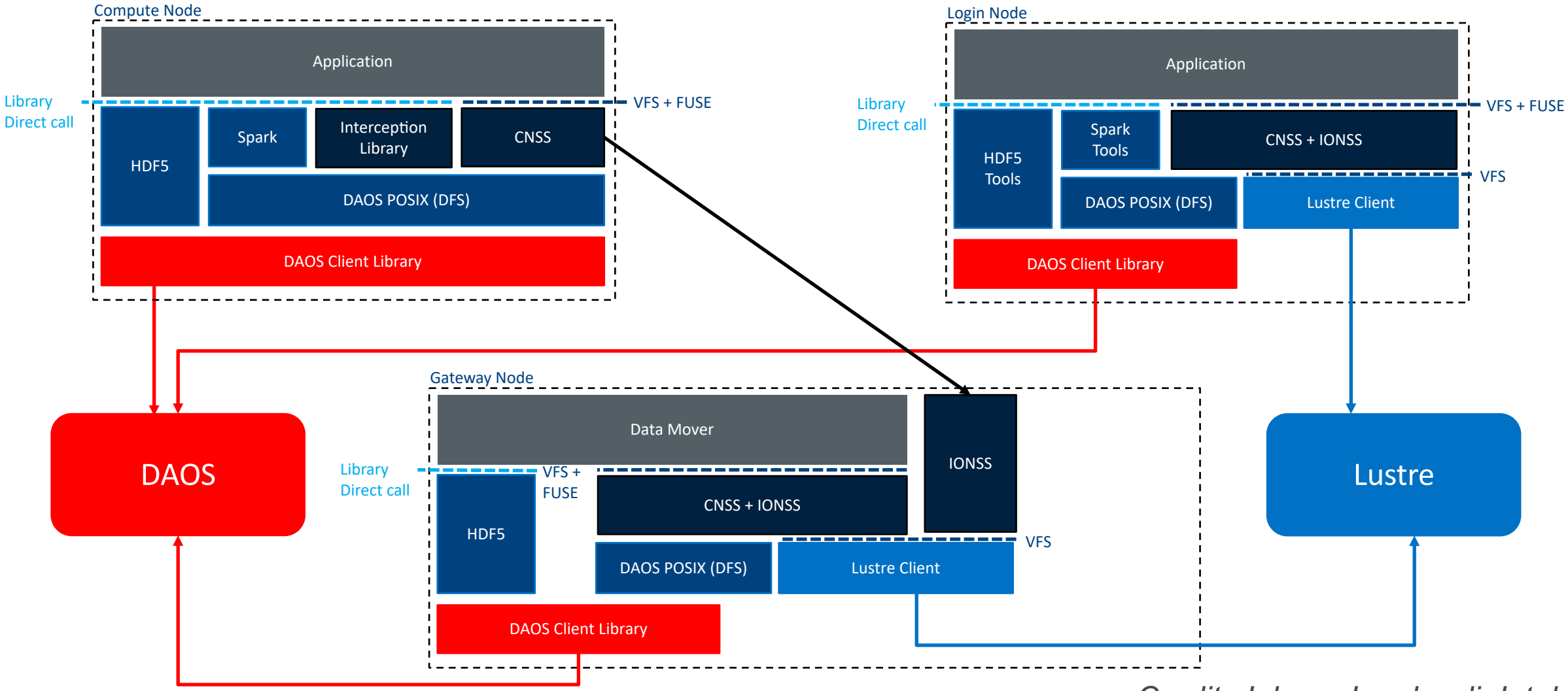
LUSTRE AND DAOS INTEGRATION

Regular PFS directories & files
HDF5 Container
DAOS POSIX Container
DAOS MPI-IO Container



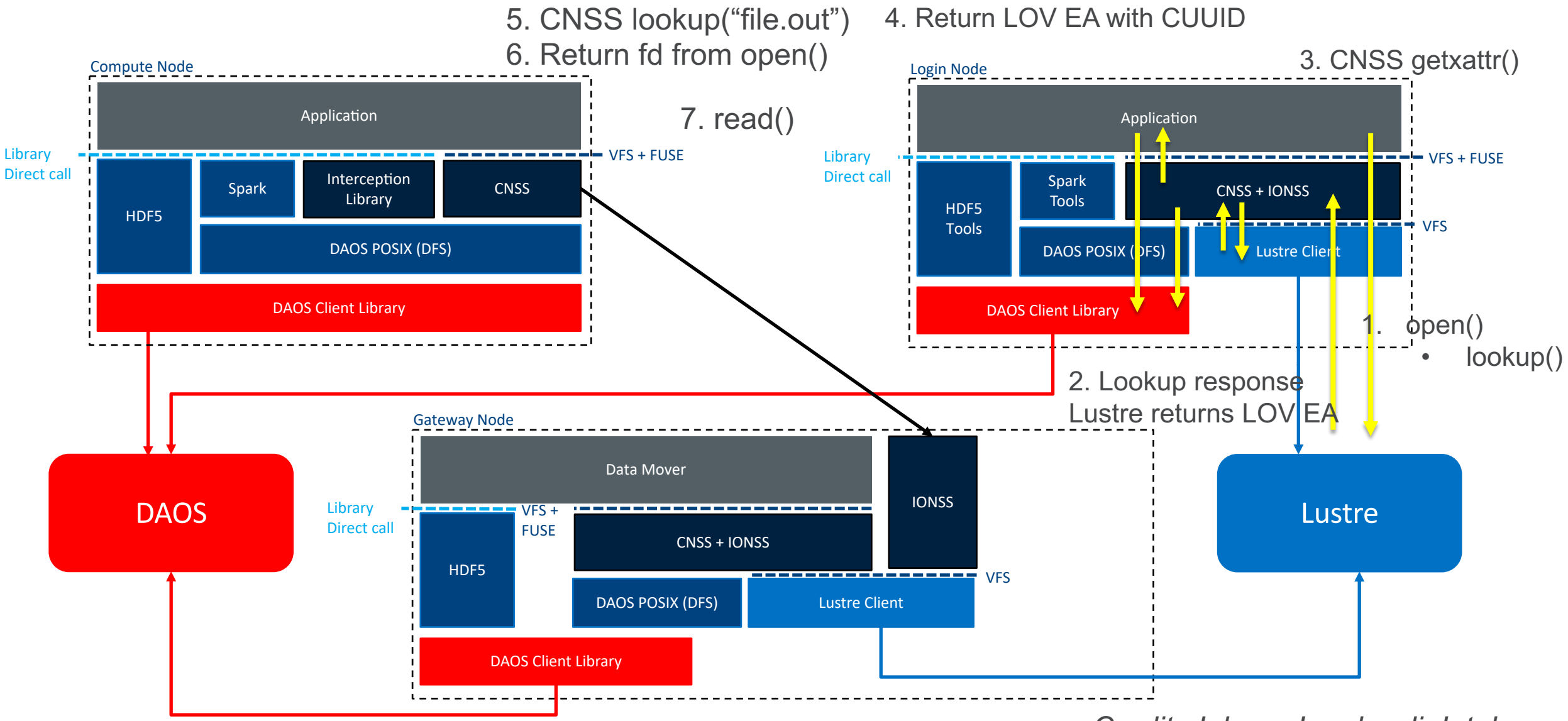
Credit: Johann Lombardi, Intel

POSSIBLE ARCHITECTURES



Credit: Johann Lombardi, Intel

POSSIBLE ARCHITECTURES



Credit: Johann Lombardi, Intel

CITATIONS

- [1] Snyder, Shane, Philip Carns, Jonathan Jenkins, Kevin Harms, Robert Ross, Misbah Mubarak, and Christopher Carothers. "A case for epidemic fault detection and group membership in hpc storage systems." In *International Workshop on Performance Modeling, Benchmarking and Simulation of High Performance Computer Systems*, pp. 237-248. Springer, Cham, 2014.
- [2] Matthieu Dorier, Philip Carns, Kevin Harms, Robert Latham, Robert Ross, Shane Snyder, Justin Wozniak, Samuel K. Gutie´rrez, Bob Robey, Brad Settlemyer, Galen Shipman, Jerome Soumagne, James Kowalkowski, Marc Paterno, and Saba Sehrish. “Methodology for the Rapid Development of Scalable HPC Data Services.“ In *2018 IEEE/ACM 3rd International Workshop on Parallel Data Storage & Data Intensive Scalable Computing Systems (PDSW-DISCS)*
- [3] P. Carns, K. Harms, J. Jenkins, M. Mubarak, R. Ross and C. Carothers, "Impact of data placement on resilience in large-scale object storage systems," *2016 32nd Symposium on Mass Storage Systems and Technologies (MSST)*, Santa Clara, CA, 2016, pp. 1-12.

ACKNOWLEDGEMENTS

- Rob Ross and the Mochi team
- Johann Lombardi and the Intel team
- This funded by the Argonne Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC02-06CH11357.