CROSS-TIER UNIFIED NAMESPACE

Johann Lombardi, Extreme Storage Architecture & Development, Intel

LAD'18, France

NOTICES AND DISCLAIMERS

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration.

No computer system can be absolutely secure.

Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. For more complete information about performance and benchmark results, visit http://www.intel.com/benchmarks .

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit http://www.intel.com/benchmarks .

Intel® Advanced Vector Extensions (Intel® AVX)* provides higher throughput to certain processor operations. Due to varying processor power characteristics, utilizing AVX instructions may cause a) some parts to operate at less than the rated frequency and b) some parts with Intel® Turbo Boost Technology 2.0 to not achieve any or maximum turbo frequencies. Performance varies depending on hardware, software, and system configuration and you can learn more at http://www.intel.com/go/turbo.

Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice.

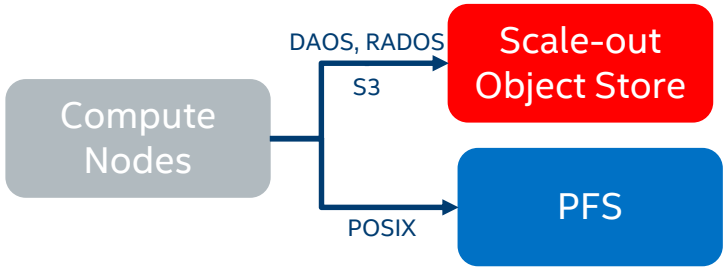
Cost reduction scenarios described are intended as examples of how a given Intel-based product, in the specified circumstances and configurations, may affect future costs and provide cost savings. Circumstances will vary. Intel does not guarantee any costs or cost reduction.

Intel does not control or audit third-party benchmark data or the web sites referenced in this document. You should visit the referenced web site and confirm whether referenced data are accurate.
Intel, the Intel logo, and Intel Xeon are trademarks of Intel Corporation in the U.S. and/or other countries.
*Other names and brands may be claimed as property of others.

Agenda

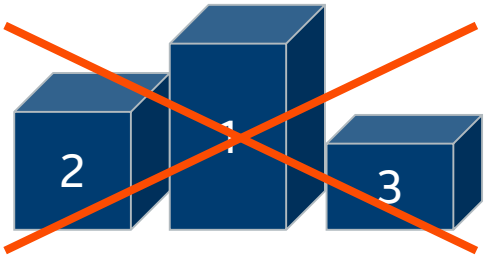
What this talk **is** about

- Multi-tier integration
  - Scale-out object store / DAOS
  - Parallel File System (PFS) / Lustre
- Expose unified namespace to end users
- Efficient dataset migration

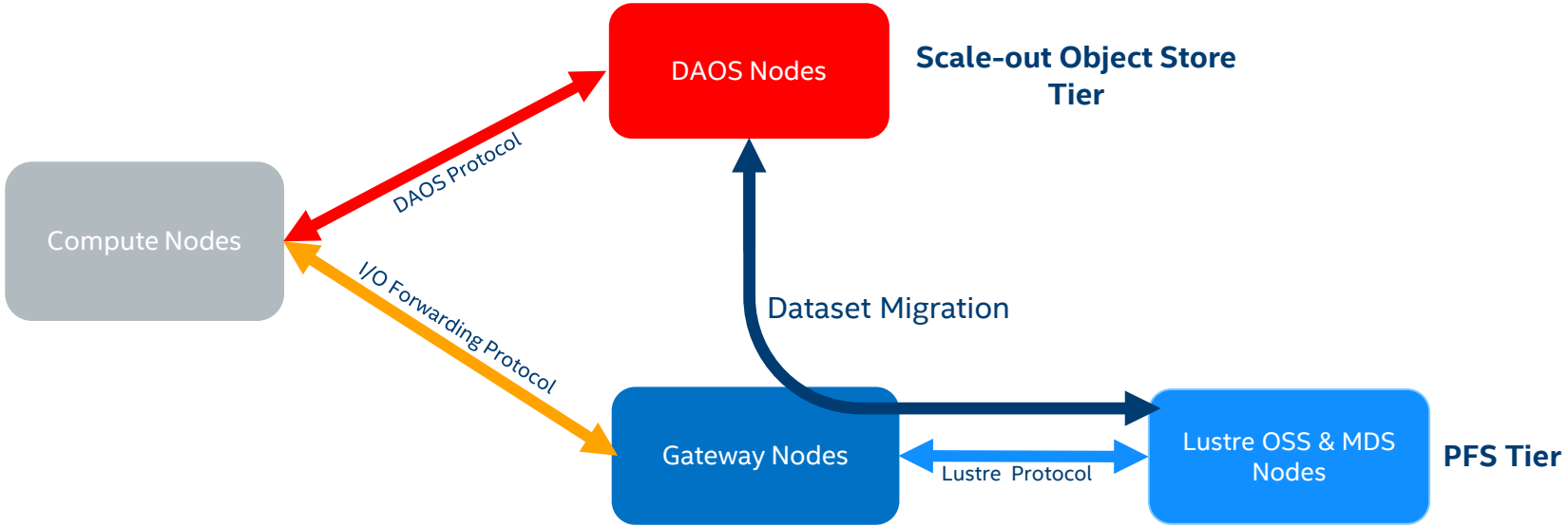


What this talk **is not** about

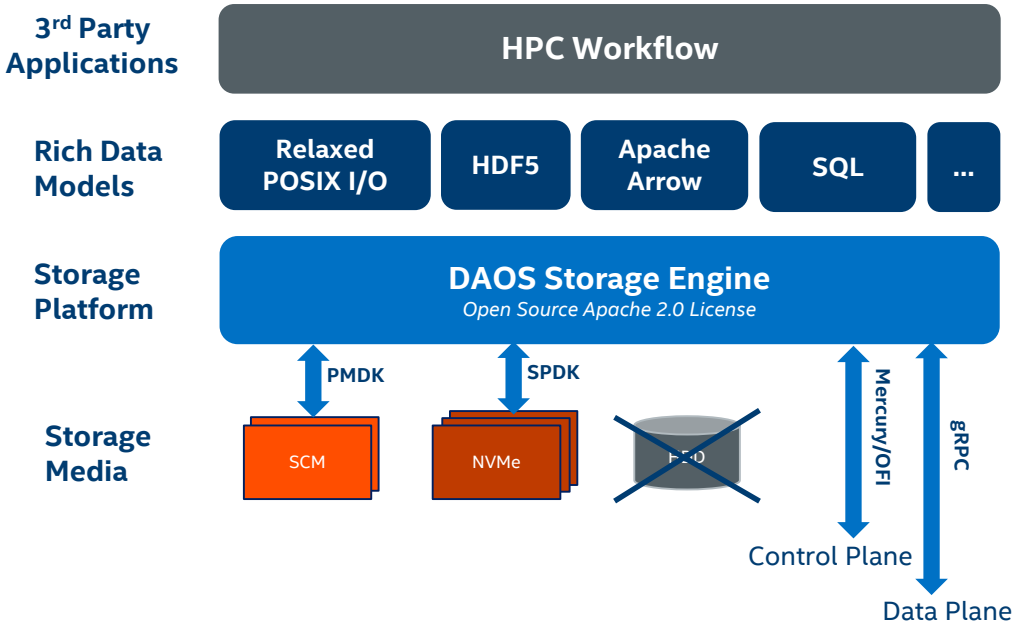
- Burst buffers or transparent caching
- DAOS internals
  - Ping me separately if you are interested in the open-source DAOS project
- A comparison between Lustre and DAOS



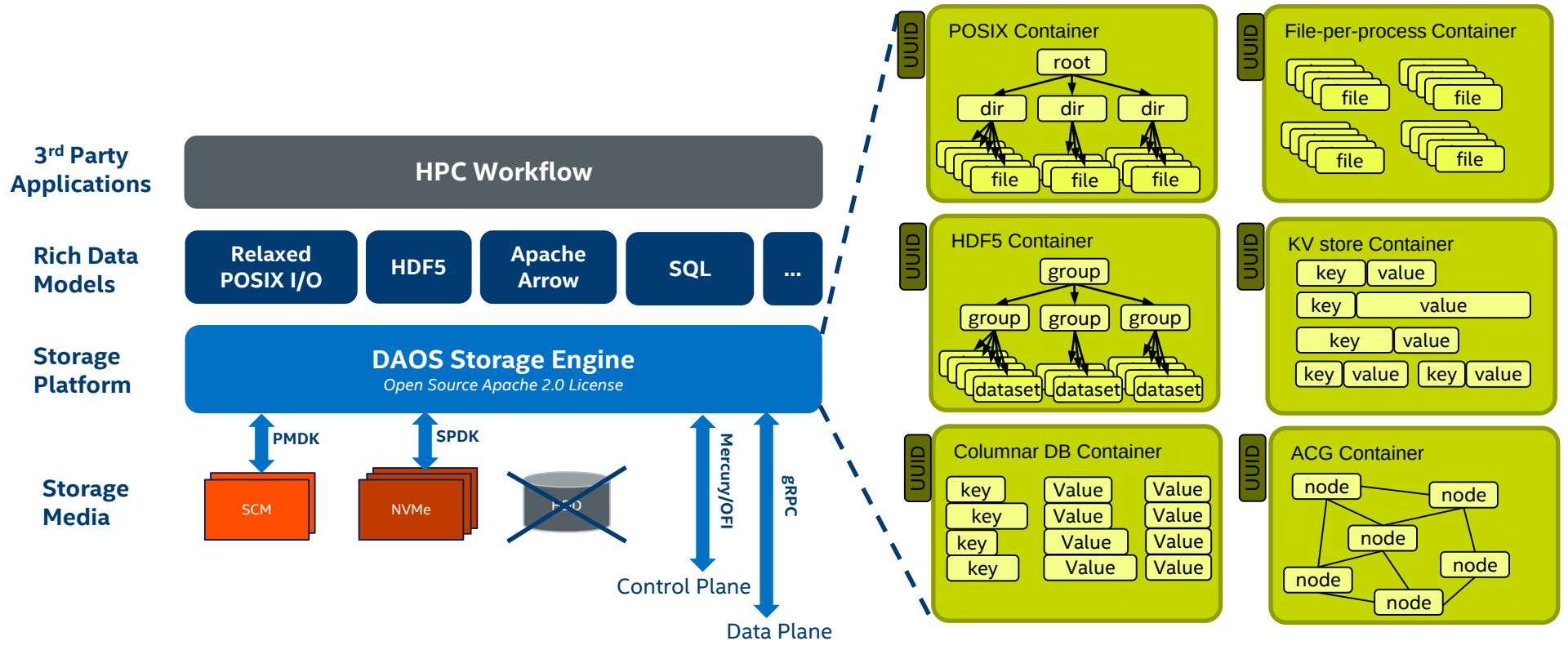
Targeted Storage Architecture



Distributed Async Object Storage

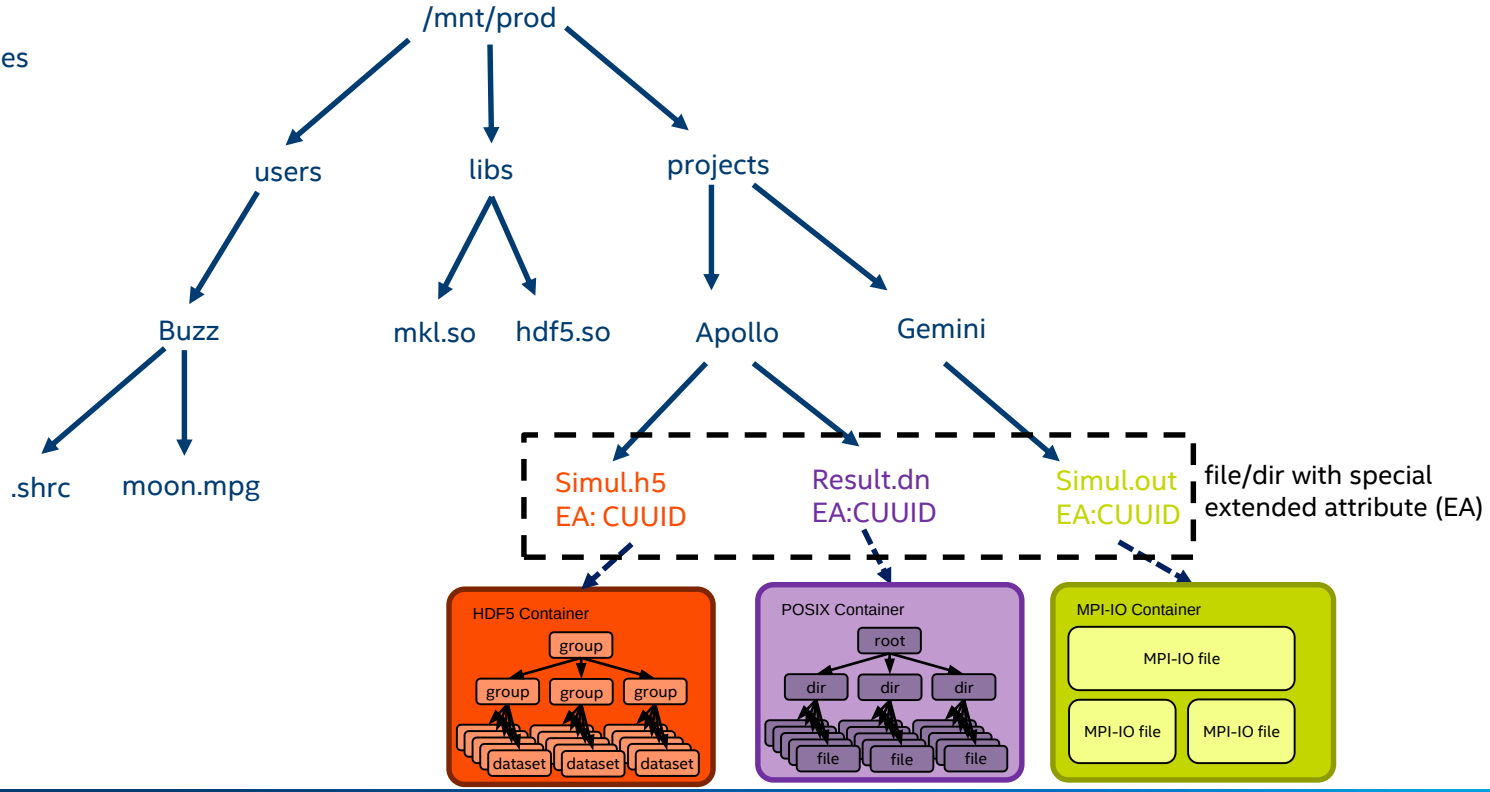


Distributed Async Object Storage



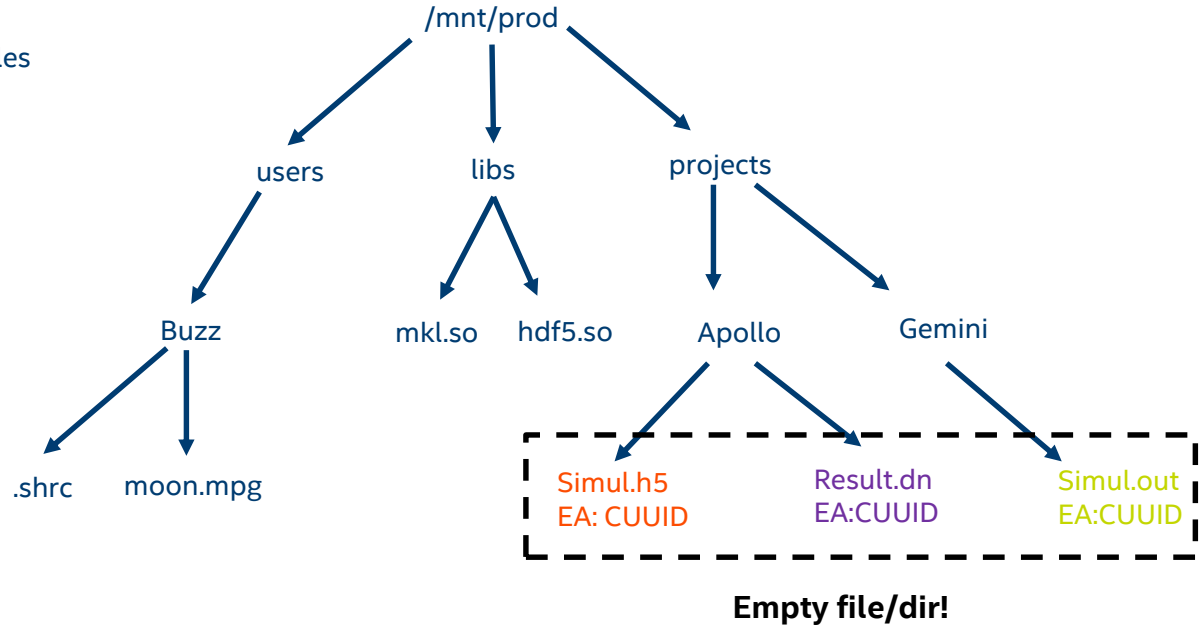
Unified Namespace Concept

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

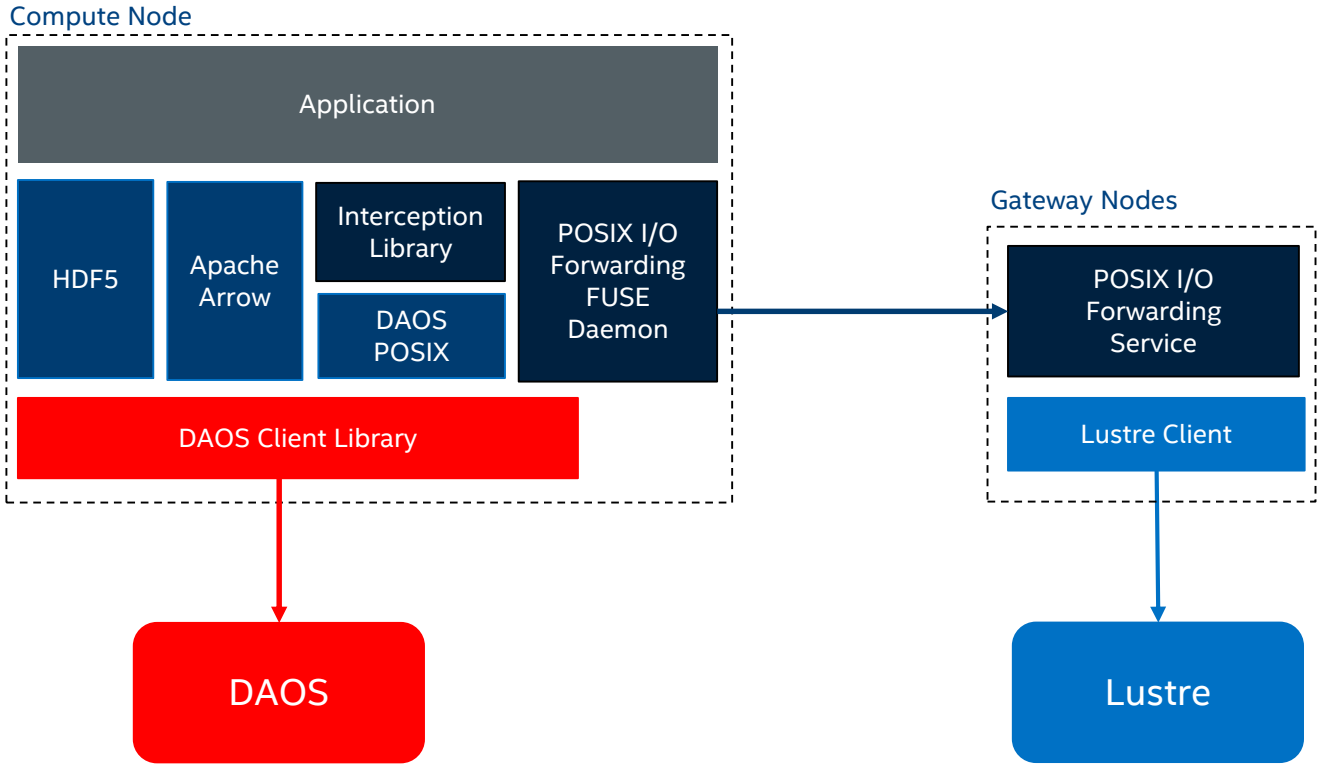


What's really stored in the PFS?

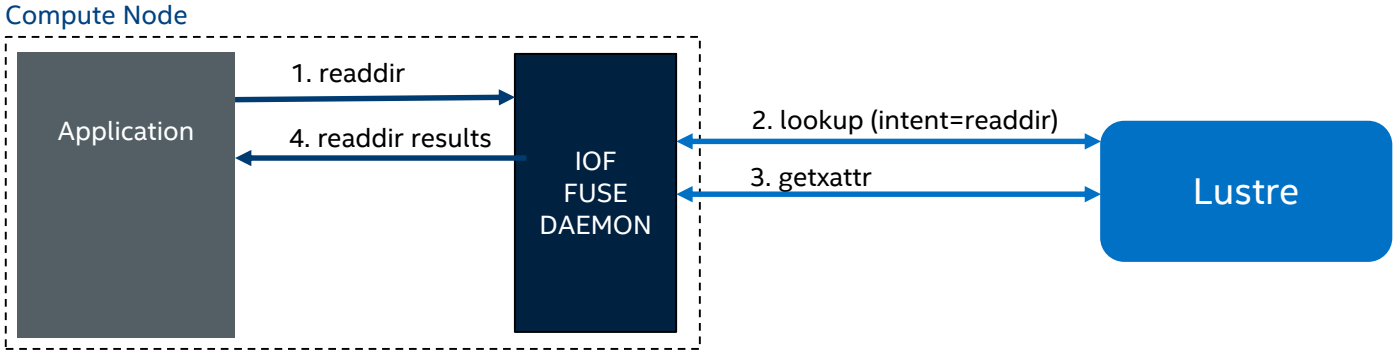
Regular Lustre directories & files
HDF5 Container
DAOS POSIX Container
DAOS MPI-IO Container



Unified Namespace Implementation – POSIX IOF



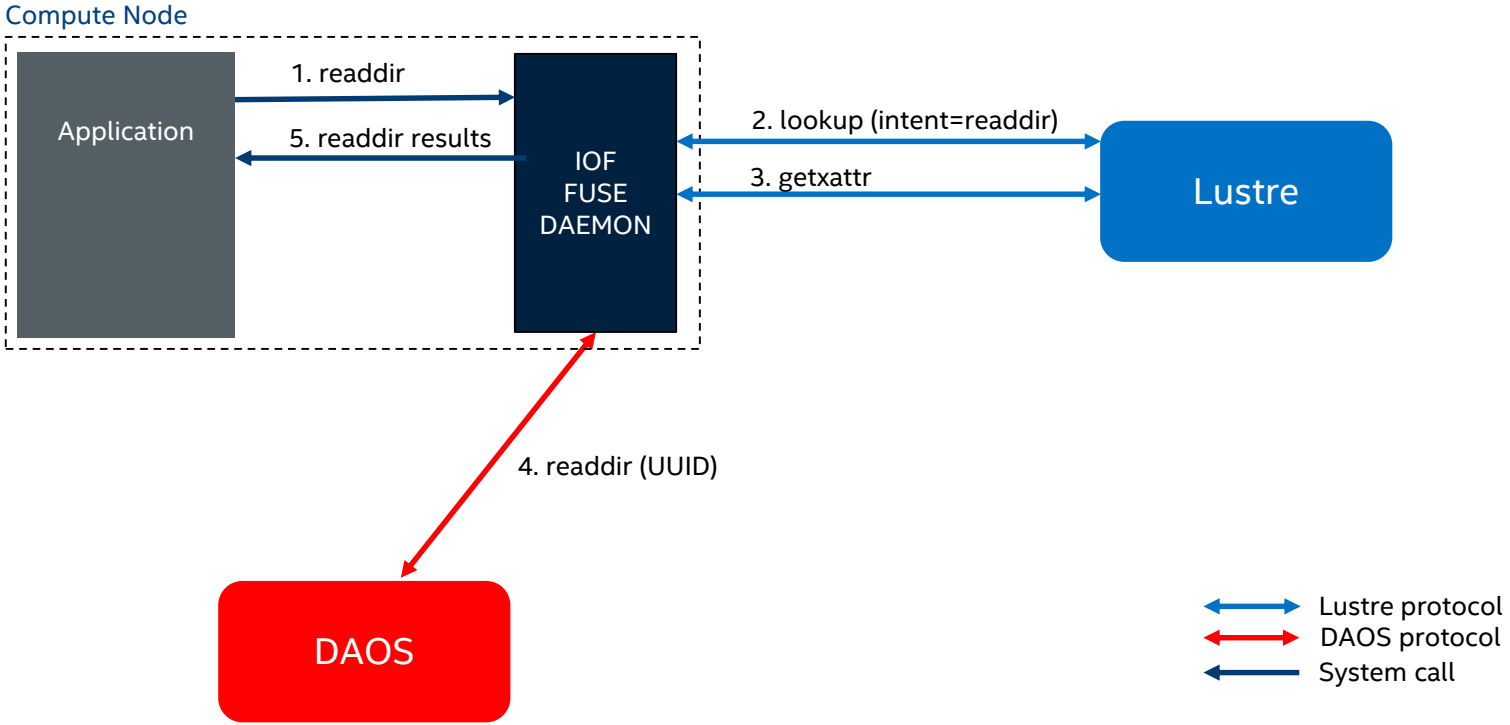
Use Case: Readdir Lustre Directory



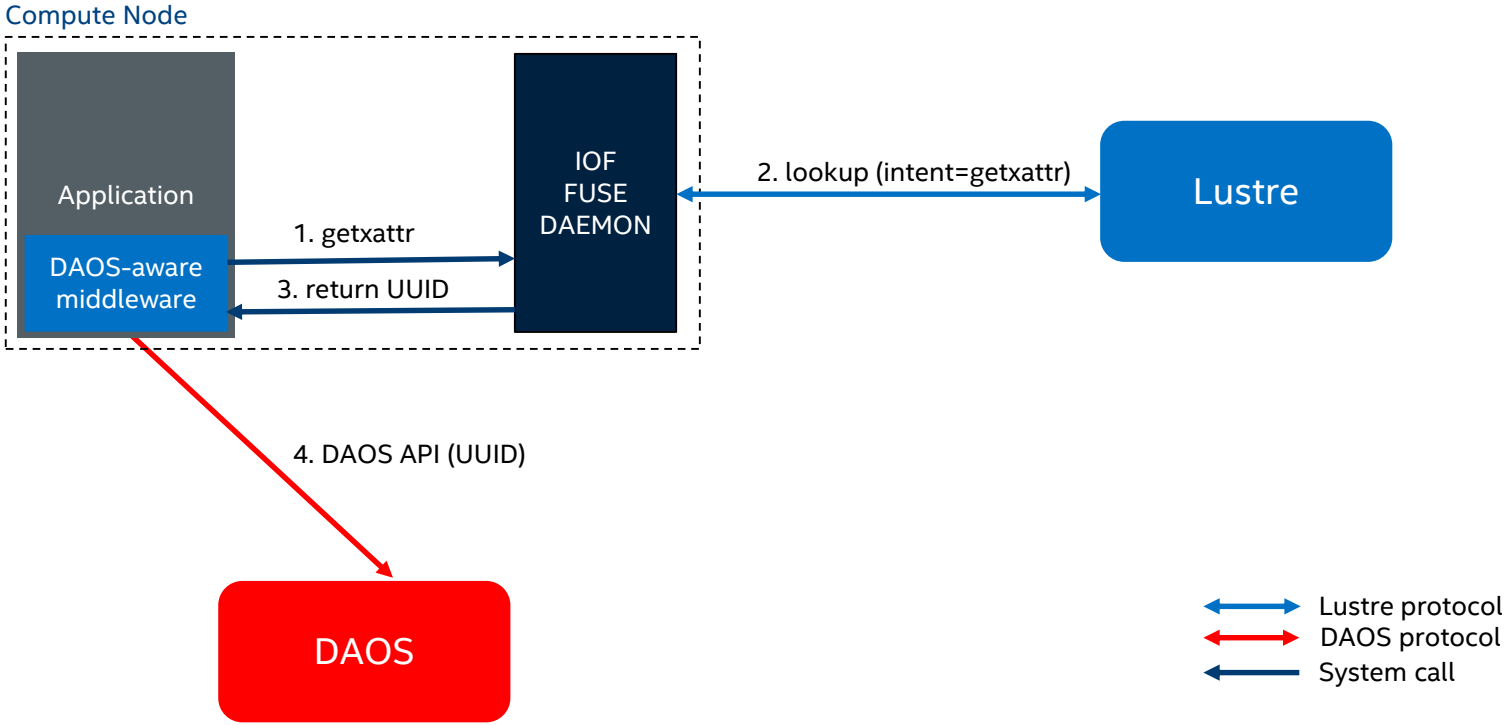
DAOS

Lustre protocol
DAOS protocol
System call

Use Case: Readdir POSIX Container



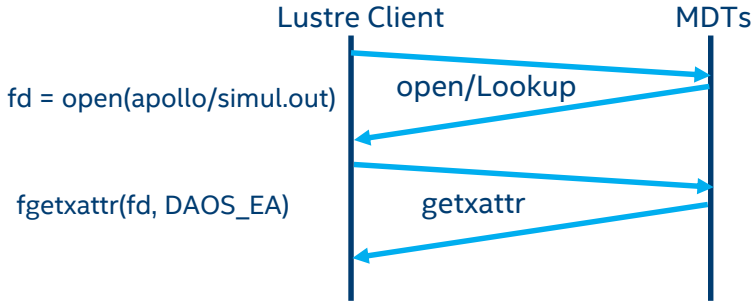
Use Case: DAOS-aware I/O Middleware



Special File/Dir Representation

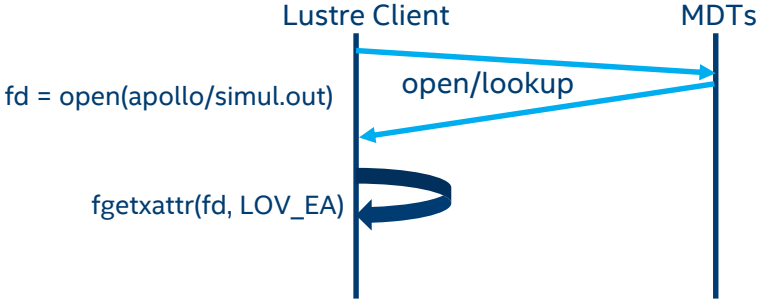
Regular Extended Attribute (EA)

- Portable
- Performance Impact
  - Extra EA fetch on every lookup
- Can't prevent Lustre file/dir from being created under the special directory

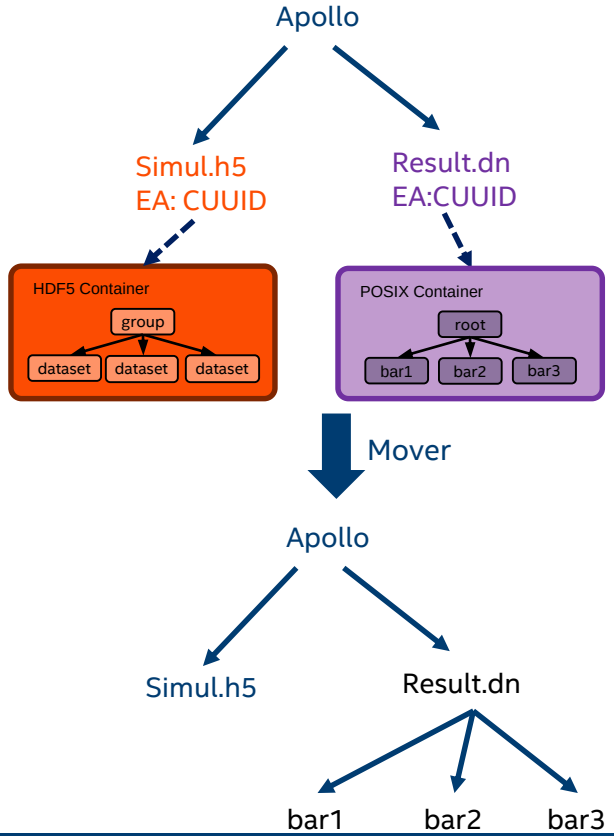


Special LOV EA

- Not Portable
- Minimal Performance Impact
  - No extra RPC
- Prohibit regular file/dir creation



Dataset Migration



Specific data mover

- Format conversion
  - Middleware-dependent
  - Middleware-agnostic
- Explore how to use layout swap functionality

Integration with Lustre Client Container Image (CCI)

- Local ldiskfs image mounted transparently on Lustre client
  - Written back to OSTs
  - High IOPS per client since MDTs not involved
- Accelerate migration of POSIX containers

Summary

Lustre change proposal

- Extend LOV EA
  - New layout type to point at external tier
  - Generic feature based on UUID
  - Can be integrated with any scale-out object stores
  - Opportunity to leverage layout swap functionality for cross-tier migration
- Effort tracked in LU-11376
  - Goal is to merge feature upstream
  - Feedback is welcomed!

Resources

- POSIX I/O Forwarding
  - https://github.com/daos-stack/iof
- DAOS
  - http://daos.io
  - https://github.com/daos-stack/daos
- Contacts
  - johann.lombardi@intel.com
  - bruno.faccini@intel.com
  - riaux.jb@intel.com


intel®
experience
what's inside™