ZFS Advancements for Modern Hardware

Adapting ZFS for large enclosures and NVMe

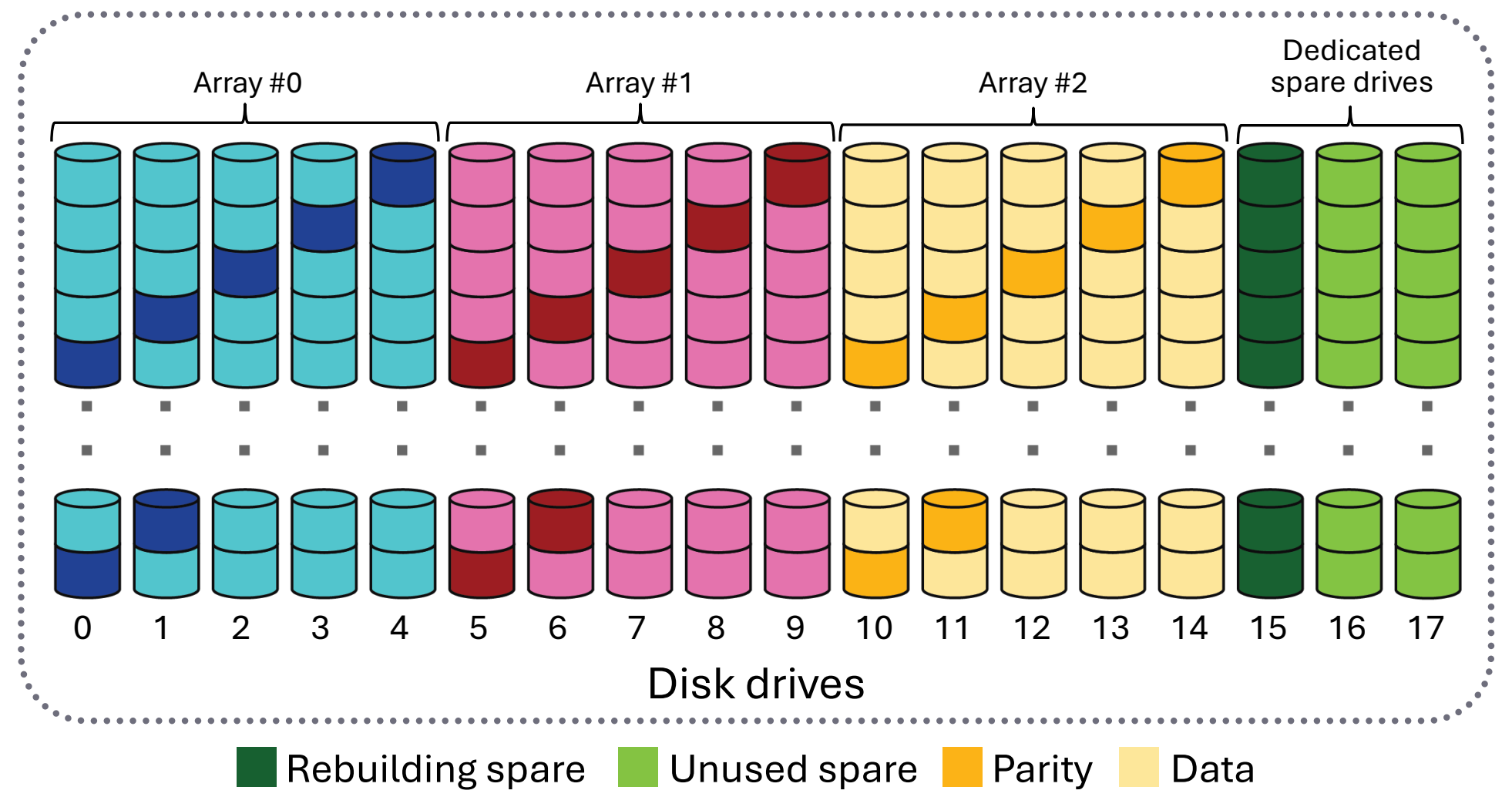
Olaf Faaland (LLNL)

Modern data storage hardware is moving to enclosures with more and larger hard disk drives (HDDs), and NVMe (solid state, very low latency devices). Both require major changes to the storage software stack.

LC collaborated with two national labs and four companies to accomplish the changes needed in Zettabyte File System (ZFS).

DRAID Improves Storage Resilience

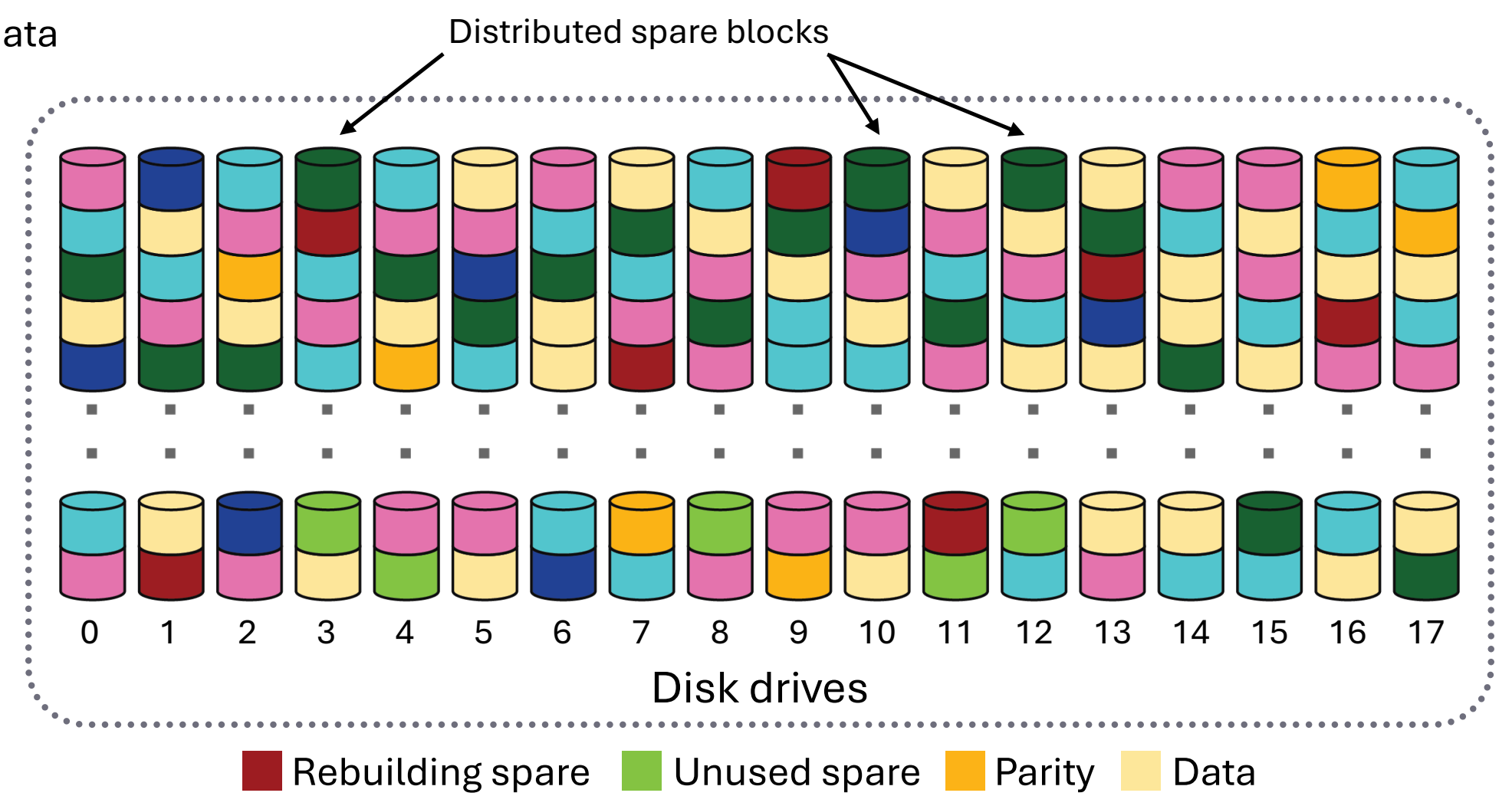
El Capitan's file system contains over 22,000 16TB hard drives, 90 drives per enclosure. Within one enclosure, failure of three drives before the first failure can be recovered results in data loss. **DRAID reduced the vulnerable window from 24 hours to 6 hours, reducing lifetime risk to acceptable levels.**



Without DRAID, all rebuild writes are to the spare drive, limited to about 250 MB/sec.

DRAID keeps unused space distributed over all disks for the "spare."

Rebuild writes are spread among all drives at GB/sec.



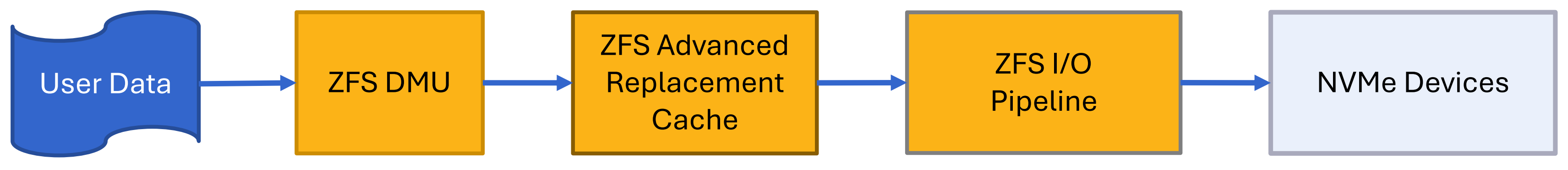
Collaborative Development Process

- Intel developed a working proof-of-concept.
- LC reworked the code to bring it to production quality.
- HPE and Oak Ridge National Laboratory tested and reported bugs.
- Delphix provided extensive code review.

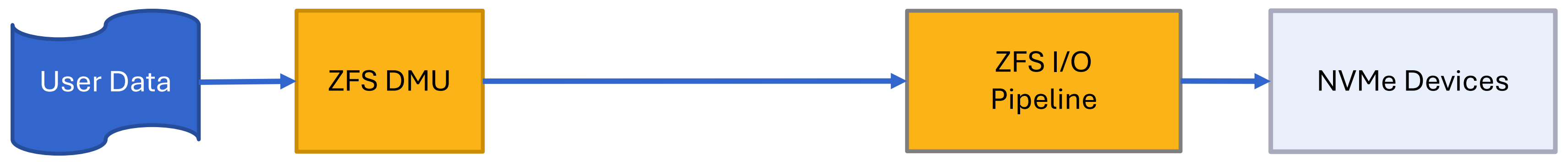
DirectIO Reduces I/O Latency

ZFS and the Linux kernel were designed to hide the latency of HDDs. NVMe is so fast that memory copies for caching and aggregation slow the data rate. DirectIO avoids those copies and provides a low-latency path to such storage.

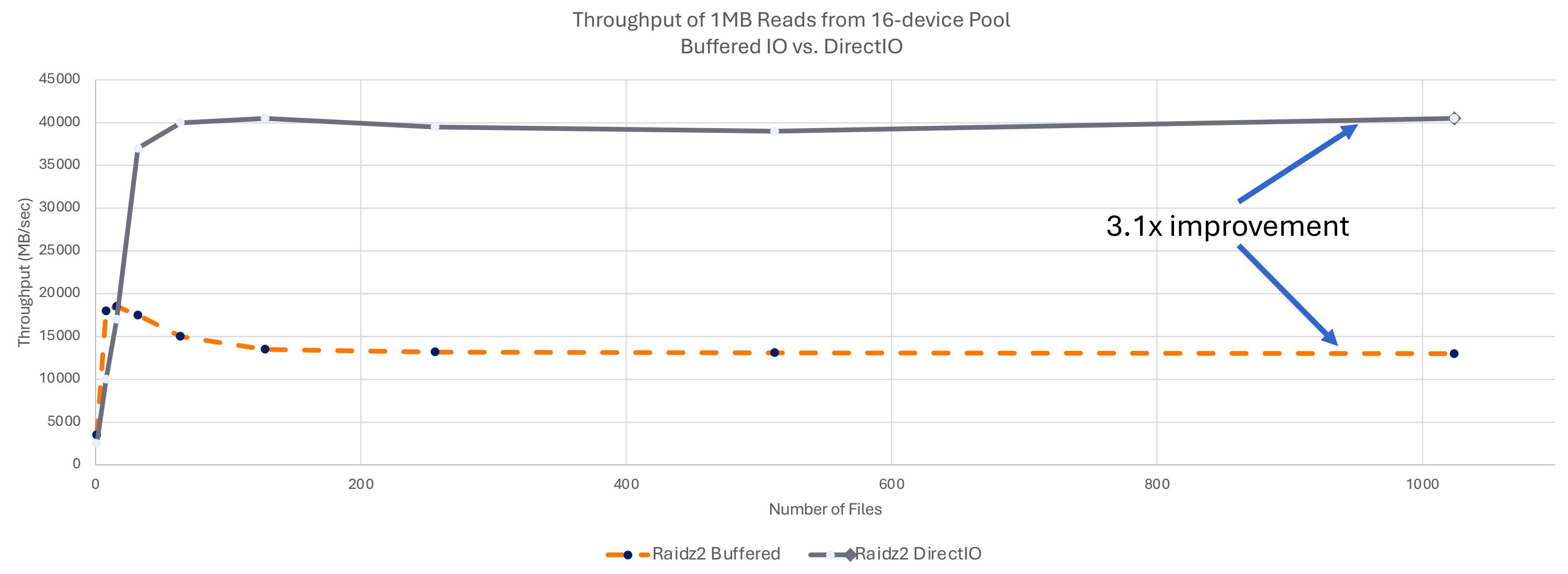
Key challenges include mixed I/O handling and deep ARC (cache) integration.



Buffered I/O: Data path passes through the cache



DirectIO: Data path avoids the copy



3.1x improvement

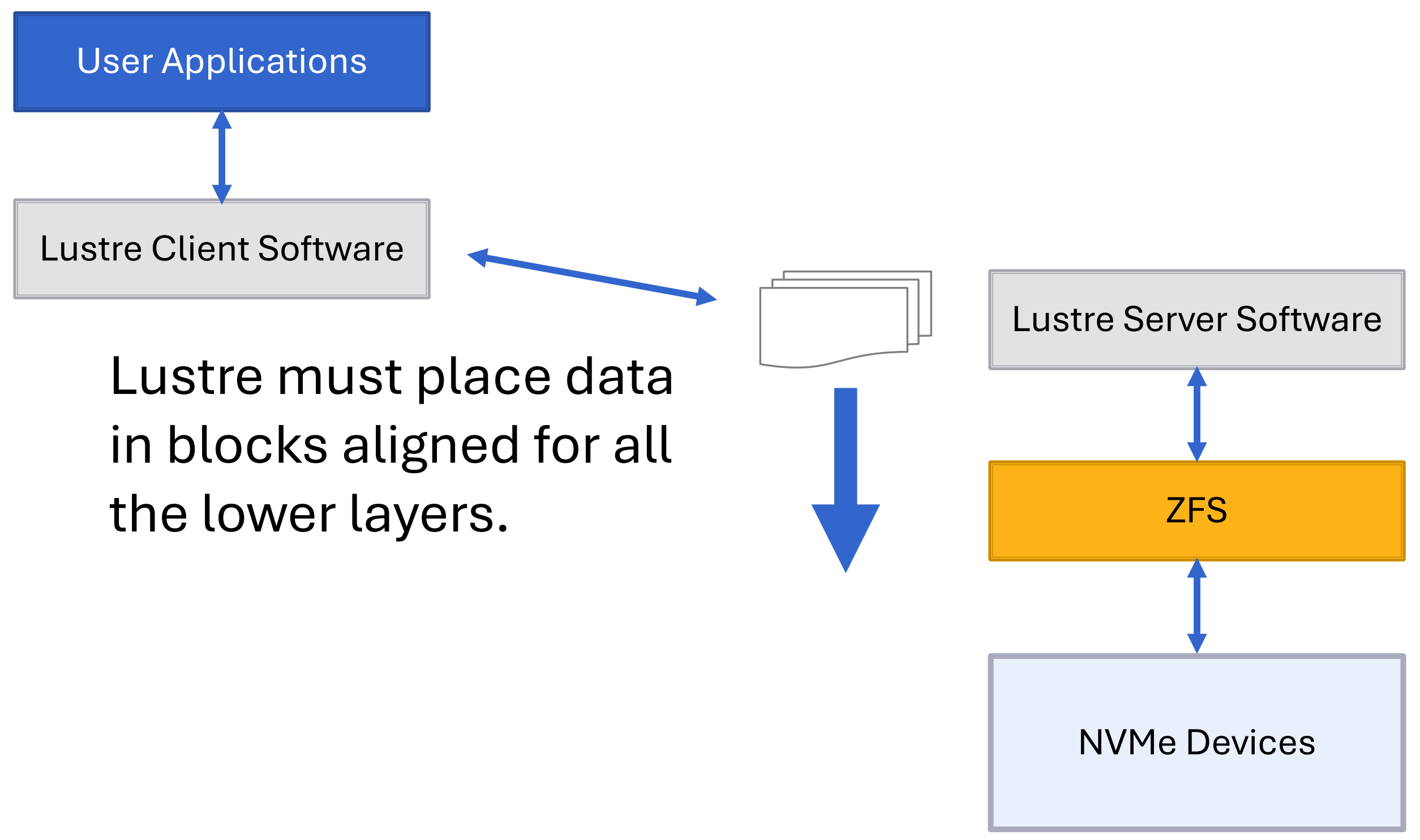
Throughput improvement with DirectIO

Collaborative Development Process

- LC established desired semantics and provided architecture advice, code review, and testing.
- Los Alamos National Laboratory developed the code.
- HPE and Oak Ridge National Laboratory tested and reported bugs.
- Delphix and TrueNAS provided extensive code review.

DirectIO Future Work

Lustre must now be modified to utilize DirectIO. This will benefit both hybrid and all-flash file systems.



Lustre must place data in blocks aligned for all the lower layers.

Deployments

- DRAID is deployed on El Capitan (#1 on Top500) and Frontier (#2) as well as many other systems across LC and the national labs.
- DirectIO will be deployed at LLNL when Lustre integration is complete.

Conclusion

- LC continues to build on our original work porting ZFS to Linux and integrating it with Lustre.
- Our reputation and current expertise allow us to influence ZFS work.
  - We enforce high software quality standards.
  - We leverage other developers through collaboration.
  - We benefit from testing and production bug reports by others.
- LC is well positioned to adapt to new storage technologies such as NVMe.

Two 5-year projects funded by ASC at <2 FTE/year plus significant ZFS community effort.

Collaborating on production-quality data storage software to achieve world-class performance, features, and reliability

Lawrence Livermore National Laboratory

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. LLNL-POST-2003864

National Nuclear Security Administration