Lawrence Livermore National Laboratory

ZFS on Linux for Lustre
SC10
November 16, 2010

Brian Behlendorf, Christopher Morrone

Lawrence Livermore National Laboratory, P. O. Box 808, Livermore, CA 94551

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

LLNL-PRES-461795

ZFS Overview

- Developed by Sun (now Oracle) on Solaris
- Combined filesystem, logical volume manager, RAID
- Copy-on-write
- Built-in data integrity
- Intelligent scrubbing and resilvering
- Very large filesystem limits

Lustre Overview

- Massively parallel distributed file system
- Major Components
  - Metadata Server (MDS)
    - Servers Metadata Target (MDT)
  - Object Storage Servers (OSS)
    - Server Objects Storage Targets (OST)
  - Clients

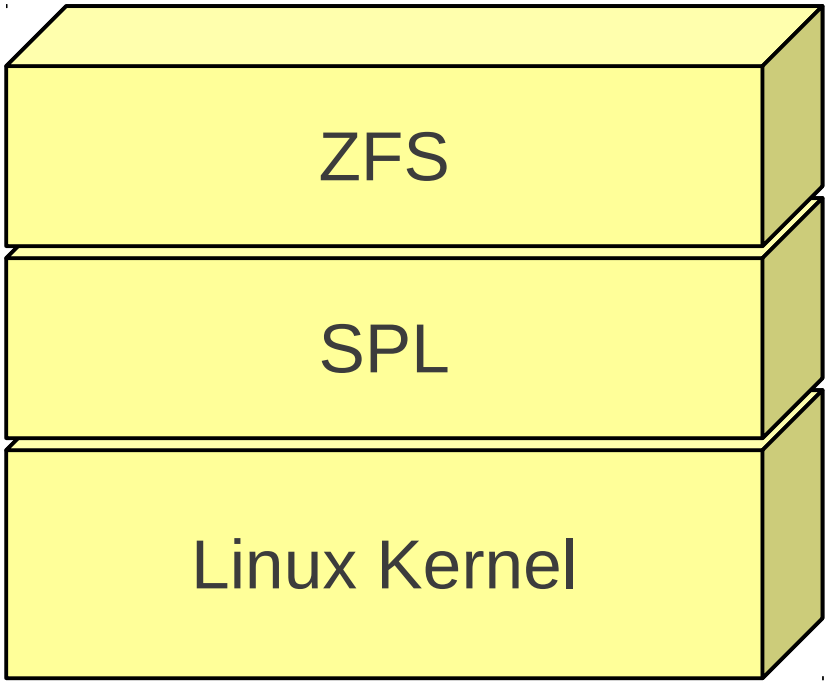
LLNL's Reasons for porting ZFS

- Lustre servers currently use ext4 (ldiskfs)
  - Random writes bound by disk IOPS rate, not disk bandwidth
  - OST size limits
  - fsck time is unacceptable
  - Expensive hardware required to make disks reliable
- Late 2011 requirement:
  - 50PB, 512GB/s – 1 TB/s
  - At a price we can afford

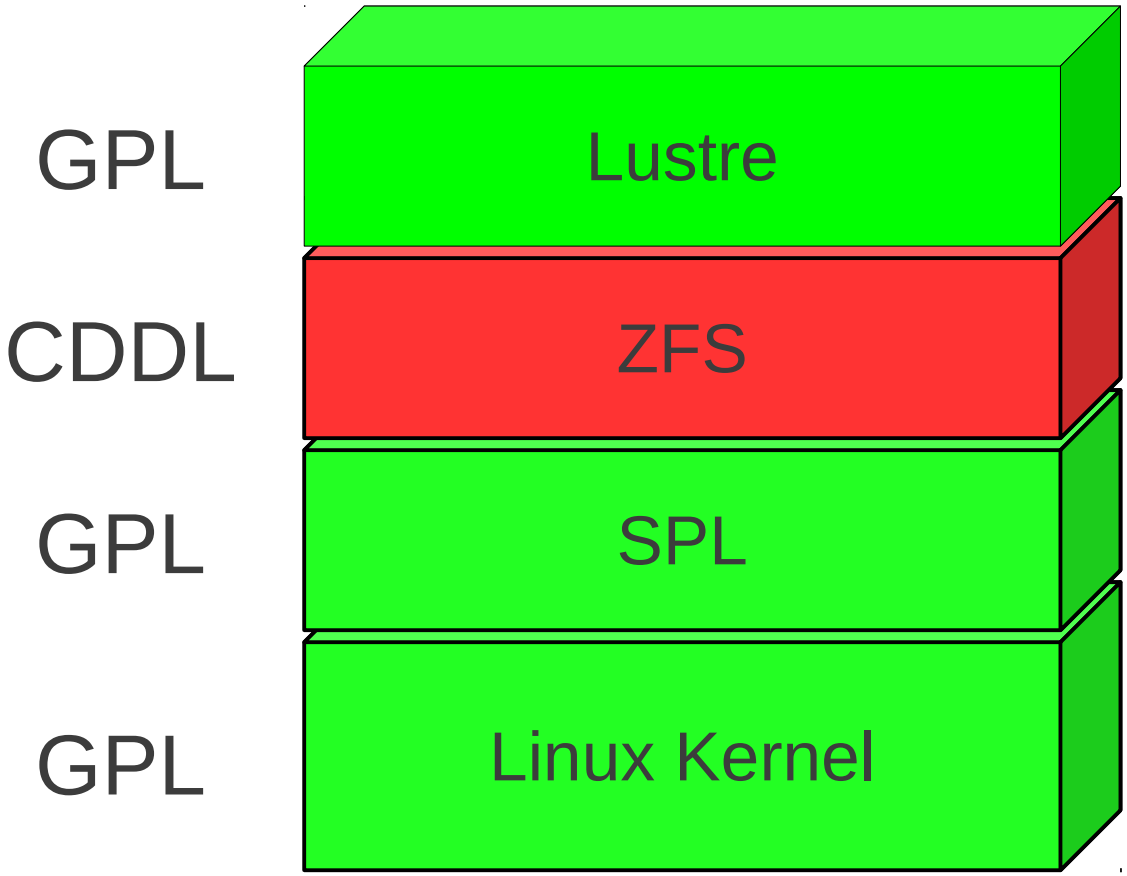
ZFS's Benefits

- COW sequentializes random writes
  - No longer bound by drive IOPS
- Single volume size limit of 16 EiB
- Zero fsck time. On-line data integrity and error handling
- Expensive RAID controllers are unnecessary

Solaris Porting Layer
Linux/ZFS Glue

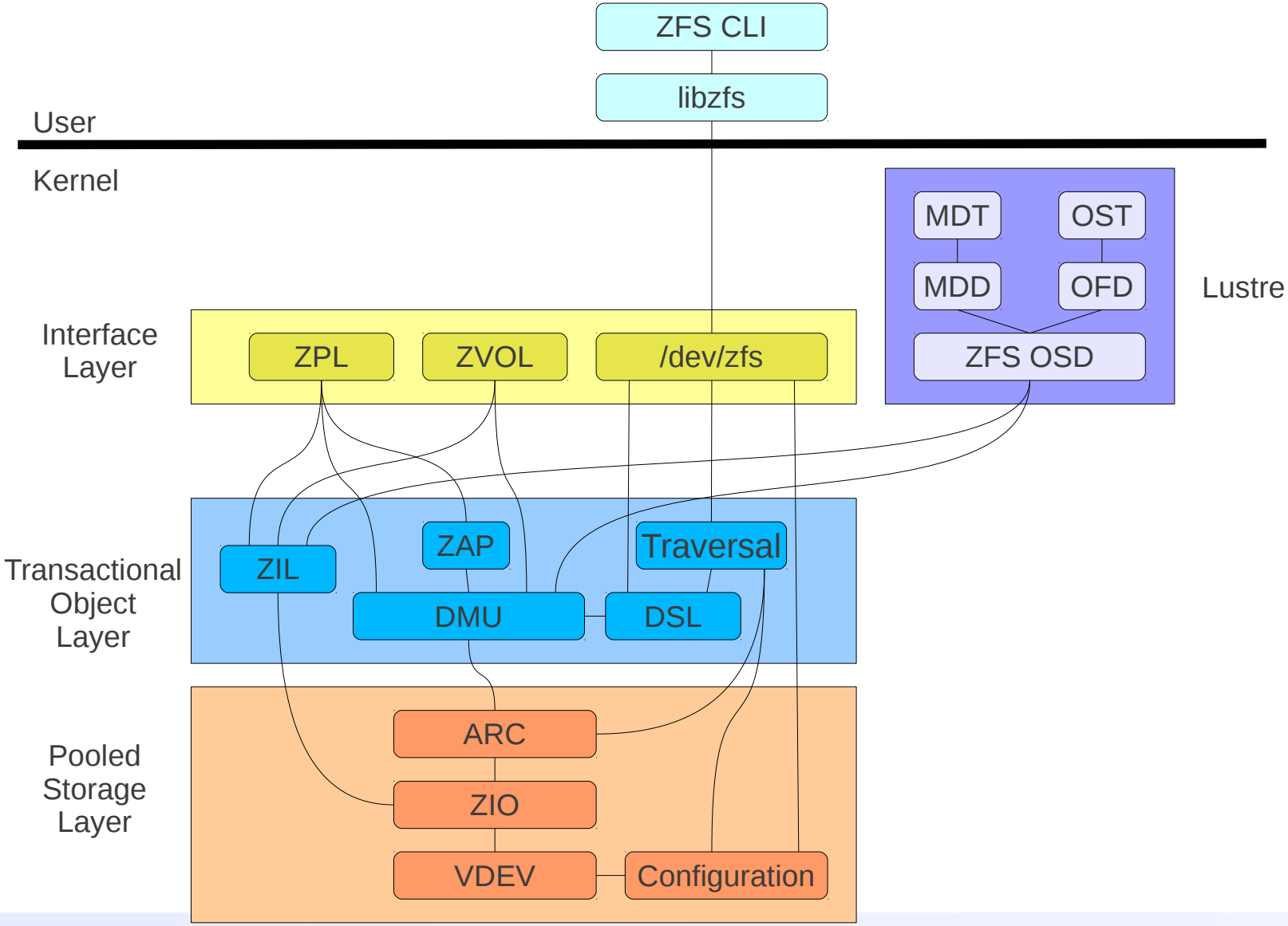


Licensing Issues

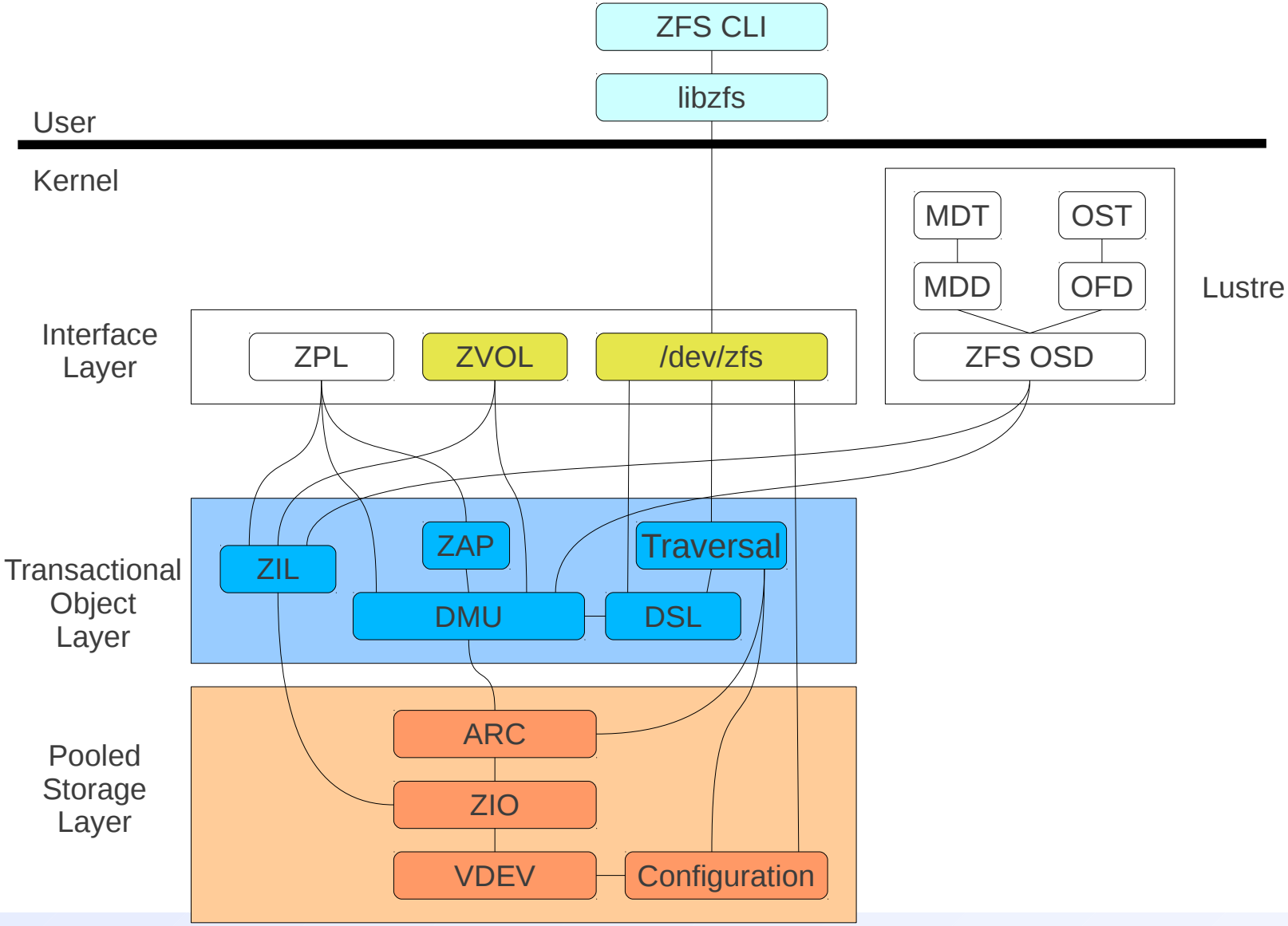


CDDL = Common Development and Distribution License
GPL = (Gnu) General Public License

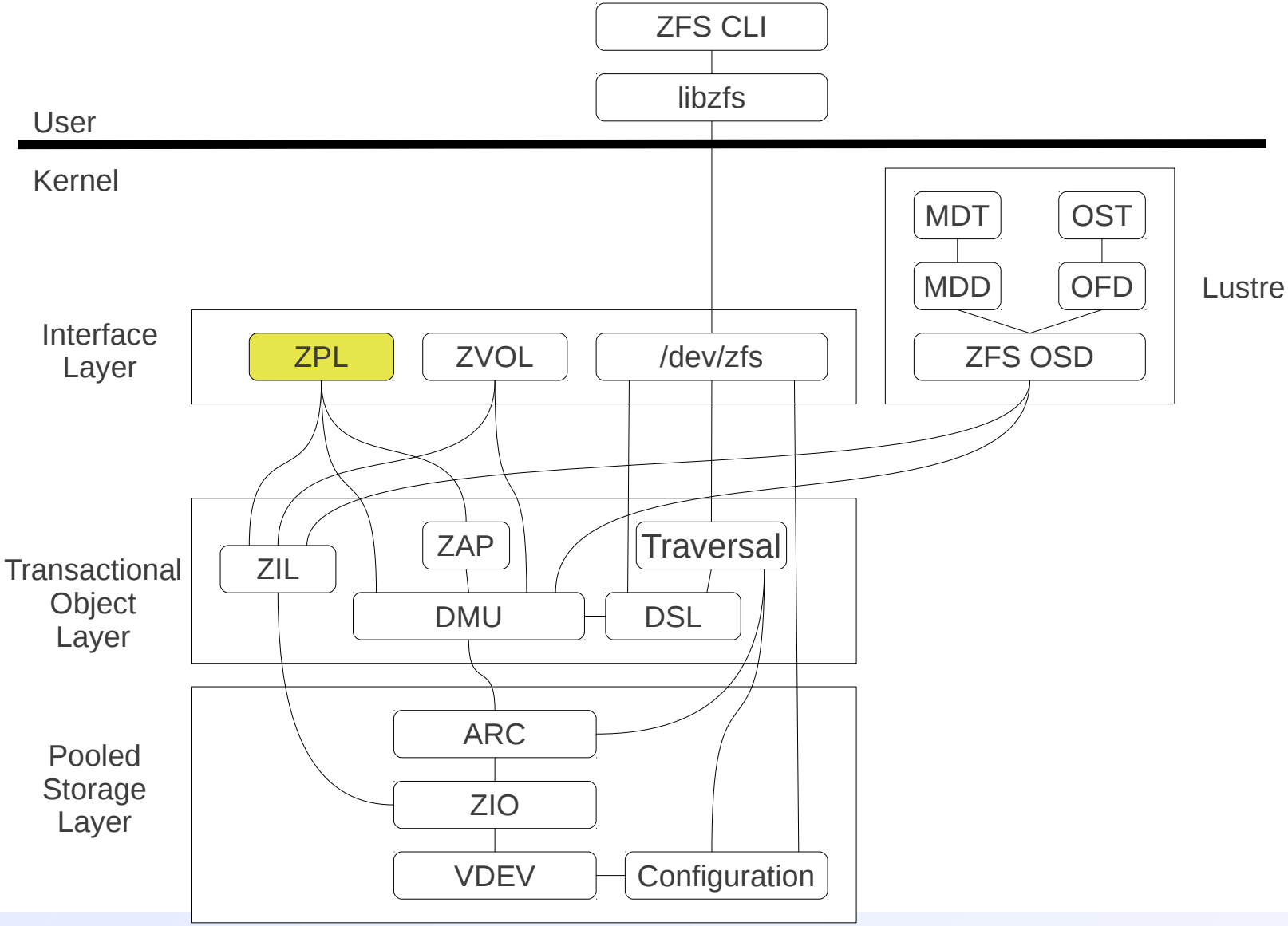
ZFS and Lustre Components



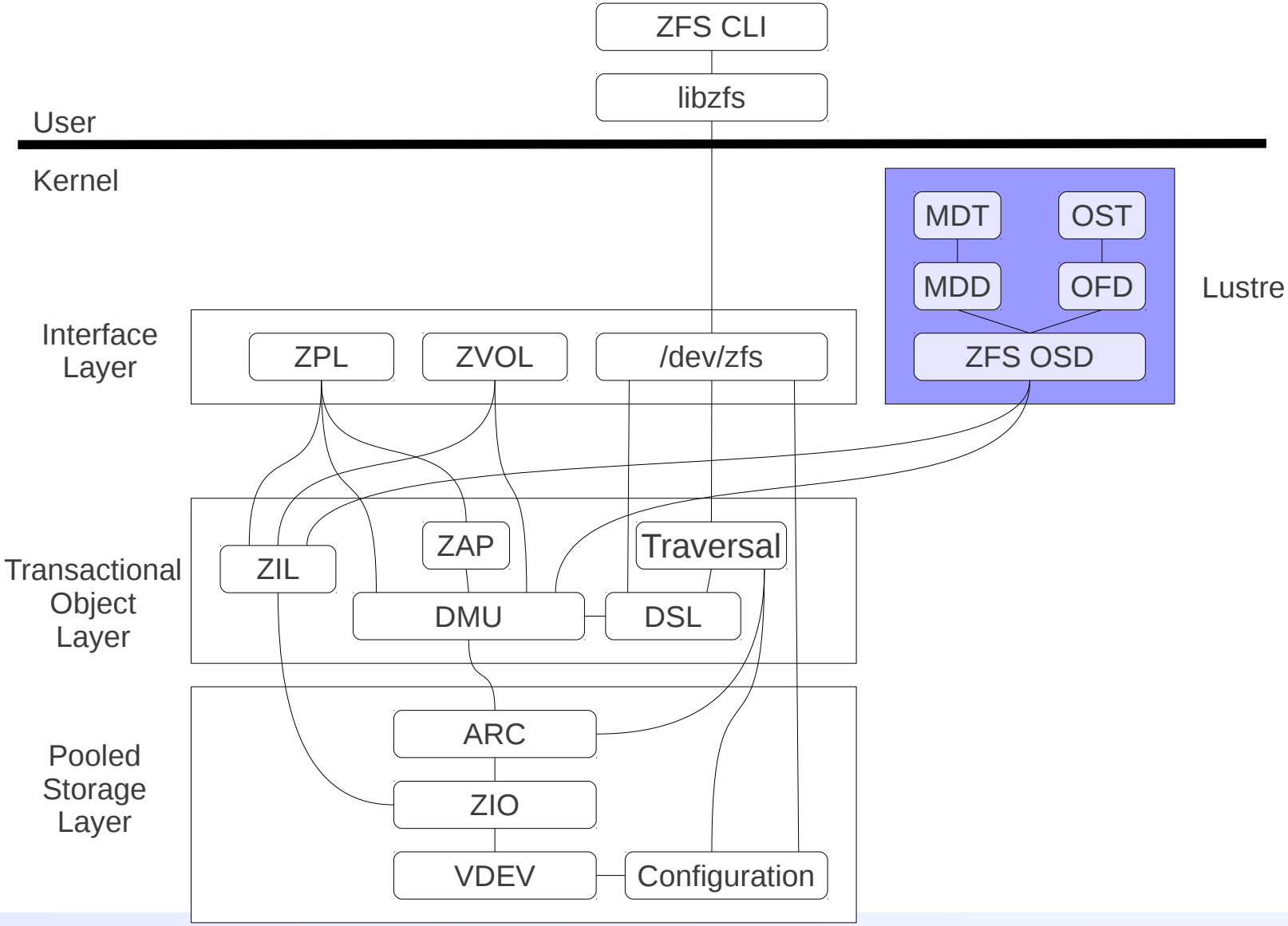
Ported by LLNL



Beta Port by KQInfotech



Under Development by Oracle



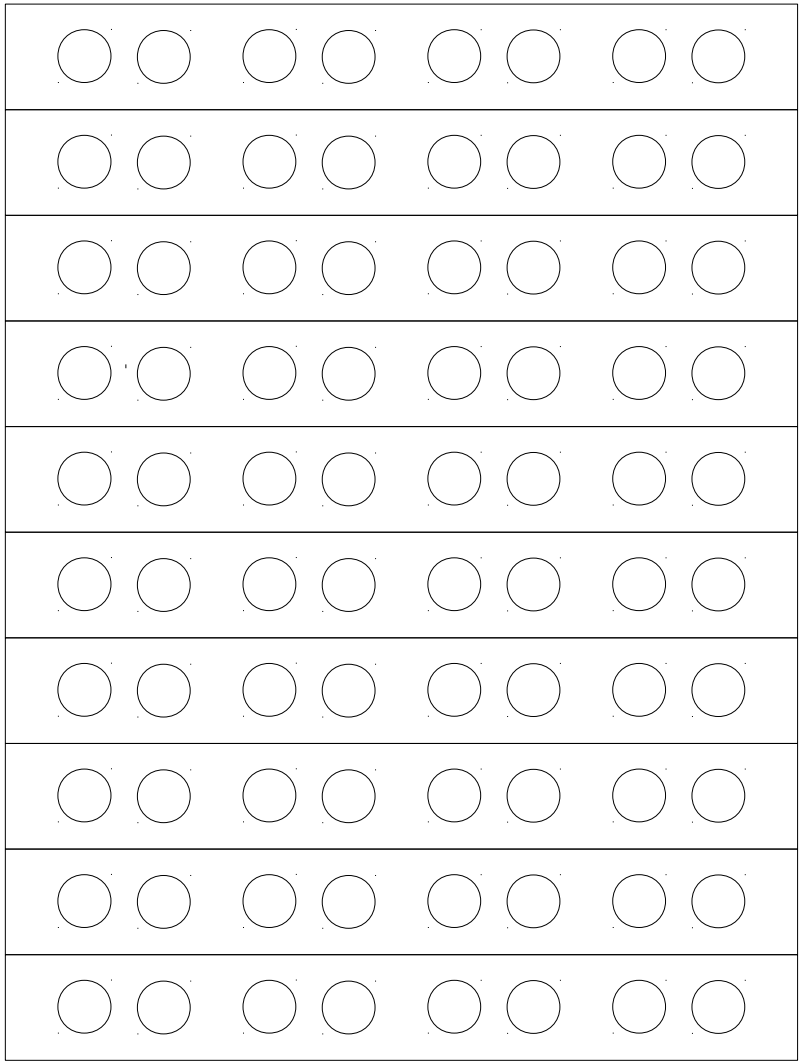
Proposed OSS Hardware

- “Scalable Unit” is 1½ racks
- 10x 4U disk enclosure (40U total)
  - 60 disks in each enclosure
- 8x 2U Linux server nodes (20U total)

SAS Link Count

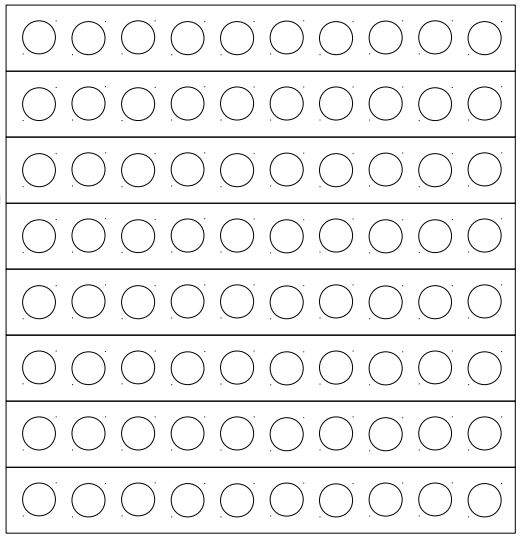
80 4-lane SAS Links
96 GB/s

10 60-bay disk enclosures
8 SAS Ports Each



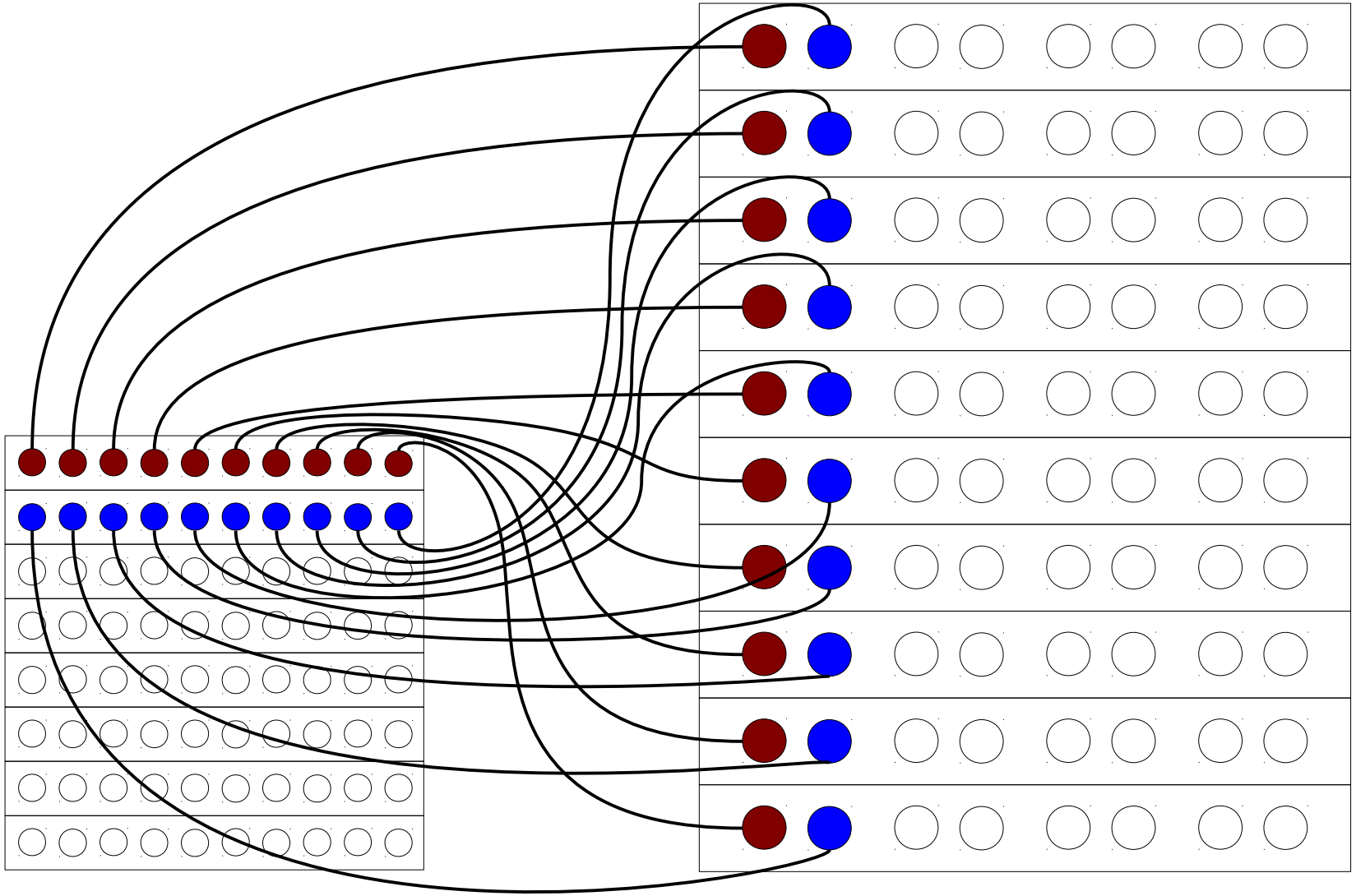
40U
(4U each)

8 OSS Nodes
10 SAS Ports Each

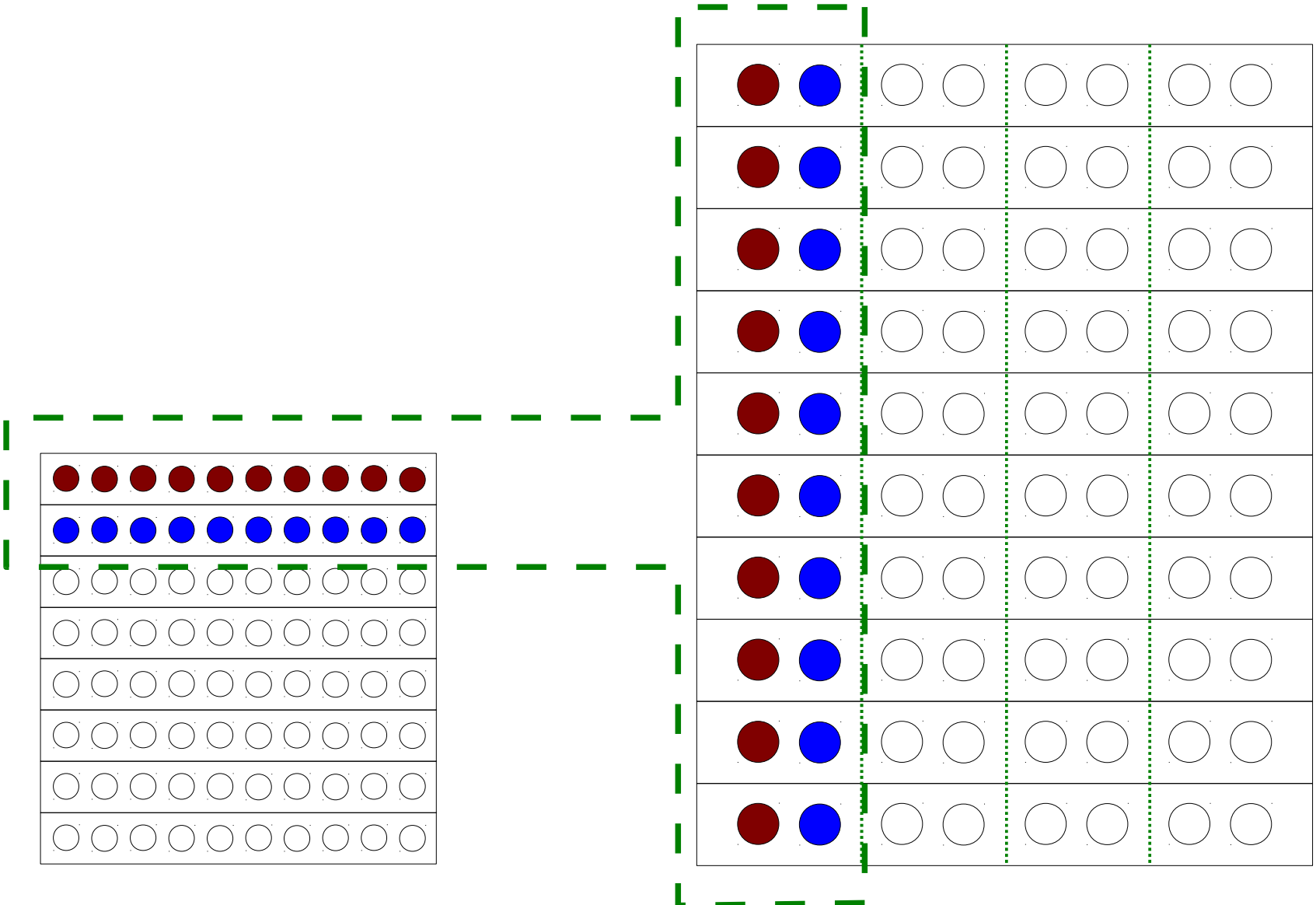


20U
(2U each)

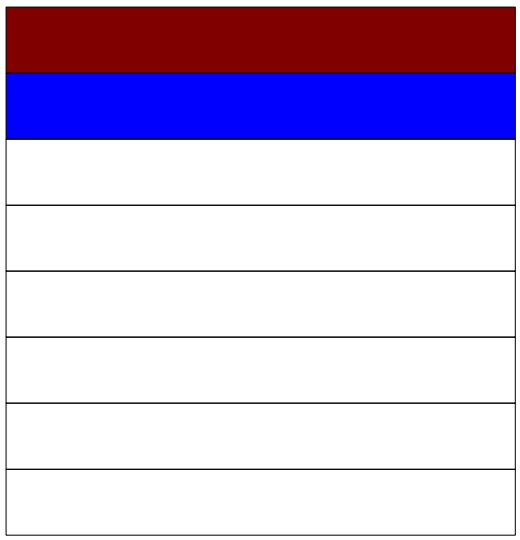
SAS Links from first two OSSs

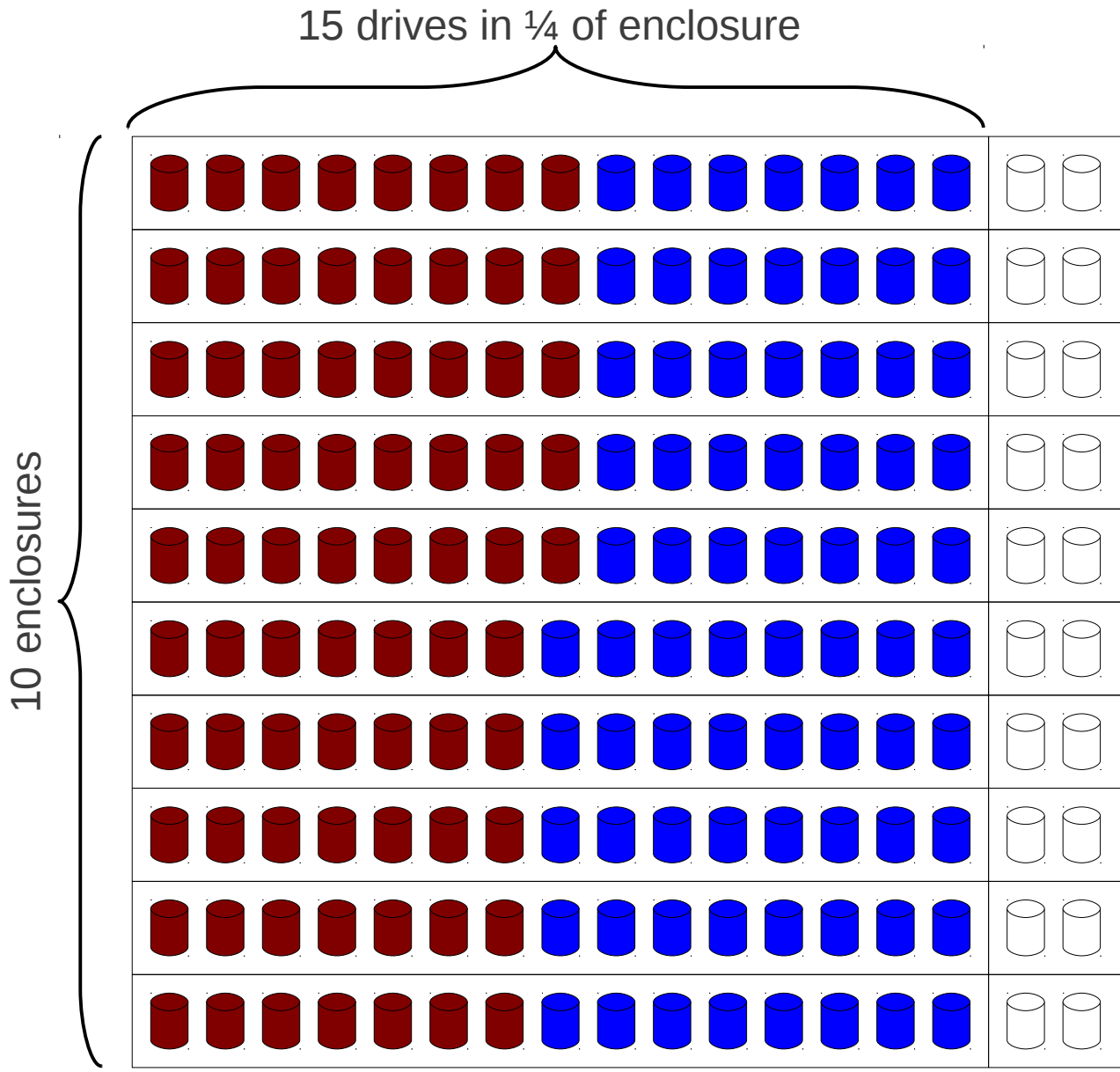


Each pair of OSS split ¼ of disks. OSS can serve as failover for partner.

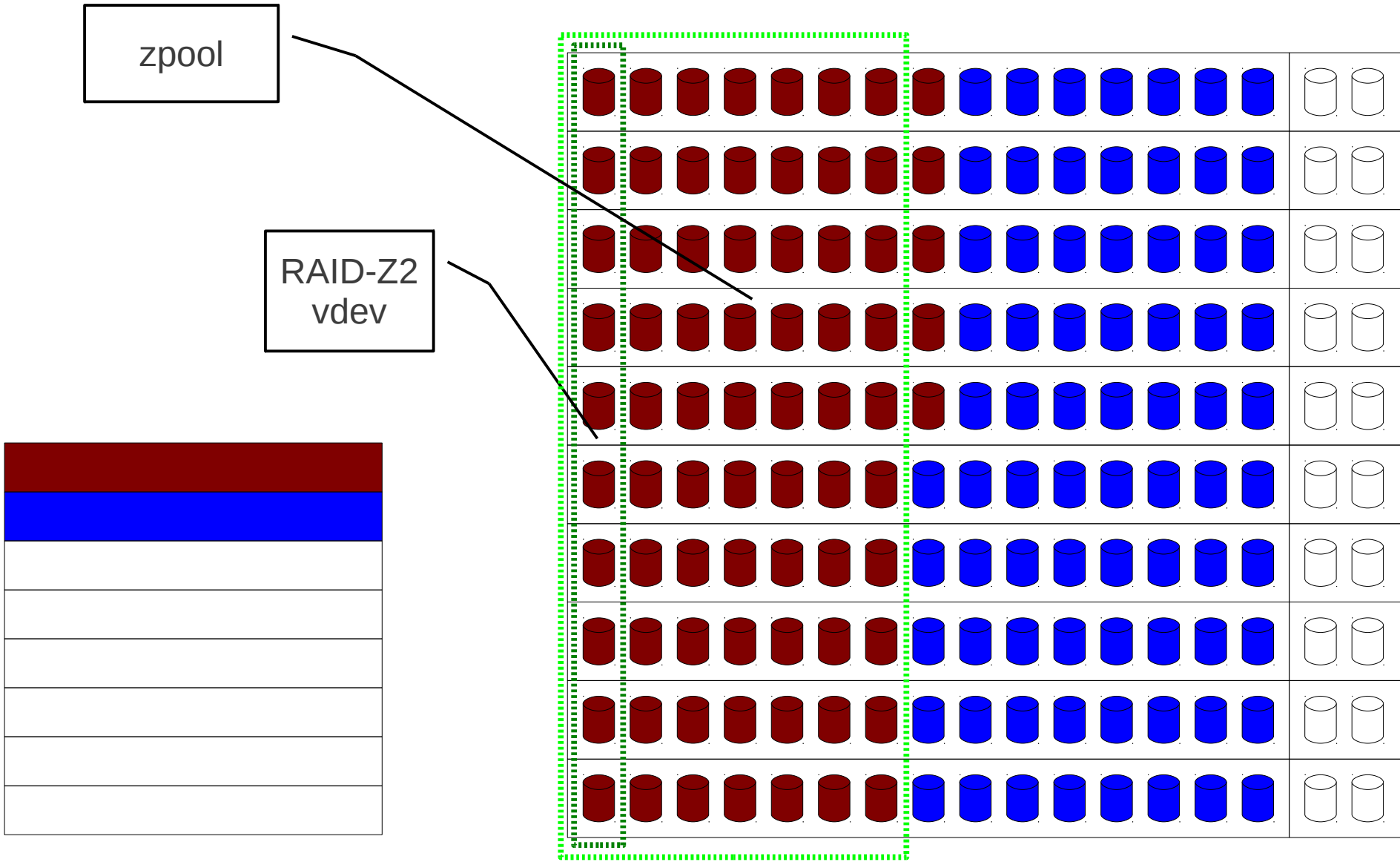


Each OSS sees 150 drives, uses 75

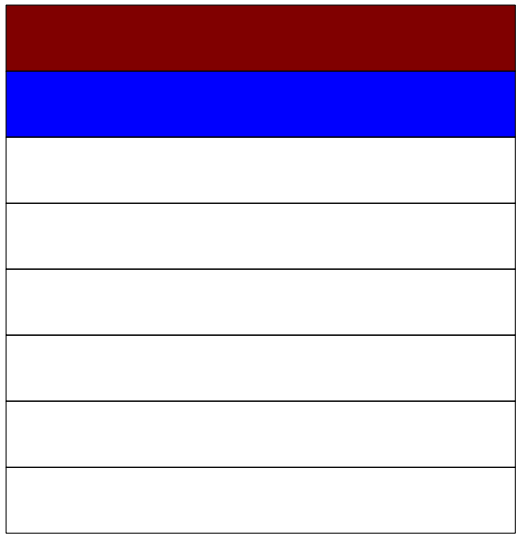


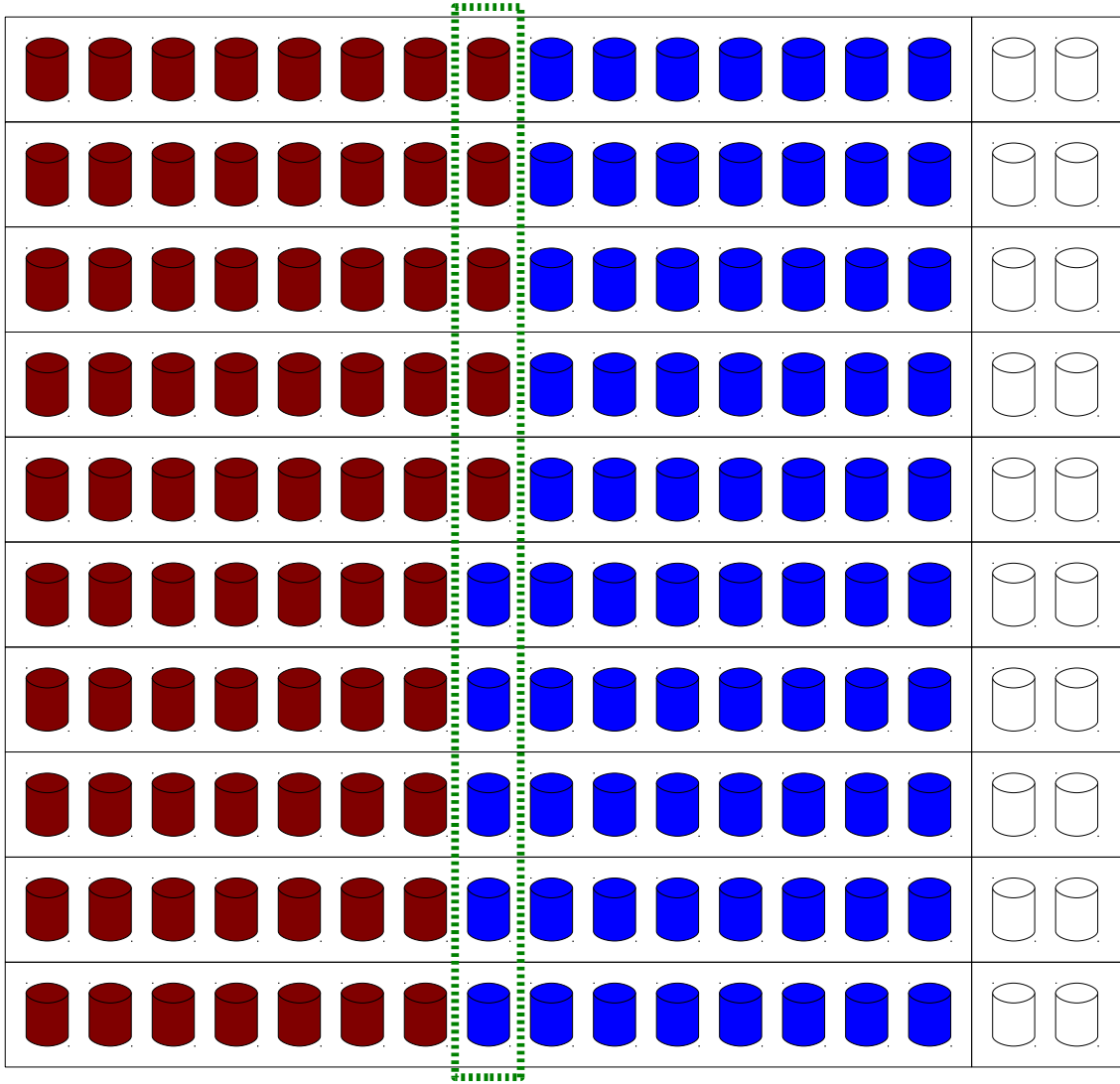


Each OSS has 7 10-disk zpools
70 disks each for normal use



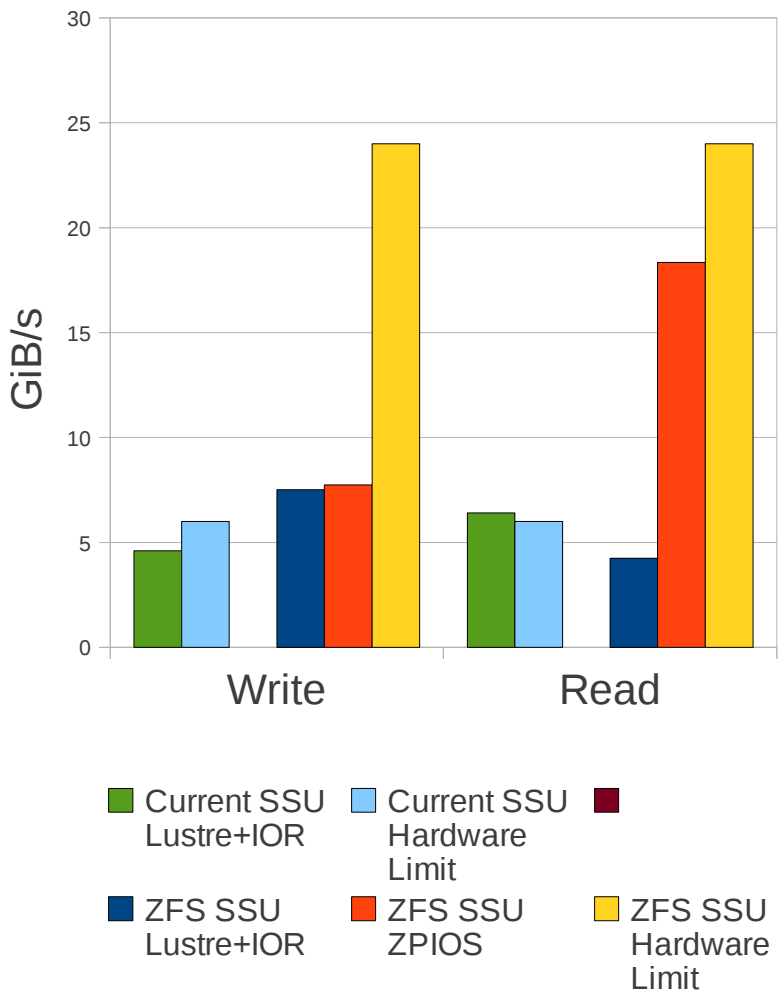
5 disks left over for each OSS
SSDs – ZIL? L2ARC?





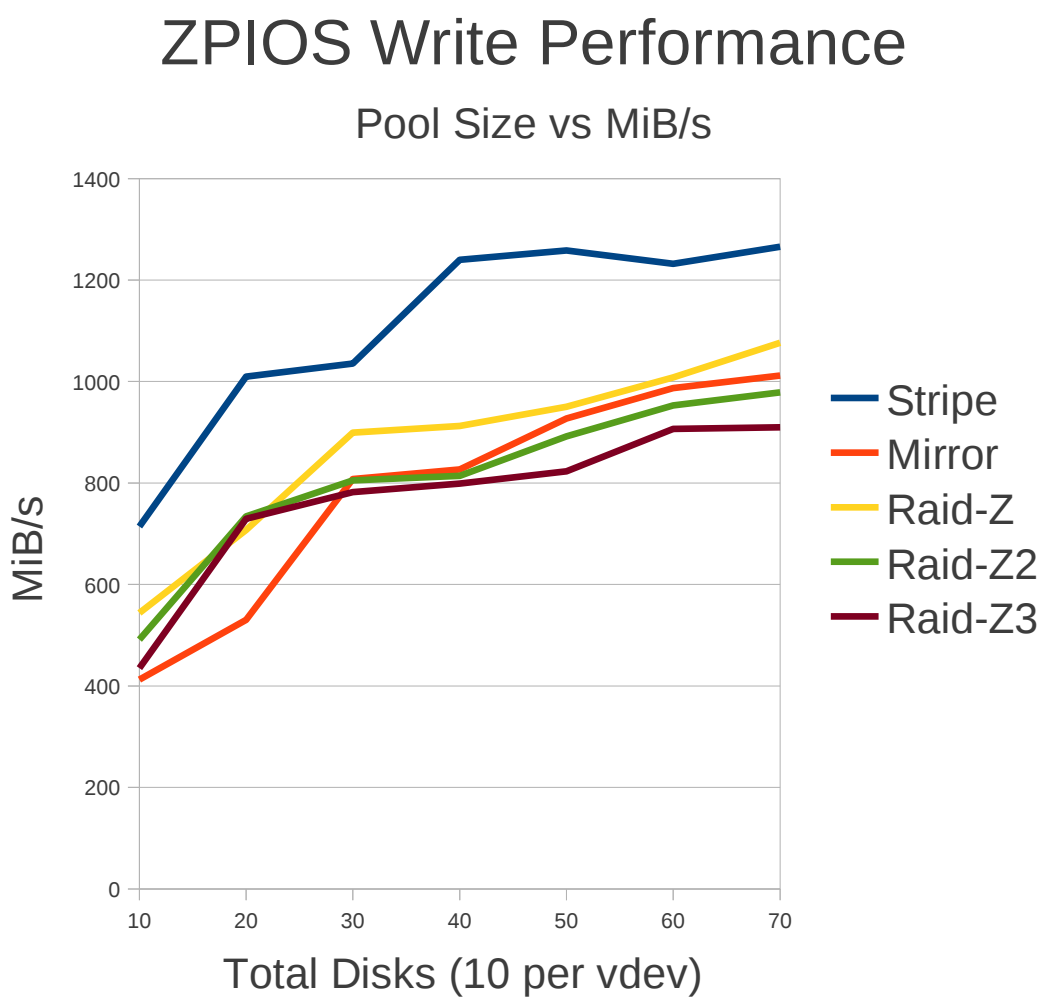
Prototype ZFS/Lustre Filesystem

ZFS Performance Comparison



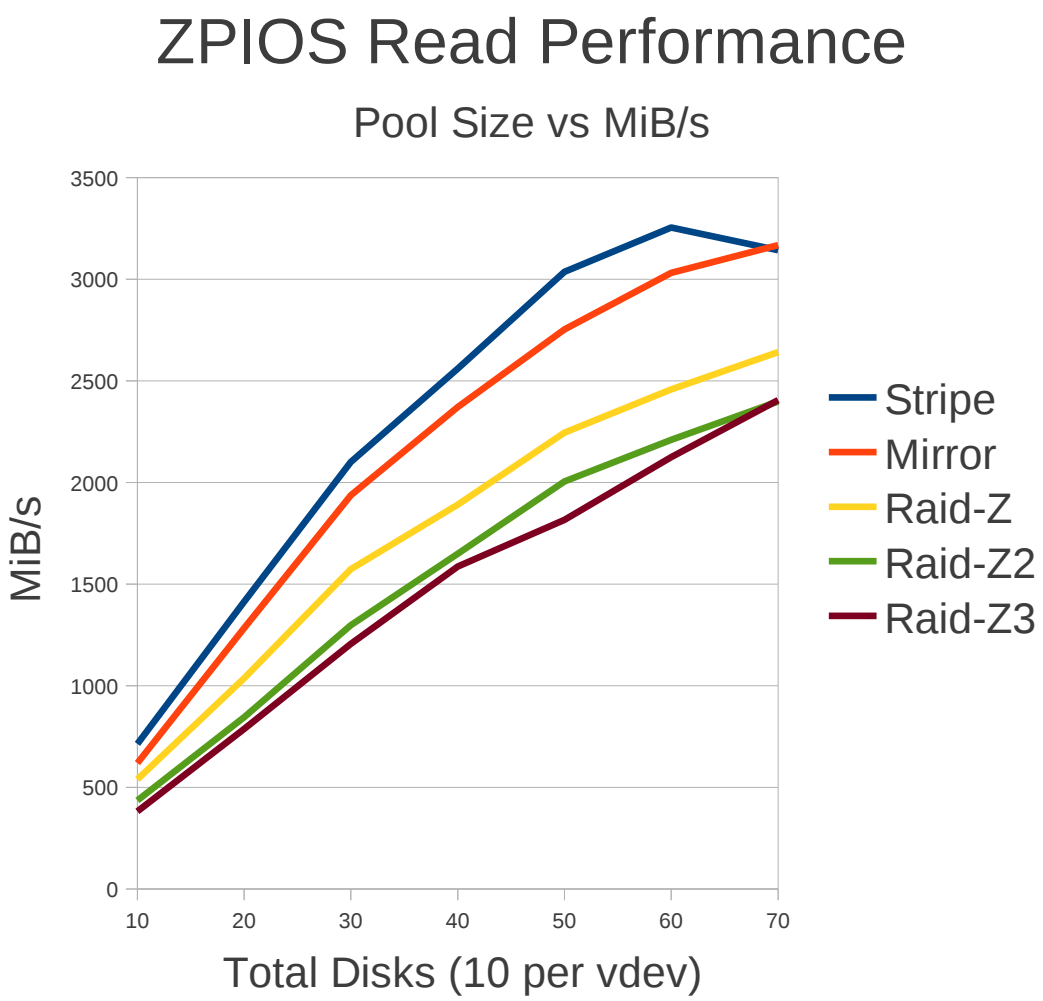
- Same number of drives
- SATA vs SAS disk
- RAID-Z2 vs RAID-6
- Write Performance is Limited by the ZFS Port
- Read Performance is Limited by Lustre/CPU
- ZFS is unoptimized, this can all be improved!

Single Node Write Performance



- Random 1MiB writes, from 128 threads, to 4096 objects
- 60 MiB/s per disk for small pools (10 disks)
- Limited by taskq when scaled up
- This is fixable

Single Node Read Performance

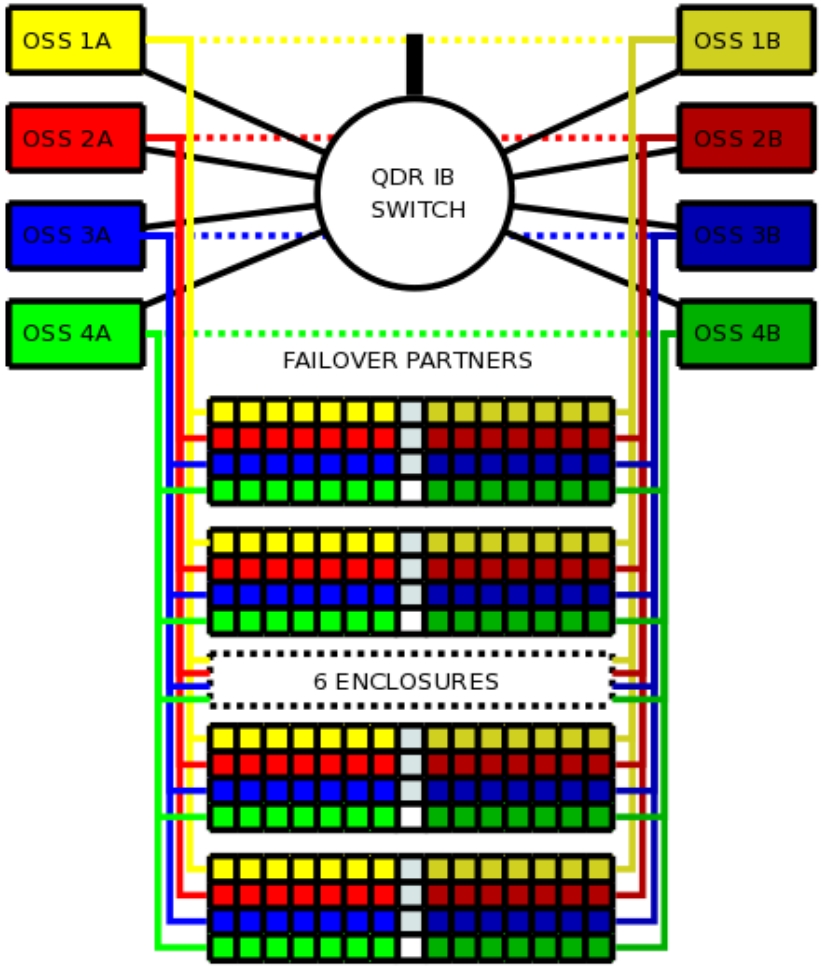


- Random 1MiB reads, from 128 threads, to 4096 objects
- Prefetch disabled
- Scales very well
- 50-60 MiB/s per disk even for large pools
- >90% CPU utilization when using 70 disks
- Can be optimized

More Information

- ZFS & SPL
  - http://zfsonlinux.org
    - Mailing Lists
    - Download software
    - Documentation
- Lustre support for ZFS (a.k.a., “KDMU branch”)
  - LLNL testing internally
  - Publicly available late 2011

OSS SSU



- 8 Linux hosts
- 10 60-bay SAS enclosures
- 560 2TB SATA drive
- 40 SSDs (ZIL, L2ARC)
- 8 QDR Infiniband adapters
- 8 dual-port 10-GigE adapters