CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

ETH zürich

Burst Buffer IME

CSCS-ETHZ

Hussein N. Harake

Agenda

- About CSCS
- DDN Burst Buffer (IME)
- Benefits
- System Layout
- Benchmark tools
- Results
- Next Steps

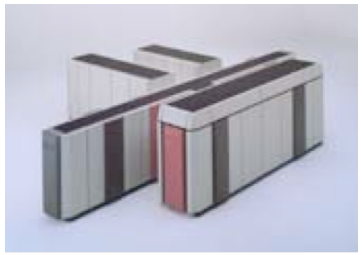
© CSCS 2015

CSCS (Swiss National Supercomputing Centre)

- Founded in 1991
- Enables world-class research with a scientific user lab
- Available to domestic and international researchers through a transparent, peer-reviewed allocation process.
- Open to academia and are available as well to users from industry and the business sector.
- Operated by ETH Zurich and is located in Lugano.

© CSCS 2015

24 years of supercomputers at CSCS



1991 NEC SX3
5.5 GF Adula

1996 NEC SX4
10 GF Gottardo

1999 NEC SX5
64 GF Prometeo

2002 IBM SP4
1.3 TF Venus

2005 Cray XT3
5.8 TF Palu

2006 IBM P5
4.5 TF Blanc

2009-12 Cray XE6
402 TF Monte Rosa

2012-13 Cray XC30
7.7 PF Piz Daint

2014 XC30 **1.25 PF**
Piz Daint extension

Data Center

- 2000 sq.m Machine Room
- 20 MW of power and Cooling capacity
- Lake Water cooling
  - 700 Liters/s

© CSCS 2015

DDN Burst Buffer (IME)

What is IME Burst Buffer?

- Infinite Memory Engine
- A caching layer that sits between applications and file system
- Library that grants applications access to the fast cache devices

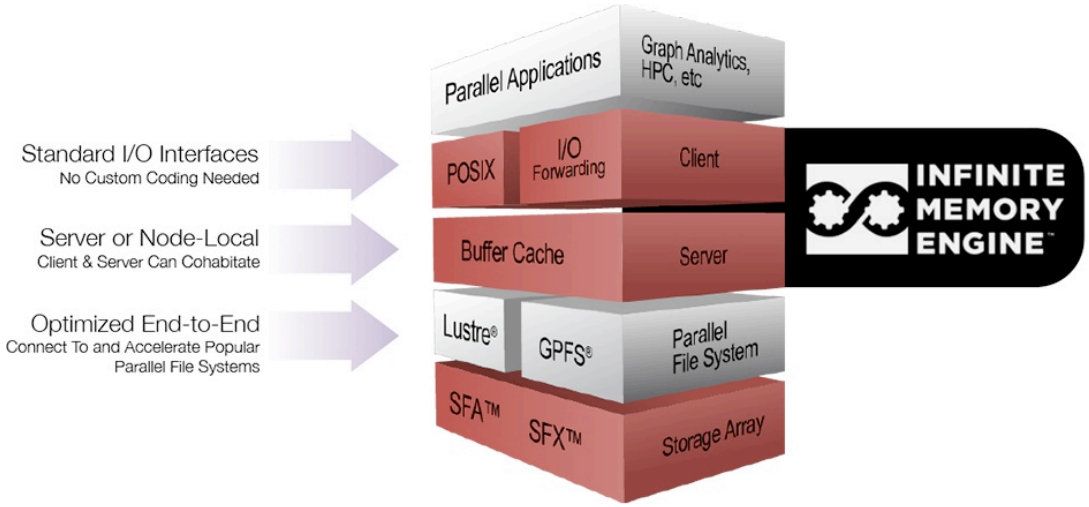
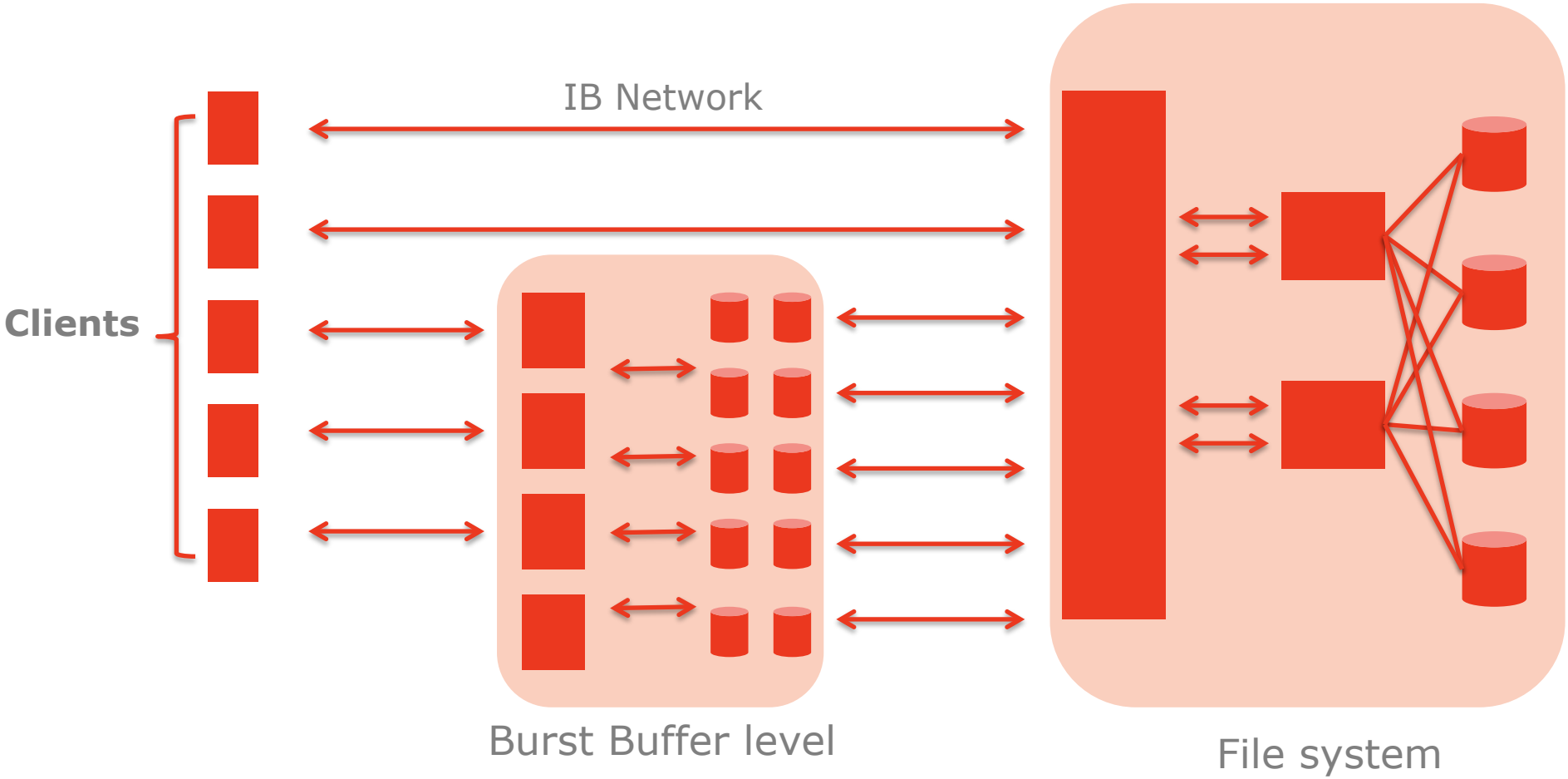


Image courtesy of DDN

DDN Burst Buffer (IME)

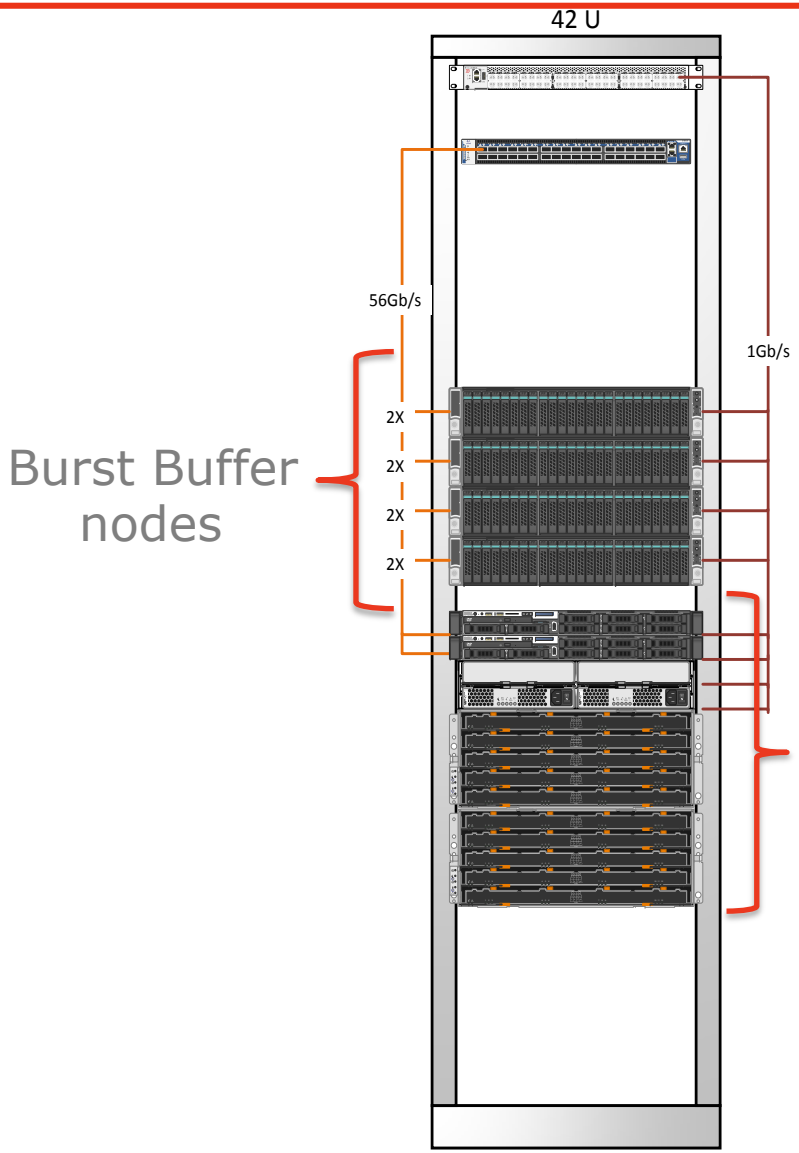


Benefits

Some of the benefits

- Accelerate I/O
- Cache data for fast access
- Reorganize the IO method of writing data to file-system
- Low Latency and High IOPs
- POSIX and non-POSIX access to the cache area

Test System Layout



Burst Buffer nodes

Lustre

- 4 * 2 sockets IO servers
- 4 FDR IB ports per server
- 24 SSDs per server
- 2 LSI / Intel controllers per server
- 2 * IO servers filesystem
- 1 dual controller
- 2 enclosures 120 disks

Filesystem Hardware Capability

Lustre Filesytem:

- 2 OSSs
- 1 MDS
- 2 * enclosures 120 Disks
- 1 Dual Controller

On the underlying storage infrastructure Lustre delivers

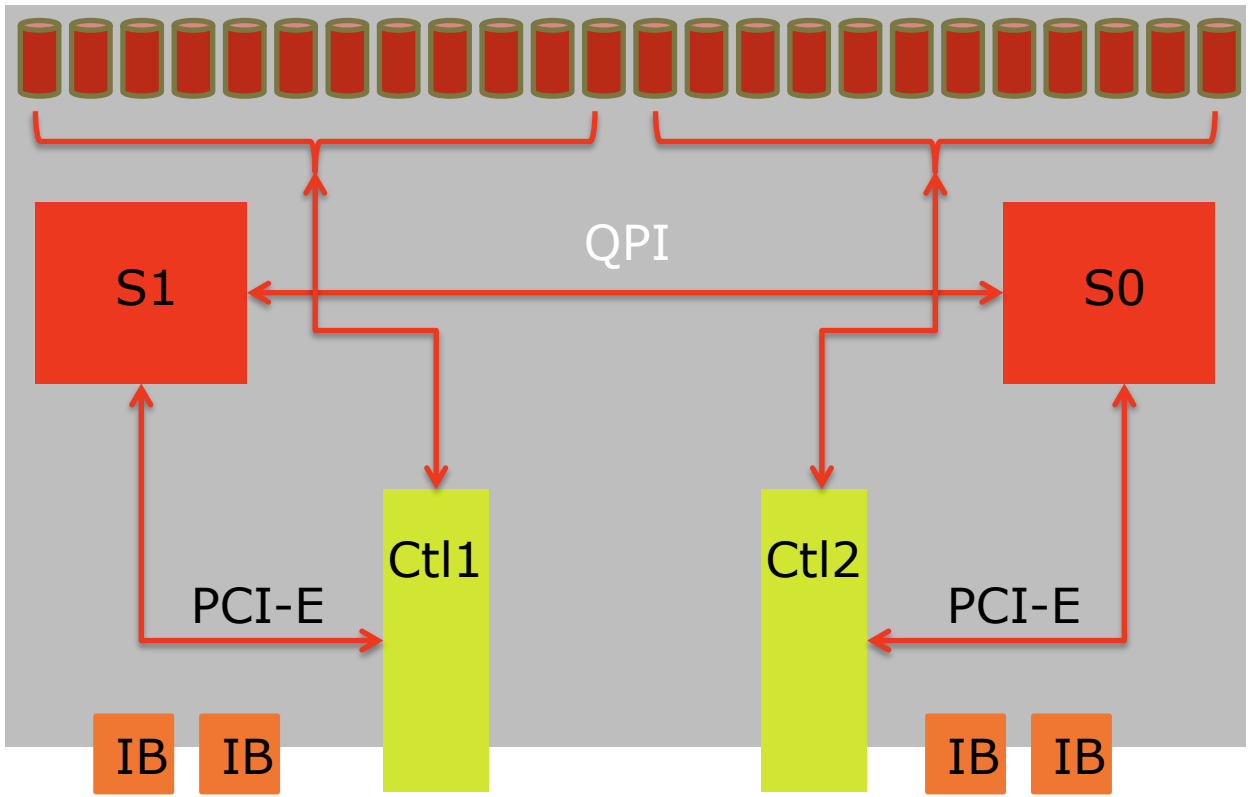
- 3GB/s on sequential write with 1MB block size
- Deliberately under-provisioned

Burst Buffer Hardware Capability

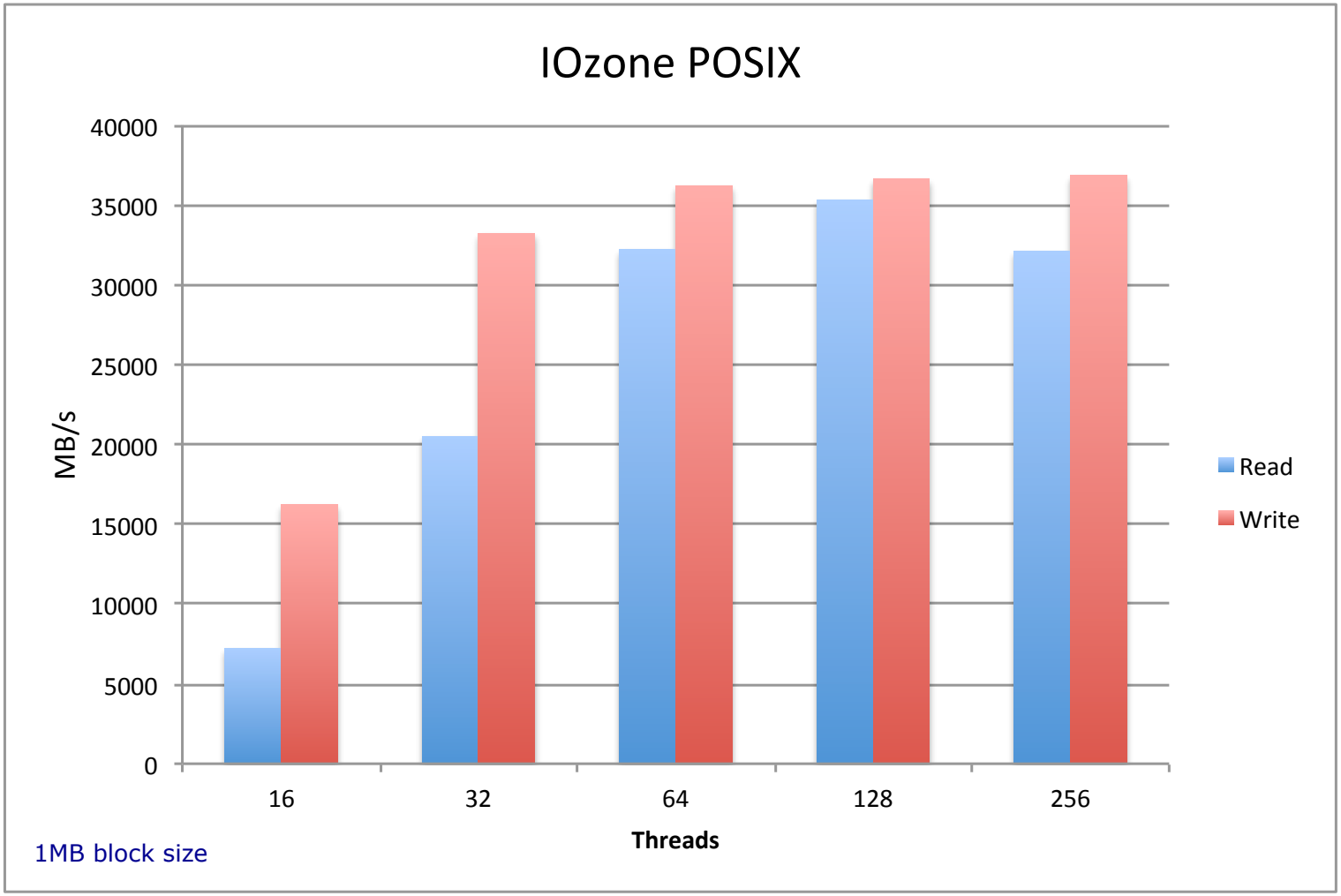
- 4 * Dual socket servers
- 24 SSD per server
- Two * IB HCA FDR
- 128GB of memory

- Each server delivers 10GB/s peak performance (40GB/s overall)

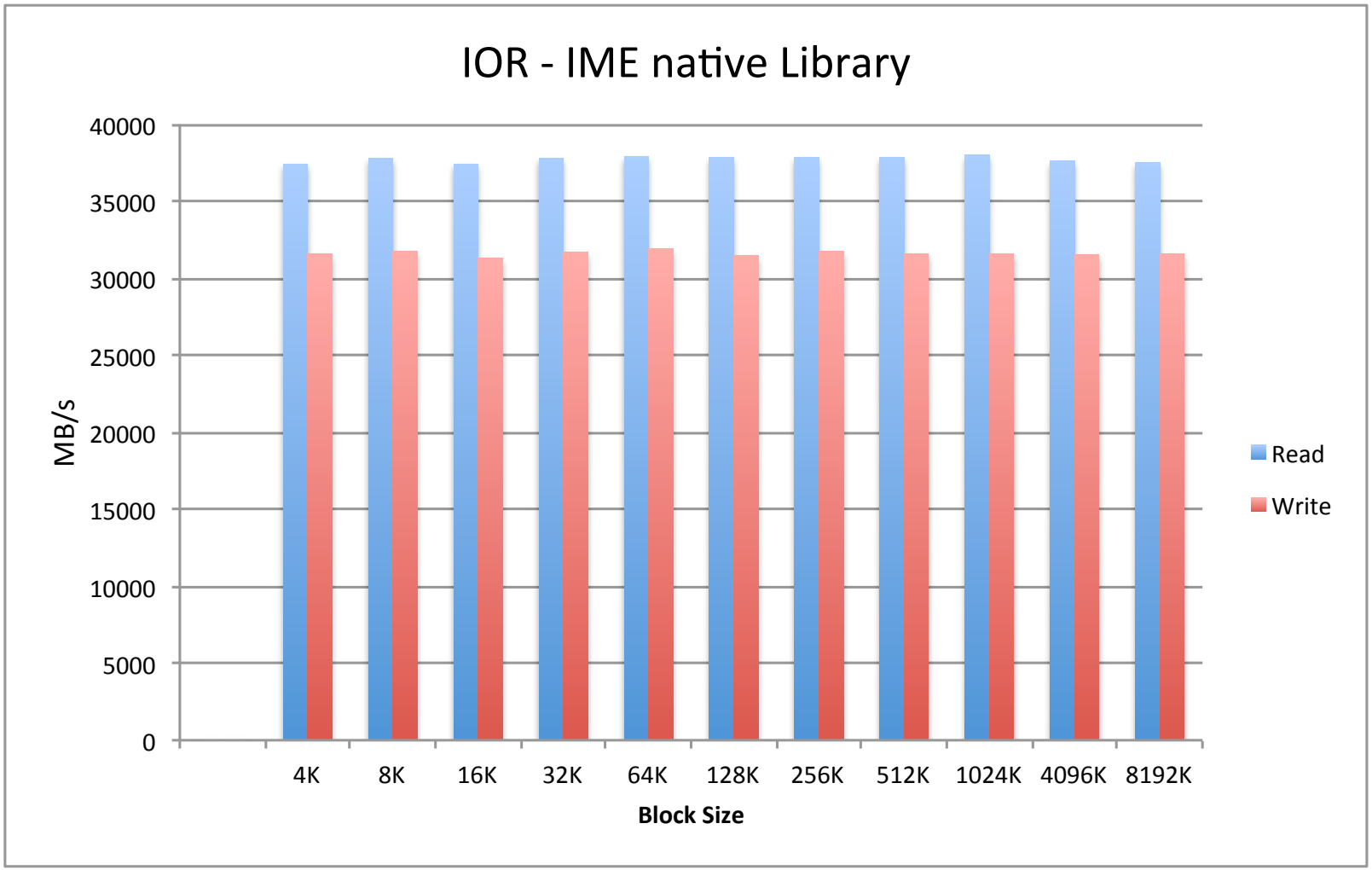
IME cache server layout



IOzone results on IME using POSIX

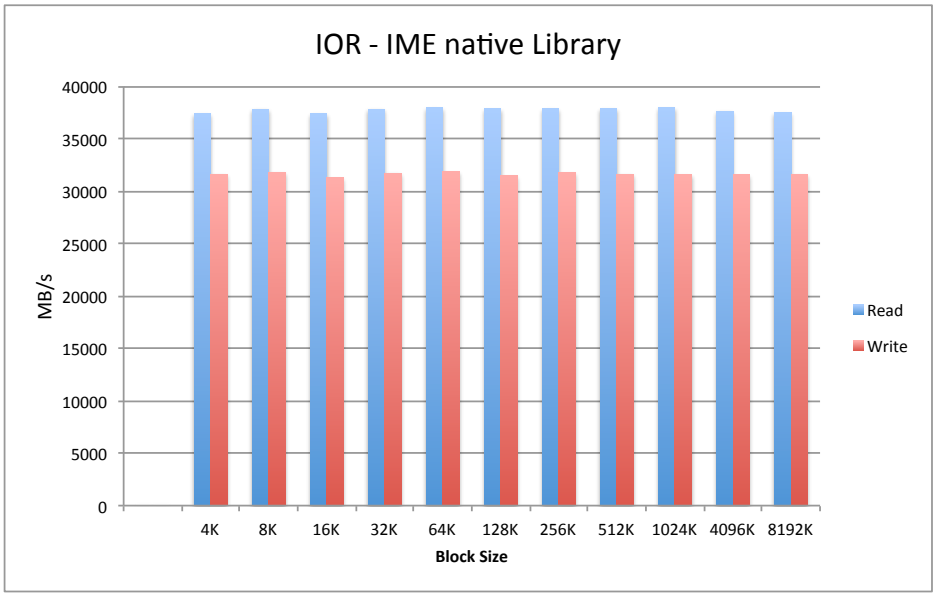


IOR – IME Native Library Results



IOPs

- 37GB at 4K block size ~9.6M IOPs
- 96 SSDs 30K IOPS per SSD 2.8M IOPs
- But how??!



Block resizing

```
greina31: CML Device Stats
greina31:  ID  Name                Read IOs     MBs  Write IOs     MBs  Unmap IOs
greina31:   0        /dev/sg24        3018   395.7          0     0.0          0
greina31:   1        /dev/sg25        2851   373.8          0     0.0          0
greina31:   2        /dev/sg26        2970   389.4          0     0.0          0
greina31:   3        /dev/sg27        2944   386.0          0     0.0          0
greina31:   4        /dev/sg28        2956   387.5          0     0.0          0
greina31:   5        /dev/sg29        2852   373.9          0     0.0          0
greina31:   6        /dev/sg30        2963   388.5          0     0.0          0
greina31:   7        /dev/sg31        2866   375.7          0     0.0          0
greina31:   8        /dev/sg32        2953   387.1          0     0.0          0
greina31:   9        /dev/sg33        2947   386.4          0     0.0          0
greina31:  10        /dev/sg34        2989   391.9          0     0.0          0
greina31:  11        /dev/sg35        2989   391.9          0     0.0          0
```

4K blocks get resized to 128K when get written to SSDs

Next Steps

- Migration data from and to cache
- Multi-rails implementation
- Job scheduler integration
- Quota management
- Ethernet support
- Data management and policy engine
- 3rd party server support

Thanks for your attention.