Optimizing FUSE for Cloud Storage

Parallels™

Profit from the Cloud

Maxim Patlasov

Linux Kernel Developer, Parallels Inc.

Vault 2015

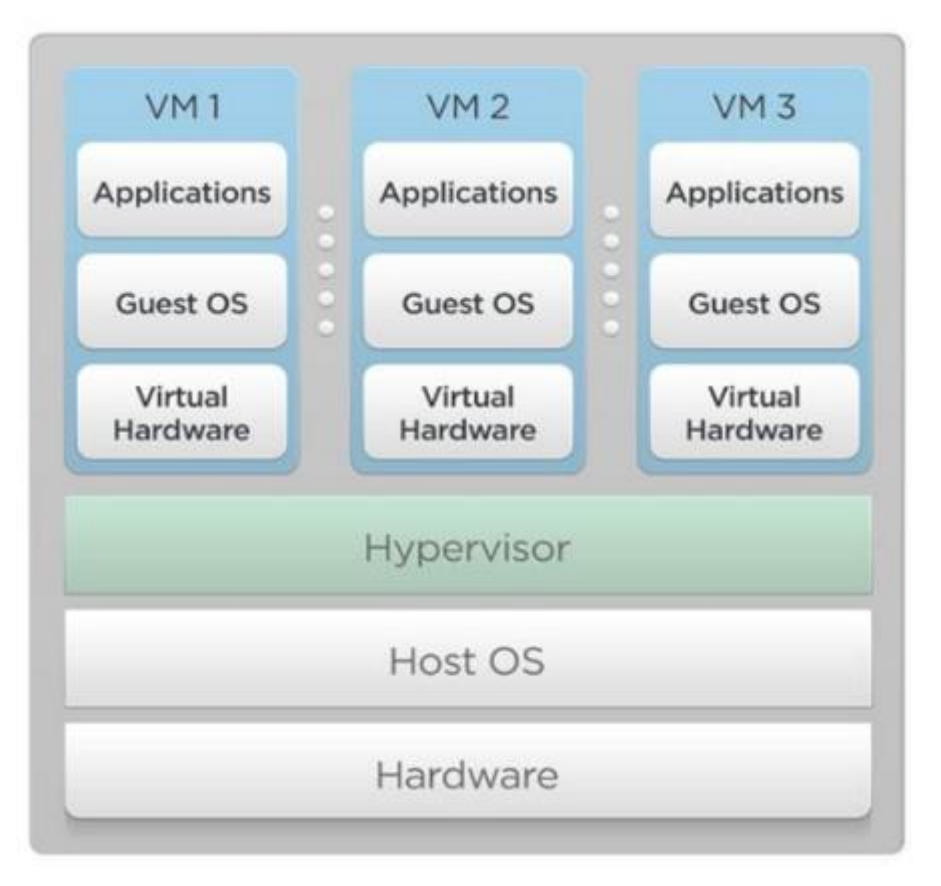
Agenda

1. Parallels Cloud Storage
2. FUSE concept
3. FUSE optimizations
4. Performance achieved
5. Future improvements

Parallels Cloud Storage

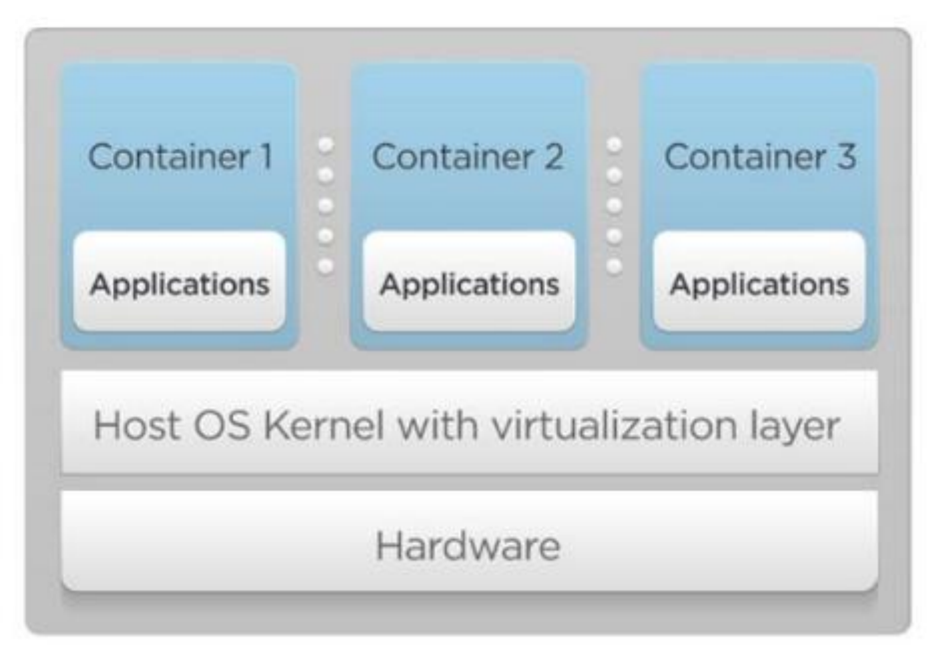
Parallels Hypervisor Virtualization

- OS flexibility
- HW emulation
- Bare metal installation



Parallels Containers

- More efficient memory management
- More efficient caching reduces I/O
- CT resource management
- Easy migration
- Easy backups&snapshots



Storage requirements

Key requirements for VM and Container's needs:

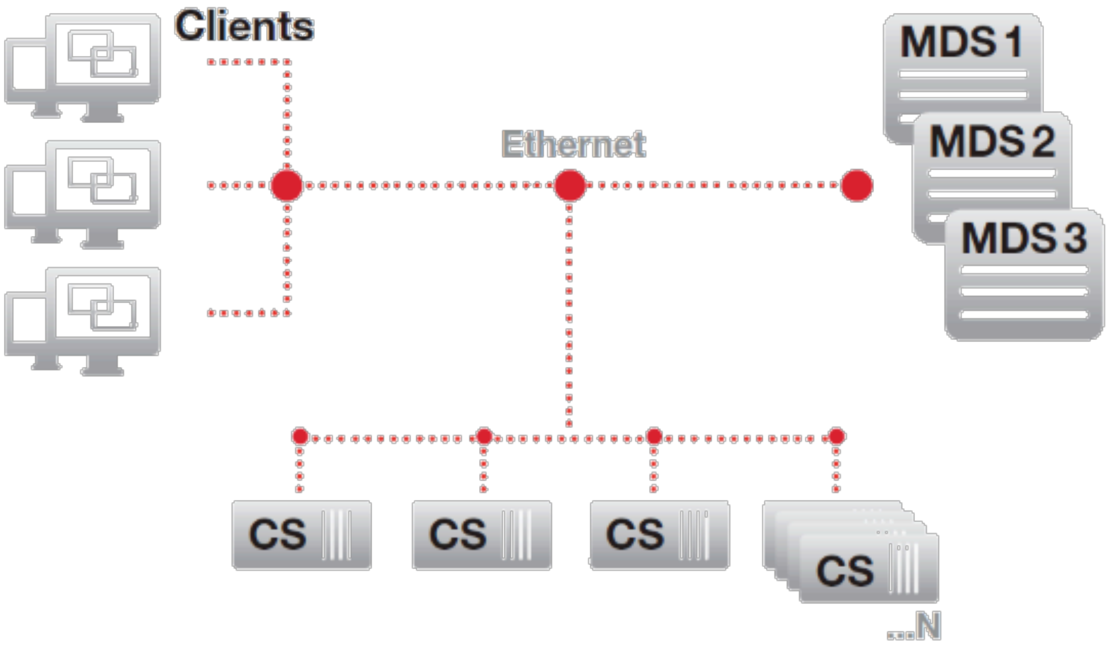
- Strong consistency
- High performance
- Fault tolerance
- Fast recovery
- Address all space from any node
- Commodity hardware
- In-flight reconfiguration and update

Parallels Cloud Storage solution

Key decisions made:

- Optimize for big files
- Union of all local storages
- Replication for fault tolerance
- Keep data and metadata separately
- Multiple metadata servers

PStorage architecture



Meta Data Server (MDS)

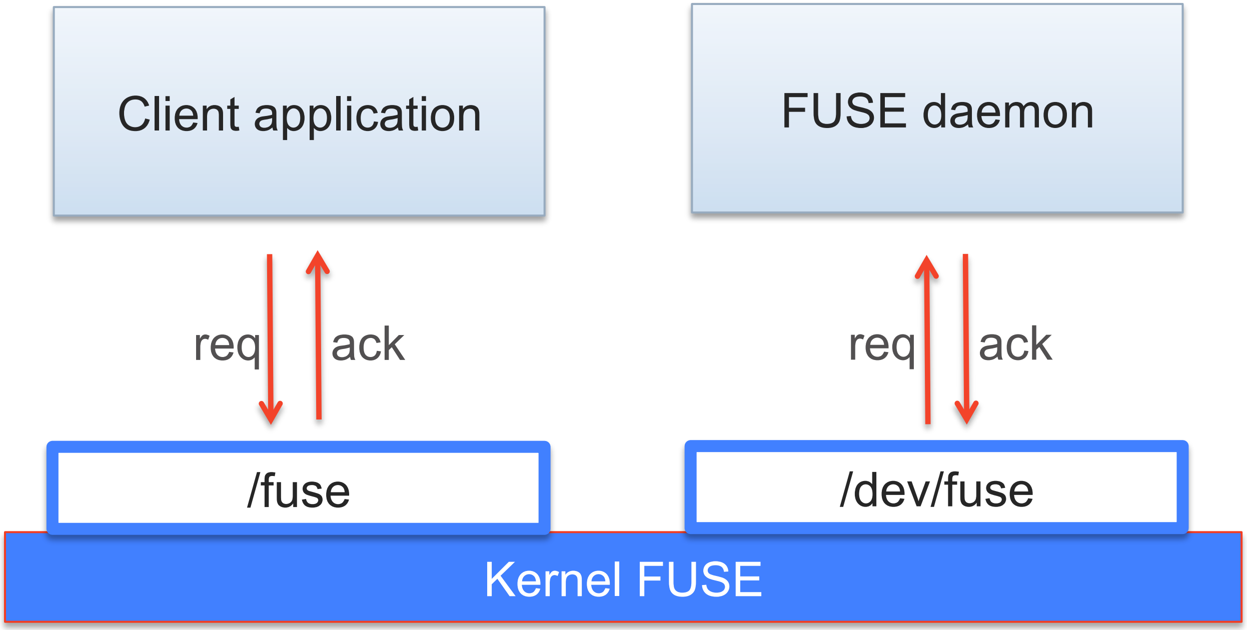
- Stores metadata in memory
- Tracks data chunks and their versions
- Is highly available
- Can run on the same server as the chunk server and client

Chunk Server (CS)

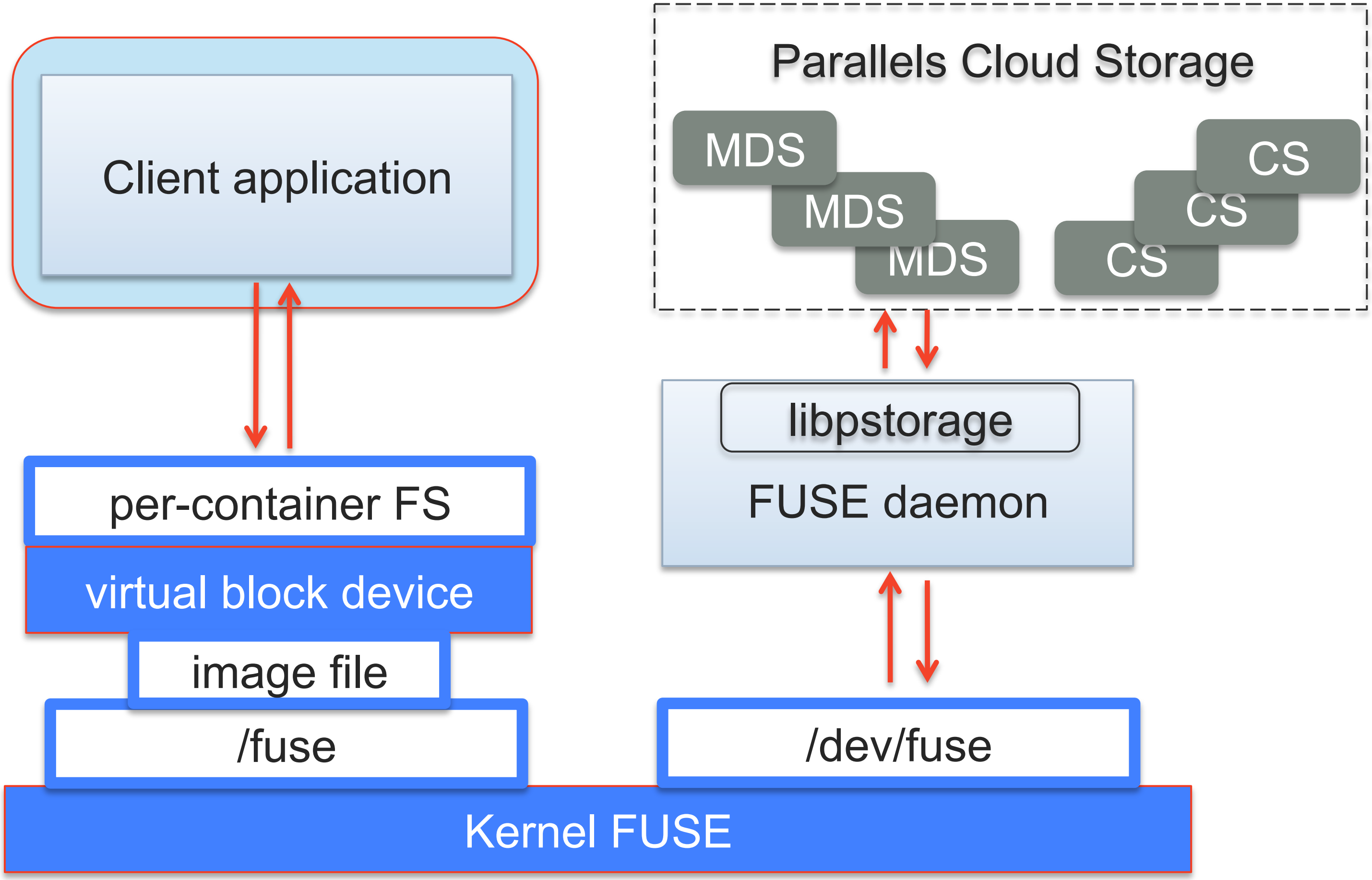
- Stores data chunks
- Manages data chunks
- Performs read/write operations on data chunks
- Can run on the same server as the client

FUSE

FUSE framework



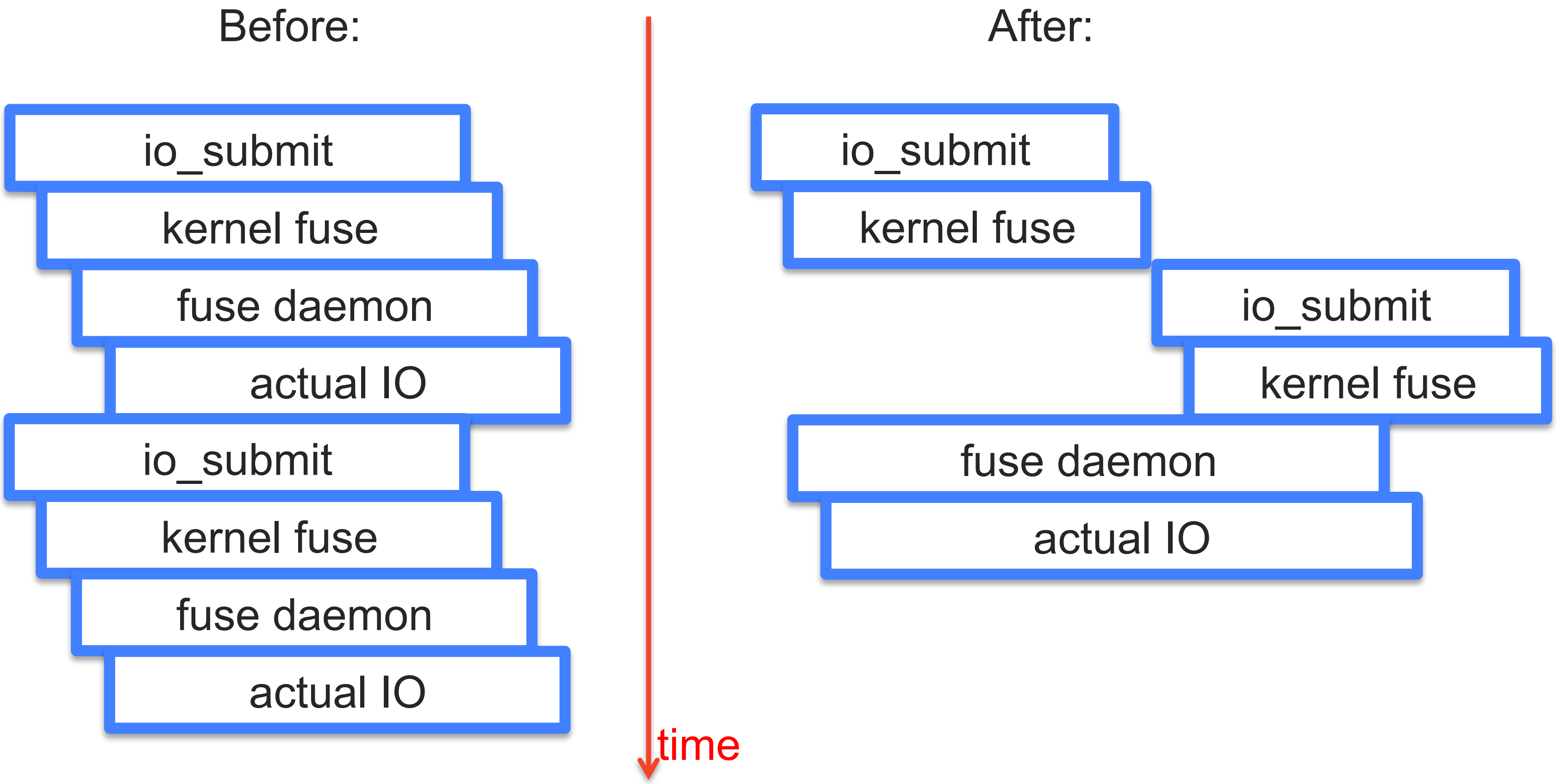
FUSE: Containers on PStorage



FUSE optimizations

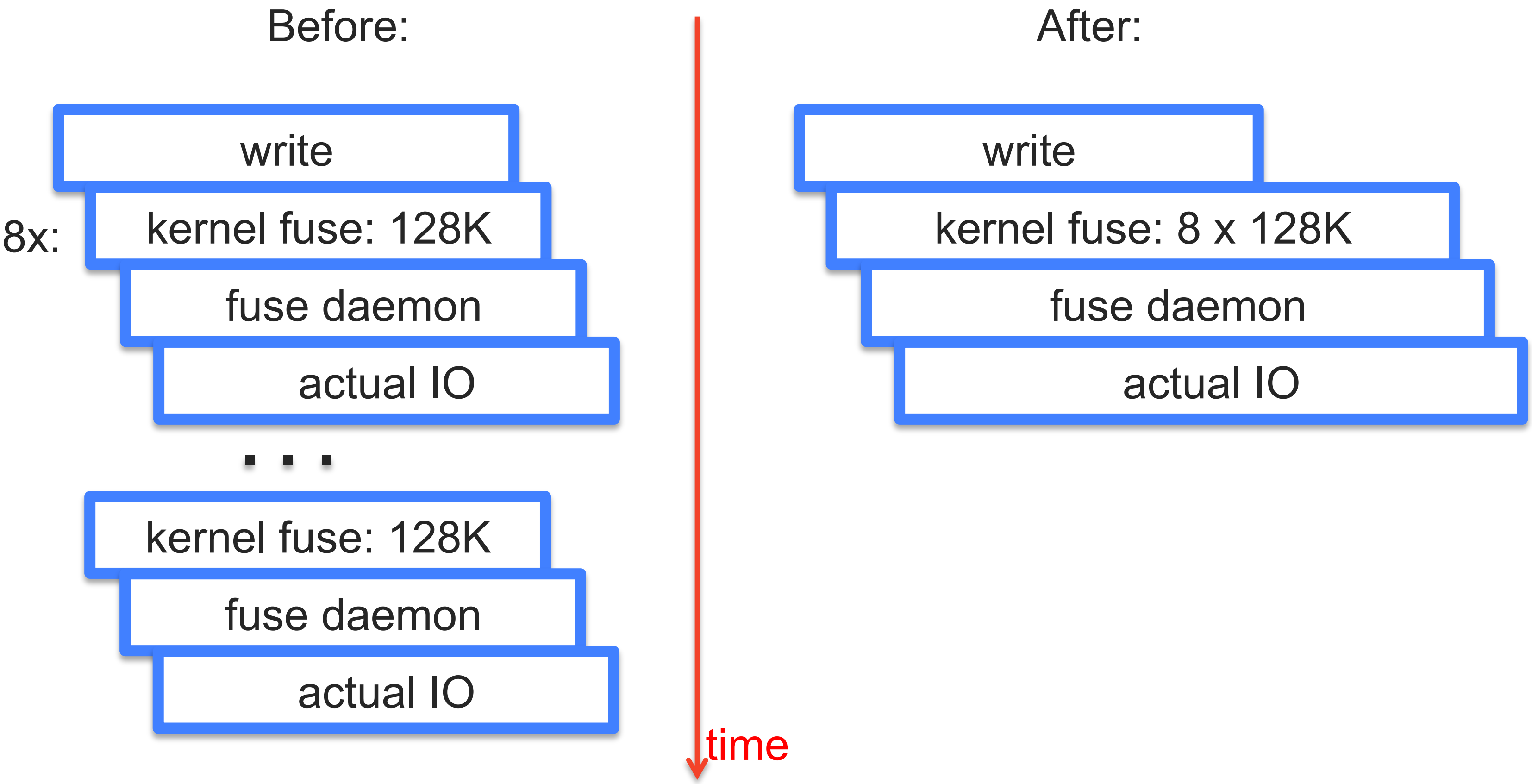
Asynchronous direct IO

Application: io_submit(&iocb1); io_submit(&iocb2);



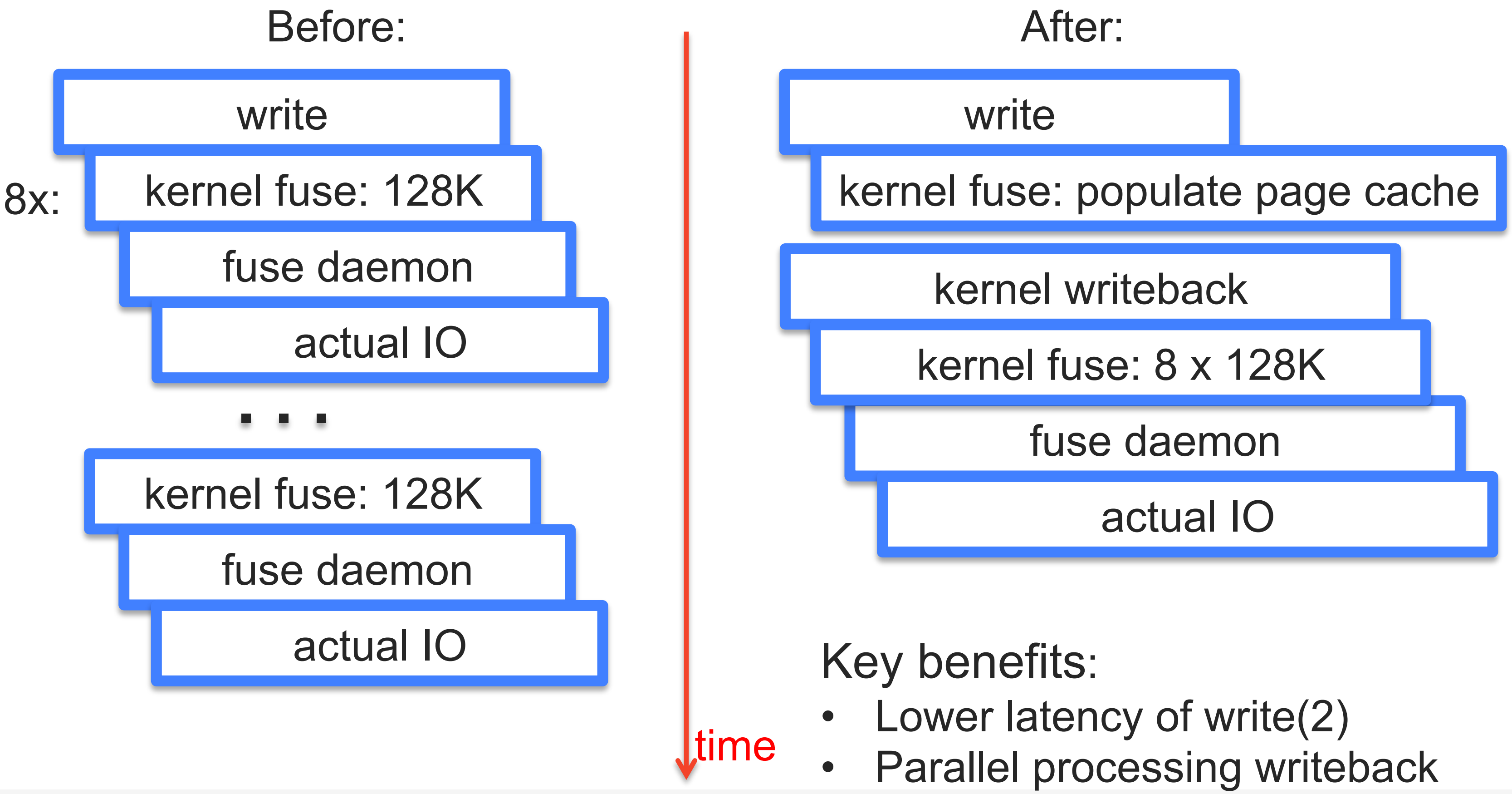
Synchronous direct IO

Application: fd = open(O_DIRECT); write(fd, buf, 1<<20);



Writeback cache

Application: buffered write(fd, buf, 1<<20);



Key benefits:

- Lower latency of write(2)
- Parallel processing writeback

Performance Comparison :: HW

iSCSI SAN Storage
DELL EqualLogic PS6510E

x1 HW SAN EQL PS6510E

48 SATA Disks: 1TB 7200rpm
(Seagate ST31000524NS)

Network: **10Gbit**
(Dell Force10 S4810)

vs.

Parallels Cloud Storage
(FUSE based)

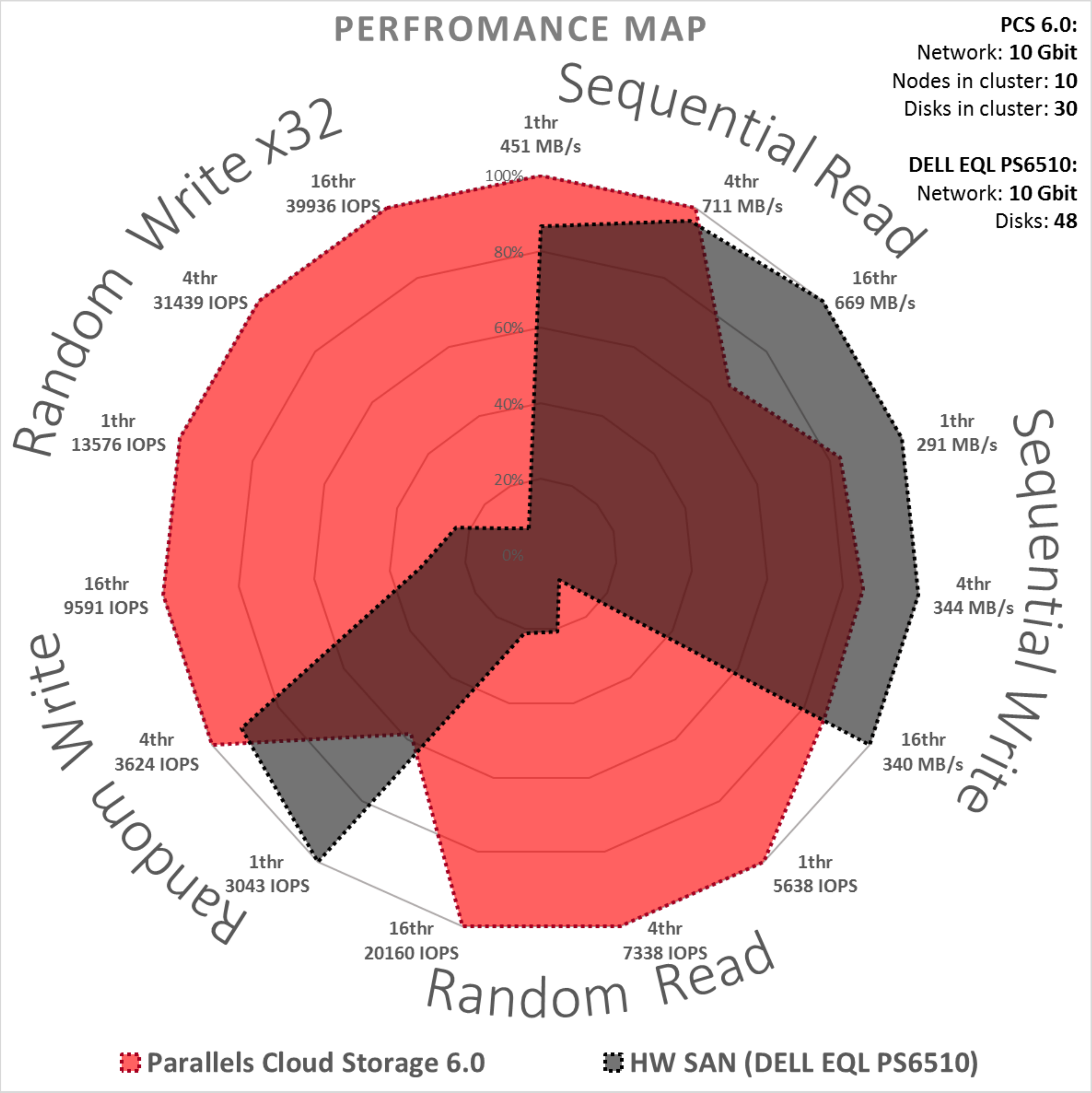
x10 compute nodes

30 SATA Disks: 2TB 7200rpm
(Seagate ST2000DM001)
+ 10 SSD for caching
(Intel SSD 520)

Network: **10Gbit**
(Brocade FastIron SuperX SX-F42XG)

PCS FASTER
than
HW SAN

Just 10 nodes PCS cluster faster than DELL EQL SAN ($97000) in most workloads



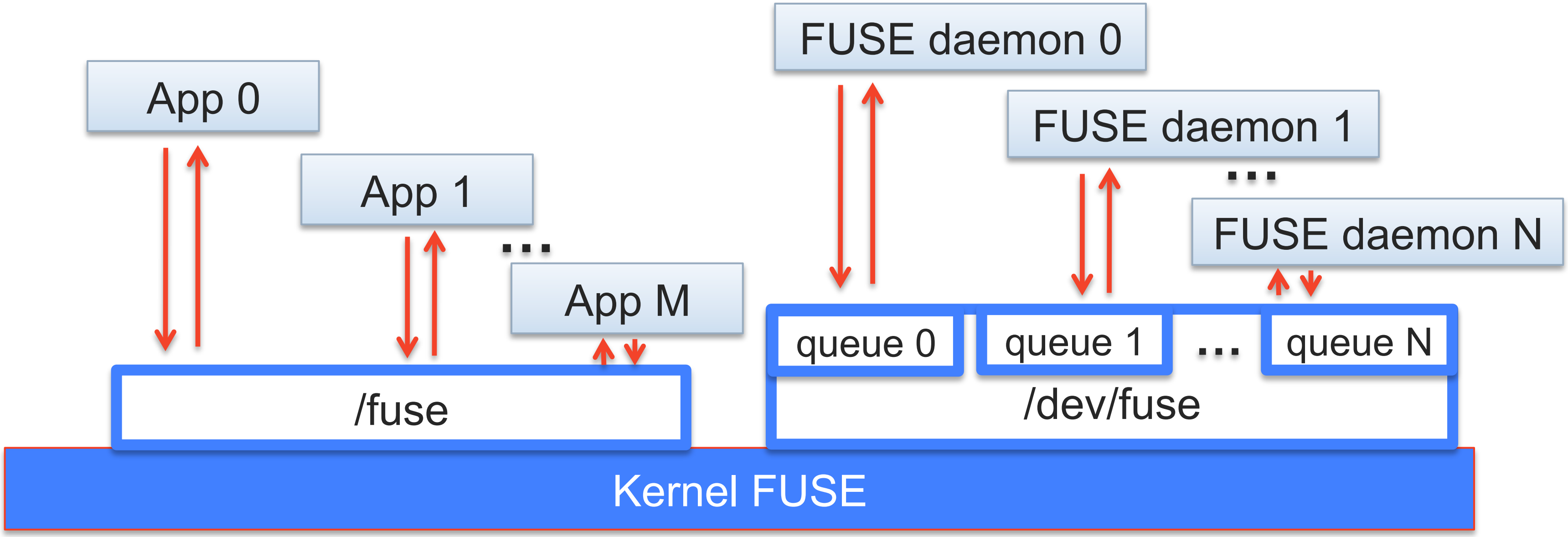
PCS 6.0:
Network: **10 Gbit**
Nodes in cluster: **10**
Disks in cluster: **30**

DELL EQL PS6510:
Network: **10 Gbit**
Disks: **48**

FUSE: what's next?

FUSE: future improvements

- Variable message size (currently 128K)
- Eliminate global lock
- Multi-queue
- CPU and NUMA affinity



Q&A