The ONE MILLION IOPS

TU Dresden Lustre FS

21-03-2016

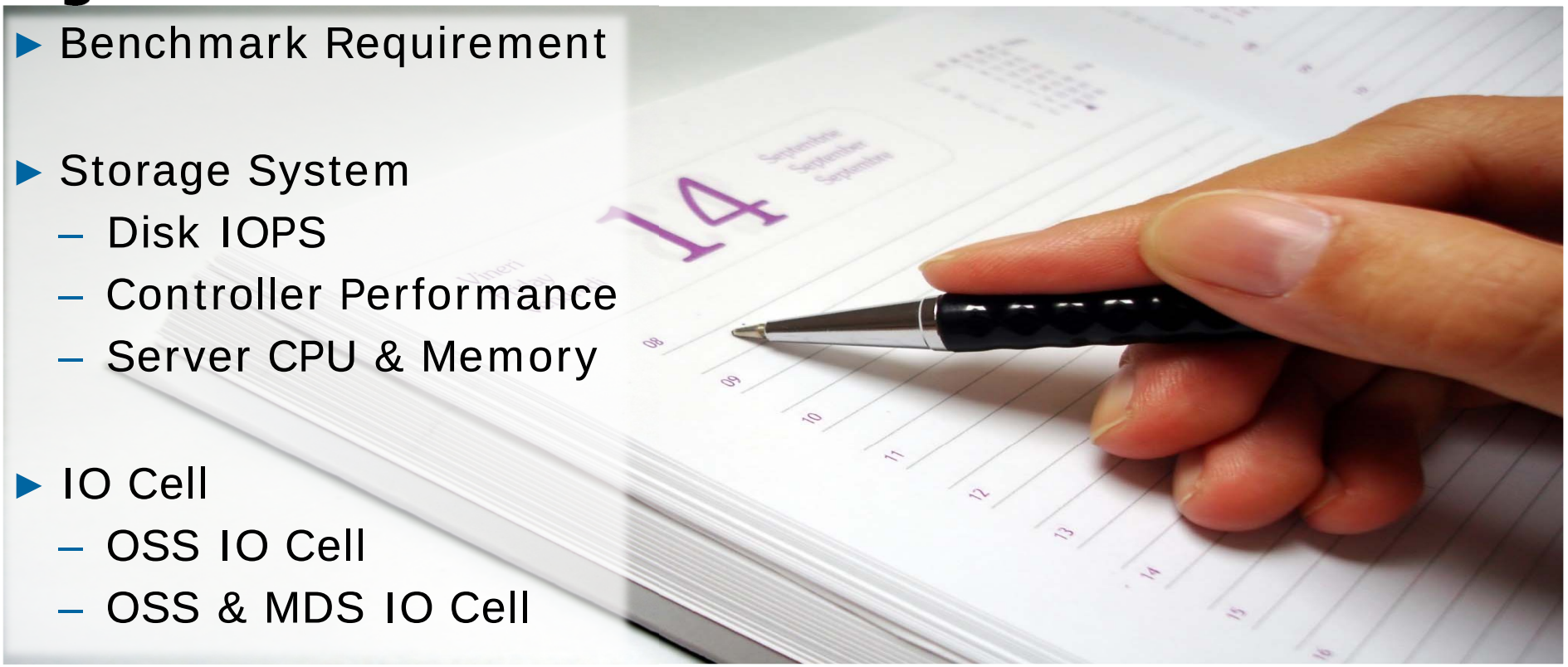
Your business technologists. **Powering progress**

ZIH
Zentrum für Informationsdienste und Hochleistungsrechnen

Atos

>> Agenda

Agenda



- Benchmark Requirement
- Storage System
  - Disk IOPS
  - Controller Performance
  - Server CPU & Memory
- IO Cell
  - OSS IO Cell
  - OSS & MDS IO Cell

>> Benchmark Requirement

Benchmark Requirement

- Commited to :
  - 1 Miops/s Random Read
- Measurement on FAT :
  - 1,3Miops Random Read
- What we have done to reach it :
  - N=609
  - ppn=24
  - n=14616
  - Runtime=266s
  - Score=1,320,137 iops

Dear All,

I'm proud to announce you that we finally reached our commitments for the randomops benchmark.

JOBID	N	n	Time	# Ops / task	IOPS
12501	609	14616	266	20000	**1,320,137**
12504	609	14616	37	2000	**1,342,102**
12515	609	14616	37	2000	**1,353,712**

Regards,

Diego Moreno

>> Storage System

Storage System Chosen

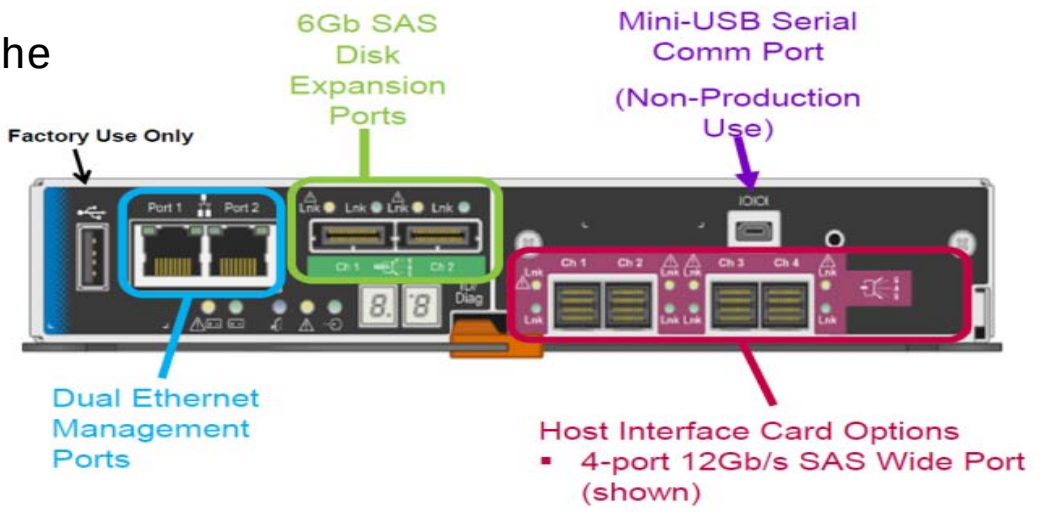
Iops Performance of Disk Drives

- Toshiba PX02SMF080 800GB 2.5 SAS
  - Random Read
    - 120 000 Iops 4K
  - Random Write
    - 25 000 Iops 4K
  - Sequential Read
    - 900 MB/s
  - Sequential Write
    - 400 MB/s

Storage System Chosen

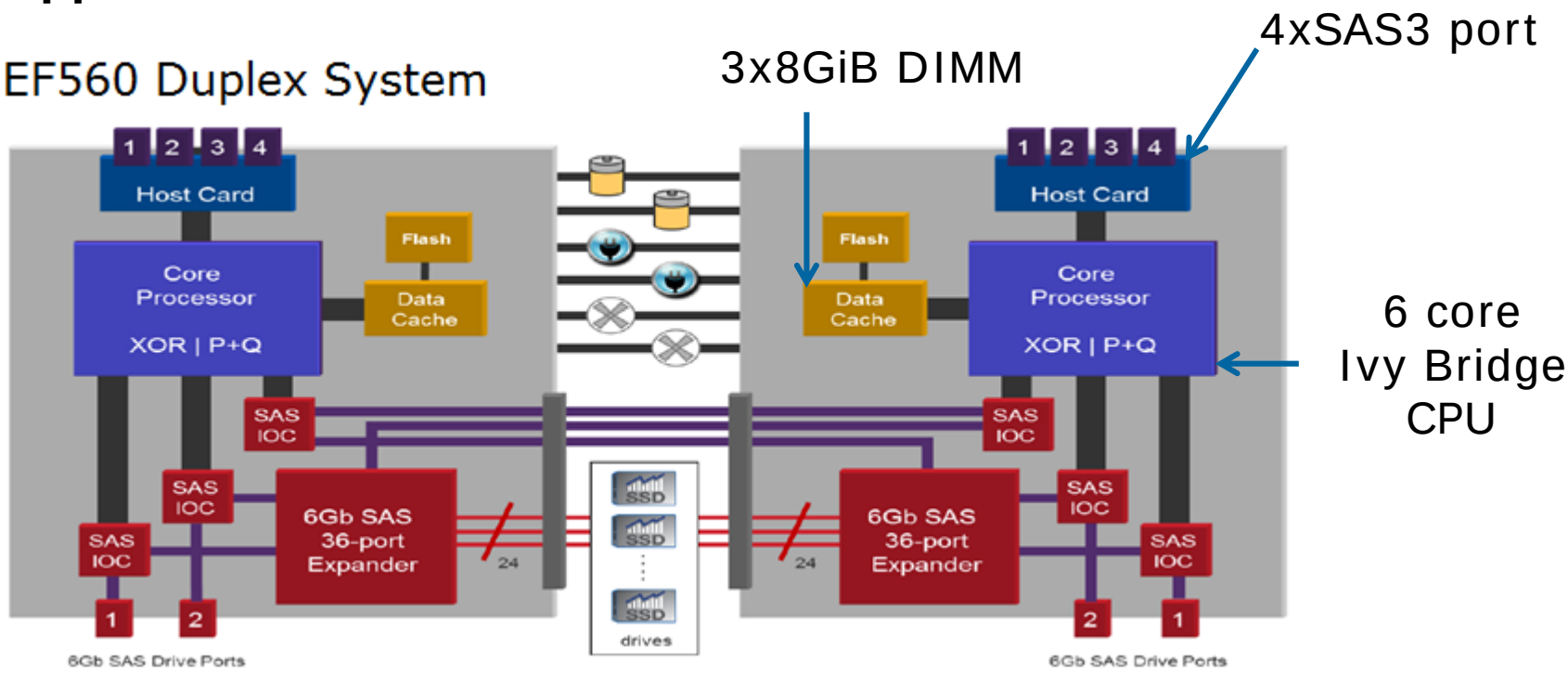
Netapp EF560

- Per controller
  - 1xIvy Bridge 6 Core
  - 4 port SAS3 12Gb/s
  - 24 Gib of Ram
  - 12 GiB of Mirrored data cache
- For each dual controller
  - 20xSAS SSD Toshiba



Storage System Chosen

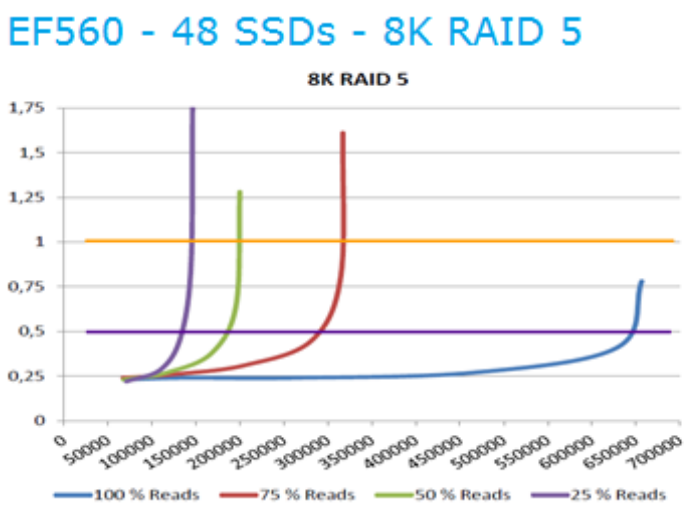
Netapp EF560



Storage System Chosen

Netapp EF560 controller Iops and Sequential Performance

- Random Read
  - 825Kiops 8KiB Raid5
  - 900Kiops cached 4k
- Random Write
  - 130Kiops 4KiB
- Sustained Read 512KiB
  - 12 000MB/s
- Sustained Write 512KiB
  - 6 000MB/s CME
  - 9 000MB/s CMD



X IOPs Mix read/write
Y latency in ms

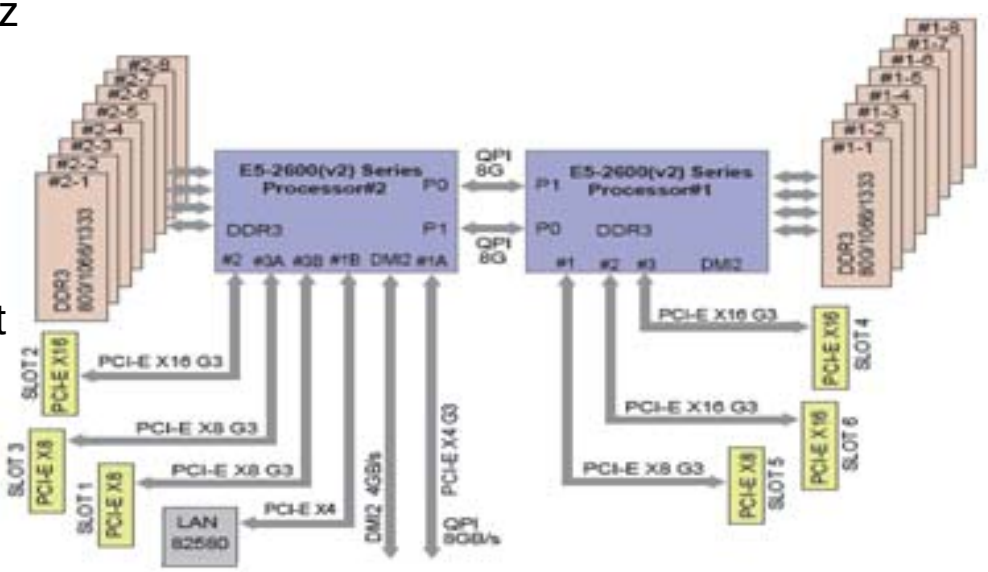
% Reads	Under 1ms	Under 0.5ms
100%	650,000	628,000
75%	314,000	281,000
50%	195,000	168,000
25%	144,000	133,000
0%	117,000	102,000

Tests were run with IOmeter & Toshiba drives

Storage System Chosen

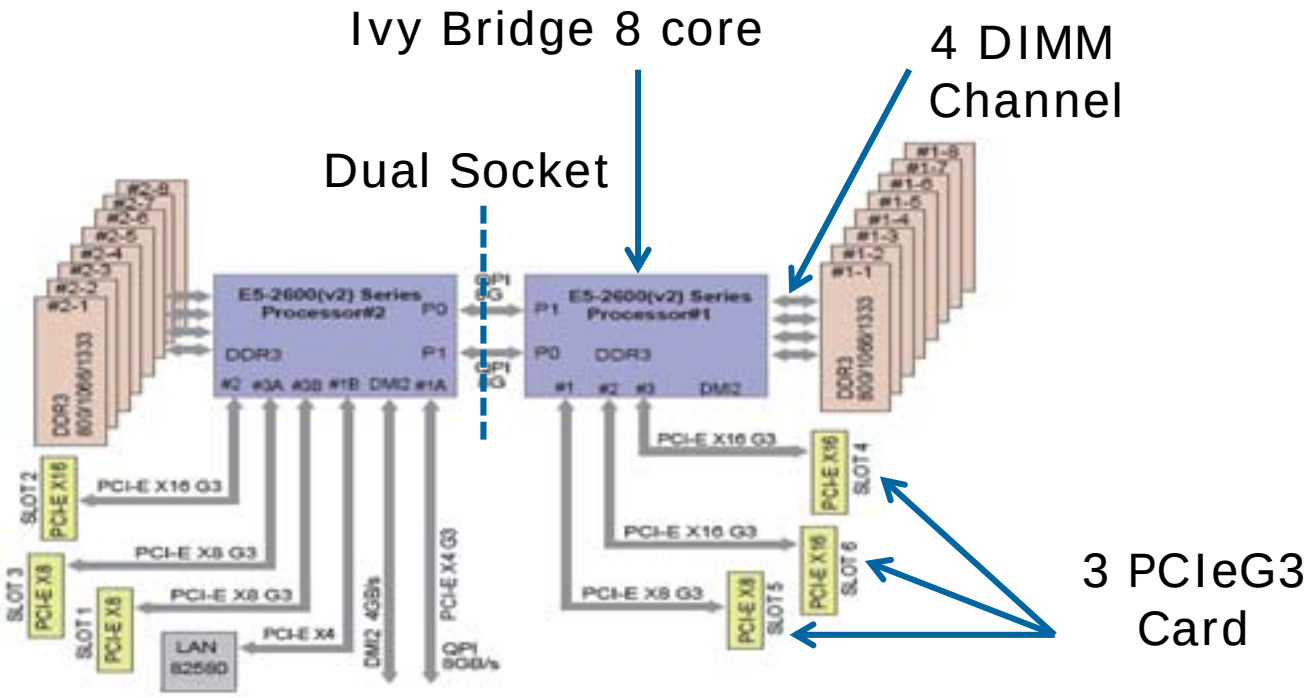
R423e3 IO Server CPU and Memory

- **CPU**
  - 2xIvy Bridge E5-2650v2@2,6GHz
  - 8 core no HT
- **Ram**
  - 2x4x8GiB@1600MT/s
- **Infiniband**
  - 2xIB FDR Card 1xCard by Socket
  - 6GB/s fullduplex
- **SAS Card**
  - 4xSAS3 Card 2xCard by Socket
  - 2.4GB/s fullduplex



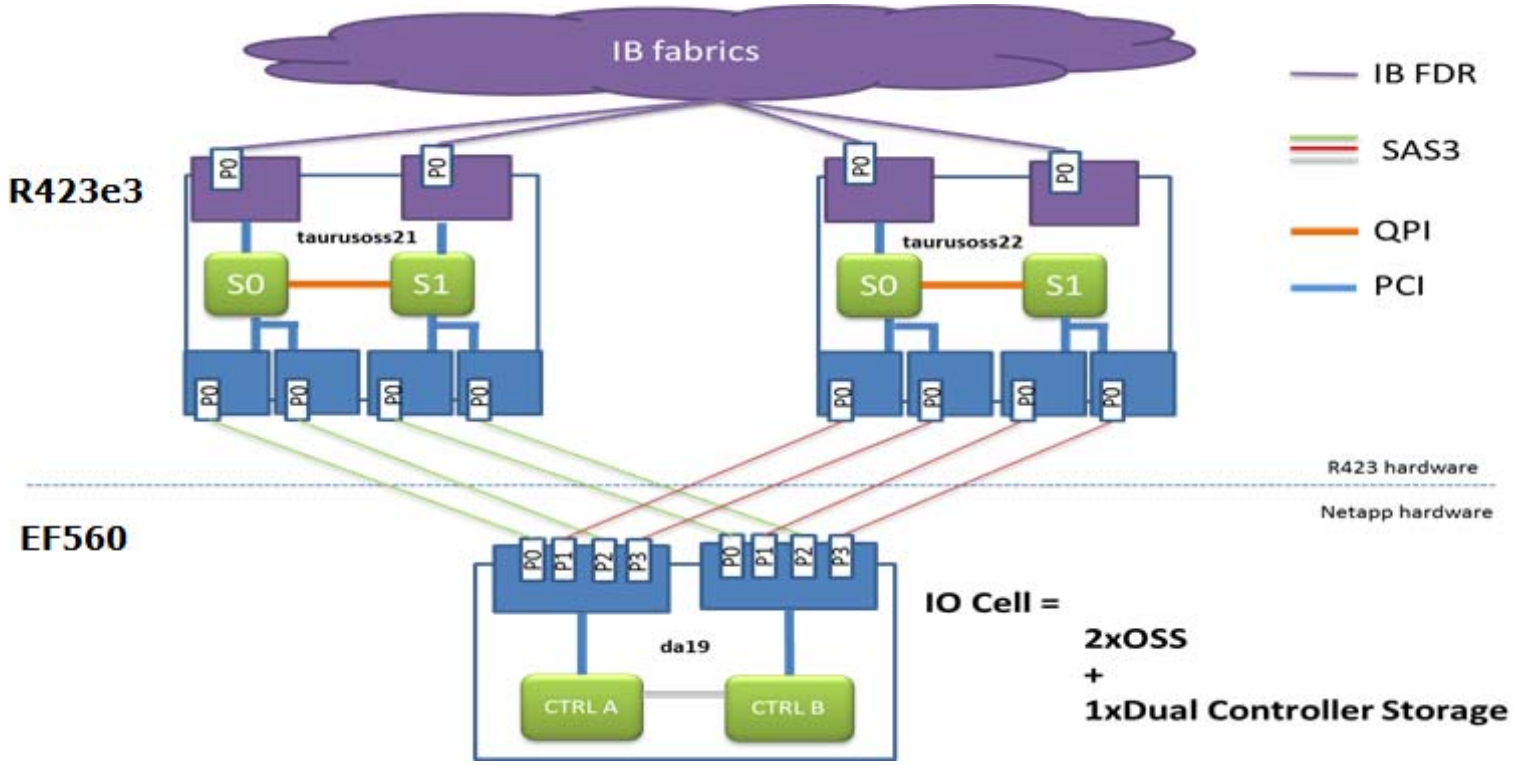
Storage System Chosen

R423e3 IO Server CPU and Memory



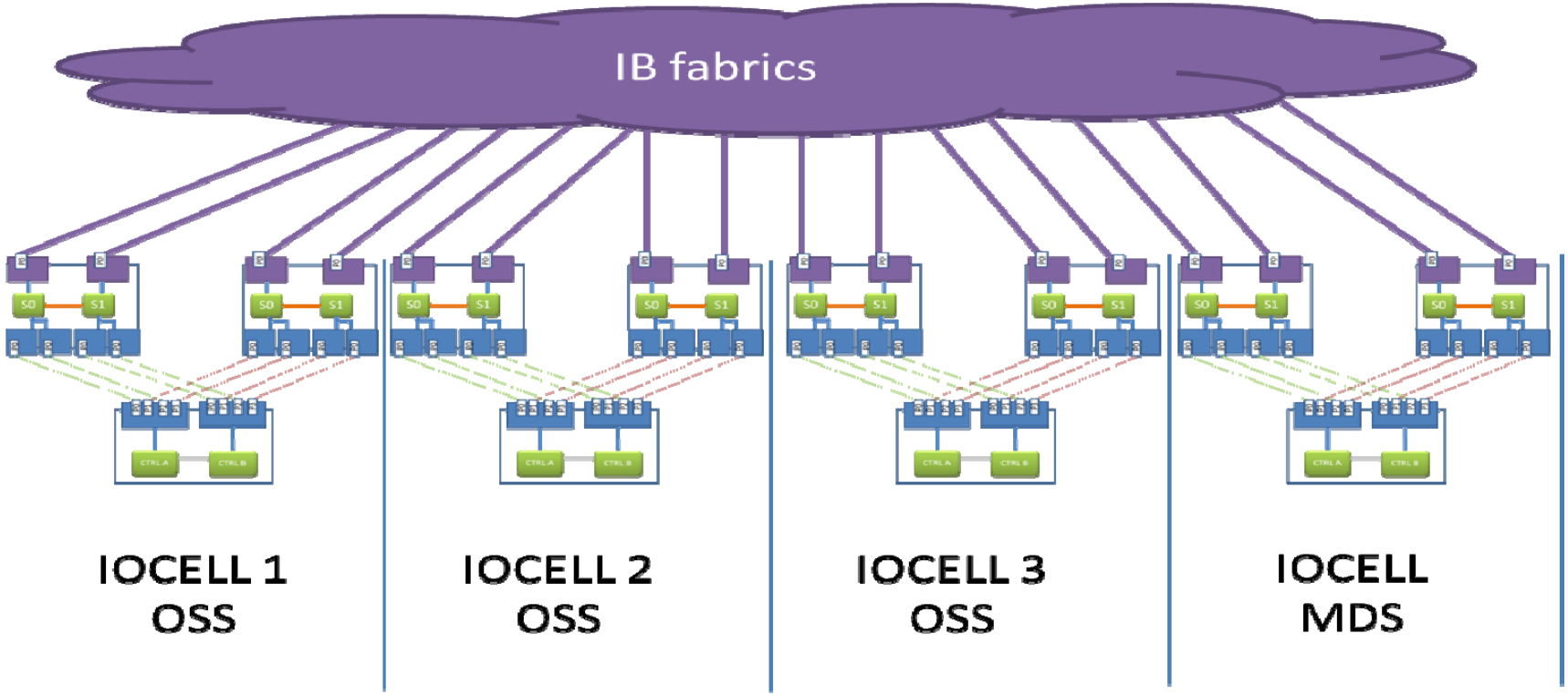
>> IO Cell

IO Cell
OSS IO Cell



IO Cell

Three OSS IO Cell & One MDS IO Cell



Thanks

For more information please contact:

johann.peyrard@atos.net

21-03-2016

TECHNISCHE UNIVERSITÄT DRESDEN
Center for Information Services and High Performance Computing (ZIH)

Performance Measurements Of a Global SSD Lustre File System

Lustre User Group 2016, Portland, Oregon

Zellescher Weg 12
Willers-Bau A 207
Tel. +49 351 - 463 – 34217

Michael Kluge (michael.kluge@tu-dresden.de)

Measurement Setup

- no read caches on the server side:

```
root@taurusadmin3:~> pdsh -w oss[21-26] lctl get_param
obdfilter.*.read_cache_enable

oss26: obdfilter.highiops-OST0009.read_cache_enable=0

…
```

- how files are opened

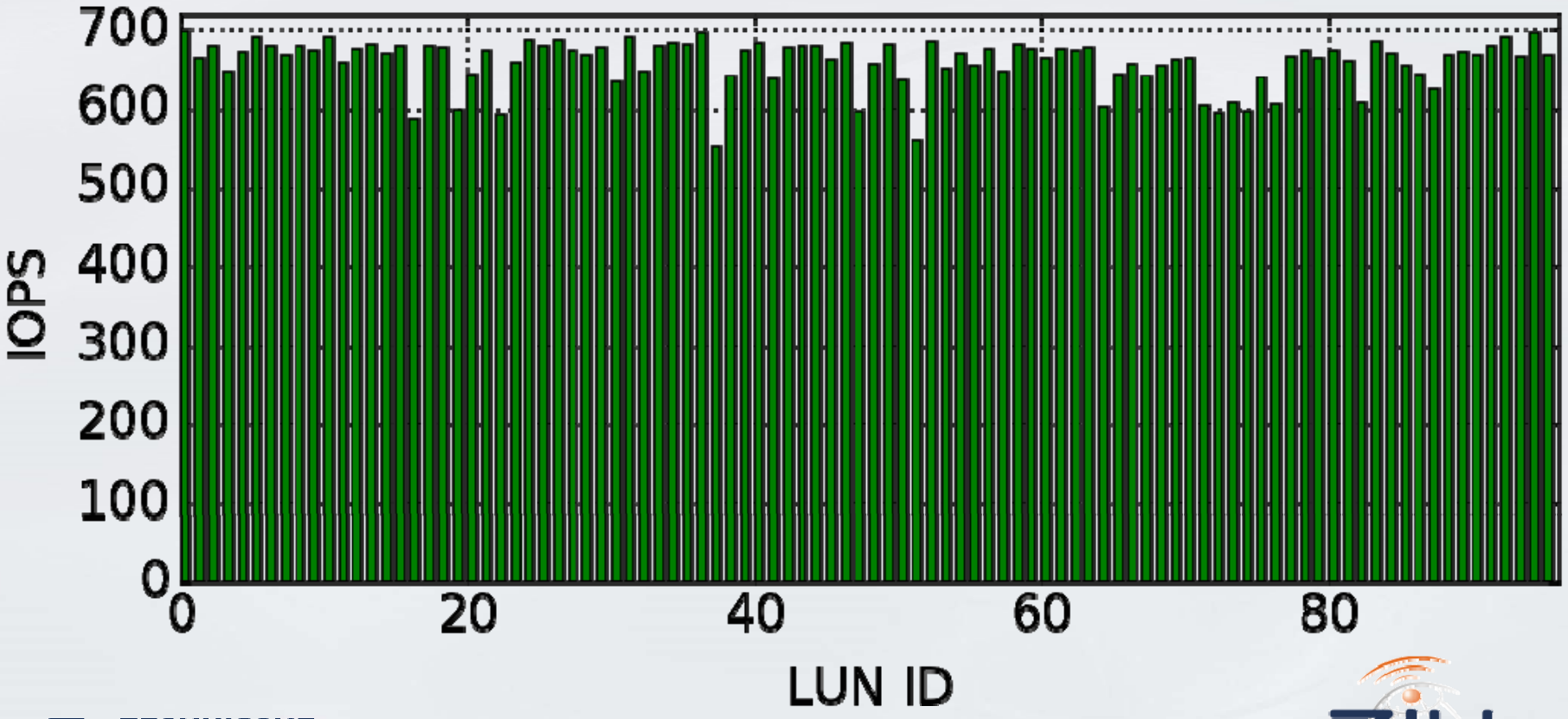
```
file_fd = open( filename, O_RDWR|O_CREAT, 0600 );

posix_fadvise( file_fd, 0, 0,
    POSIX_FADV_RANDOM | POSIX_FADV_NOREUSE | POSIX_FADV_DONTNEED);
```

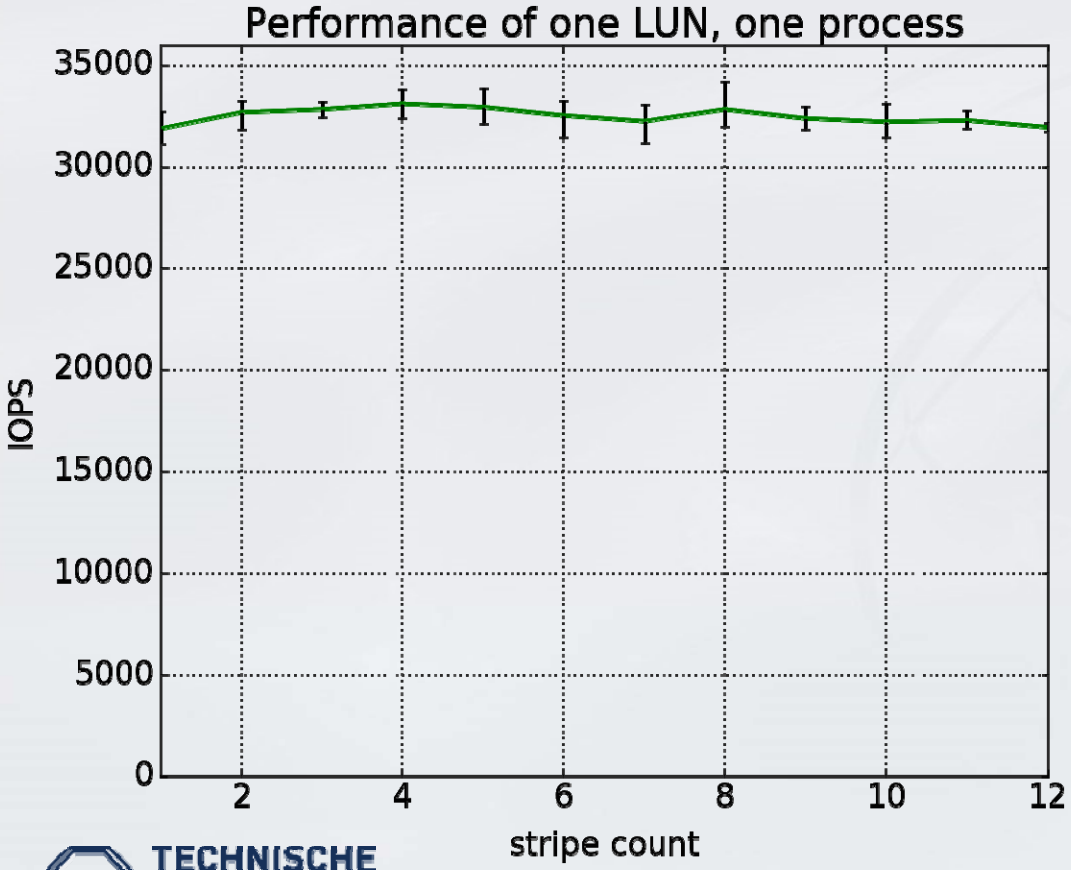
Measurement Setup

- Size on disk at least 4x size of server RAM
- Rotation of MPI ranks
- Data written before the test
- Always started at LUN 0
- size of one IOP: 4 KiB to 1 MiB
- Data collected NOT in exclusive mode
- Data presented as maximum at least three measurements
- Each run was about 5 minutes
- 1 individual file per process, always used pread/pwrite

Single Process / Single LUN for Our SATA Disks (writes)

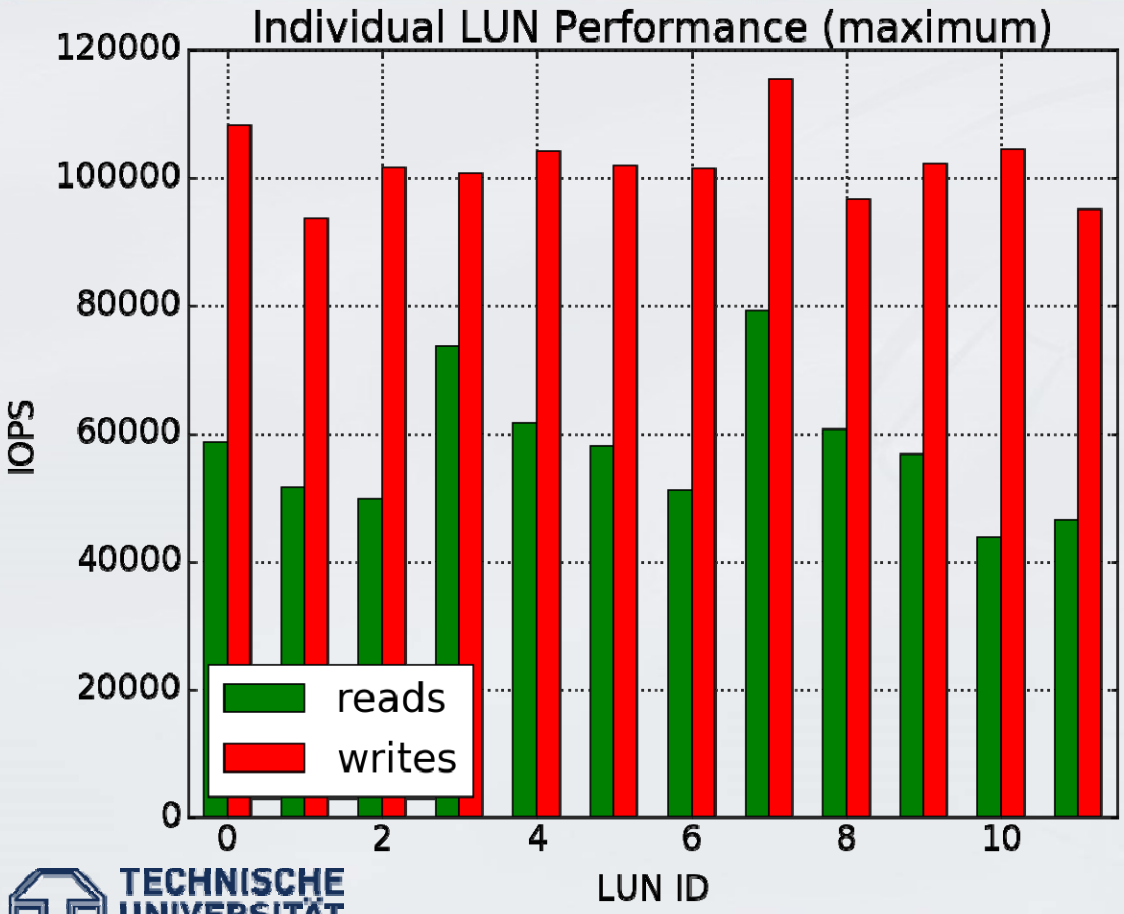


Single Process / 1 - 12 LUNs



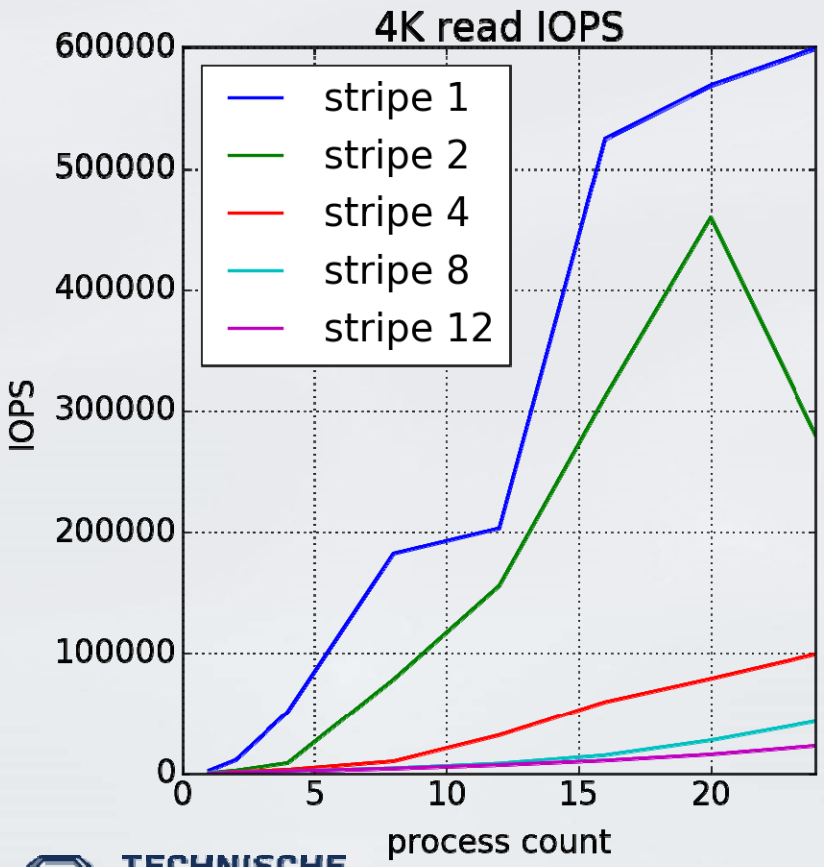
- Single process becomes I/O bound immediately

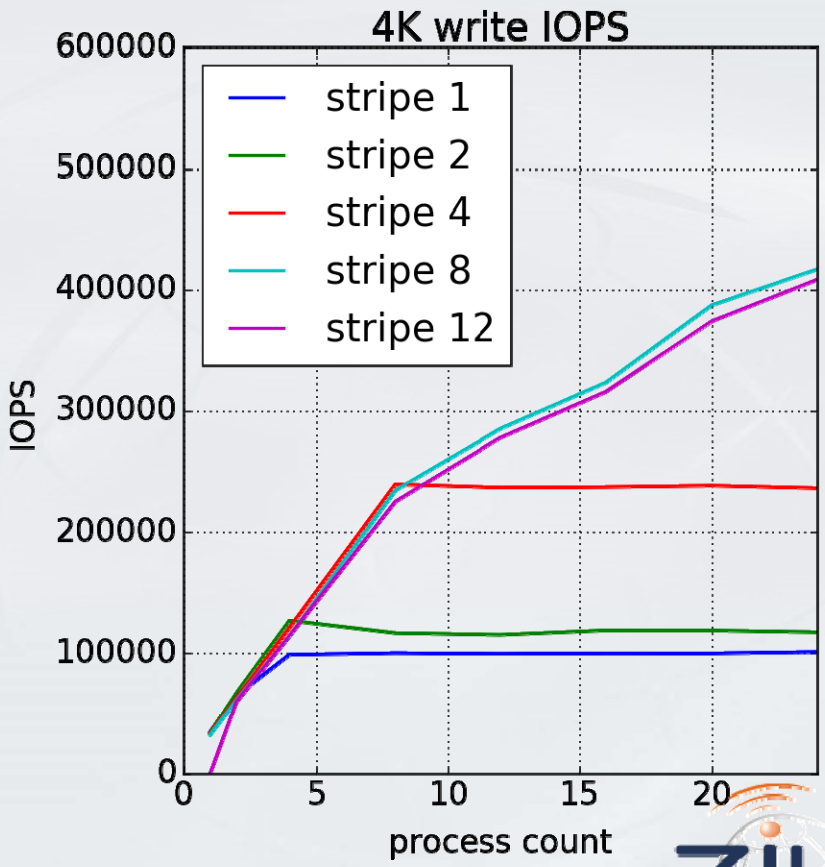
Testing Alls LUNs Of The SSD File System



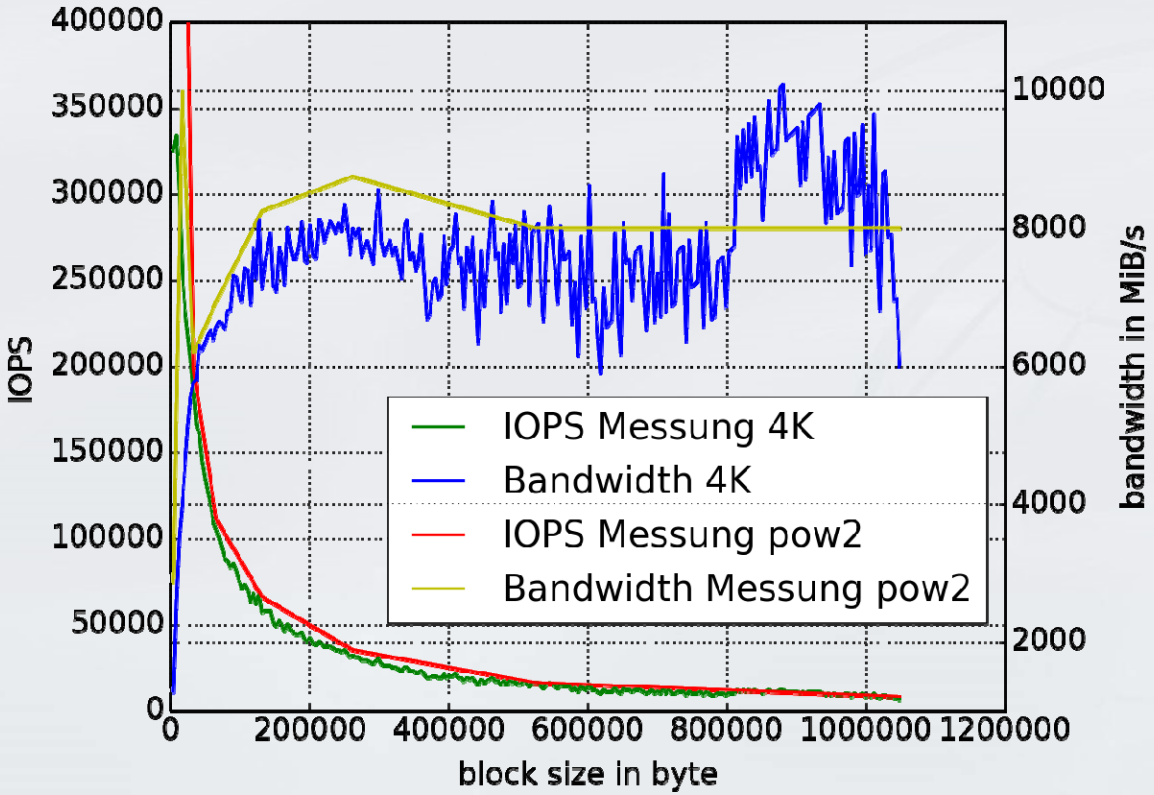
- Utilized all cores
- Writes to a single LUN are faster than reads
- Performance variations > 40%

One Node / Many LUNs / Different Stripes / Reads+Writes



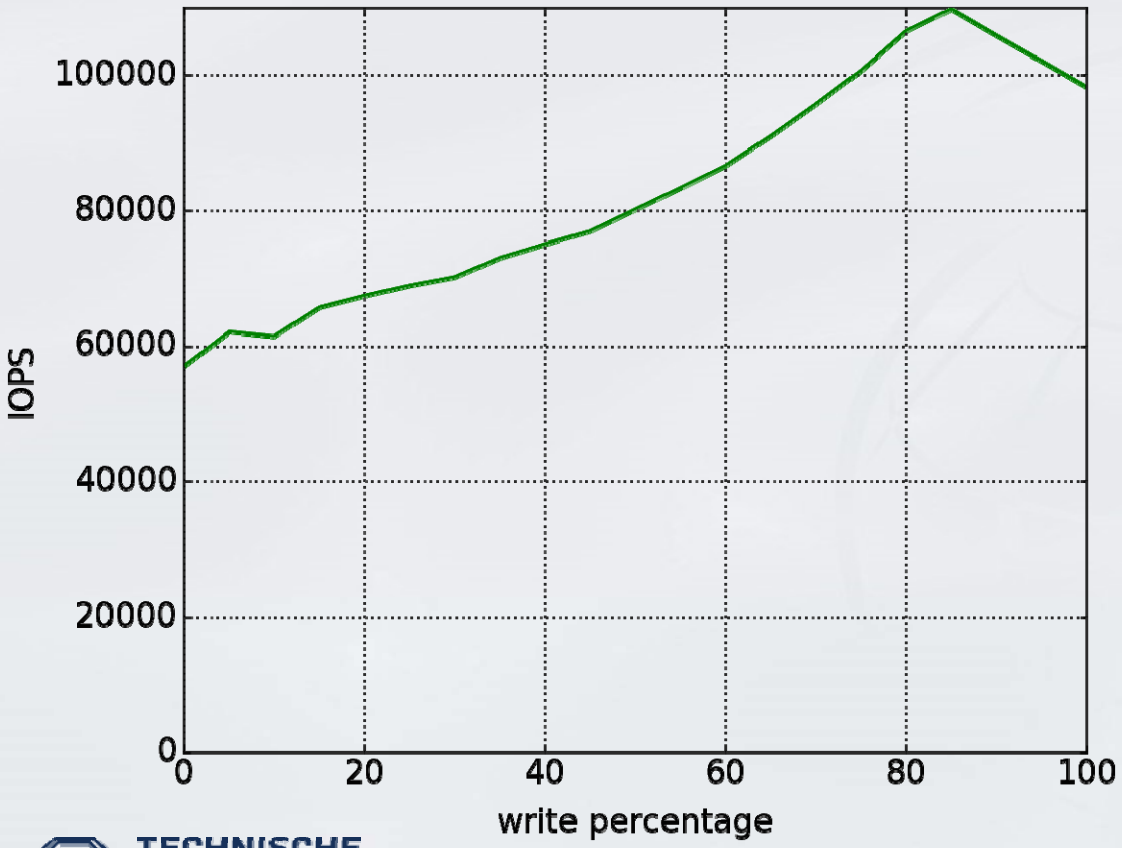


One Node / Different Block Size (Writes)



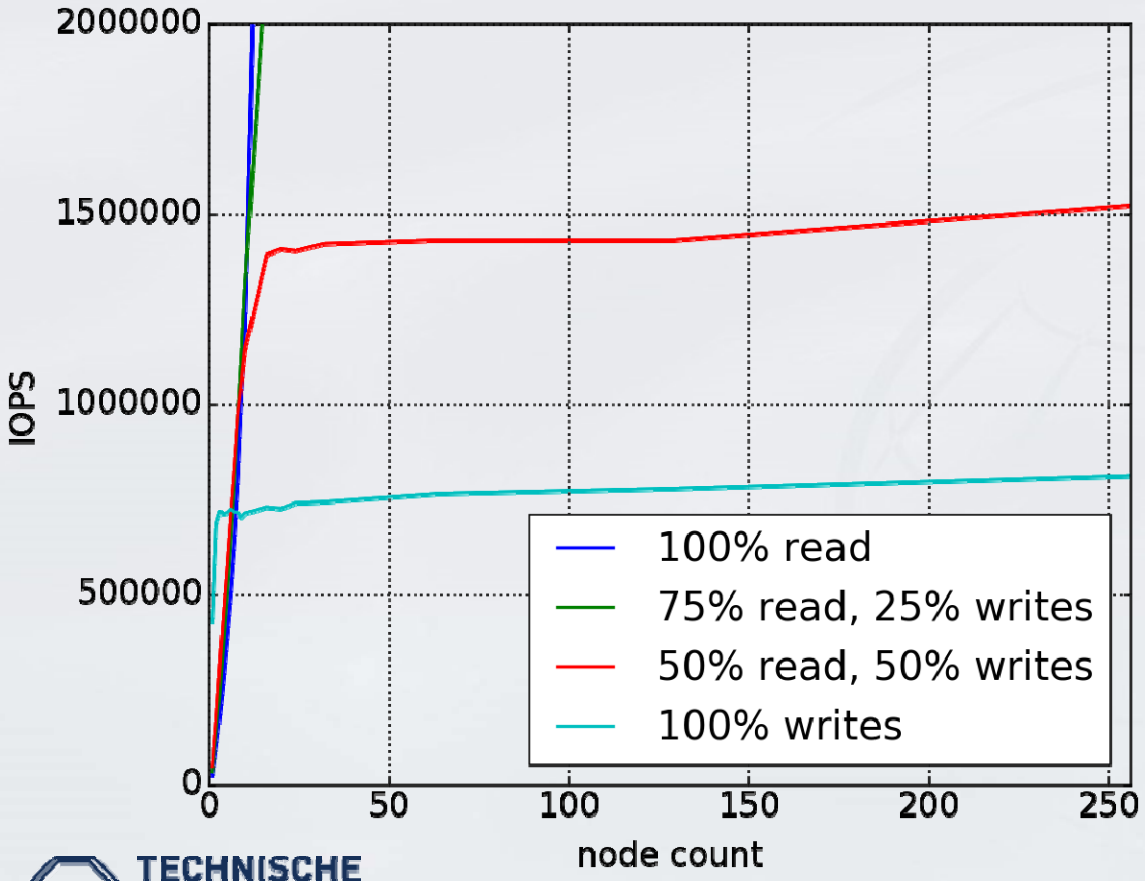
- Different block sizes from 4 KiB to 1 MiB
- Measurement done twice
  - 4K steps
  - Only powers of 2
- Peak IOPS still the same
- Peak bandwidth at 10 GB/s
- There is some pattern in the bandwidth curve

One Node / Ratio between Reads and Writes



- Changing the mixture between reads and writes
- From 0% to 100% writes in steps of 5%
- For 24 processes there is a sweet spot …

Many Nodes / Many LUNs



- Stripe 12 (all files have one stripe on all SSDs)
- Up to 256 nodes (out of ~1500)
- 24 processes/node
- Reads are still cached?
- 1 Mio. IOPS was measured with:
  - 40 TB data
  - 1500 nodes, 24 ppn
  - reads only

What to take home

- Single process can issue about 30.000 IOPS (CPU bound)
- One node can issue > 100.000 write IOPS
- (close to) Peak IOPS of the file system can be reached with only a few nodes
- Performance remains stable as node numbers increase
- Writes appear to be faster as long as the performance capacity of the underlying hardware is not maxed out (writes on most SSDs are generally slower)


QUESTIONS
ANSWERS