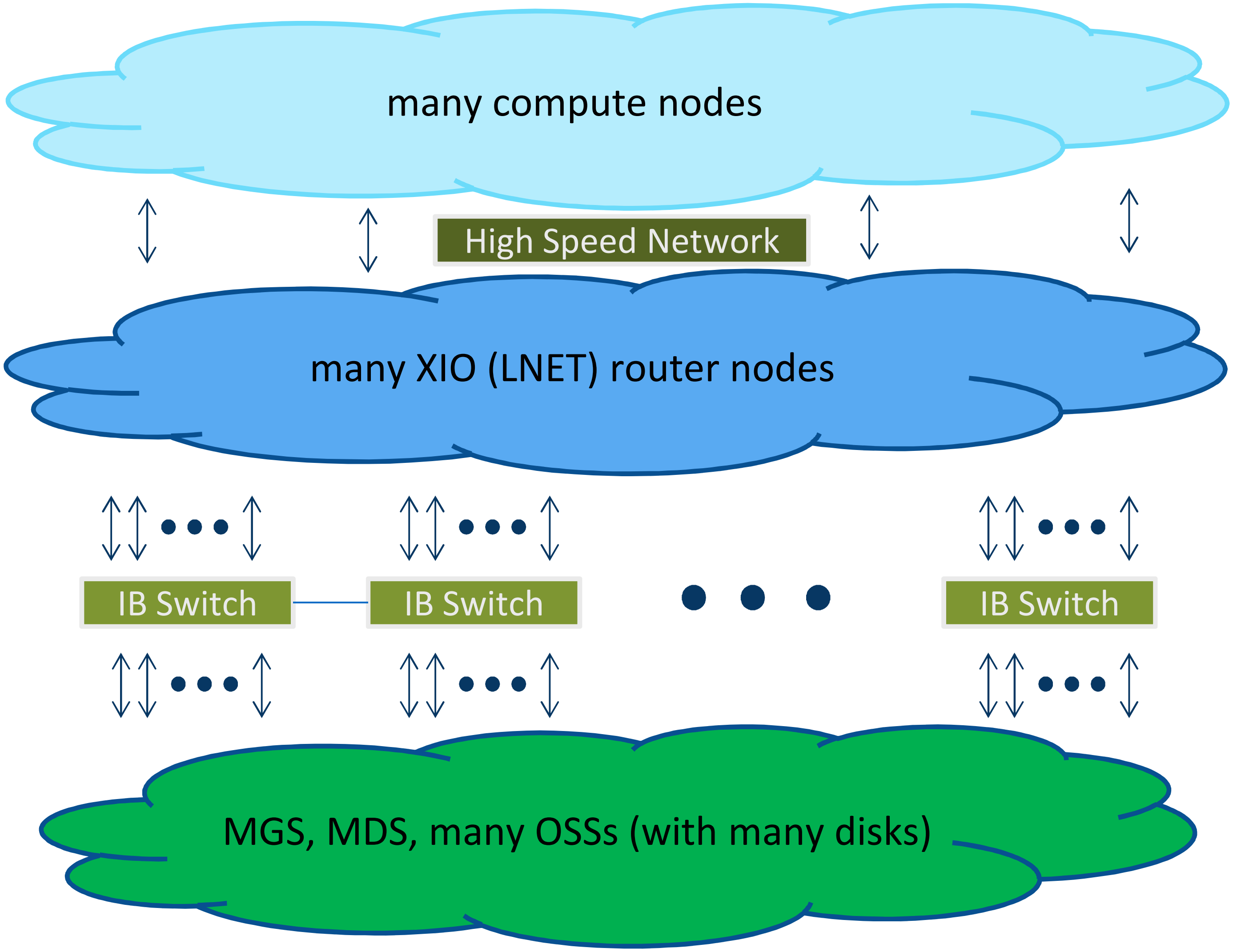
Instrumenting IOR to Diagnose Performance Issues on Lustre File Systems

Doug Petesch
Mark Swan
Cray, Inc.

Agenda

- **Background**
  - Lustre components
  - Measuring I/O performance
  - IOR basics
- **Examples of imperfections**
  - Distribution of files on OSTs
  - OSTs (disk position)
  - OSSs (IB cable connection, failover)
  - LNET router (node failure)

Components of an external Lustre file system


many compute nodes
High Speed Network
many XIO (LNET) router nodes
IB Switch
IB Switch
IB Switch
MGS, MDS, many OSSs (with many disks)

Measuring I/O Performance

$$\text{I/O RATE} = \frac{\text{DATA}}{\text{TIME}}$$

Application vs. File System performance

$$\text{I/O RATE} = \frac{\text{DATA}}{\text{TIME}}$$

Application view:

- Fixed amount of data to move
- Measure time to complete

File System view:

- Run for a fixed time
- Measure data moved

Reasons to use IOR

- Scales from a single thread to thousands of nodes
- Can generate a wide variety of I/O patterns
- Can be run by unprivileged users
- Often specified as official measurement method

- Easy to modify
  - Record time stamp of each transfer
  - Each rank print timings to own file
  - Scripts automatically generate plots with gnuplot

Fixed Data vs. Fixed Time

- **“Fixed data” is default for IOR**
  - Rate determined by slowest file system component
  - Does not keep whole file system busy all the time

- **“Fixed time” IOR options:**

```
# posix file per process, O_DIRECT, 8 MiB records
OPTIONS="-E -B -F -e -g -b 48g -t 8m”
# write for 3 minutes then read for 2 minutes
aprun -n $RANKS IOR $OPTIONS -w -D 180 -k
aprun -n $RANKS IOR $OPTIONS -r -D 120
```

- **Ideally equivalent**
  - But only under perfect conditions

Sample IOR command line and output

```
aprun –n 100 IOR -C -B -F -t 4m -b 4g –k

Summary:
api                 = POSIX
test filename       = testdir/IOR_POSIX
access              = file-per-process
pattern             = segmented (1 segment)
ordering in a file = sequential offsets
ordering inter file=constant task offsets=1
clients             = 100 (4 per node)
repetitions         = 1
xfersize            = 4 MiB
blocksize           = 4 GiB
aggregate filesize = 400 GiB

Max Write: 6015.63 MiB/sec (6307.84 MB/sec)
Max Read:  3046.21 MiB/sec (3194.19 MB/sec)
```

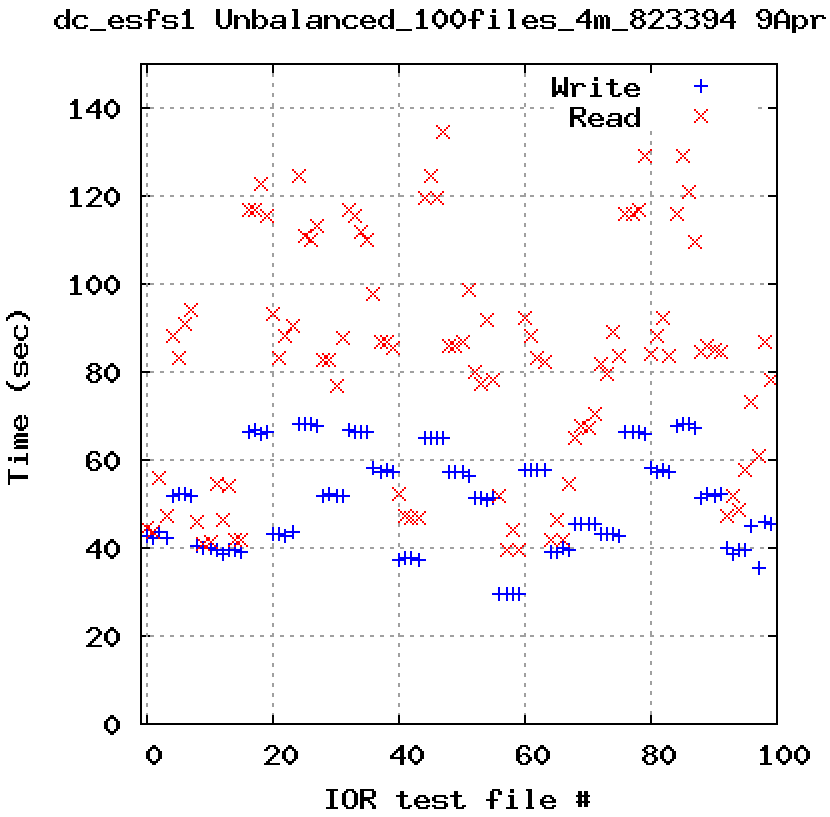
Output from IOR -vvv (verbose=3)

```
Test 0: Iter=0, Task=0, Time=1365558598.489247, write open start
Test 0: Iter=0, Task=0, Time=1365558598.489978, write open stop
Test 0: Iter=0, Task=0, Time=1365558598.496538, write start
Test 0: Iter=0, Task=0, Time=1365558641.157996, write stop
Test 0: Iter=0, Task=0, Time=1365558666.575858, write close start
Test 0: Iter=0, Task=0, Time=1365558666.576329, write close stop
Test 0: Iter=0, Task=0, Time=1365558666.597461, read open start
Test 0: Iter=0, Task=0, Time=1365558666.597855, read open stop
Test 0: Iter=0, Task=0, Time=1365558666.599108, read start
Test 0: Iter=0, Task=0, Time=1365558754.811135, read stop
Test 0: Iter=0, Task=0, Time=1365558801.056288, read close start
Test 0: Iter=0, Task=0, Time=1365558801.056823, read close stop
```

Plot time to write and read each file

```
aprun –n 100 IOR -C -B -F -t 4m -b 4g –k
```

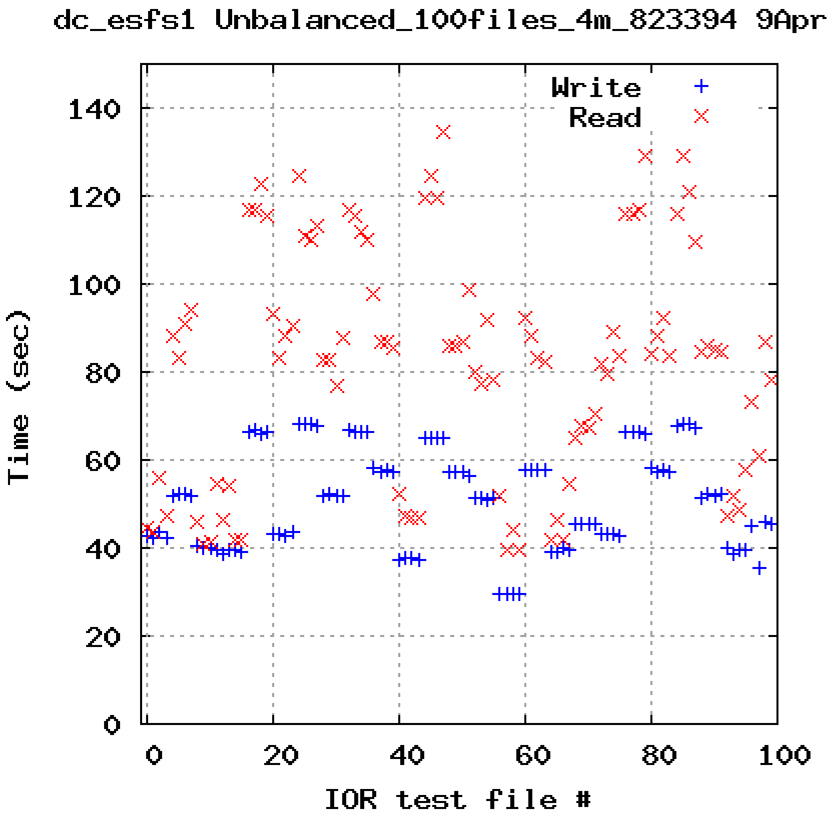
NetApp E5400 file system with 18 OSTs

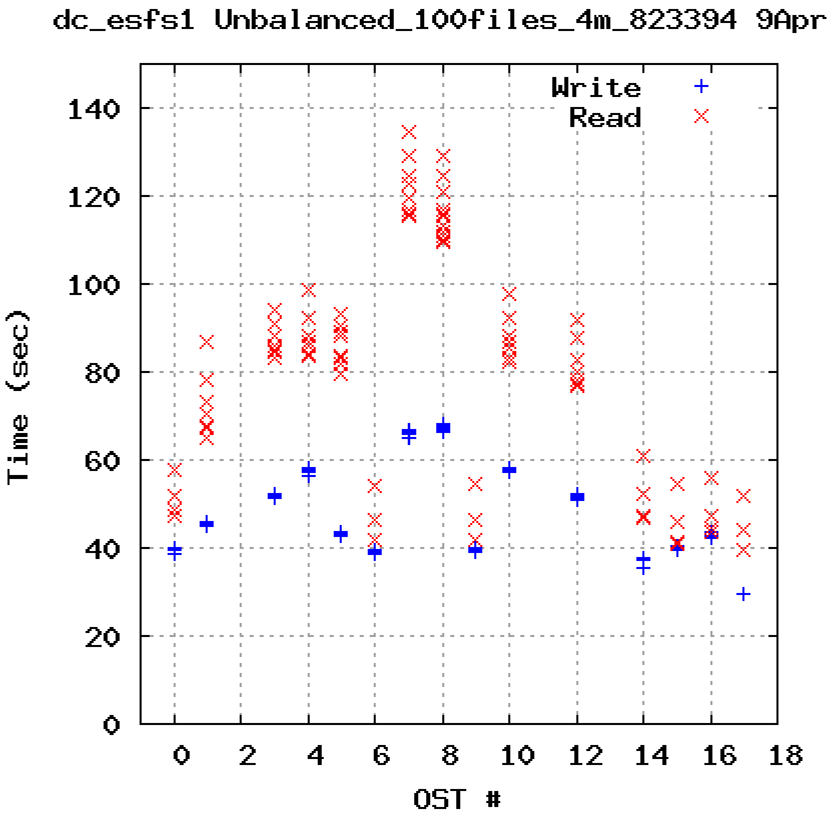


Files not spread evenly across OSTs

```
aprun –n 100 IOR -C -B -F -t 4m -b 4g –k
```

NetApp E5400 file system with 18 OSTs



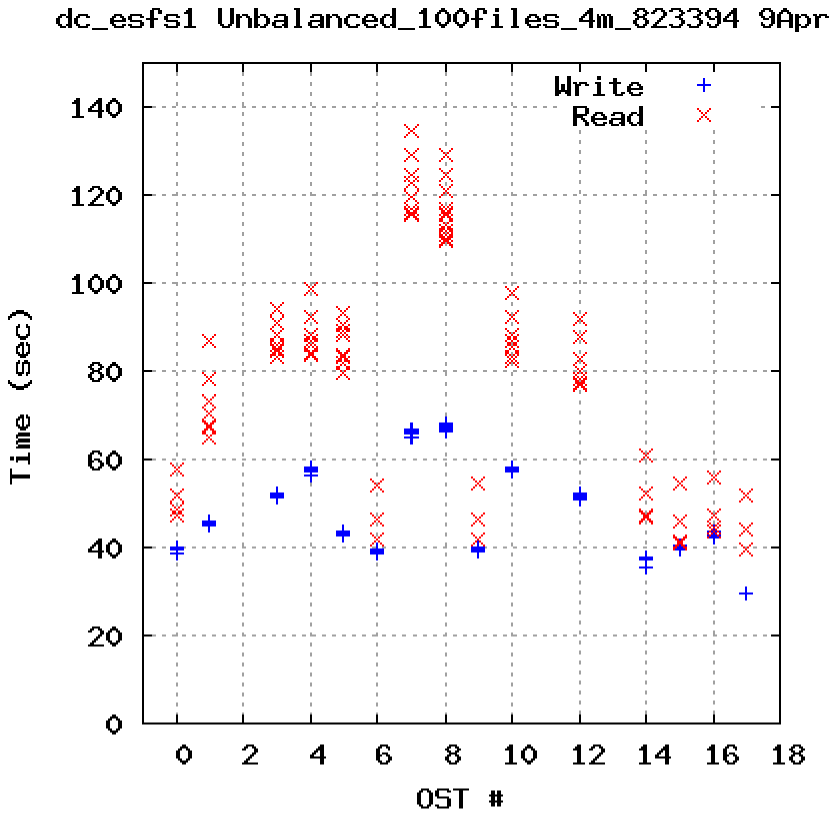


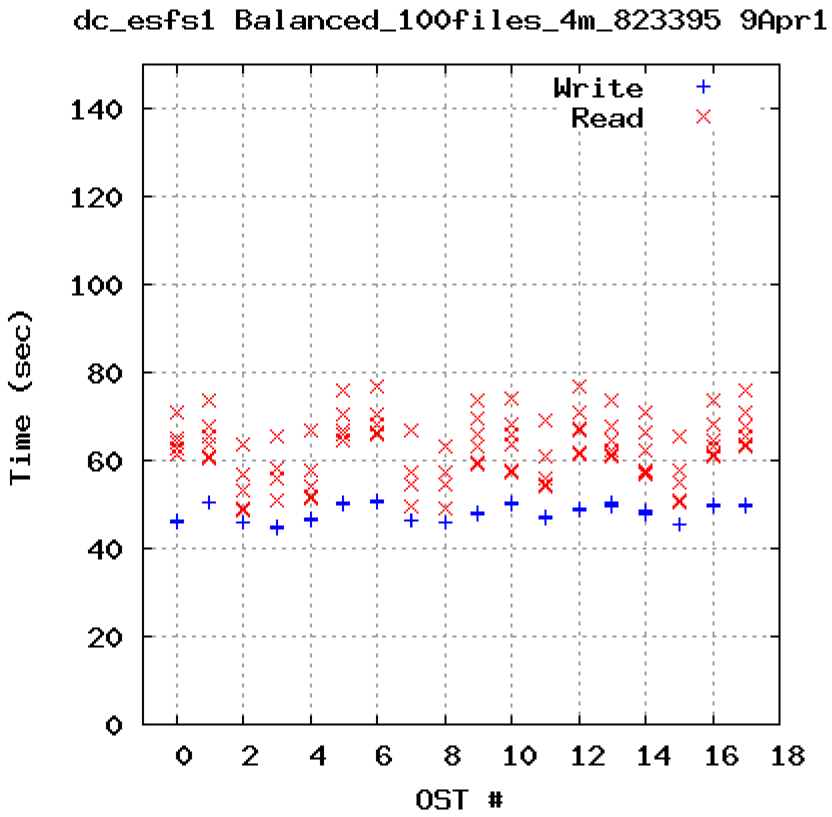
Better Balance = Better Performance

Still 100 files on 18 OSTs

Write: 6308 MB/sec
Read: 3194 MB/sec

Write: 8419 MB/sec
Read: 5594 MB/sec

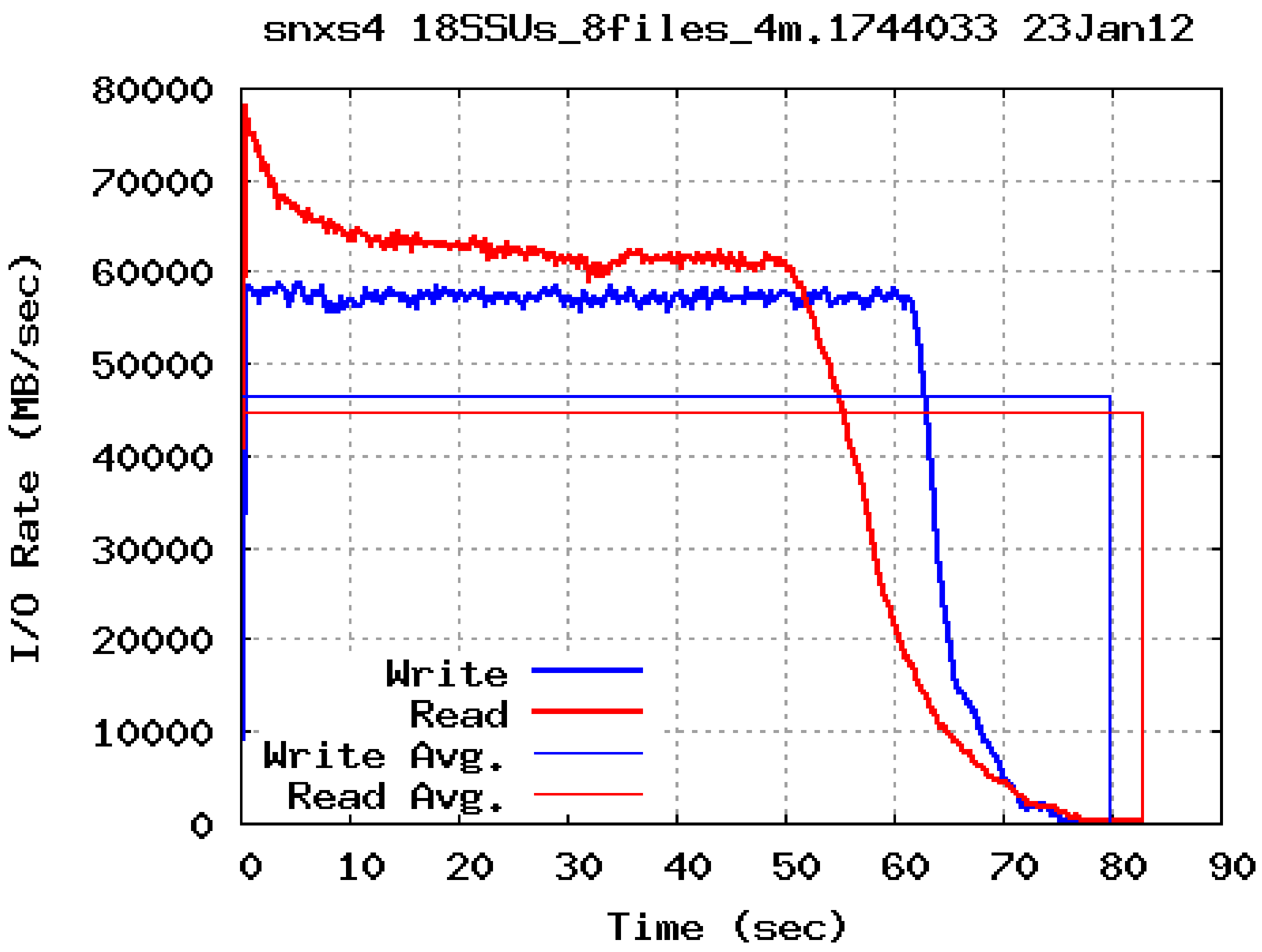




qos_threshold_rr=100

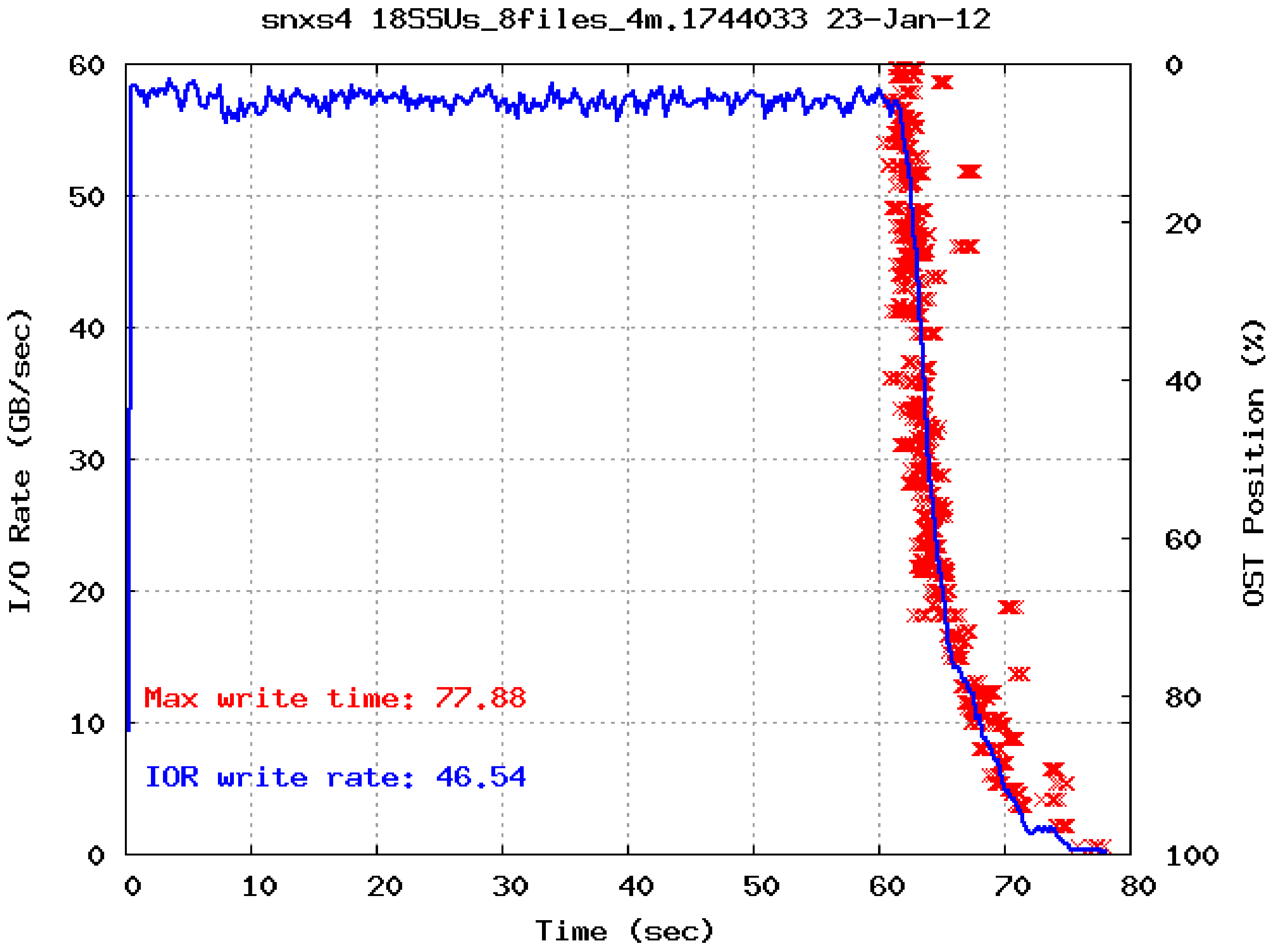
Plot rates over time

Cray Sonexion 1300, 18 SSUs, 144 OSTs, 1152 (3 GiB) files



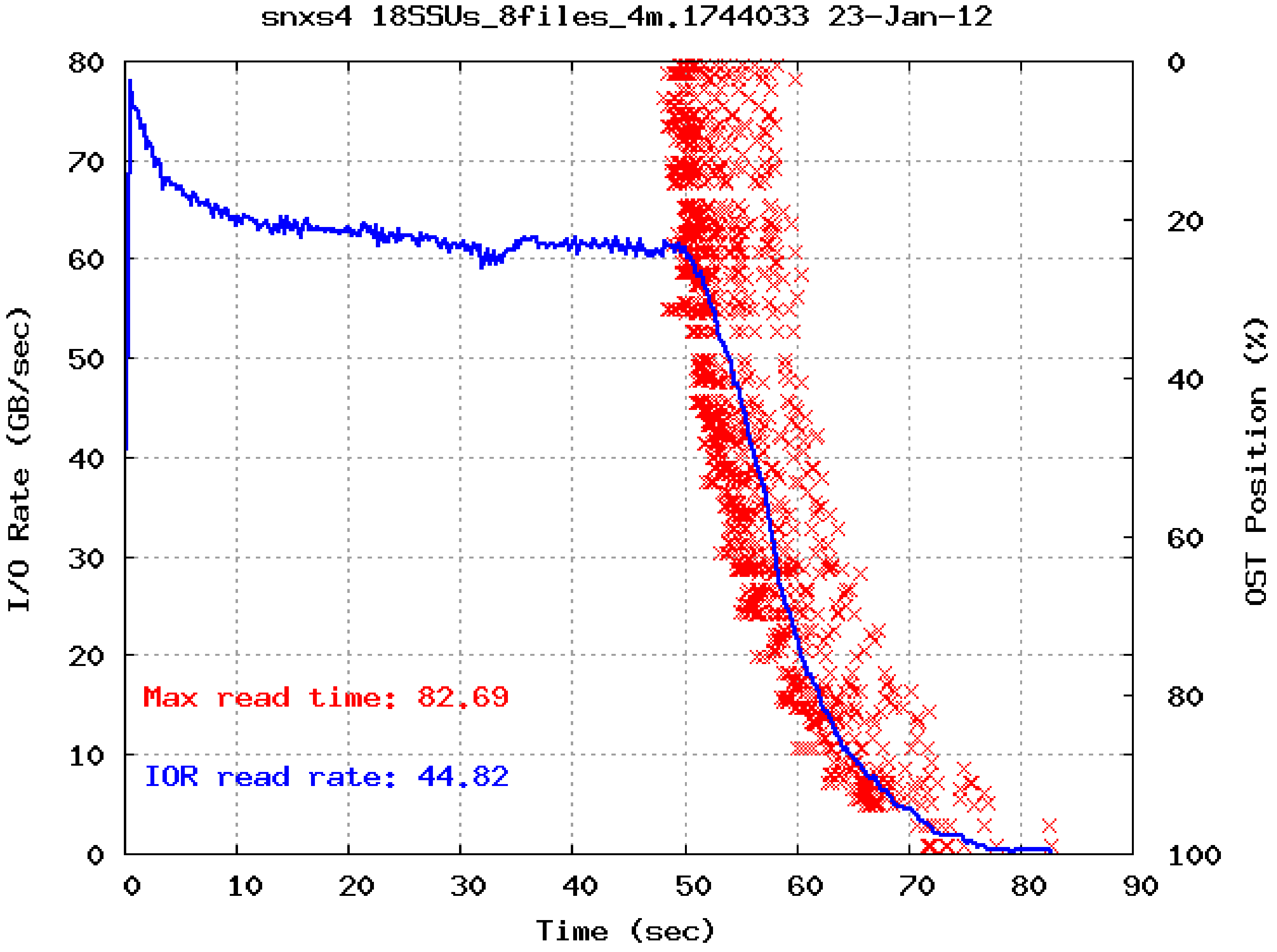
Write rate vs. disk position

Cray Sonexion 1300, 18 SSUs, 144 OSTs, 1152 (3 GiB) files

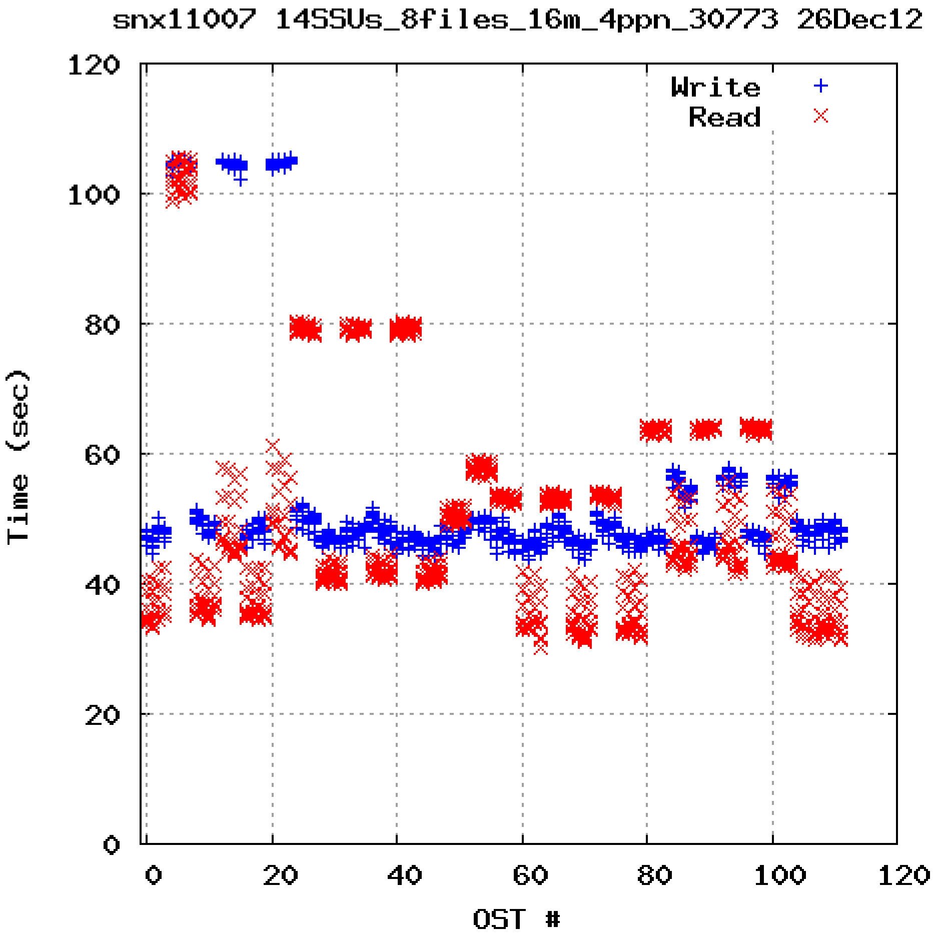


Read rate vs. disk position

Cray Sonexion 1300, 18 SSUs, 144 OSTs, 1152 (3 GiB) files



One IB link at SDR speed



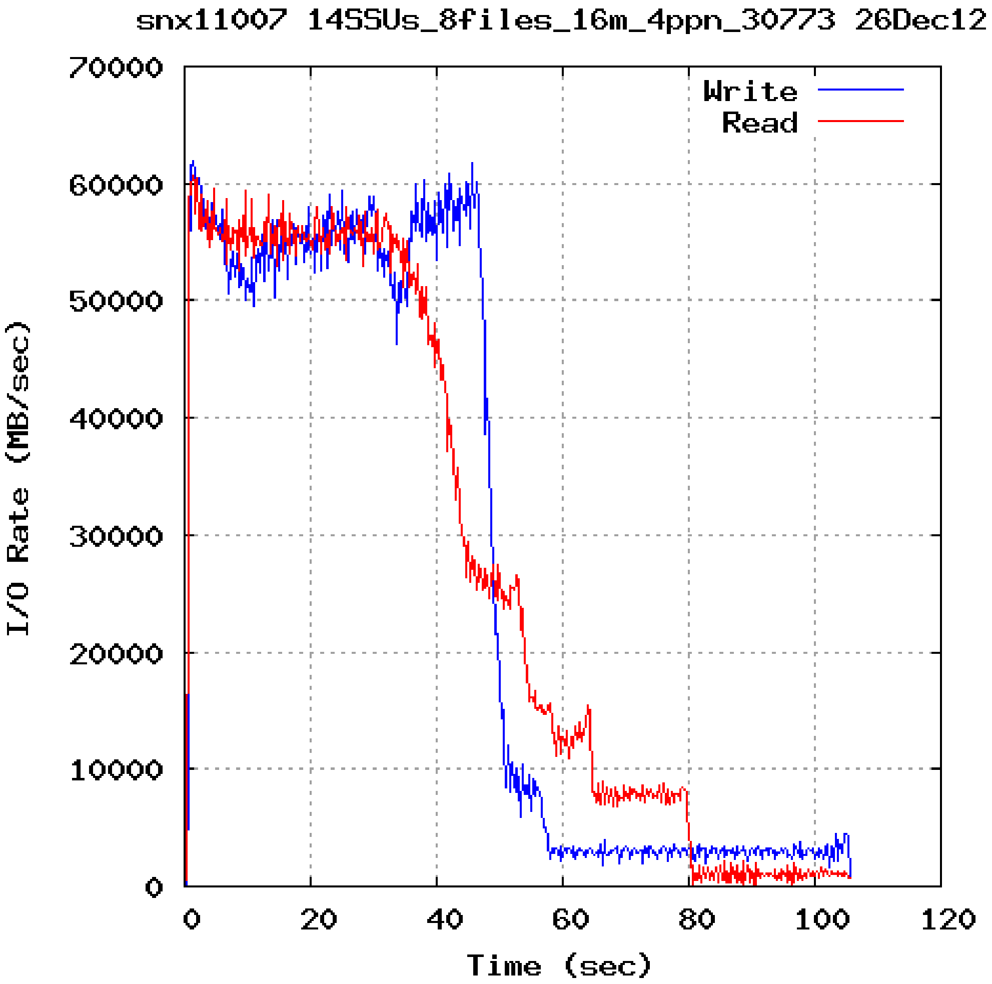
48 cabinet Cray XE
14 SSU Sonexion 1600
112 OSTs, 896 files
224 nodes, ~5% of total

1 OSS cable at SDR rate:
(96 GiB)/(103 sec) = 1 GB/sec

- Affects writes for FGR group
- Affects reads just for 1 OSS

Other FGR group effects due to job placement in torus.

One IB link at SDR speed

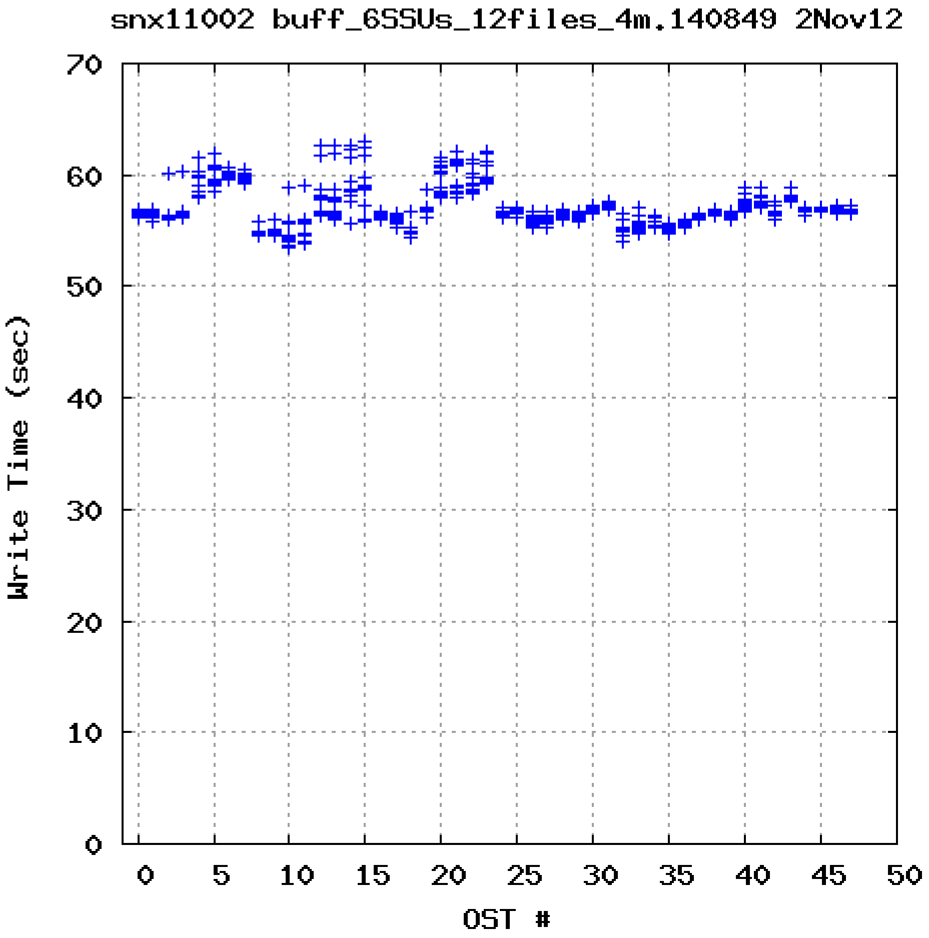


48 cabinet Cray XE
14 SSU Sonexion 1600
112 OSTs, 896 files
224 nodes, ~5% of total

1 OSS cable at SDR rate:
(96 GiB)/(103 sec) = 1 GB/sec

- Affects writes for FGR group
- Affects reads just for 1 OSS

Failed LNET router



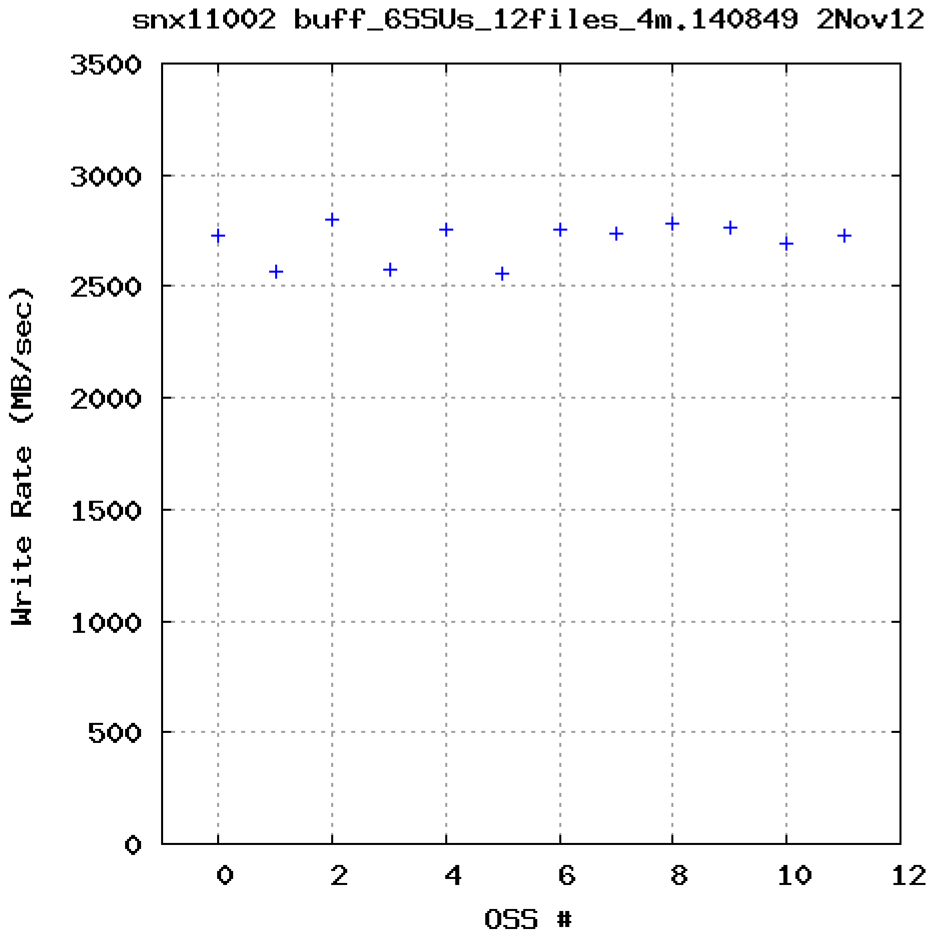
Cray XE
6 SSUs of Sonexion 1600
48 OSTs, 12 OSSs
4:3 router:OSS ratio

XE router: 2.6 GB/sec
OSS potential: 3 GB/sec

Group with 3 routers is slower

Time for fixed data

Failed LNET router



Cray XE
6 SSUs of Sonexion 1600
48 OSTs, 12 OSSs
4:3 router:OSS ratio

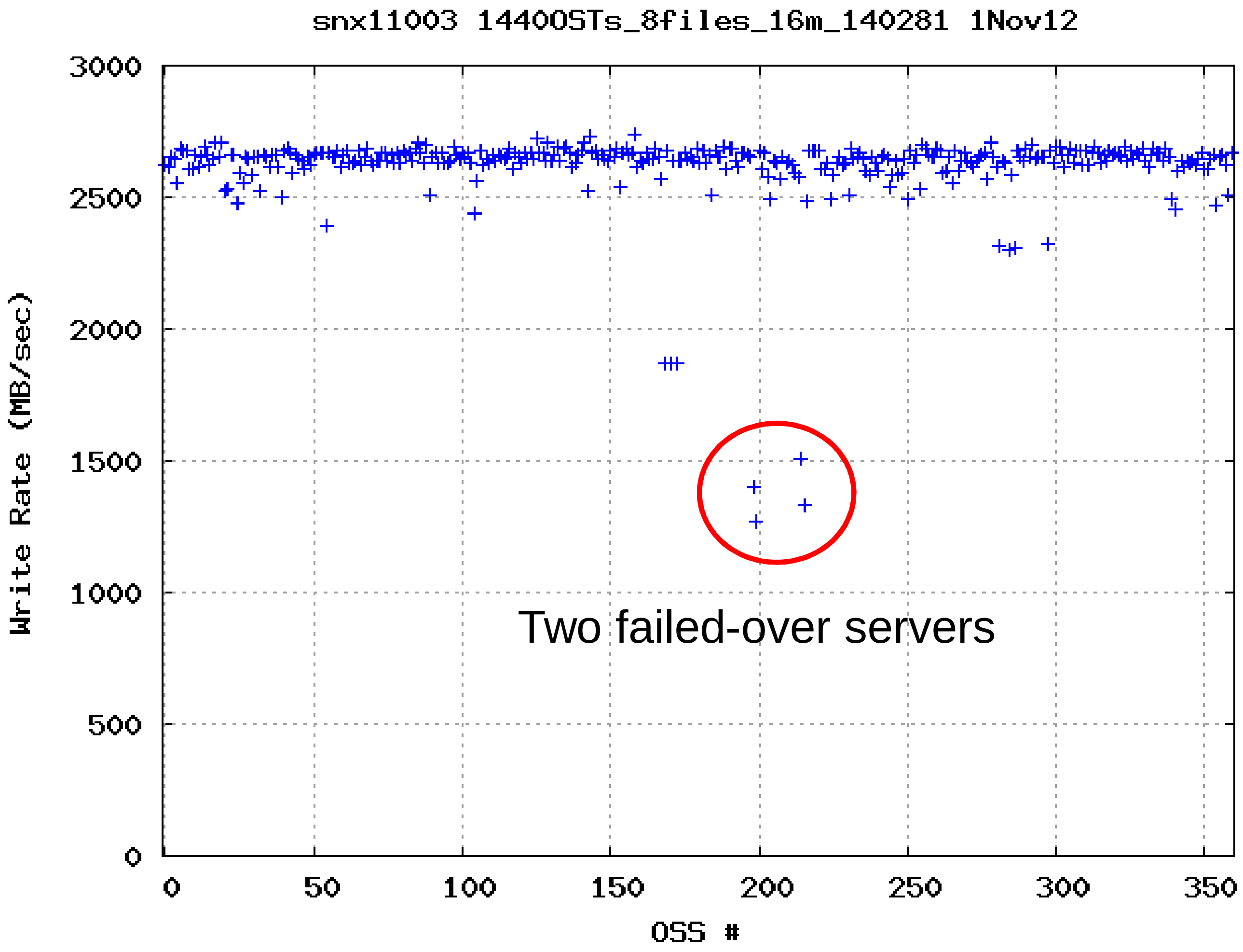
XE router: 2.6 GB/sec
OSS potential: 3 GB/sec

Group with 3 routers is limited to 2.6 GB/sec per OSS

Rate over fixed time

Server write rate during fail-over

Cray Sonexion 1600, 180 SSUs, 360 OSSs, fixed time



Thank You

Questions?