PAUL SCHERRER INSTITUT

PSI

WIR SCHAFFEN WISSEN – HEUTE FÜR MORGEN

Alvise Dorigo :: Science IT :: Paul Scherrer Institut

AFM Experience @PSI

Spectrum Scale User Group - London, 8-9th May 2019

PSI

“The Paul Scherrer Institute, PSI, is the largest research institute for natural and engineering sciences within Switzerland. We perform world-class research in three main subject areas: Matter and Material; Energy and the Environment; and Human Health. By conducting fundamental and applied research, we work on long-term solutions for major challenges facing society, industry and science.”

PSI operates various Large Scale Facilities:

- SLS: Swiss Light Source synchrotron
- SINQ: Spallation Neutron Source
- SμS: Swiss Muon Source
- SwissFEL: Swiss Free Electron Laser

Data intensive research

- SwissFEL scientists produce **~1 PBytes** of data (images) per **year**
  - projections foresee a **doubling**

- These data needs to be promptly (ideally in real time) transferred from online (where they are produced) to offline storage (where they stay for long time)

- Online storage is like a fast-access read/write **CACHE**

- Offline is a long-term storage used for data analysis and to store its artifacts

- We give the possibility of preliminary analysis of data in the online (computing cluster is connected to the CACHE)

The storage+AFM @SwissFEL facility

Lenovo DSS-G220, 1.2PB net

- 1x xCAT node + 2x I/O servers, 2x FDR-56G connections per node
- Spectrum Scale + GNR 4.2.3-7 (Lenovo dss-g-2.0a)
- RH 7.4, OFED 4.2
- 2x Filesystem with 8M-blocksize (called 'RAW’ and 'RES'), 8+2p
- Max write speed (aggregated): 9-9.5 GB/s (writing files ~100GB)

2x AFM gateways

- HPE ProLiant DL380 Gen9 + HPE ProLiant DL380 Gen10
- 256GB RAM each node
- 2 x E5-2687Wv4 @3.00GHz (24 cores), HT OFF
- 2 x Gold 6130 @2.10GHz (32 cores), HT OFF
- 1 FDR InfiniBand connection 56G each node
- RH 7.6, OFED 4.5
- GPFS 5.0.2.3 + *efix4 (*issue: *uid not correctly transferred to Home)*
- NO Protocol, only gateway+perfmon

NFS export node

Same hardware as AFM/Gen9, RH7.5 OFED 4.4

Data workflow overview @SwissFEL

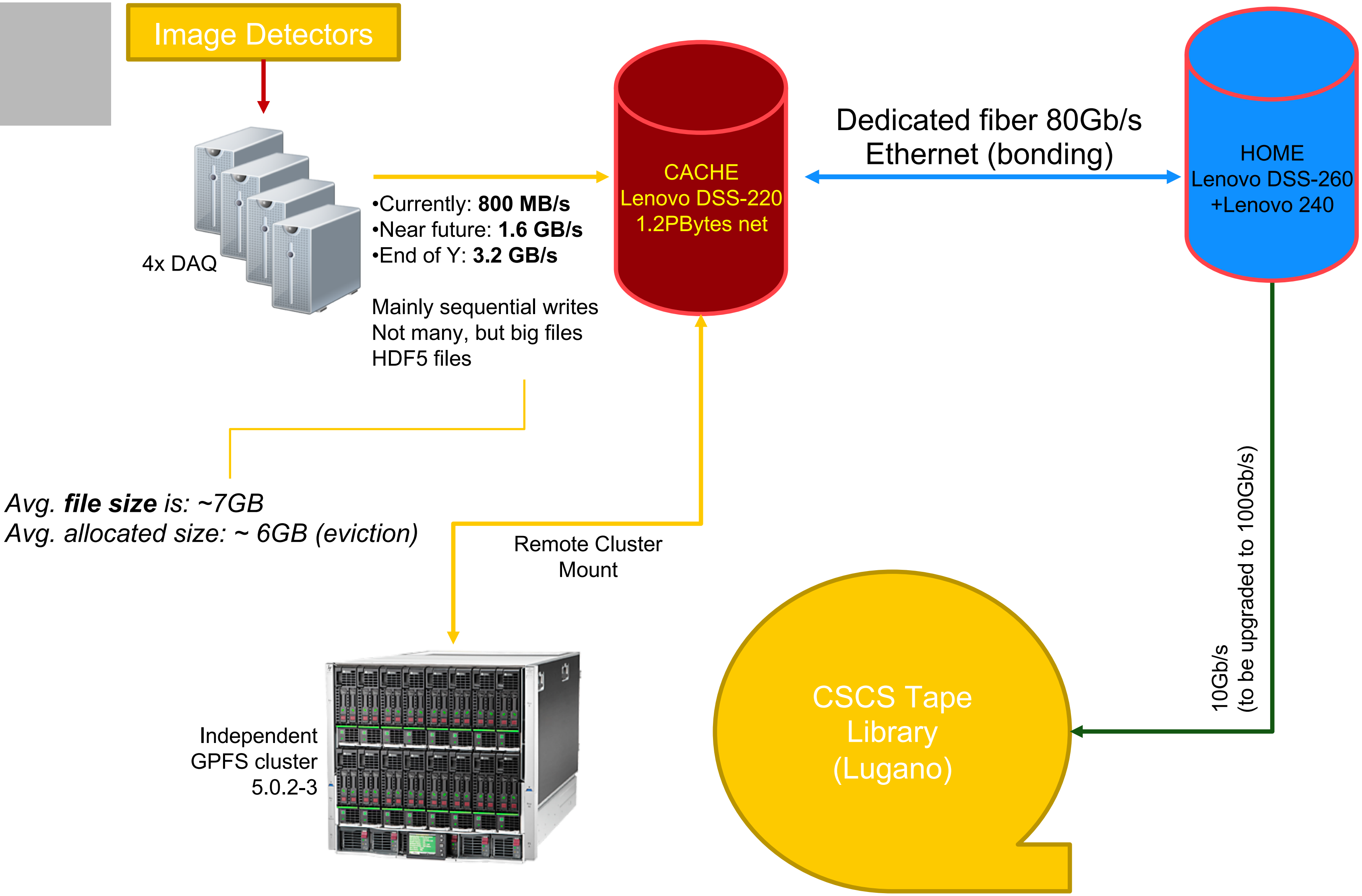


Image Detectors

4x DAQ

- Currently: **800 MB/s**
- Near future: **1.6 GB/s**
- End of Y: **3.2 GB/s**

Mainly sequential writes
Not many, but big files
HDF5 files

*Avg. **file size** is: ~7GB*
Avg. allocated size: ~ 6GB (eviction)

CACHE
Lenovo DSS-220
1.2PBytes net

Dedicated fiber 80Gb/s
Ethernet (bonding)

HOME
Lenovo DSS-260
+Lenovo 240

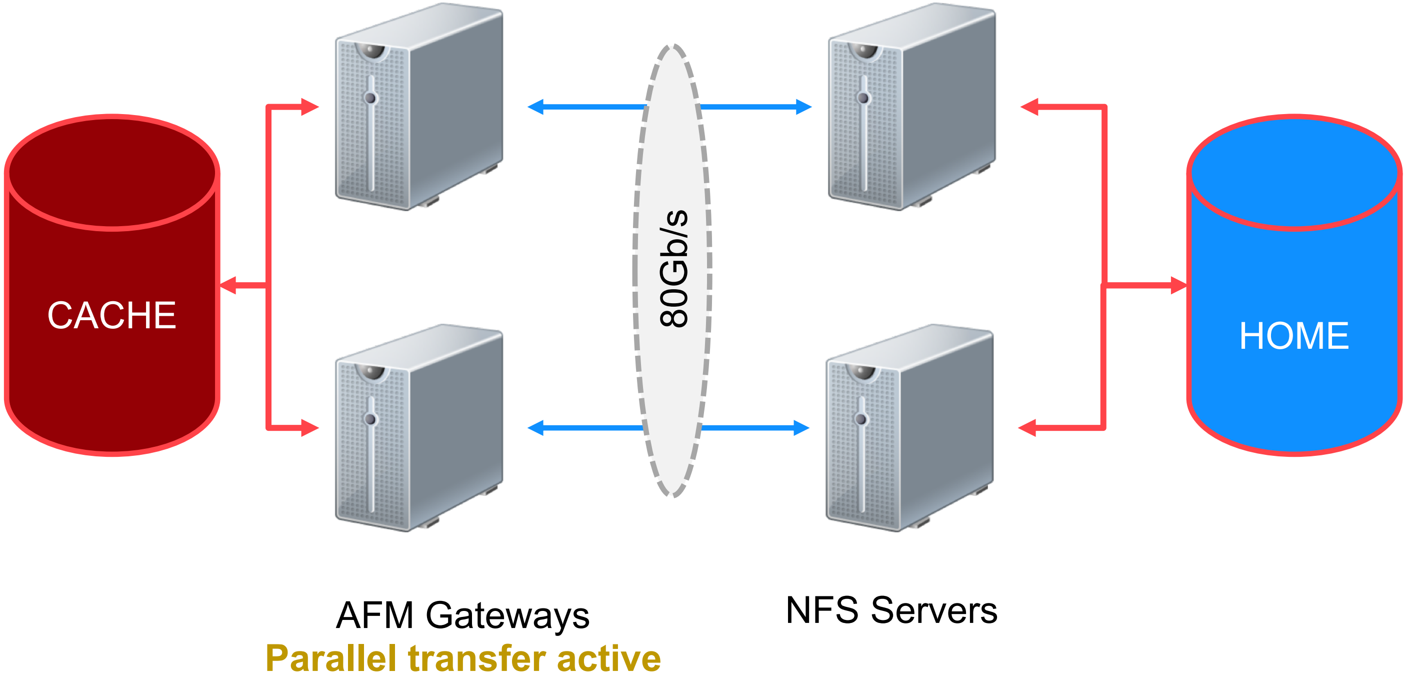
Remote Cluster
Mount

Independent
GPFS cluster
5.0.2-3

CSCS Tape
Library
(Lugano)

10Gb/s
(to be upgraded to 100Gb/s)

Gateway/NFS

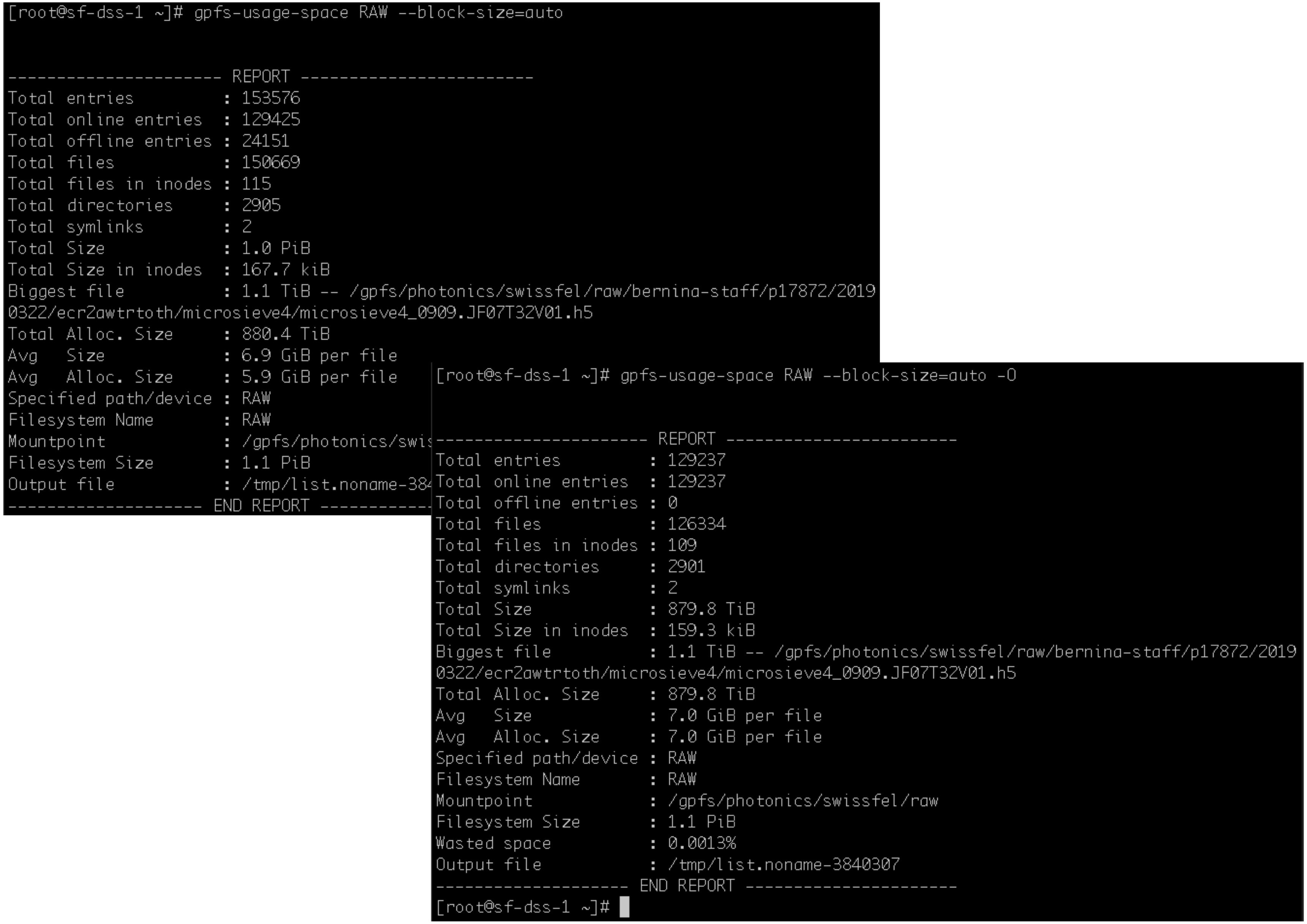


NFS was chosen instead of native protocol
because of the long distance (~1 Km)

AFM Mode

- **Single Writer**
- Home (part of Offline storage) is a R/O backup copy (users can read to produce analysis's artifacts)
- Cache has **eviction-enabled**, to virtually extend its real fast-access space (1.2PB) to almost 2.5 PB…
  - to be expanded with a new Lenovo-240 to 5 PB
- Eviction is automatic and based on **filesets-level quota**
- Possible evaluation in future of eviction by mean of callbacks based on entire FS occupancy (already implemented @ETHZ)

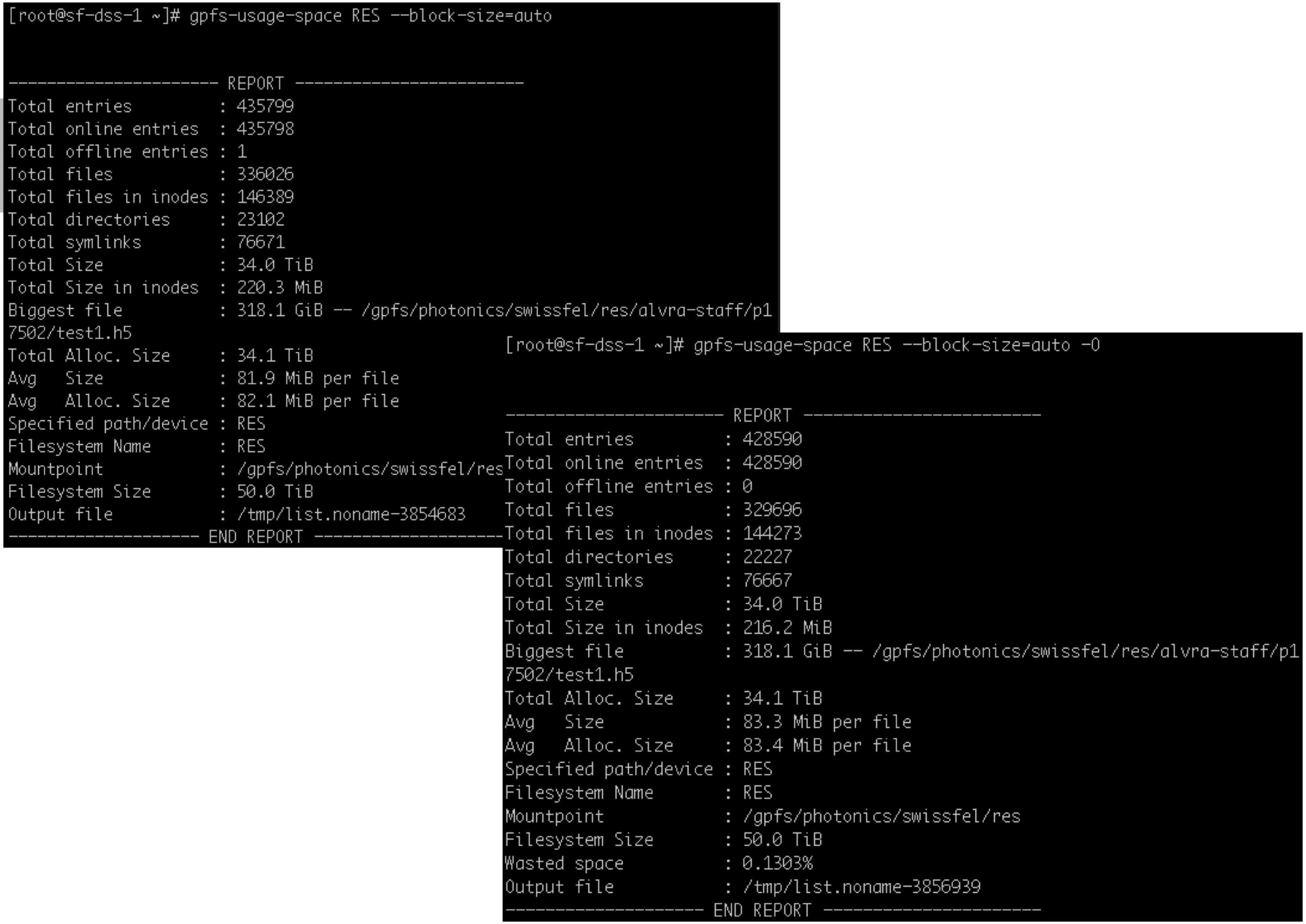
RAW (DAQ) Filesystem stats



```
[root@sf-dss-1 ~]# gpfs-usage-space RAW --block-size=auto


---------------------- REPORT ------------------------
Total entries          : 153576
Total online entries   : 129425
Total offline entries  : 24151
Total files            : 150669
Total files in inodes  : 115
Total directories      : 2905
Total symlinks         : 2
Total Size             : 1.0 PiB
Total Size in inodes   : 167.7 kiB
Biggest file           : 1.1 TiB -- /gpfs/photonics/swissfel/raw/bernina-staff/p17872/2019
0322/ecr2awtrtoth/microsieve4/microsieve4_0909.JF07T32V01.h5
Total Alloc. Size      : 880.4 TiB
Avg   Size             : 6.9 GiB per file
Avg   Alloc. Size      : 5.9 GiB per file
Specified path/device  : RAW
Filesystem Name        : RAW
Mountpoint             : /gpfs/photonics/swis
Filesystem Size        : 1.1 PiB
Output file            : /tmp/list.noname-384
-------------------- END REPORT -----------
```

```
[root@sf-dss-1 ~]# gpfs-usage-space RAW --block-size=auto -O


---------------------- REPORT ------------------------
Total entries          : 129237
Total online entries   : 129237
Total offline entries  : 0
Total files            : 126334
Total files in inodes  : 109
Total directories      : 2901
Total symlinks         : 2
Total Size             : 879.8 TiB
Total Size in inodes   : 159.3 kiB
Biggest file           : 1.1 TiB -- /gpfs/photonics/swissfel/raw/bernina-staff/p17872/2019
0322/ecr2awtrtoth/microsieve4/microsieve4_0909.JF07T32V01.h5
Total Alloc. Size      : 879.8 TiB
Avg   Size             : 7.0 GiB per file
Avg   Alloc. Size      : 7.0 GiB per file
Specified path/device  : RAW
Filesystem Name        : RAW
Mountpoint             : /gpfs/photonics/swissfel/raw
Filesystem Size        : 1.1 PiB
Wasted space           : 0.0013%
Output file            : /tmp/list.noname-3840307
-------------------- END REPORT ----------------------
[root@sf-dss-1 ~]# █
```

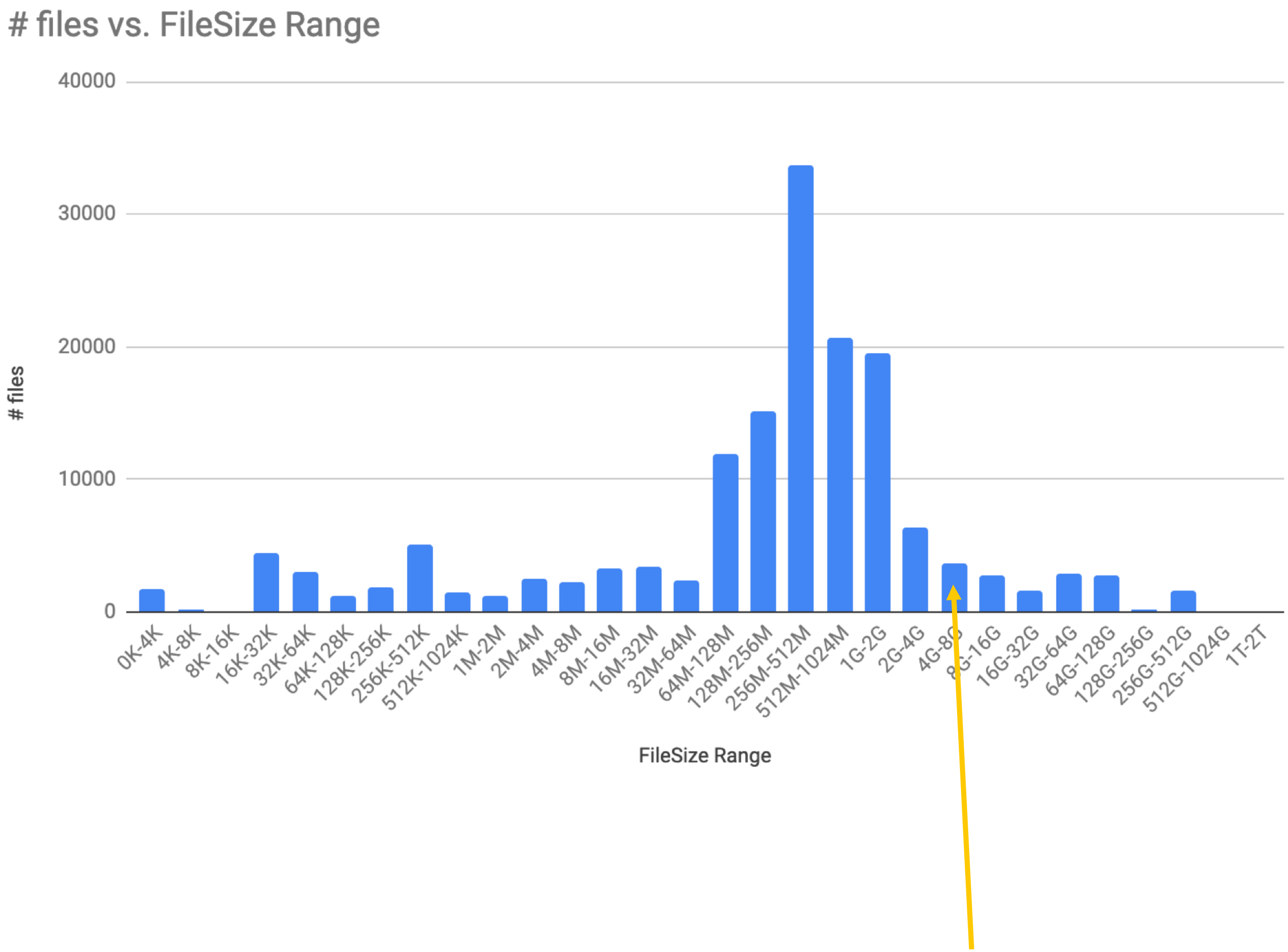
RES (RESults) Filesystem stats



```
[root@sf-dss-1 ~]# gpfs-usage-space RES --block-size=auto



---------------------- REPORT ------------------------
Total entries          : 435799
Total online entries   : 435798
Total offline entries  : 1
Total files            : 336026
Total files in inodes  : 146389
Total directories      : 23102
Total symlinks         : 76671
Total Size             : 34.0 TiB
Total Size in inodes   : 220.3 MiB
Biggest file           : 318.1 GiB -- /gpfs/photonics/swissfel/res/alvra-staff/p1
7502/test1.h5
Total Alloc. Size      : 34.1 TiB
Avg   Size             : 81.9 MiB per file
Avg   Alloc. Size      : 82.1 MiB per file
Specified path/device  : RES
Filesystem Name        : RES
Mountpoint             : /gpfs/photonics/swissfel/res
Filesystem Size        : 50.0 TiB
Output file            : /tmp/list.noname-3854683
-------------------- END REPORT --------------------
```

```
[root@sf-dss-1 ~]# gpfs-usage-space RES --block-size=auto -O



---------------------- REPORT ------------------------
Total entries          : 428590
Total online entries   : 428590
Total offline entries  : 0
Total files            : 329696
Total files in inodes  : 144273
Total directories      : 22227
Total symlinks         : 76667
Total Size             : 34.0 TiB
Total Size in inodes   : 216.2 MiB
Biggest file           : 318.1 GiB -- /gpfs/photonics/swissfel/res/alvra-staff/p1
7502/test1.h5
Total Alloc. Size      : 34.1 TiB
Avg   Size             : 83.3 MiB per file
Avg   Alloc. Size      : 83.4 MiB per file
Specified path/device  : RES
Filesystem Name        : RES
Mountpoint             : /gpfs/photonics/swissfel/res
Filesystem Size        : 50.0 TiB
Wasted space           : 0.1303%
Output file            : /tmp/list.noname-3856939
-------------------- END REPORT ----------------------
```

Filesize distribution (RAW)



Avg. filesize

NFS fine-tuning

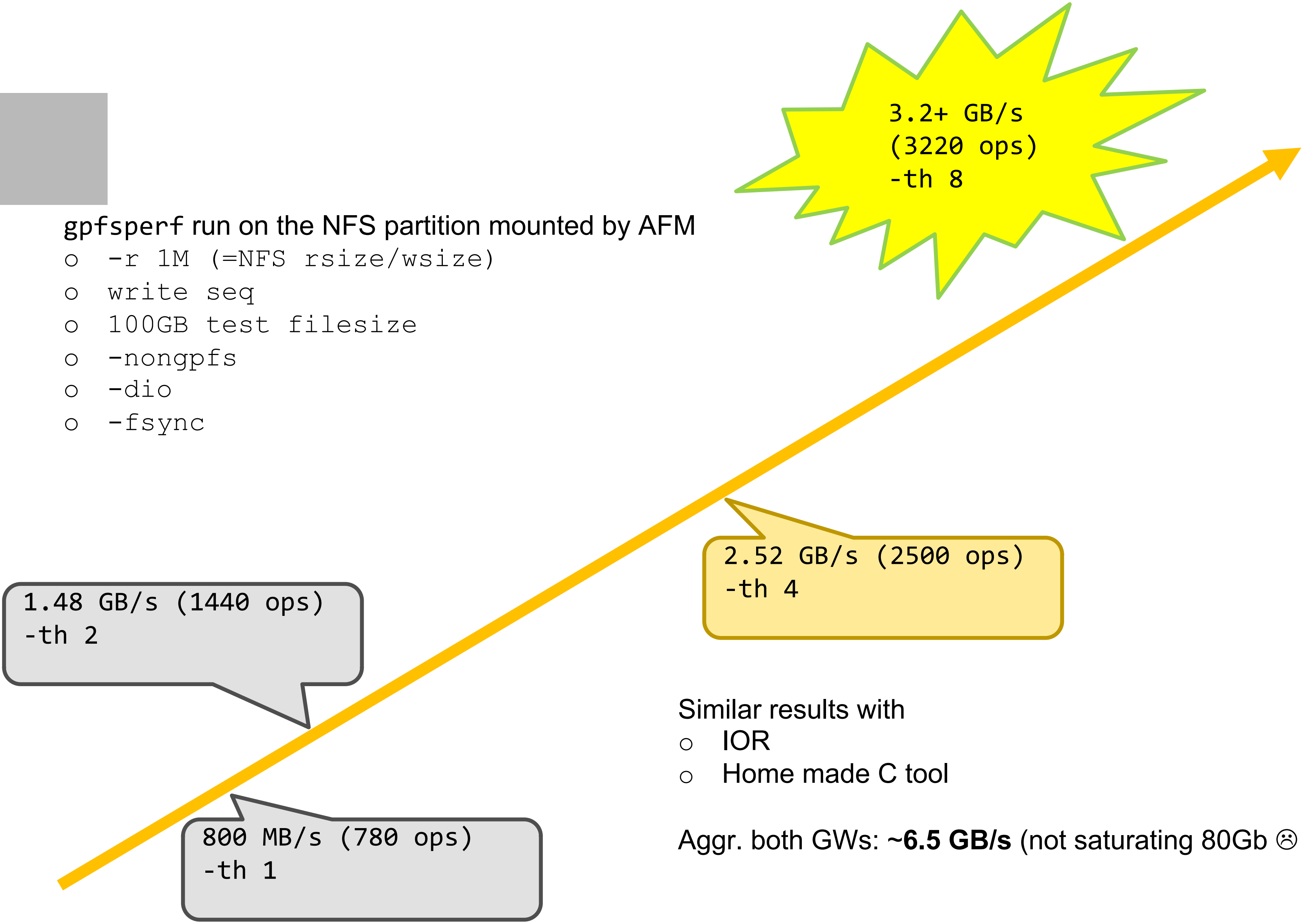
```
sysctl::values:
 net.core.rmem_max:
    value: '536870912'
 net.core.wmem_max:
    value: '536870912'
 net.core.rmem_default:
    value: '167772160'
 net.core.wmem_default:
    value: '167772160'
 net.core.optmem_max:
    value: '167772160'
 net.core.netdev_max_backlog:
    value: '250000'
 net.ipv4.tcp_rmem:
    value: '4096 87380 268435456'
 net.ipv4.tcp_wmem:
    value: '4096 87380 268435456'
 net.ipv4.tcp_mem:
    value: '4096 87380 268435456'
 net.core.netdev_budget:
    value: '600'
 net.core.netdev_max_backlog:
    value: '250000'
 net.ipv4.tcp_congestion_control:
    value: 'htcp'
```

```
sysctl::values:
 net.ipv4.tcp_mtu_probing:
    value: '1'
 net.ipv4.tcp_low_latency:
    value: '0'
 net.ipv4.tcp_sack:
    value: '1'
 net.ipv4.tcp_no_metrics_save:
    value: '1'
 net.ipv4.tcp_timestamps:
    value: '0'
 net.ipv4.tcp_slow_start_after_idle:
    value: '0'
 net.core.somaxconn:
  value: '1024'
 vm.dirty_background_bytes:
    value: '1073741824'
 vm.dirty_bytes:
    value: '2147483648'
 vm.dirty_expire_centisecs:
    value: '200'
 vm.dirty_writeback_centisecs:
    value: '400'
 sunrpc.tcp_slot_table_entries:
    value: '64'
```

NFS-pure performance (single GW node)

gpfsperf run on the NFS partition mounted by AFM

- -r 1M (=NFS rsize/wsize)
- write seq
- 100GB test filesize
- -nongpfs
- -dio
- -fsync



800 MB/s (780 ops)
-th 1

1.48 GB/s (1440 ops)
-th 2

2.52 GB/s (2500 ops)
-th 4

3.2+ GB/s
(3220 ops)
-th 8

Similar results with

- IOR
- Home made C tool

Aggr. both GWs: ~**6.5 GB/s** (not saturating 80Gb ☹)

AFM Relevant Parameters (1/3)

- afmMaxWorkerThreads 1024
- afmParallelReadThreshold 1024 (unit is MB)
- afmParallelWriteThreshold 1024 (unit is MB)
- afmParallelWriteChunkSize 128M
- afmParallelReadChunkSize 128M
- afmNumReadThreads 24
- afmNumWriteThreads 24
- afmNumFlushThreads 32
- afmHardMemThreshold 32G

AFM Relevant Parameters (2/3)

`afmDIO = 2`

"*AFM uses Direct I/O writing on NFS mounted partitions*"

Avoids high "pressure" on NFS client

Avoids saturation of physical RAM (OS's cache)

Avoids Gateways' Load raising to 1000 !

Allows to reach higher and more stable throughput

~~afmDIO=0~~ eventually hangs the entire cluster

- FS un-accessible
- `mmfsd` 100% CPU on gateway node
- System useless for many minutes

AFM Relevant Parameters (3/3)

afmMaxWriteMergeLen

It helps to "*coalesce data to be sent*" to Home

afmAsyncDelay

It helps to "*replace multiple writes to the home cluster with a single write containing the latest data*".

AFM Parameters Tuning

I've not got any theory/recipe but, according to my experiments (and my interpretation of the documentation), my best guess is:

afmMaxWriteMergeLen ~ avg file size

&&

afmAsyncDelay = afmMaxWriteMergeLen / avg_write_speed

leading to:

- Low Load (Load1=25, 24 cores), 25Khz context switch, 3.5% CPU … per GW
- Low throughput jitter
- Throughput to Home ~ write speed on cache
  - **Data stored in Home in real-time !**

Observed limit

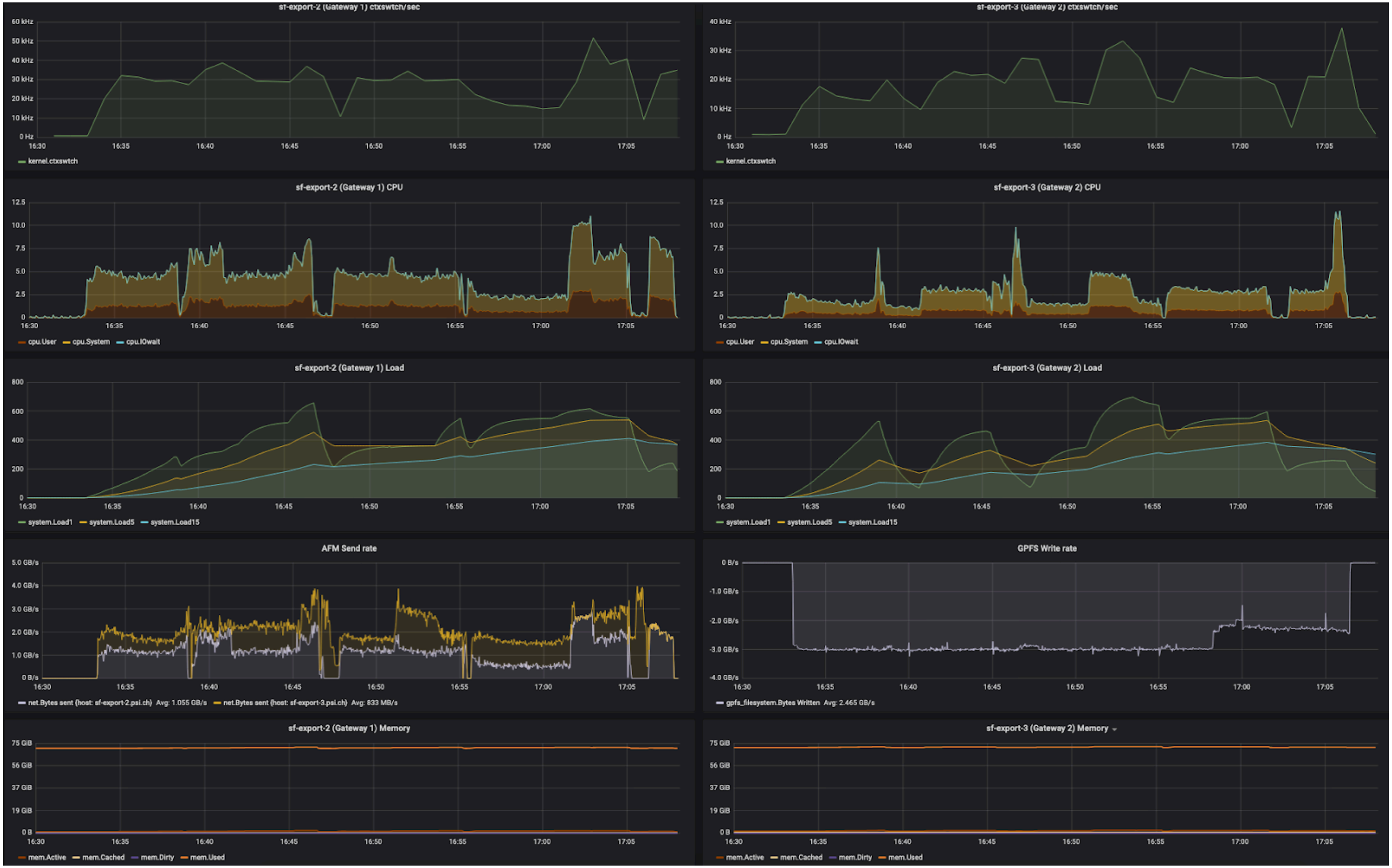
System cannot cope well with

1 TB sequential written files

I haven't found a combination of parameters that can steadily transfer @3GB/s so huge files.

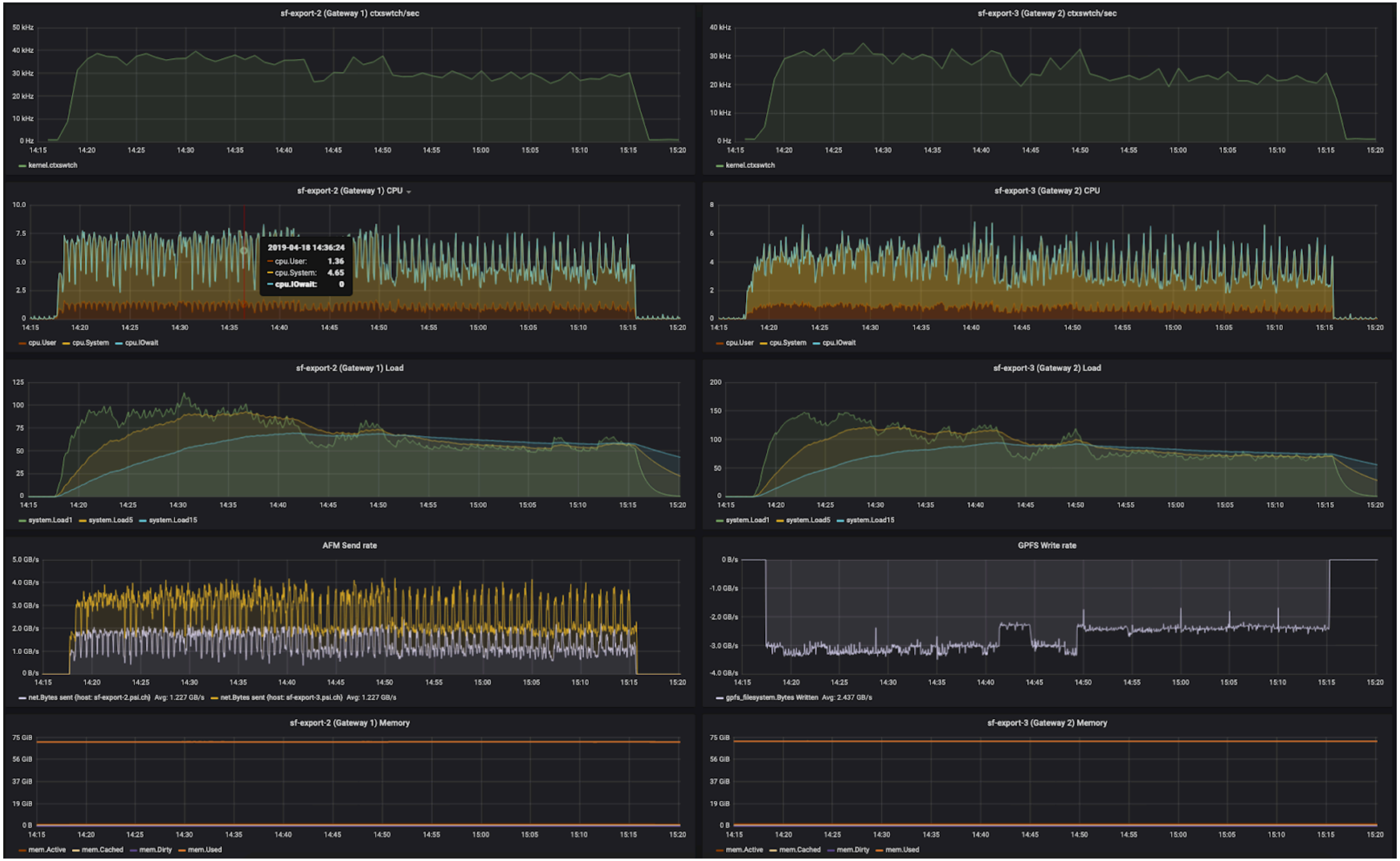
Fortunately, so far, this is not our use-case...
only one file over 150k is 1TB :-)

Test case “bad”



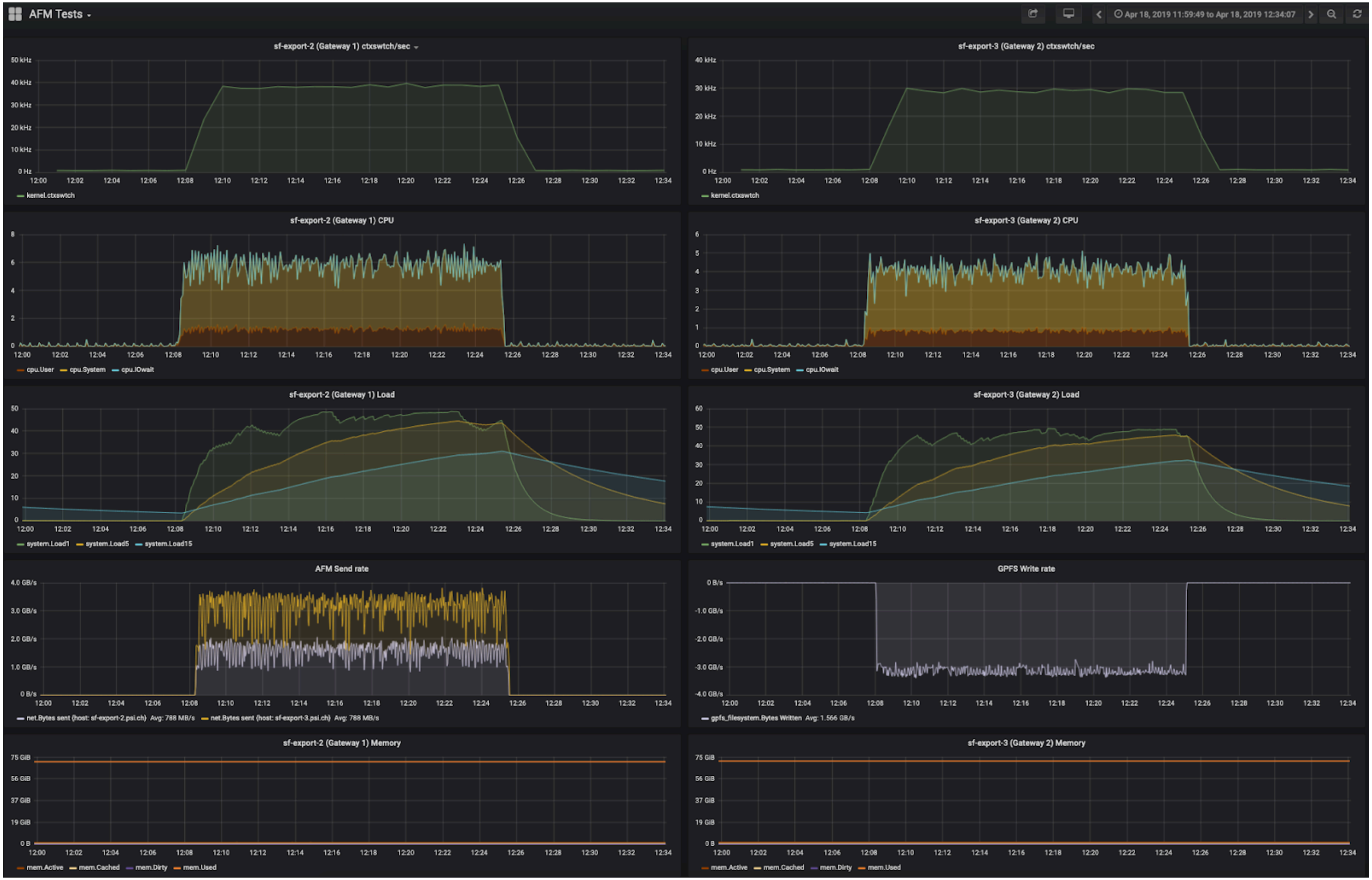
filesize	500G
MergeLen	30G
seq w. GB/s	3
AsyncDelay	15s

Test case “not 100% good”



filesize	100G
MergeLen	30G
seq w. GB/s	3
AsyncDelay	15s

Test case “100% good”



filesize	30G
MergeLen	30G
seq w. GB/s	3
AsyncDelay	15s

So far so good !

We are satisfied with AFM because:

- Direct support from AFM developers (and through PMR)
  - With interactive WebEx debugging sessions

- AFM maturity level (despite some smaller issue)
  - Even if I must admit that a "course on AFM" would be required, covering several use-cases (not only write pattern, but much more).... because of so many parameters and experience to accustomed to

- Our use case is "honored" (having data safe @Home in real time @3GB/s)

- Starting a fileset previously filled of files, AFM can reach 6.5 GB/s (NFS's limit)

- Despite our satisfaction, it is still a really complicated "beast" and **documentation should be greatly improved** with examples for parameter tuning in different use case scenarios