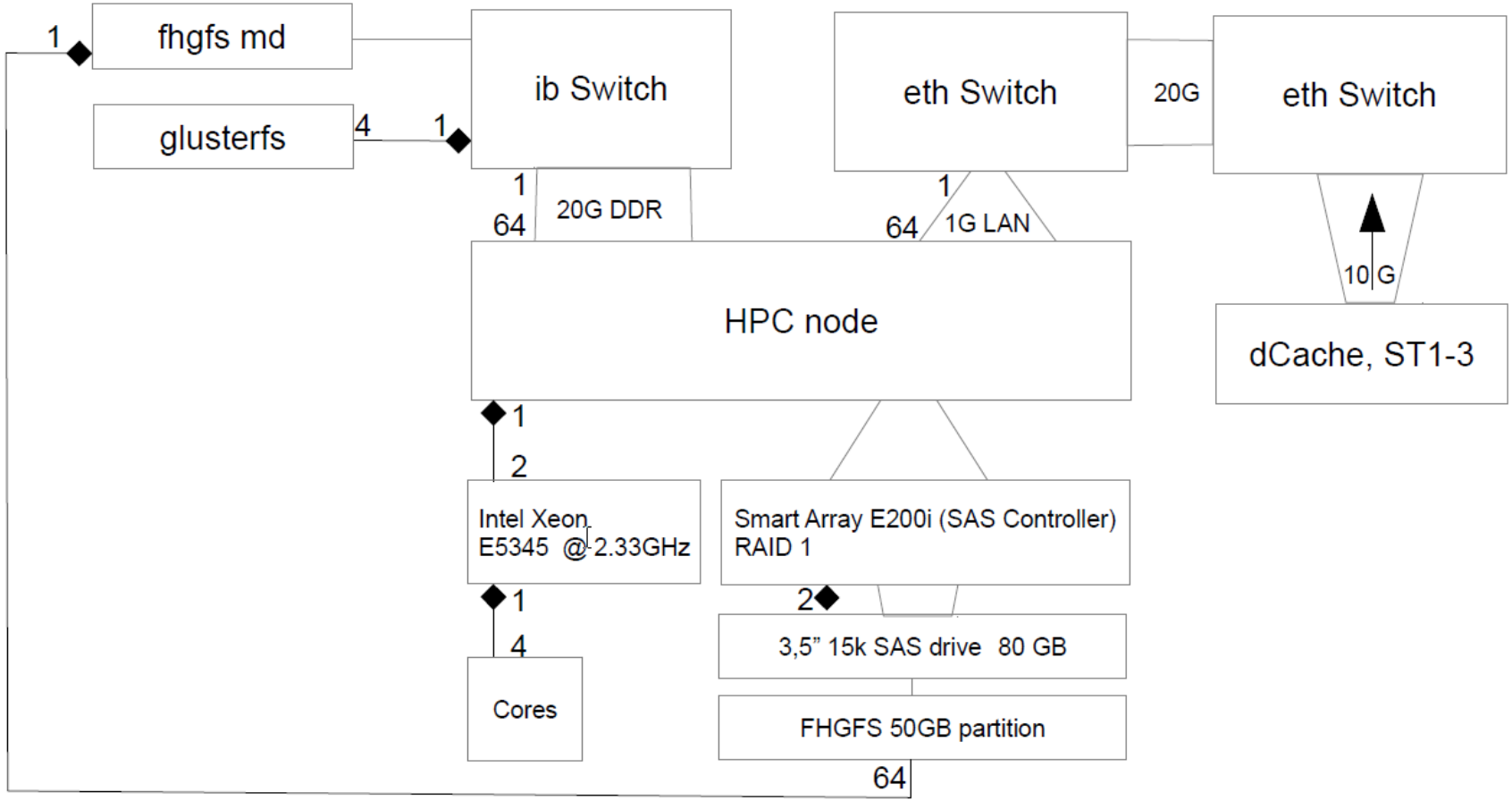
Benchmarking Storage Systems

- How to characterize the system
  - Storage
  - Network
  - Clients
- Specific benchmarks
  - iozone
  - mdtest
  - h5perf
  - Hdf5-aggregation (tiff2nexus)
  - Pilatus Detector-Simulation
  - Visualization of huge (>250GB) HDF5 datasets
- Conclusion
  - limitations

System description

- UML-like description
  - Simple annotation of redundancies/multiplicities
  - Obvious bottlenecks



System characterization

- Network
  - Basic test: iperf, ntop, ping, traceroute, etc …
  - mpi pingpong for ib as well as tcp
    - (`mpirun –mca btl ^tcp –hostfile pingpong.hosts mpitests-IMB-MPII pingpong`)

- Storage
  - Omreport et al. to report the basics (`omreport storage pdisk controller=0`)
  - iozone for basic I/O speed (`iozone -i0 -i1 -r2m -s300g -f /storage_location/file -+n`)
  - Mdtest to test basic capabilities
  - H5perf to test mpiio/phdf5 capabilities
  - Things which might have an influence …
    - Topology (of course)
      - Number of head nodes, controller, number of physical disks and speed
    - Protocol (and implementation)
    - Underlying filesystem & utilization
    - OS & Kernel; Kernel driver and modules
    - IRQ / Numa configuration.

- Clients
  - Things which might have an influence …
    - Most of the storage relevant items
    - CPU and Frontbus speed
    - Governance model
    - Bios and firmware
    - Boot options

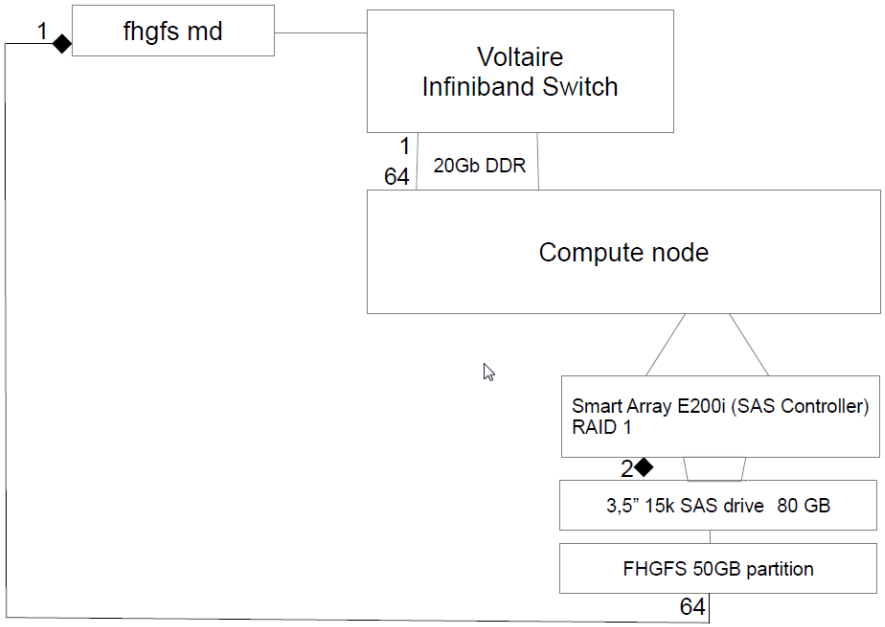
Storage characterization - fhgfs

Name:	FHGFS 2011
Vendor:	Fraunhofer
Version:	2011.04.r21
Protocol:	fhgfs client/server
Storage Size:	3.2TByte / 94% free
# of head nodes:	64
OS/Kernel:	SL 6.3 / 2.6.32-279.5.1.el6.x86_64
Disks per node:	2*80GB*0.5 SAS
Disk speed:	15k
Transfer speed:	6.00Gb/s
Raid Level:	1
Filesystem:	xfs
IRQ binding:	none
# Metadata server:	1
OS/Kernel:	SL 6.3 / 2.6.32-279.5.1.el6.x86_64
Disks per node:	2*80GB SAS
Disk speed:	15k
Transfer speed:	6.00Gb/s
Raid Level:	1
Filesystem:	ext4
IRQ binding:	none
Interconnect:	Infiniband DDR 20Gb/s
PingPong:	max. 1.000MB/s

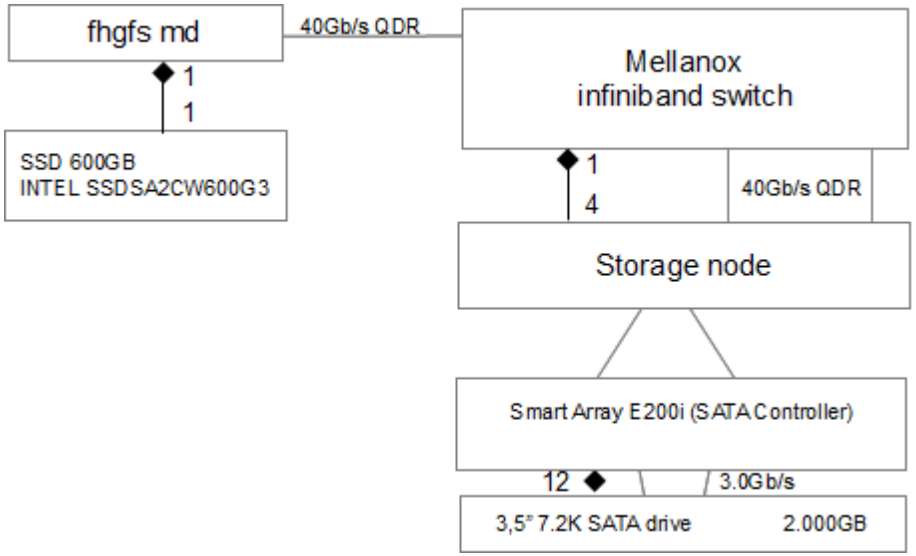
Name:	FHGFS 2012
Vendor:	Fraunhofer
Version:	2011.04.r21
Protocol:	fhgfs client/server
Storage Size:	73TByte / 99% free
# of head nodes:	4
OS/Kernel:	SL 6.3 / 2.6.32-279.5.1.el6.x86_64
Disks per node:	12*2TB SATA
Disk speed:	7.2k
Transfer speed:	3.00Gb/s
Raid Level:	5
Filesystem:	xfs
IRQ binding:	none
# Metadata server:	1
OS/Kernel:	SL 6.3 / 2.6.32-279.5.1.el6.x86_64
Disks per node:	1*600GB SSD
I/O speed:	270 MB/sec (*read*) and 220 MB/sec (*write*).
Raid Level:	5
Filesystem:	ext4
IRQ binding:	none
Interconnect:	Mellanox Infiniband QDR 40Gb/s
PingPong:	max. 2.200MB/s

Storage characterization - fhgfs

FHGFS 2011



FHGFS 2012



Storage tests

- Pilatus 6M
  - Pump 2000 images (total 18 GByte) with up to 50Hz and up to 16 concurrent streams
- Tiff2nexus
  - Convert 100.000 GISAXS images into one HDF5 file
- H5perf
  - Test posix, mpiio and pHDF5
- Iozone
  - Basic I/O test
- mdtest
  - Basic meta data operations (low I/O)
- Visualization
  - SCA3D, adaptive visualization of distributed 3d documents in open information spaces

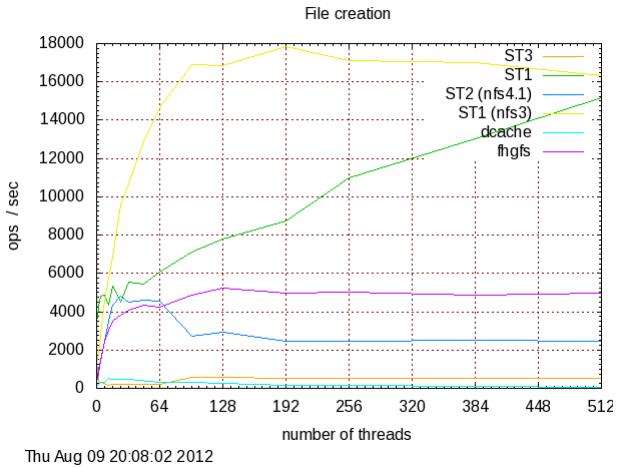
Benchmarks – mdtest

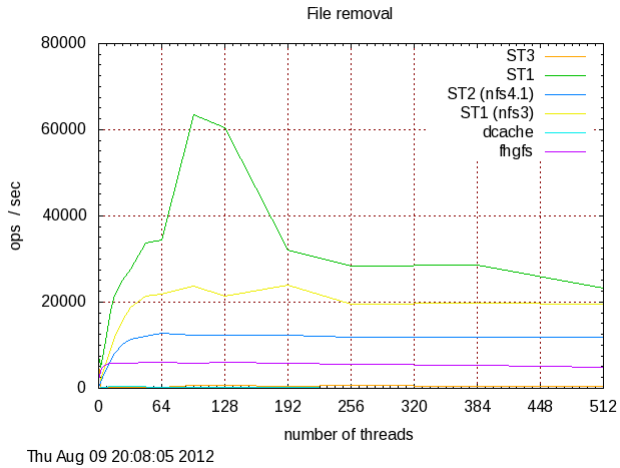
Name: mdtest
Version: 1.8.3
Url: http://sourceforge.net/projects/mdtest/
Description: mdtest is an MPI-coordinated metadata benchmark test that performs open/stat/close operations on files and
OS: Scientific Linux 6.2
Kernel: 2.6.32-279.1.1.el6.x86_64
Boot: `root=UUID=24fa1ddf-e909-4665-a612-cdef7b60abc0 rd_NO_LUKS rd_NO_LVM rd_NO_MD rd_NO_DM LANG=en_US.UTF-8 crashkernel=130M@0M`
CPU affinity: none
IRQ binding: none
Governance: unknown
MPI: openmpi 1.5.3

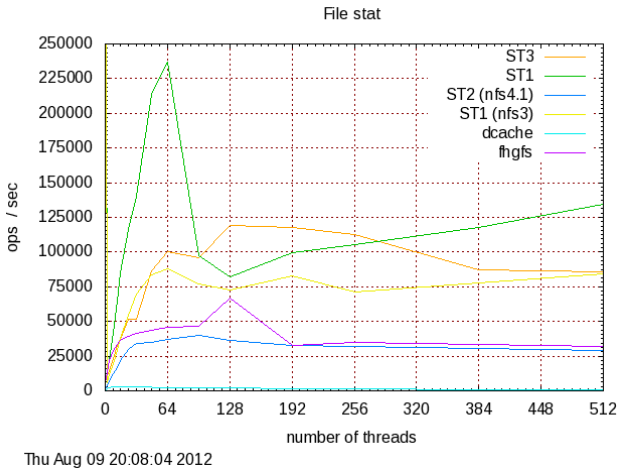
Platform: DESY-HPC 2011

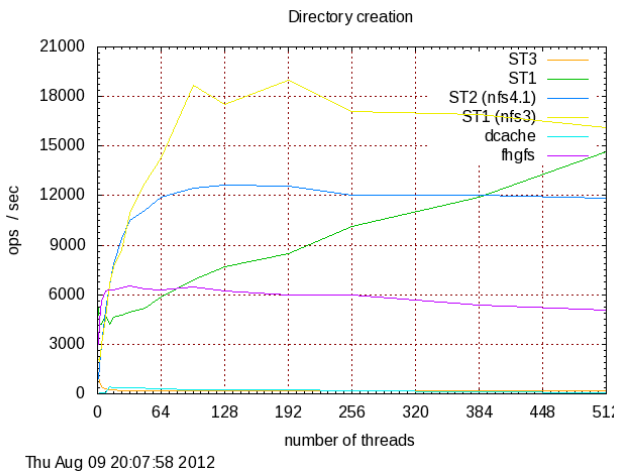
	Description	Type	Capacity / TB	Protocol
1	dCache PS 2011	dCache	10.000	NFS 4.1
2	fhgfs 2011	FHGFS (ipoib)	3.2	FHGFS 2011.04.r16
3	ST1	WAFL	20	NFS 3
4	ST2	WAFL	40	NFS 3
5	ST2	WAFL	4*10	NFS 4.1
6	ST3	GPFS	443	NFS 3
7	Fhgfs	FHGFS (ipoib)	73	FHGFS 2011.04.r19
8	Glusterfs	Glusterfs (rdma)	73	3.2.6

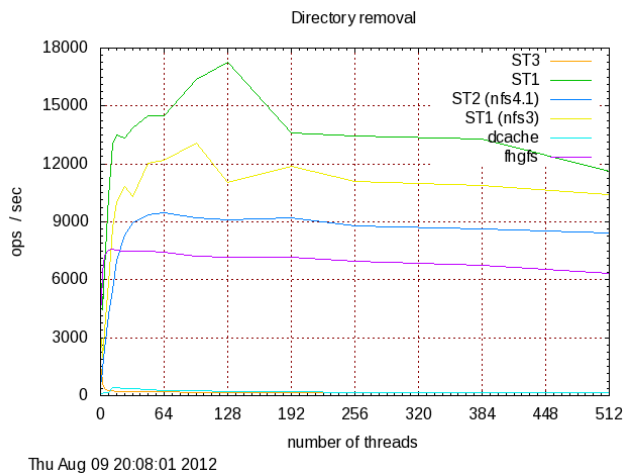
Benchmarks – mdtest

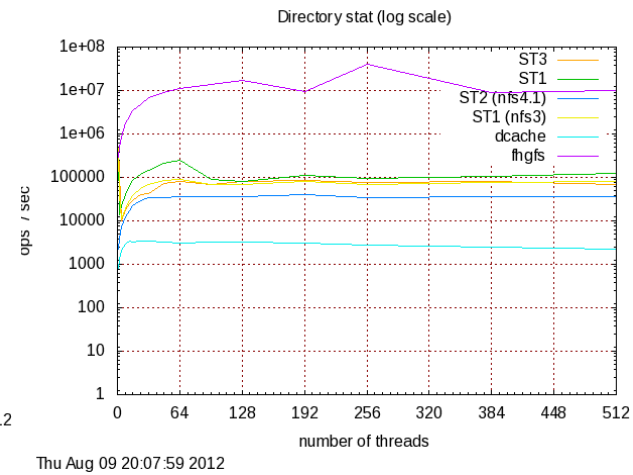




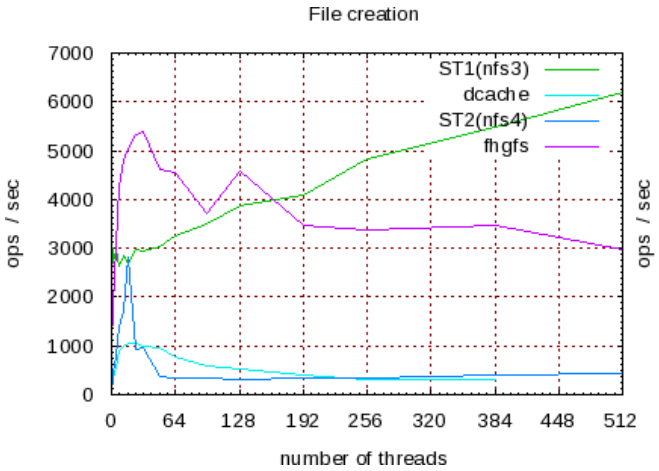


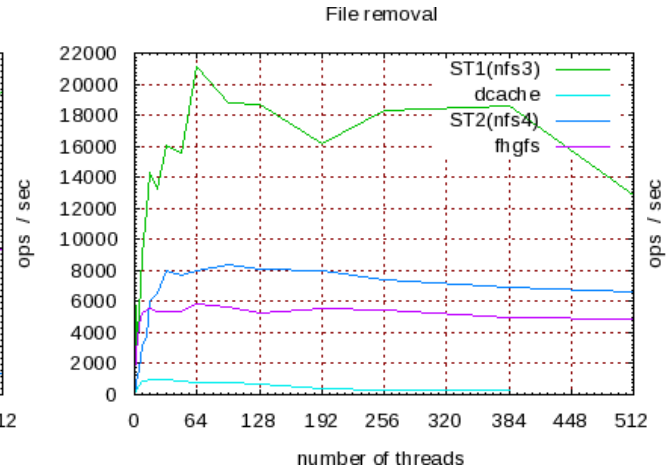


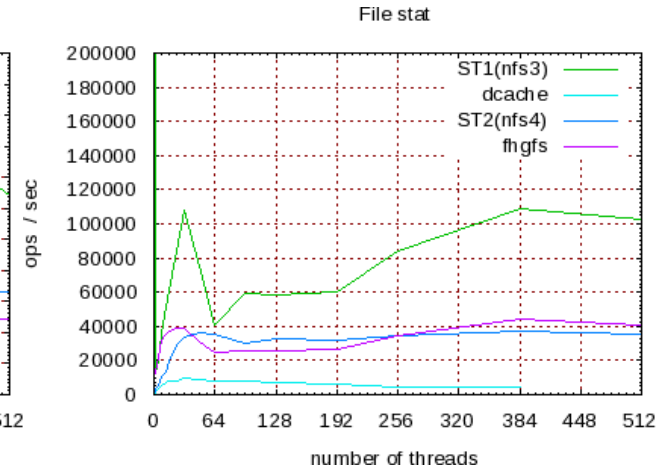


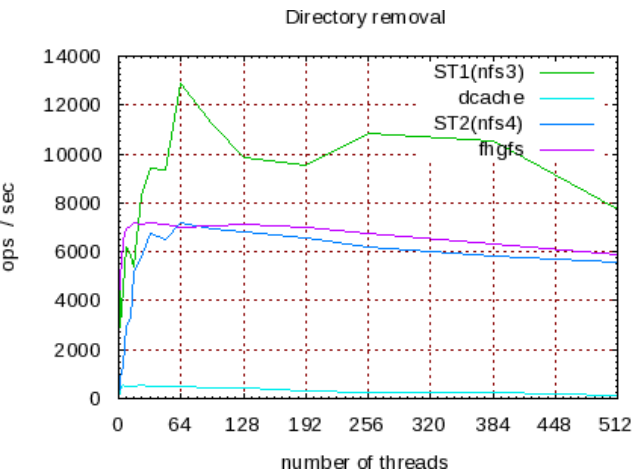


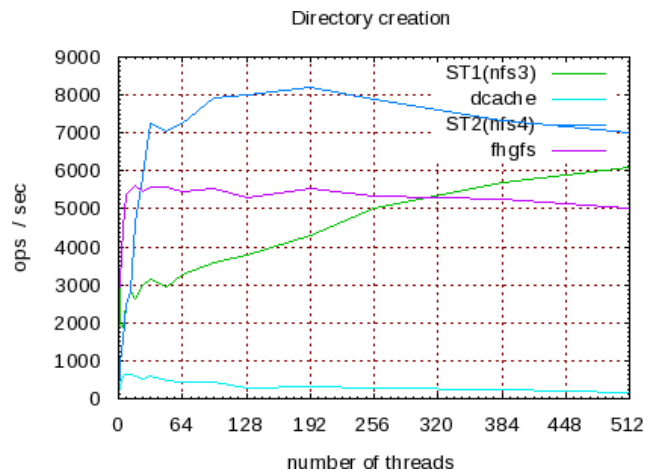
Benchmarks – mdtest

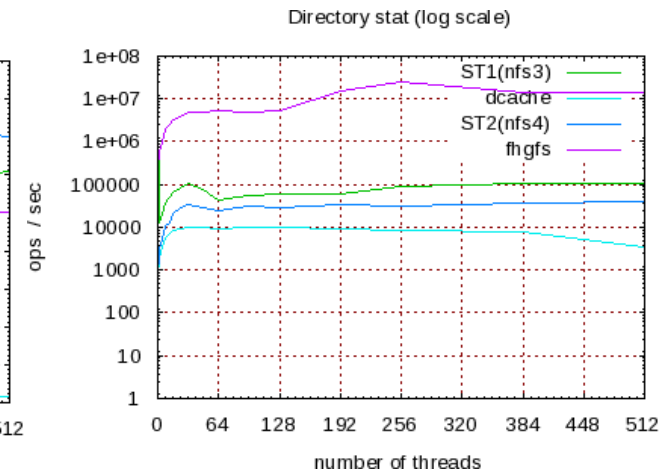










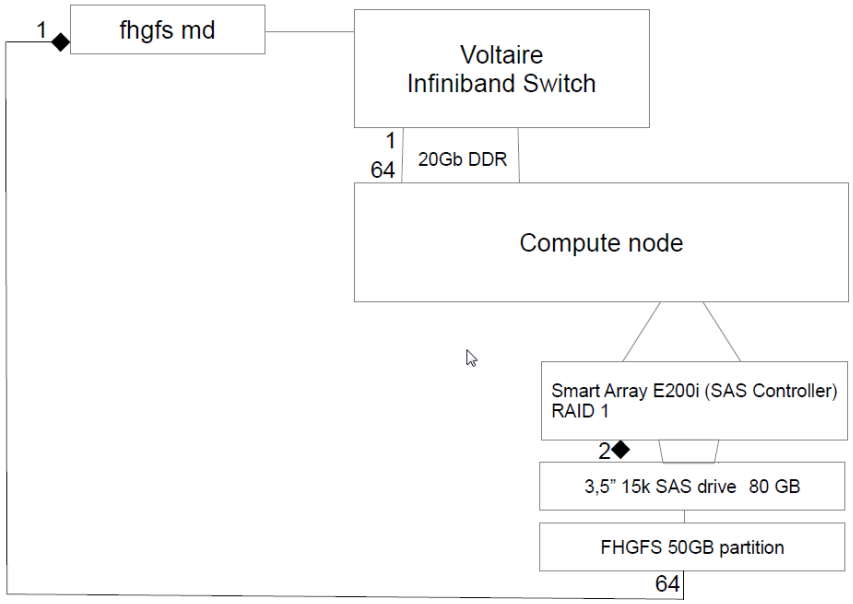


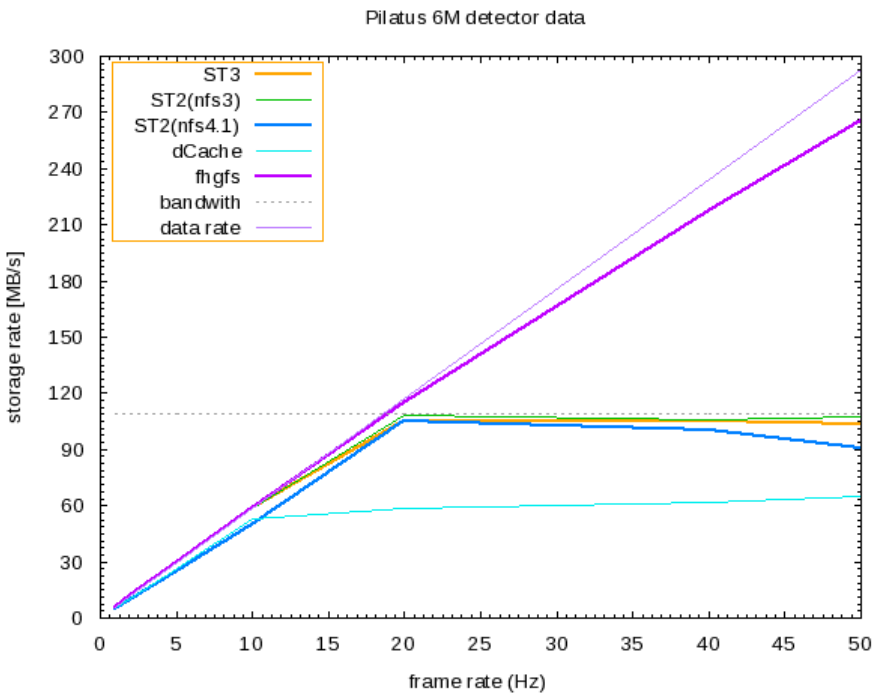
Benchmarks – Pilatus 6M

- Pilatus 6M detector simulation
  - Currently the most demanding detector at Synchrotrons and running EuroFEL lightsources
  - Can operate at ~20Hz
  - Data format either raw (tiff) or compressed (cbf)
  - Data rates @20Hz: 1Gb/s for cbf, twice as much for tiff
    - Not a challenge at all
  - Mutliple beamlines equipped with Pilatus 6M, so up to 4 parallel/concurrent streams
    - Most systems start to suffer
- Execution: `pssh -t 0 -H "host1 host2" pilatus.sh`

	Description	Type	Capacity / TB	Protocol
1	dCache PS 2011	dCache	10.000	NFS 4.1
2	fhgfs 2011	FHGFS (ipoib)	3.2	FHGFS 2011.04.r16
3	ST1	WAFL	20	NFS 3
4	ST2	WAFL	40	NFS 3
5	ST2	WAFL	4*10	NFS 4.1
6	ST3	GPFS	443	NFS 3
7	Fhgfs	FHGFS (ipoib)	73	FHGFS 2011.04.r19
8	Glusterfs	Glusterfs (rdma)	73	3.2.6

Benchmarks – Pilatus 6M

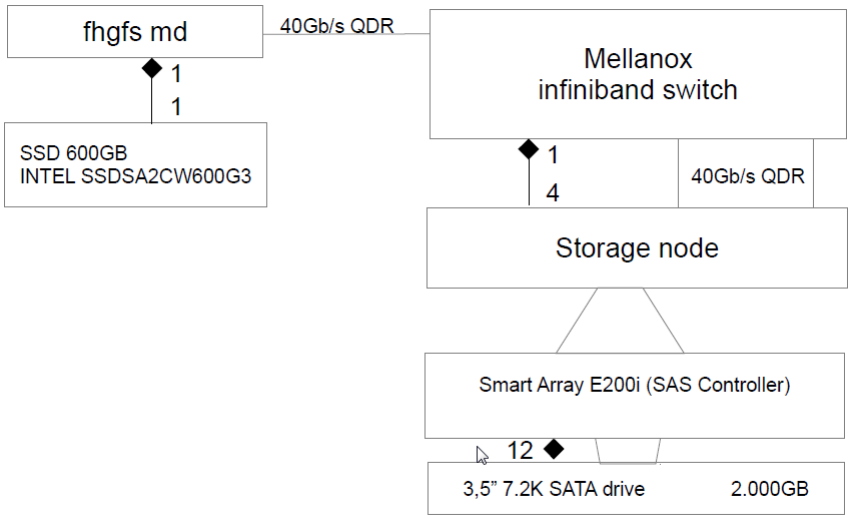


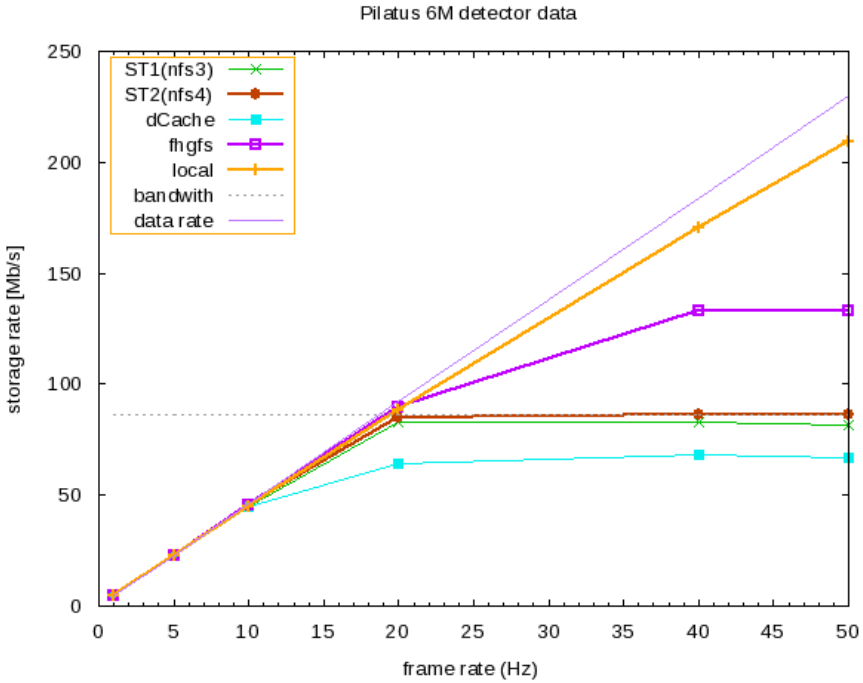


Single stream:

- 10Hz no problem
- 20Hz no problem (except for dCache)
- 50Hz no problem for fhgfs

Benchmarks – Pilatus 6M

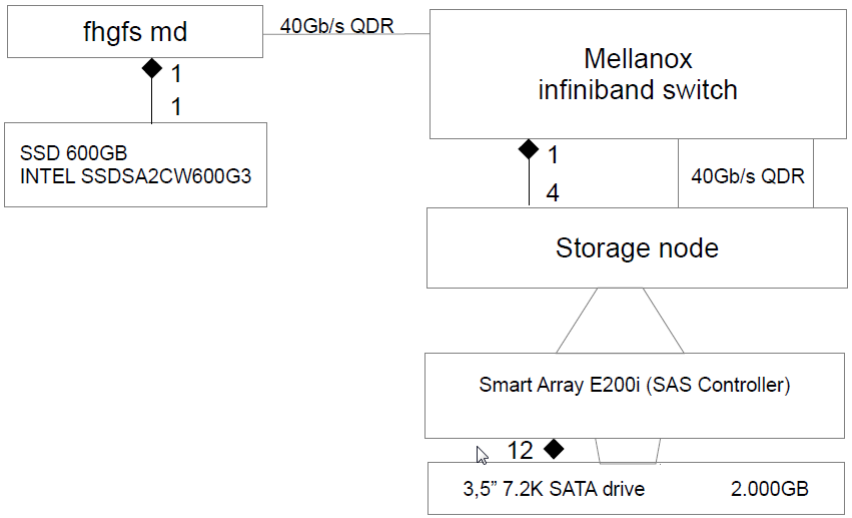


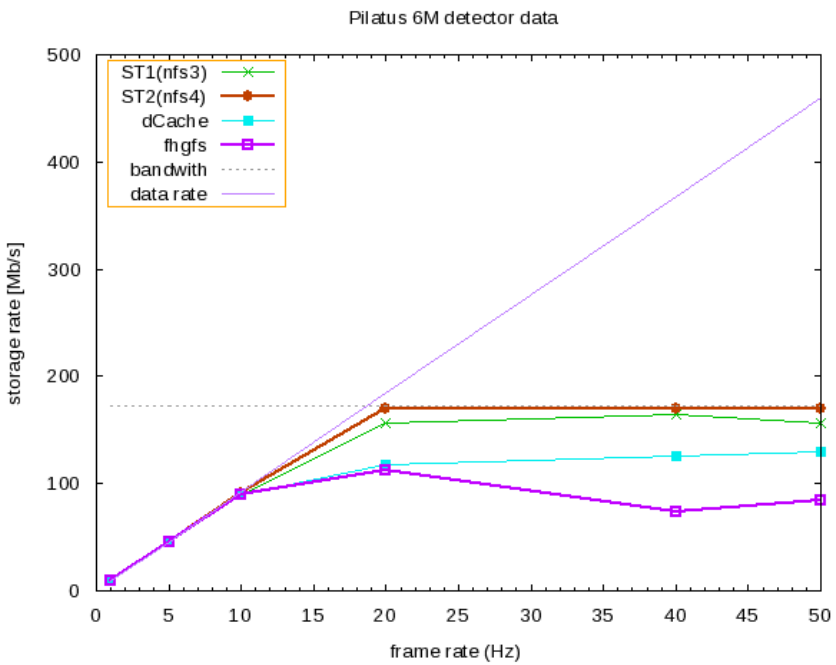


Single stream:

- 10Hz no problem
- 20Hz no major problem
- 50Hz might be a problem

Benchmarks – Pilatus 6M

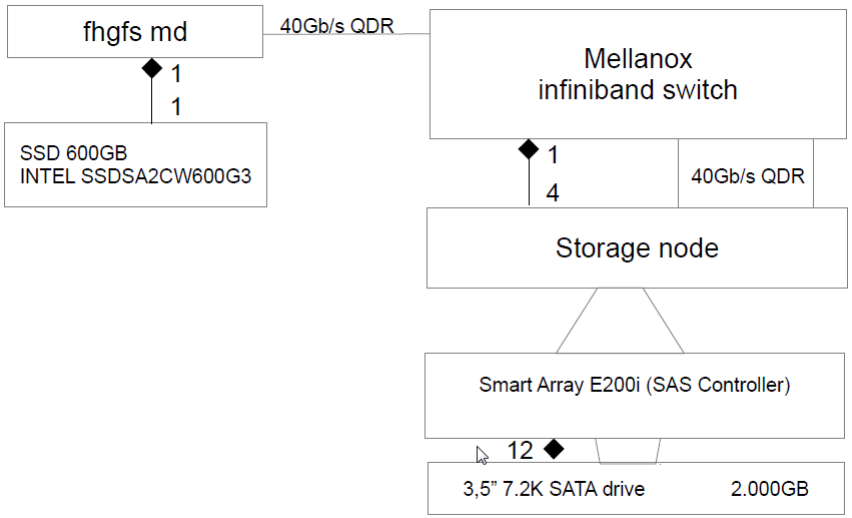


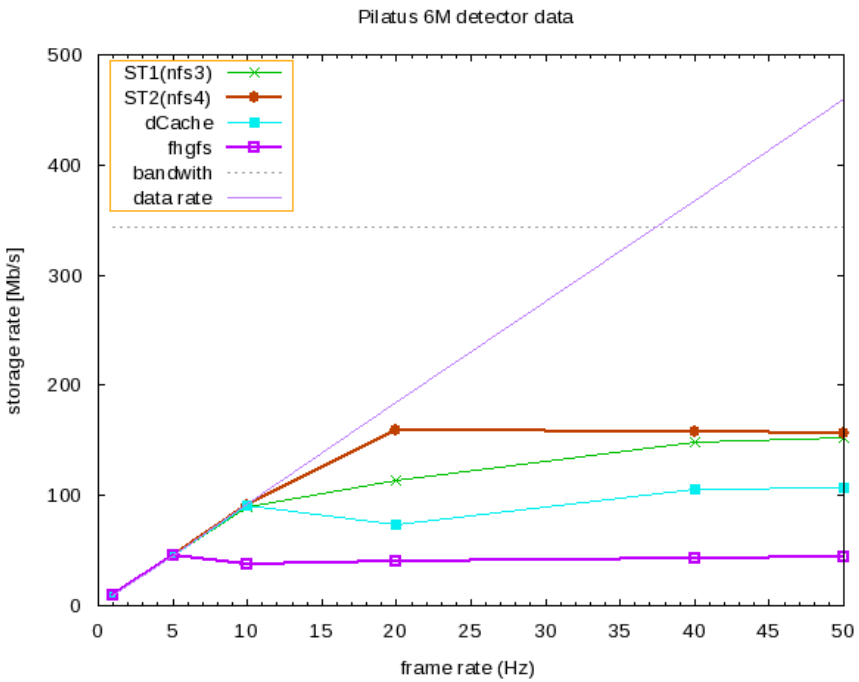


Two concurrent streams:

- 10Hz no problem
- 20Hz no major problem
- 50Hz might be a problem

Benchmarks – Pilatus 6M





Four concurrent streams:

- 10Hz no problem
- 20Hz no major problem
- 50Hz becomes a problem

Problems

- Kernel updates
  - Spoil NFS daemon for certain combinations of eth0, kernel & firmware
    - High load and memory consumption
    - No transfer of large files
  - Kernel bug
    - cpu timer elevates cpu consumption by ksoftirqd & frequent kernel cache trashing
    - Extremely high number of timer interrupts (/proc/interrupts)
- Re-configuration renders an extremely fast fhgfs into an extremely slow one
- Bios vs System governance models have a substantial influence
- Irqbalancer on/off
- Numa bindings of eth/ib
- Autotuning (e.g. defragmentation runs)
- Infiniband not homogenous
  - Some host-host speeds at 70% of the design value, some at only 50%.
  - Reason unknown. Might correlate with cable length.
  - Benchmarks depend on the choice of hosts
- Several more factor affecting the results – too many parameters not under control

in brief

- ST1 nfs3 behaves ok.
- ST2 nfs4.1 showed initially some instabilities
  - Meta data operations lack behind nfs3
  - Greatly improved meanhwile
- dCache not fast, but very stable and nfs 4.1 capable
- ST3
  - Not as fast as advertisements suggest
  - Meta data operations not convincing
- Glusterfs
  - In our configuration unusable
- FHGFS
  - Strongly depends on the configuration, but could easily outperform any of the others at a fraction of the costs.

Conclusions

- Mdtest+iozone+h5perf sufficient to test basis capabilities
  - Poor performance on tests -> poor performance in real applications
  - Good performance on tests doesn't always imply good performance in real applications
- Needs to simulate applications on real hardware to get an idea
  - Can offer access to platforms
- Qualitative figures (stable, unstable, unusable) helpful