High Speed Detectors: Problems and Solutions?

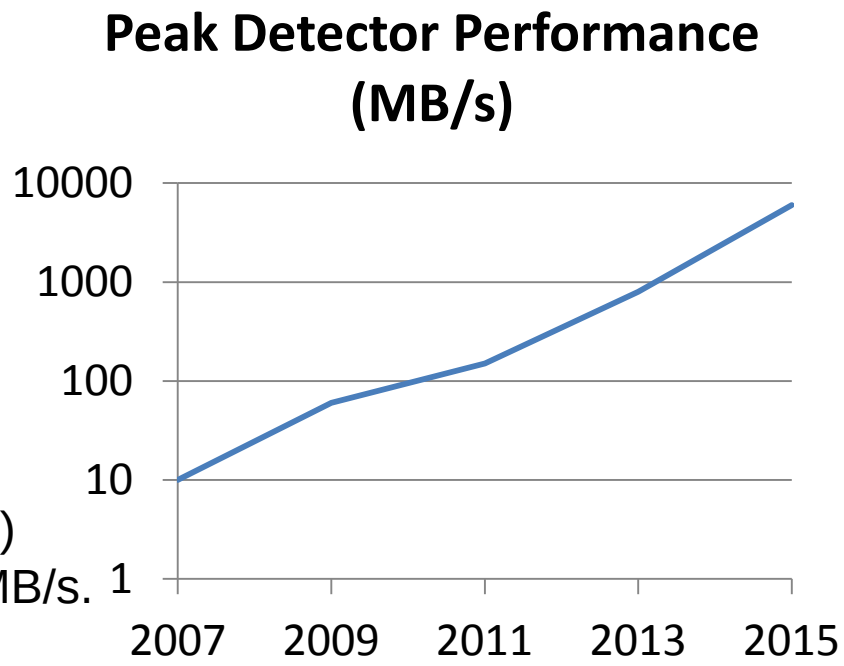
Nick Rees, Mark Basham, Frederik Ferner, Ulrik Pedersen, Tobias Richter, Jonathan Thompson
(Diamond Light Source, Oxfordshire)

History

- Early 2007:
  - Diamond first user.
  - No detector faster than ~10 MB/sec.
- Early 2009:
  - first Lustre system (DDN S2A9900)
  - first Pilatus 6M system @ 60 MB/s.
- Early 2011:
  - second Lustre system (DDN SFA10K)
  - first 25Hz Pilatus 6M system @150 MB/s.
- Early 2013:
  - first GPFS system (DDN SFA12K)
  - First 100 Hz Pilatus 6M system @ 600 MB/sec
  - ~10 beamlines with 10 GbE detectors (mainly Pilatus and PCO Edge).
- Early 2015:
  - delivery of Percival detector (6000 MB/sec).

diamond

History

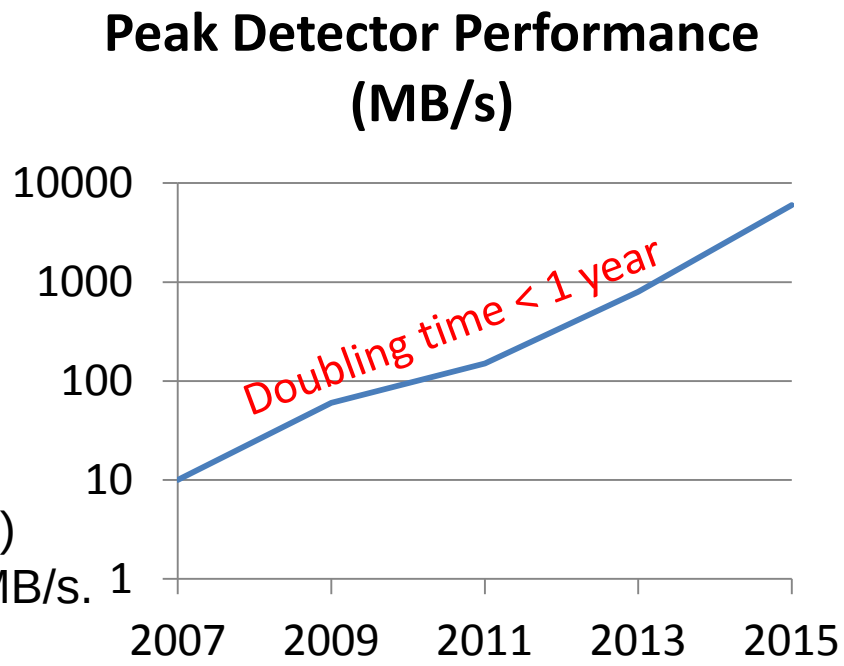


- Early 2007:
  - Diamond first user.
  - No detector faster than ~10 MB/sec.
- Early 2009:
  - first Lustre system (DDN S2A9900)
  - first Pilatus 6M system @ 60 MB/s.
- Early 2011:
  - second Lustre system (DDN SFA10K)
  - first 25Hz Pilatus 6M system @150 MB/s.
- Early 2013:
  - first GPFS system (DDN SFA12K)
  - First 100 Hz Pilatus 6M system @ 600 MB/sec
  - ~10 beamlines with 10 GbE detectors (mainly Pilatus and PCO Edge).
- Early 2015:
  - delivery of Percival detector (6000 MB/sec).

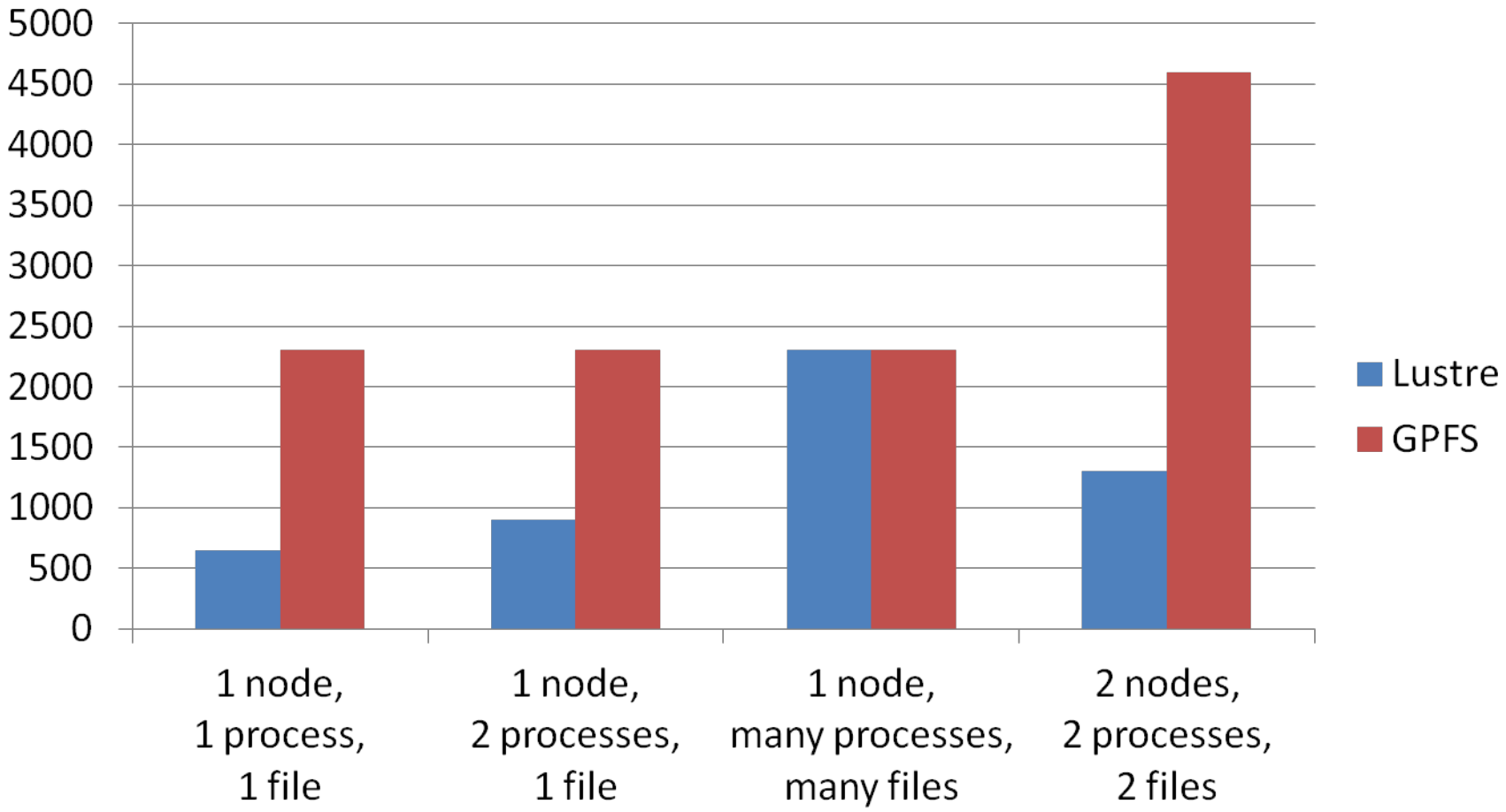
diamond

History

- Early 2007:
  - Diamond first user.
  - No detector faster than ~10 MB/sec.
- Early 2009:
  - first Lustre system (DDN S2A9900)
  - first Pilatus 6M system @ 60 MB/s.
- Early 2011:
  - second Lustre system (DDN SFA10K)
  - first 25Hz Pilatus 6M system @150 MB/s.
- Early 2013:
  - first GPFS system (DDN SFA12K)
  - First 100 Hz Pilatus 6M system @ 600 MB/sec
  - ~10 beamlines with 10 GbE detectors (mainly Pilatus and PCO Edge).
- Early 2015:
  - delivery of Percival detector (6000 MB/sec).

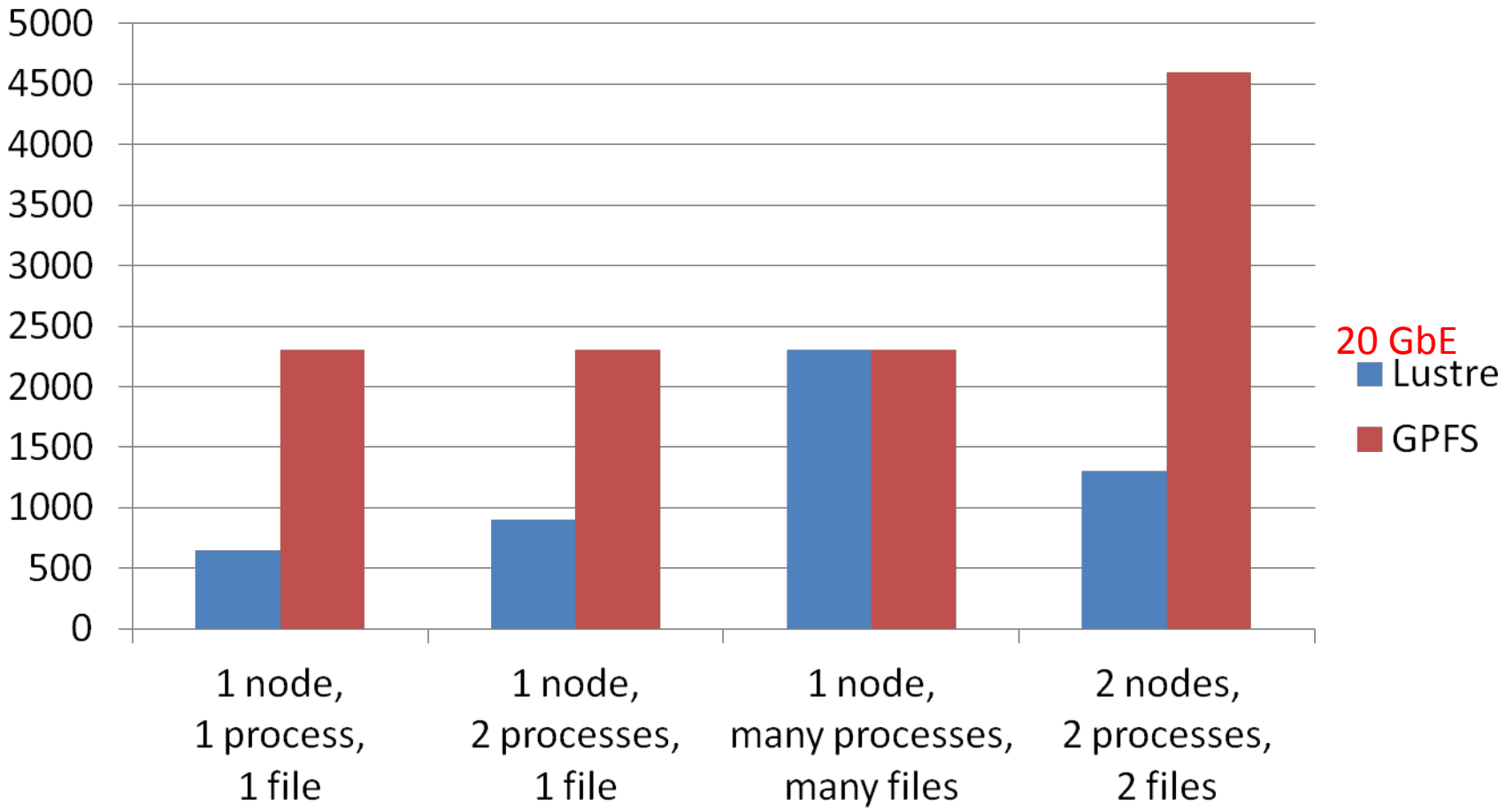


Why move to GPFS? – 20 GbE nodes



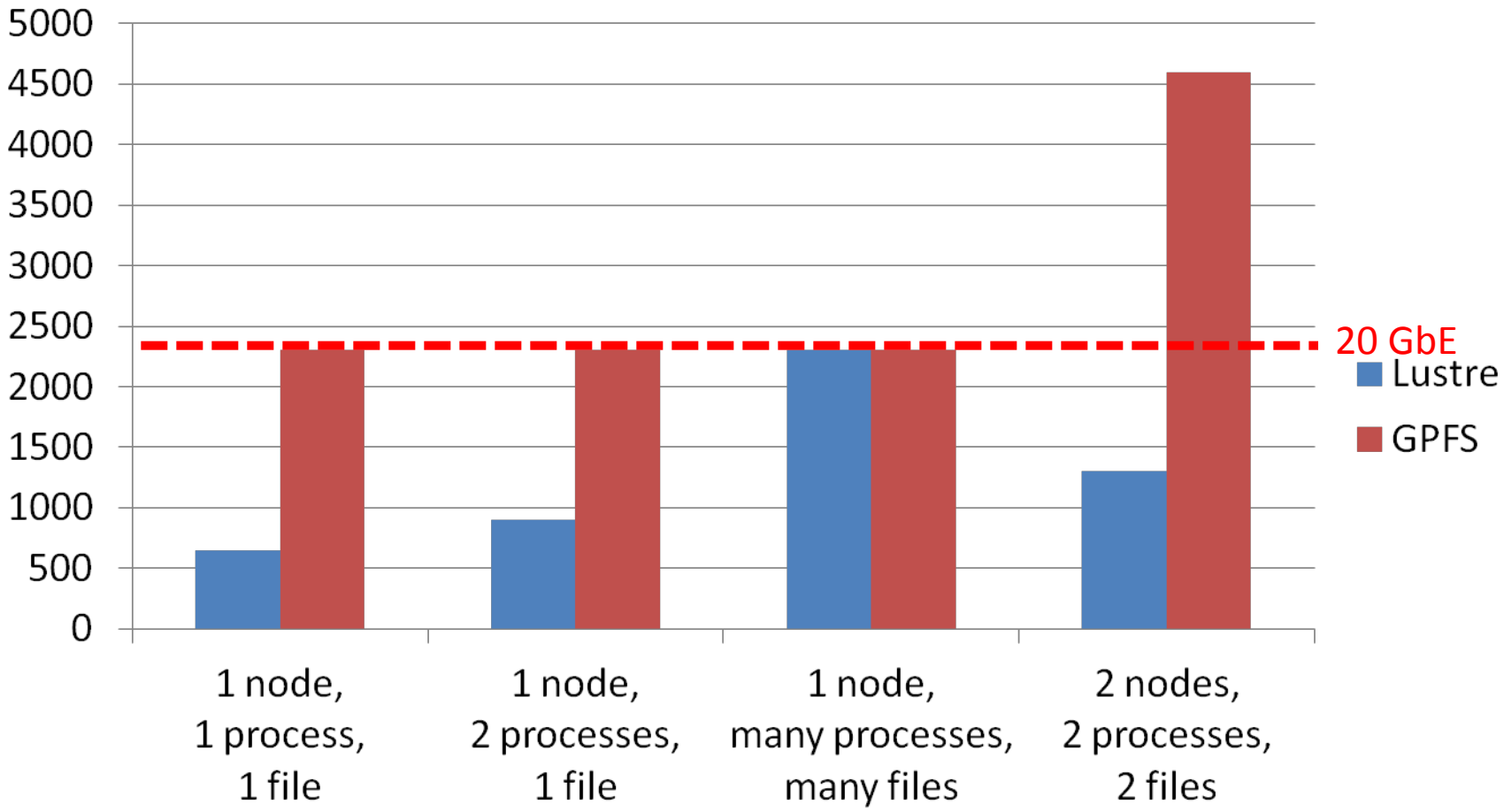
diamond

Why move to GPFS? – 20 GbE nodes



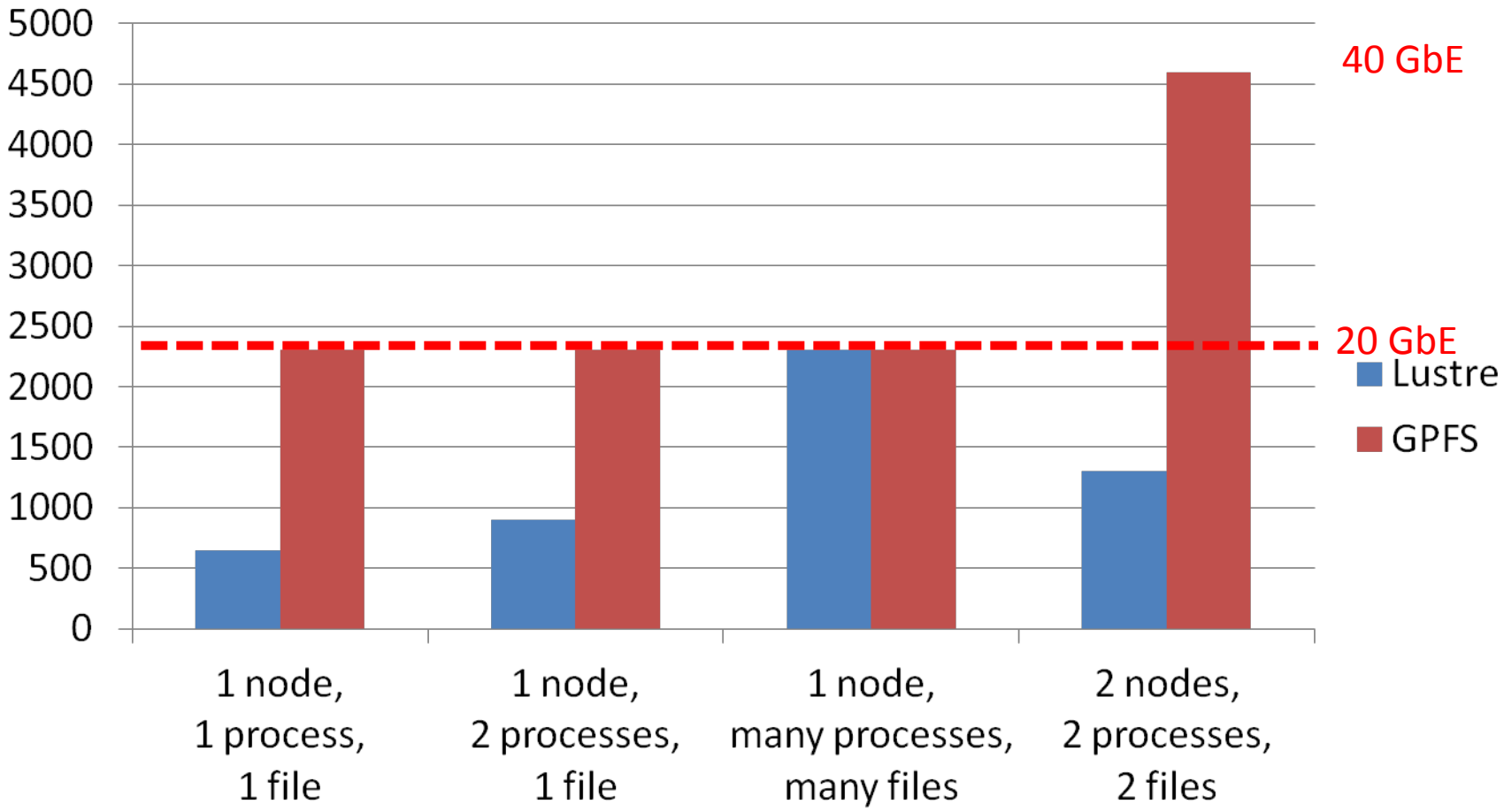
diamond

Why move to GPFS? – 20 GbE nodes



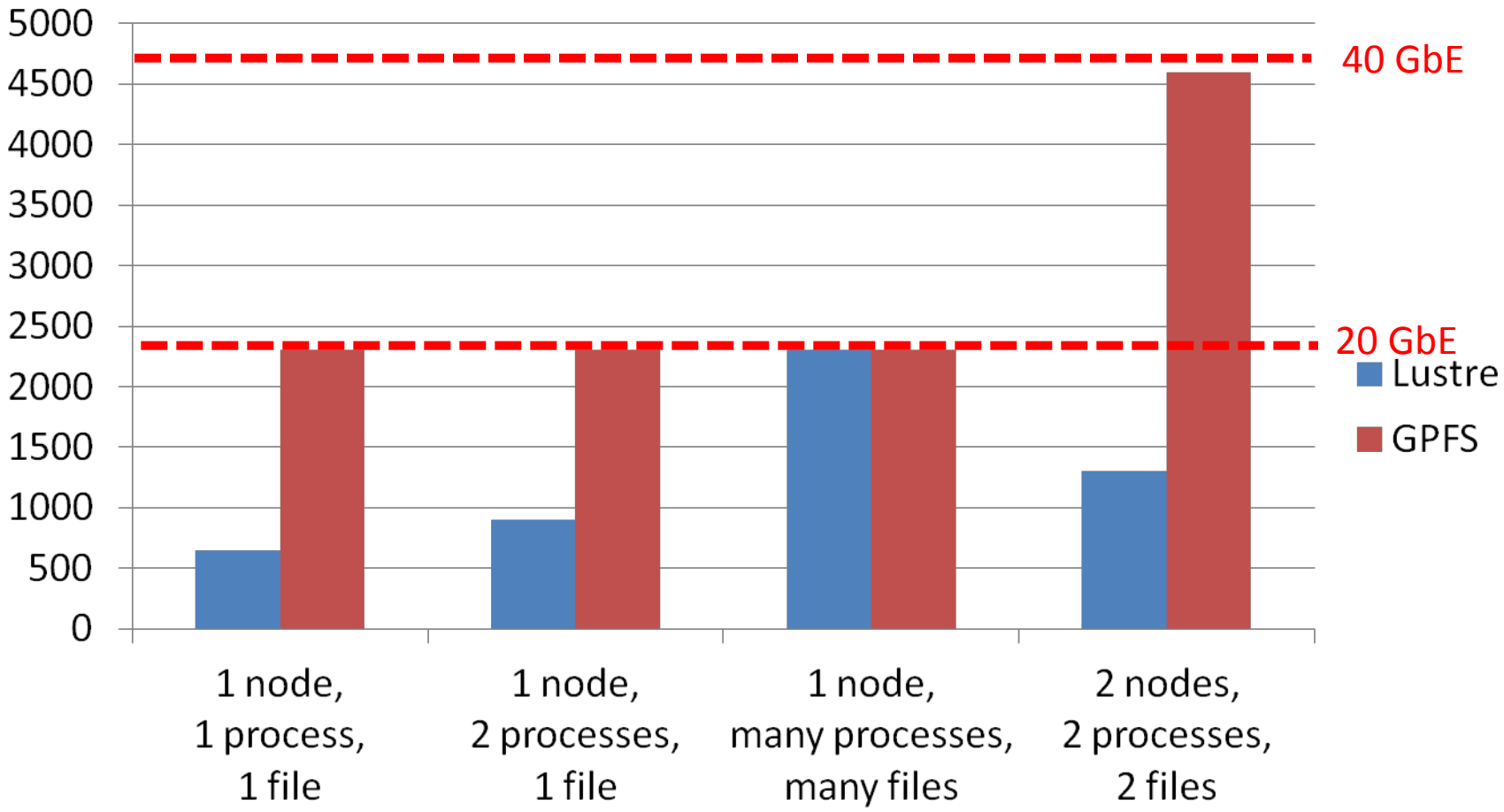
diamond

Why move to GPFS? – 20 GbE nodes



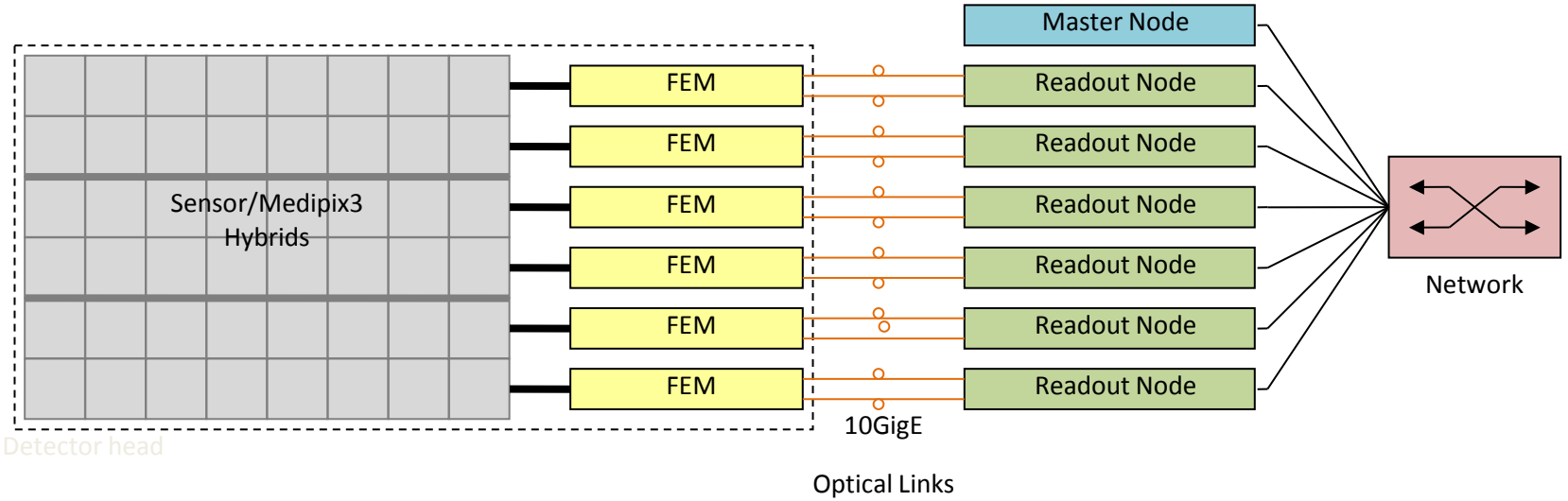
diamond

Why move to GPFS? – 20 GbE nodes

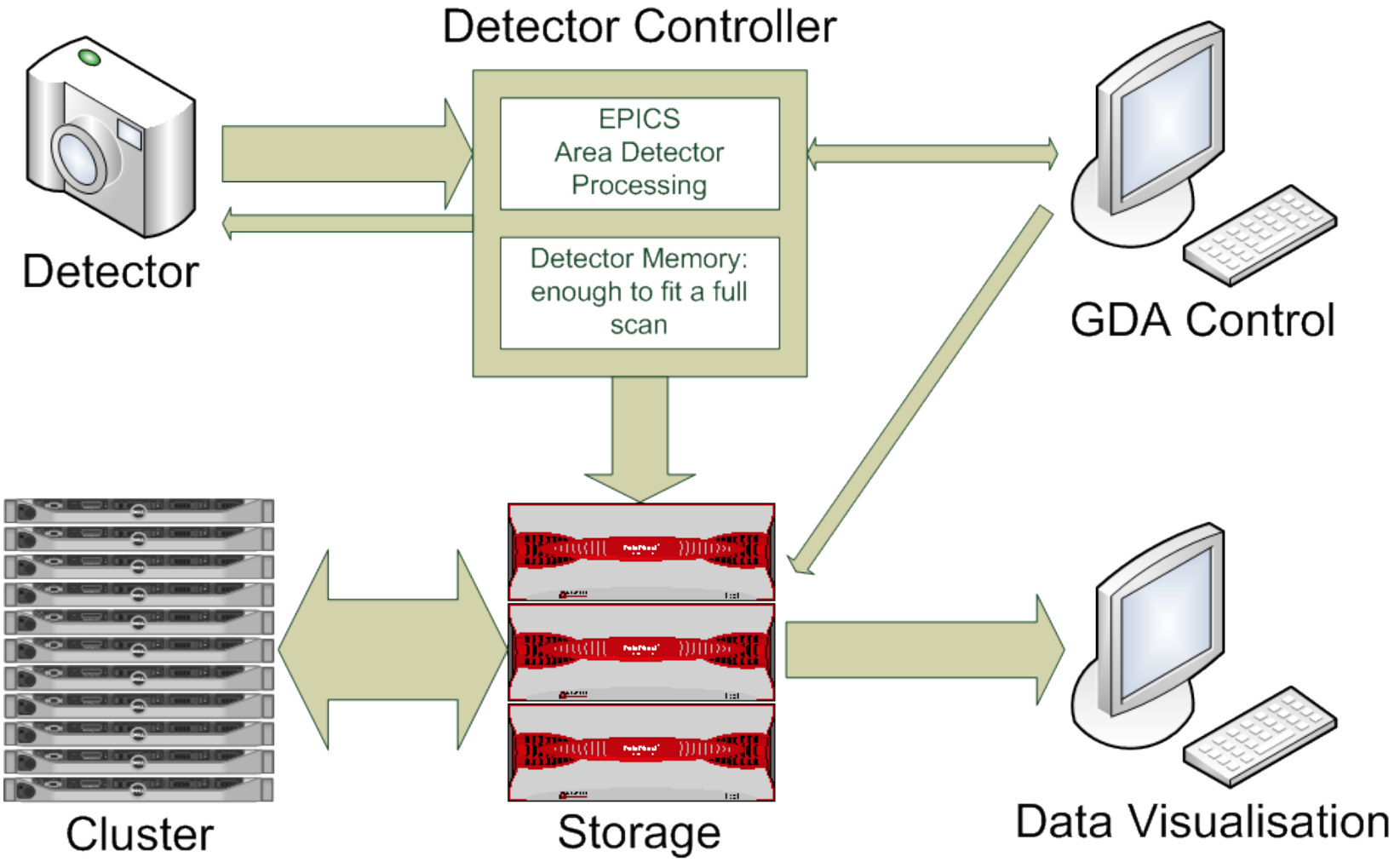


diamond

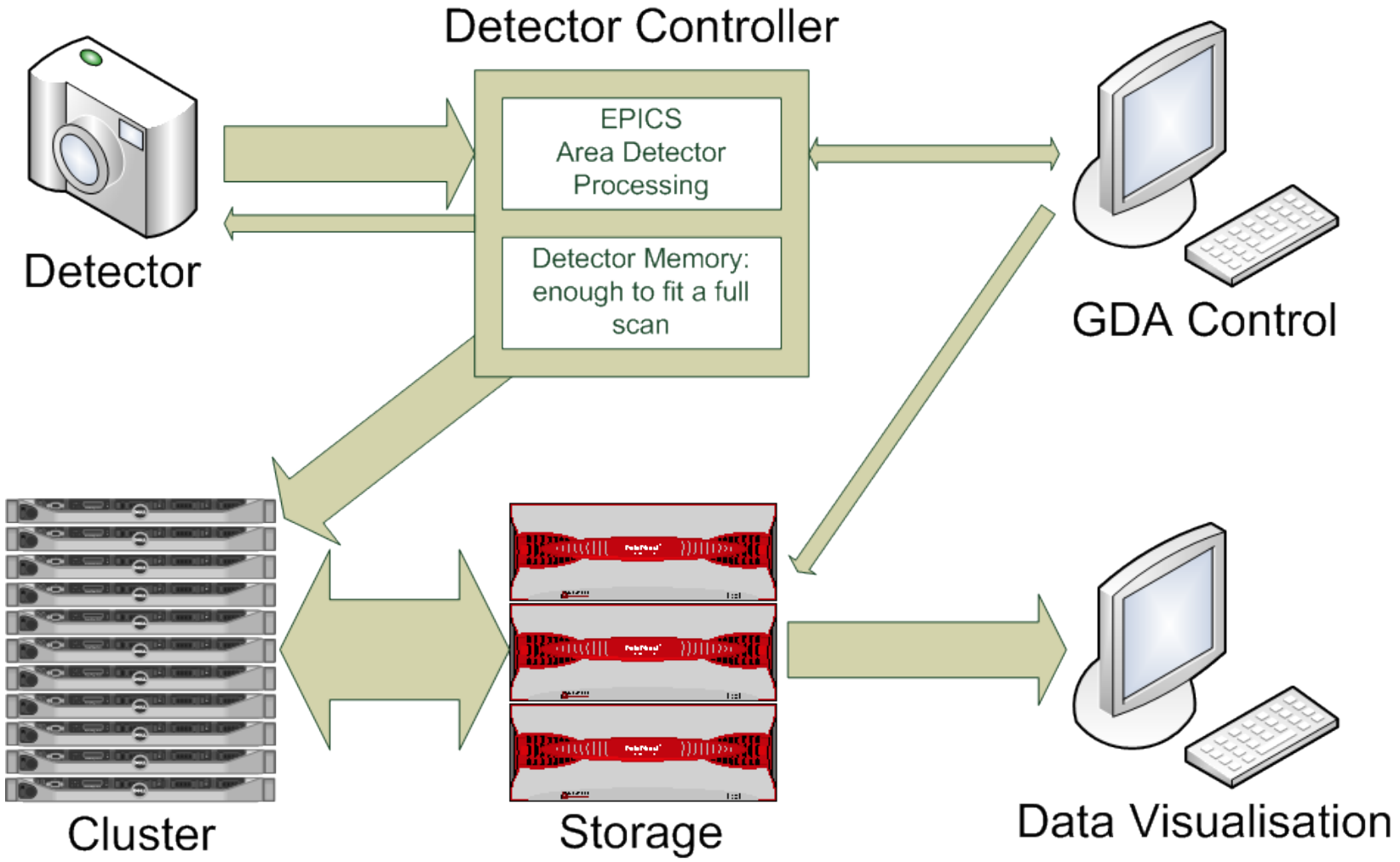
Basic Parallel Detector Design



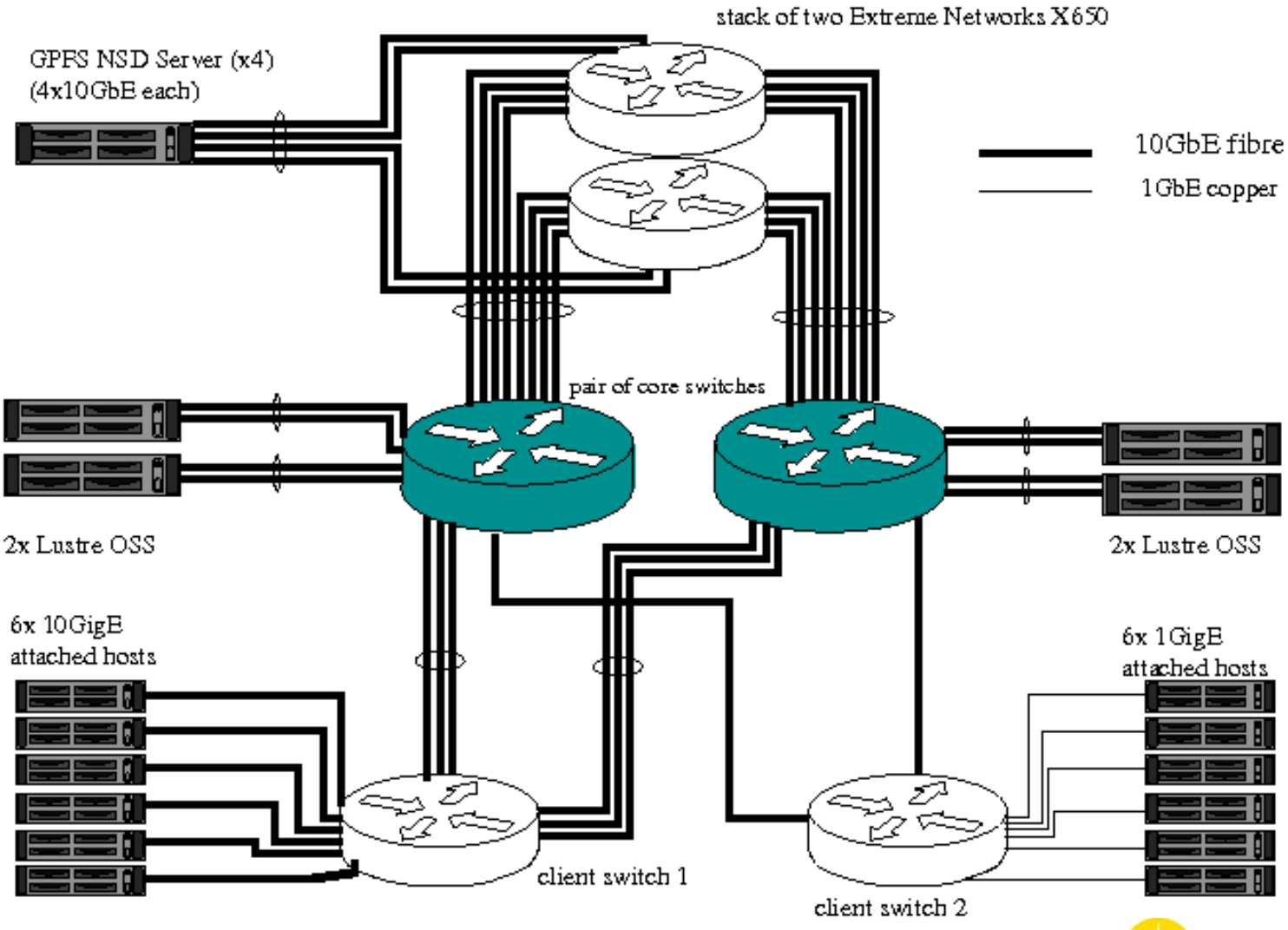
DLS Detector Model



Future High Data Rate Approach

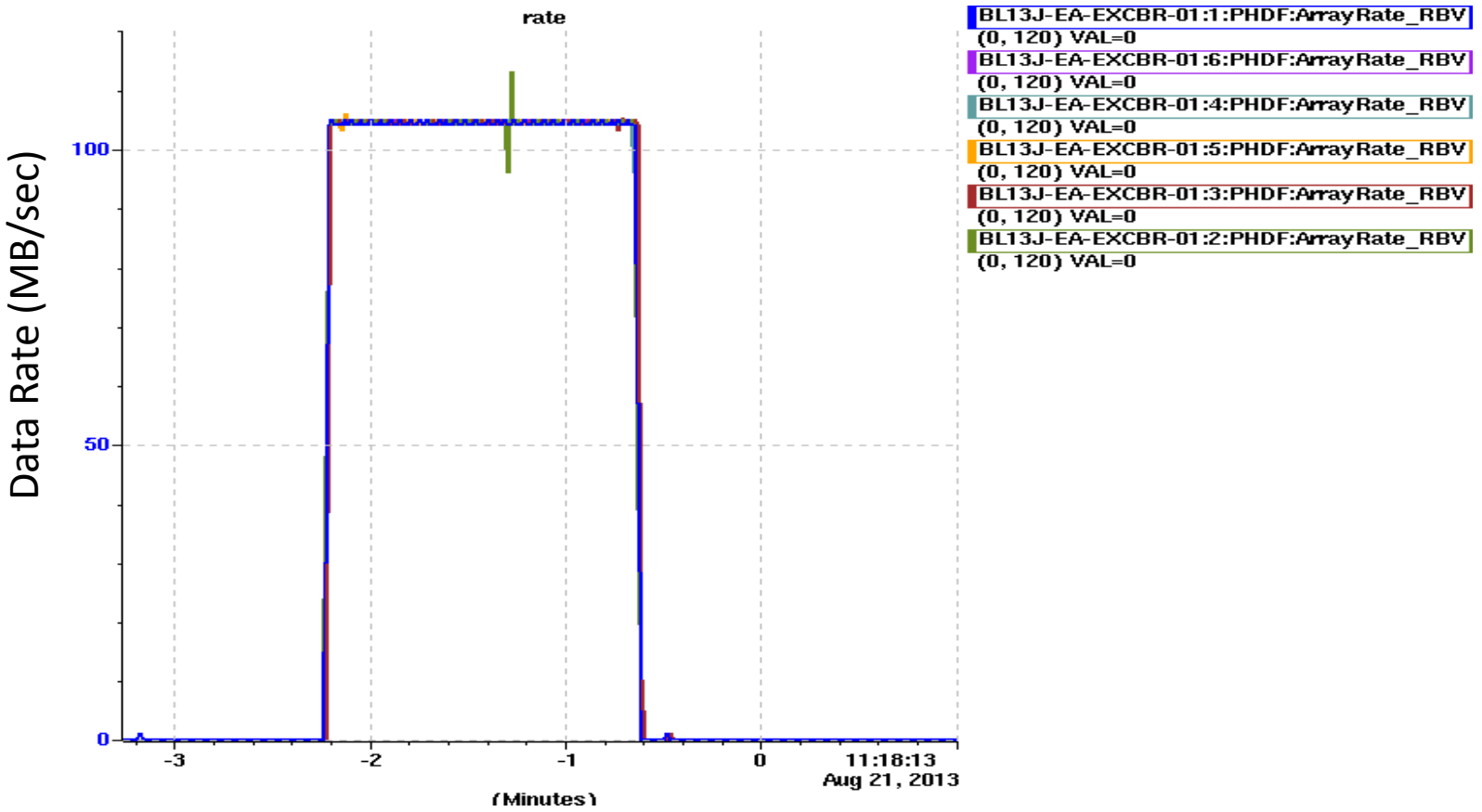


Test Network Layout

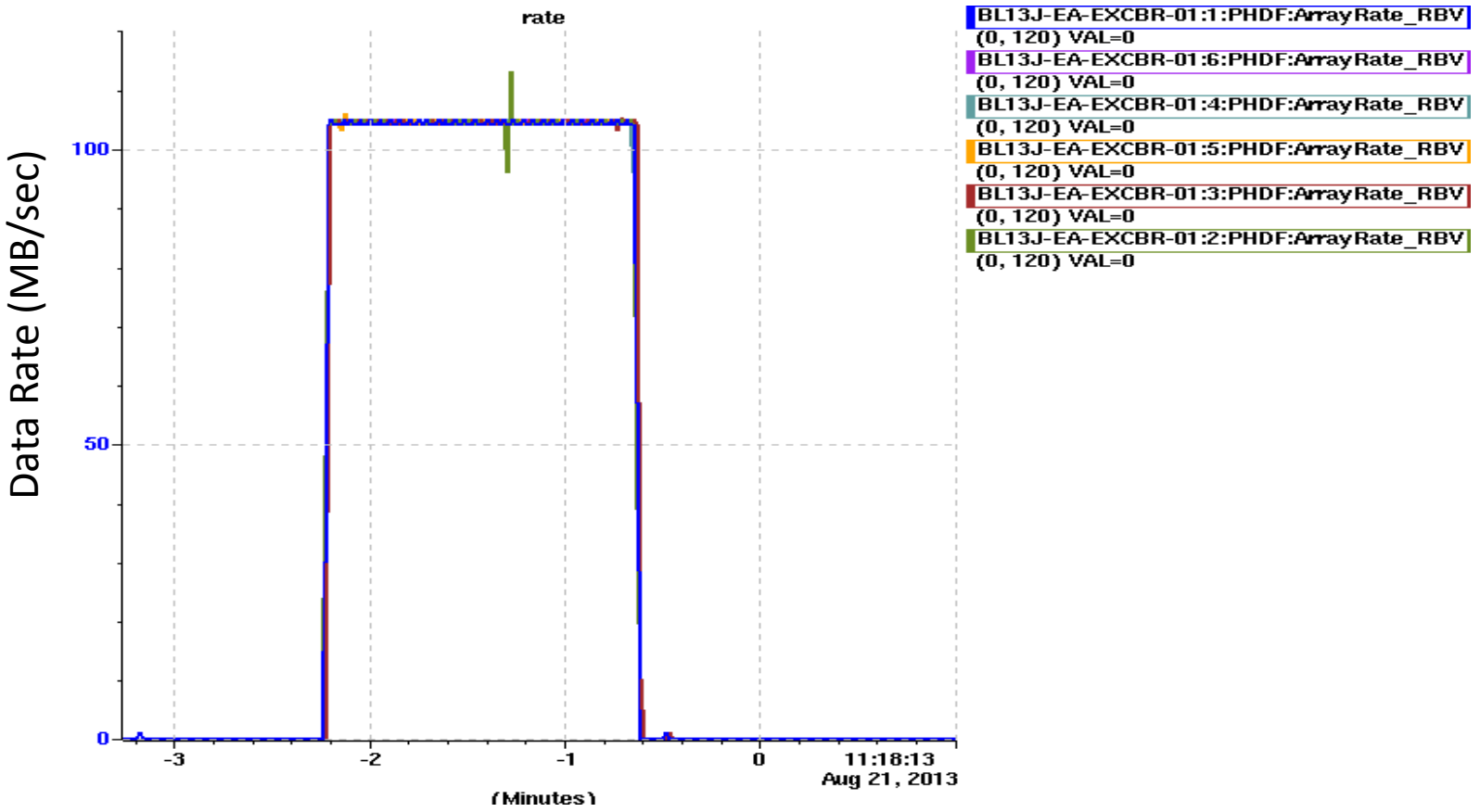


diamond

What we want – GbE nodes writing to disk



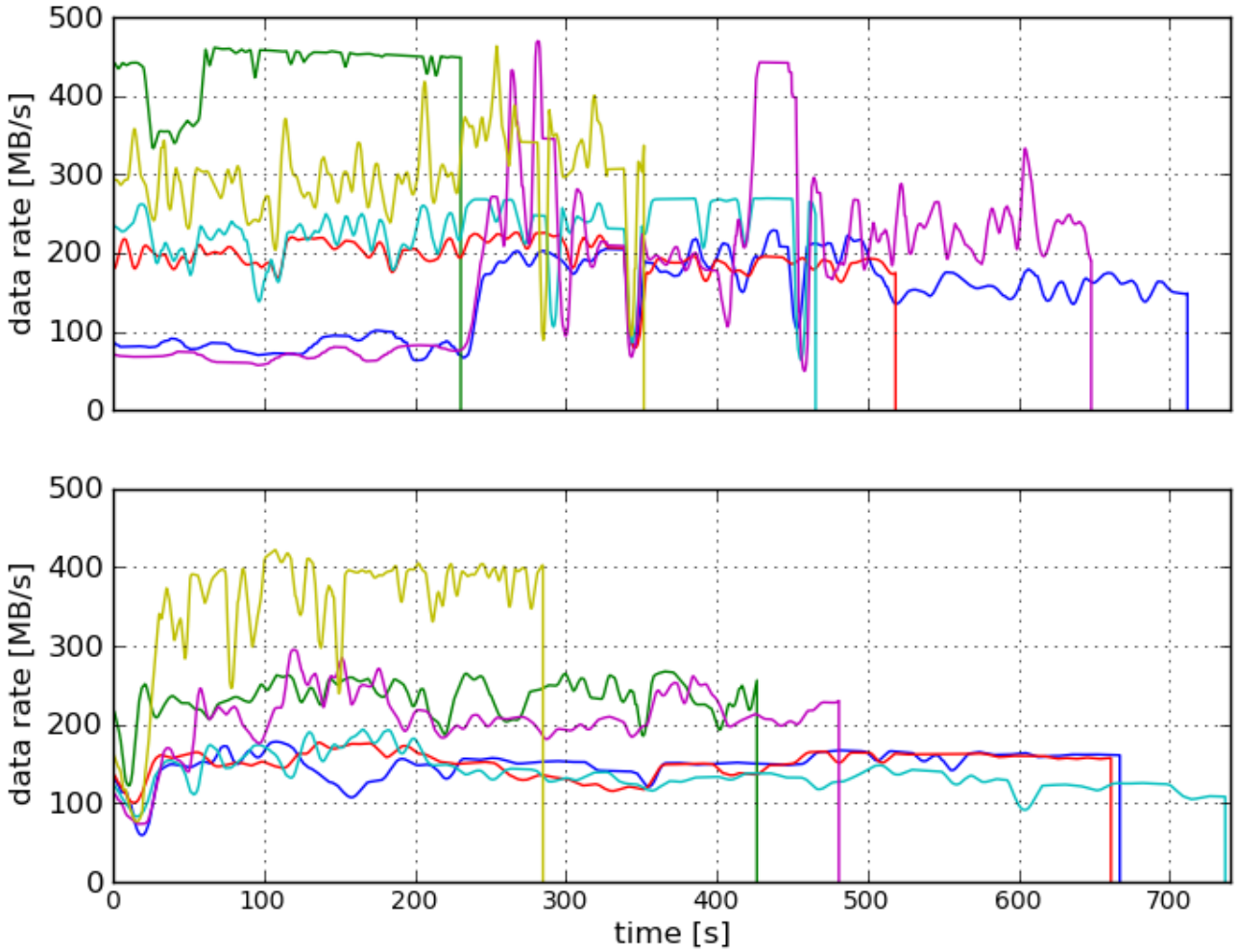
What we want – GbE nodes writing to disk



With 1 GbE links life is simple!

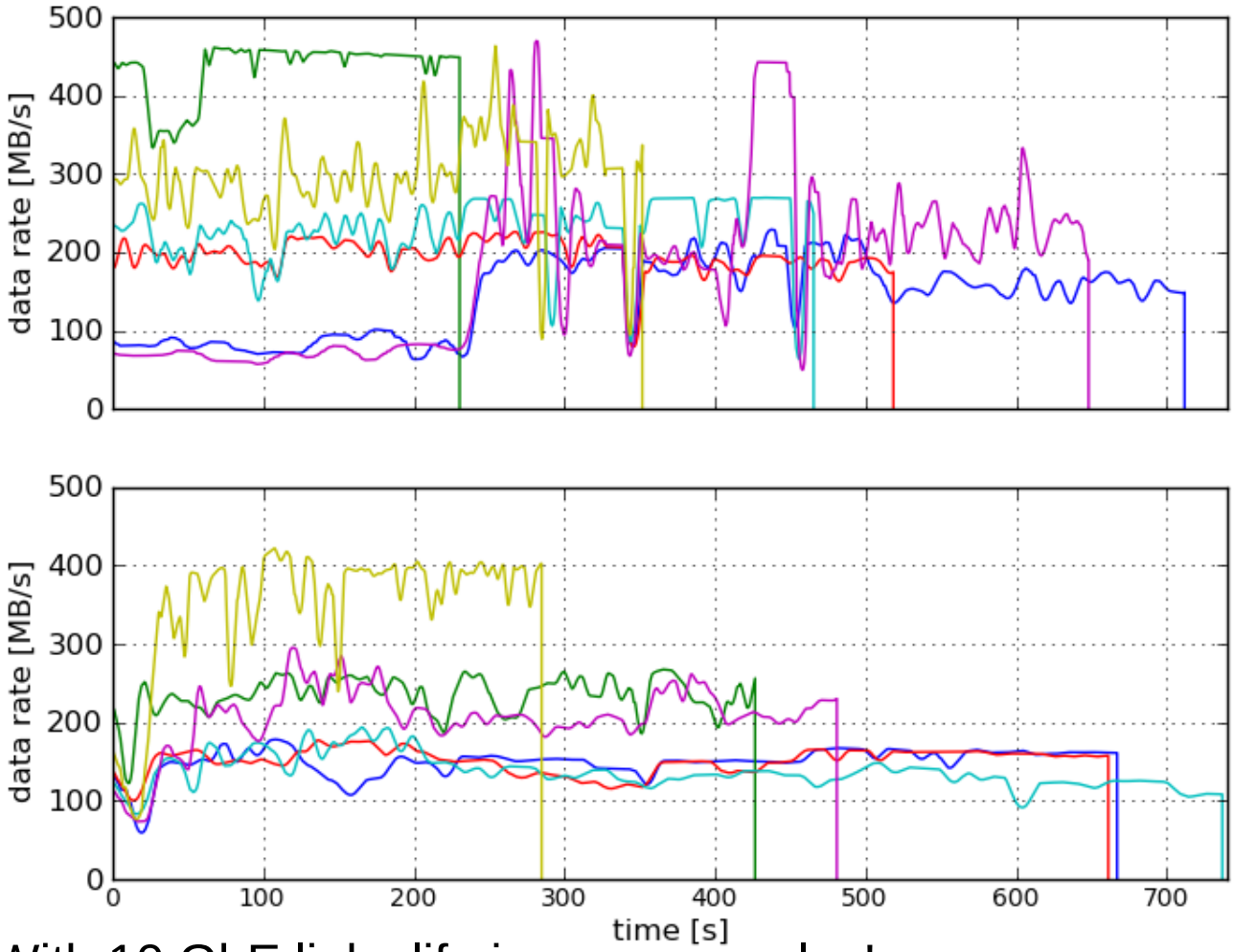
diamond

The problem – 10 GbE nodes writing to disk



diamond

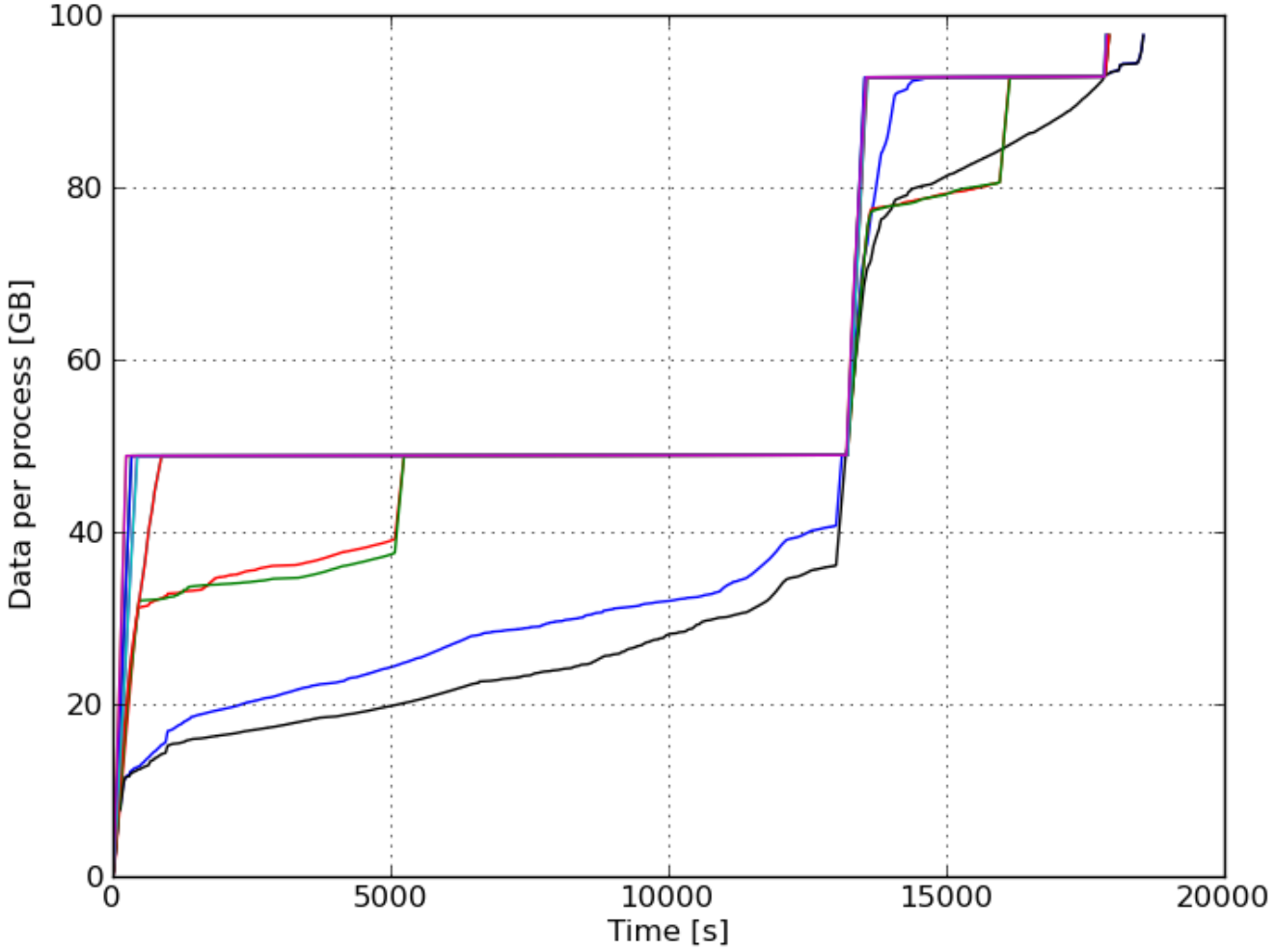
The problem – 10 GbE nodes writing to disk



- With 10 GbE links life is more complex!
- pHDF5 cannot deliver the aggregate performance of multiple independent HDF5 jobs..

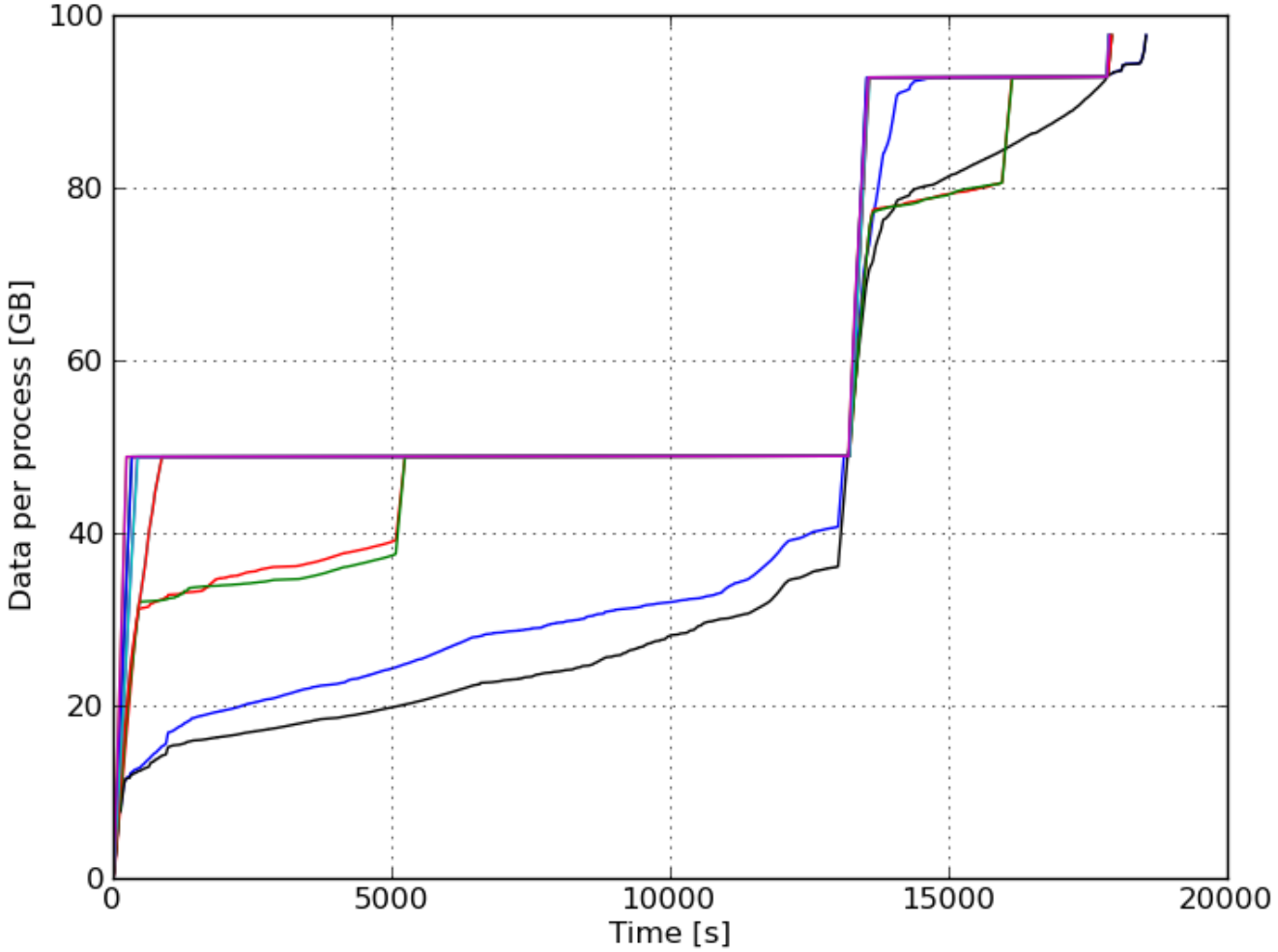
diamond

Another problem – Speeds and B-trees



diamond

Another problem – Speeds and B-trees



- pHDF5 performance can vary unexpectedly
- B-tree expansion can slow writing (v1.10 has extensible arrays for data which is unlimited in only one dimension)

diamond

HDF5 Summary

diamond

HDF5 Summary

- HDF5 is mature software that grew up in the HPC environment.
- It is a widely used standard and has the richest set of high performance functionality of any file format.
- It has some caveats we knew about:
  - HDF5 is single threaded.
  - pHDF5 relies on MPI, which doesn't happily co-exist with highly threaded architectures like EPICS.
- It has some caveats we didn't know about:
  - There are also problems with pHDF at 10 GbE speed.
  - B-Tree expansion can cause long delays

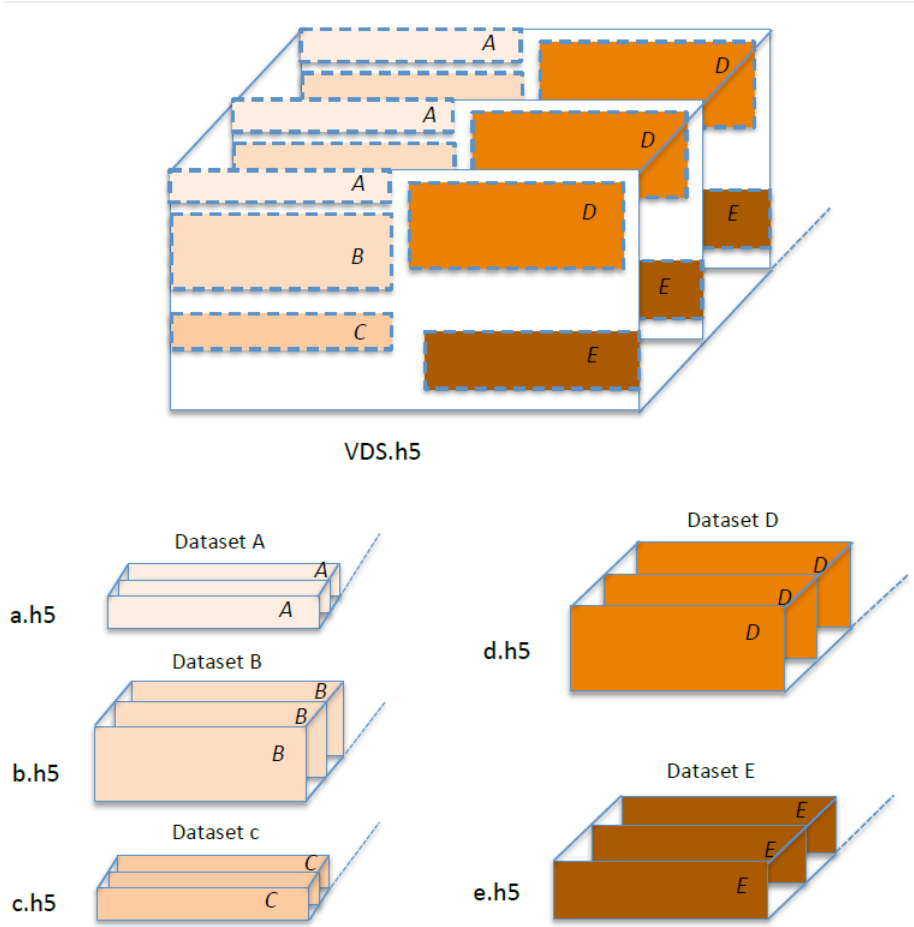
diamond

Solution: Single Writer Multiple Reader

- High speed detectors write large files to avoid the overhead of lots of small files.
- However, this means that data processing can’t start until a large amount of data has been generated.
- Single Writer Multiple Reader (SWMR) addresses this with extensions to the HDF5 file format to allow readers to have a coherent view of the file even as it is updated.
- Still requires some funds to finalise (~300k funded out of ~400k).

diamond

Solution: Virtual Dataset Concept



- Parent dataset in VDS.h5 composed of data mapped from datasets in 5 subordinate files.
- Subordinate datasets can be
  - Written independently and in parallel.
  - Compressed and chunked independently
- Parent dataset can be read as normal.
- Eliminates need for pHDF5

diamond

Summary

- HDF5 is the most fully featured high-performance file format there is.
  - But it still has issues
- GPFS has much better single process throughput than Lustre
  - But is still is slower or no faster at times.
- We are working with The HDF Group to enhance HDF5 for these applications.
  - But we would like some community help.

diamond