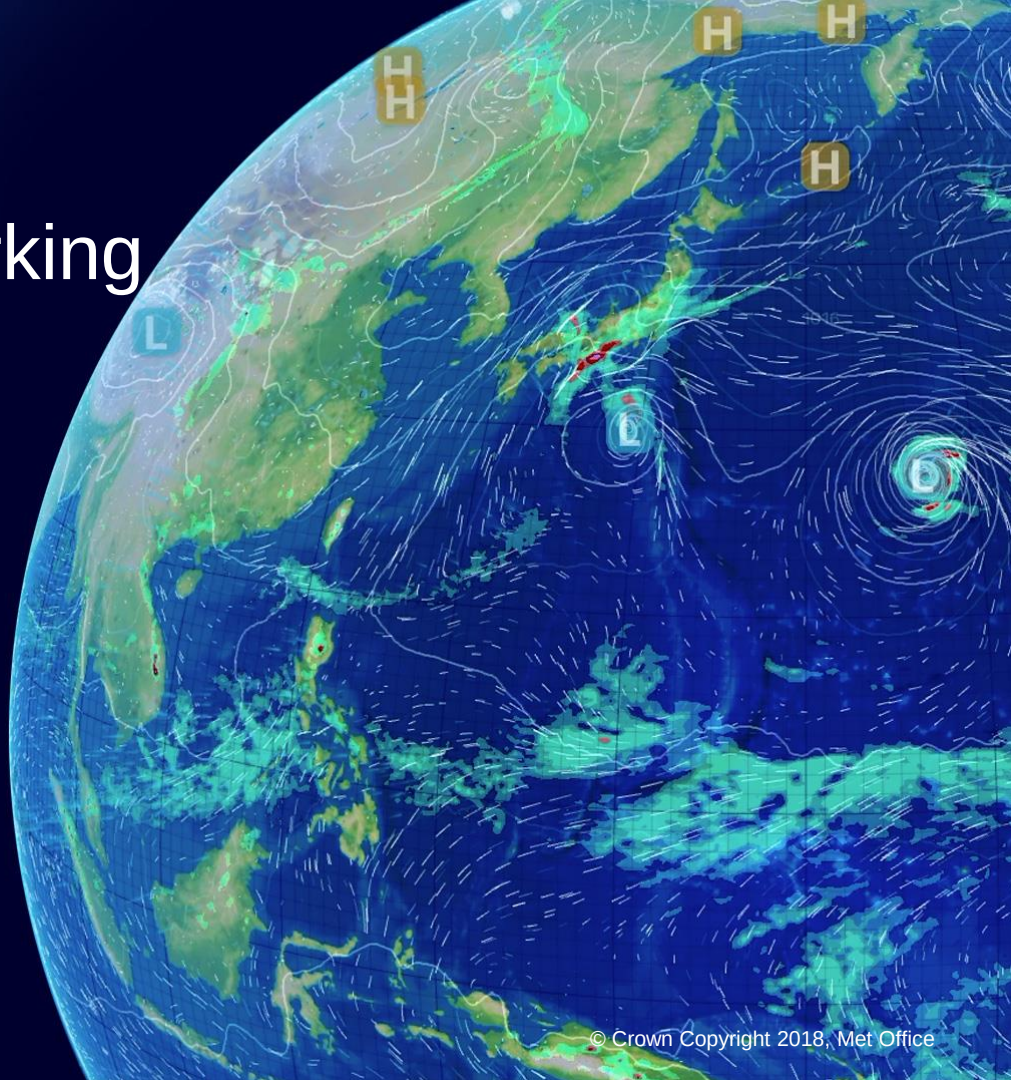
Met Office
Hadley Centre

Challenges of benchmarking HPC storage systems

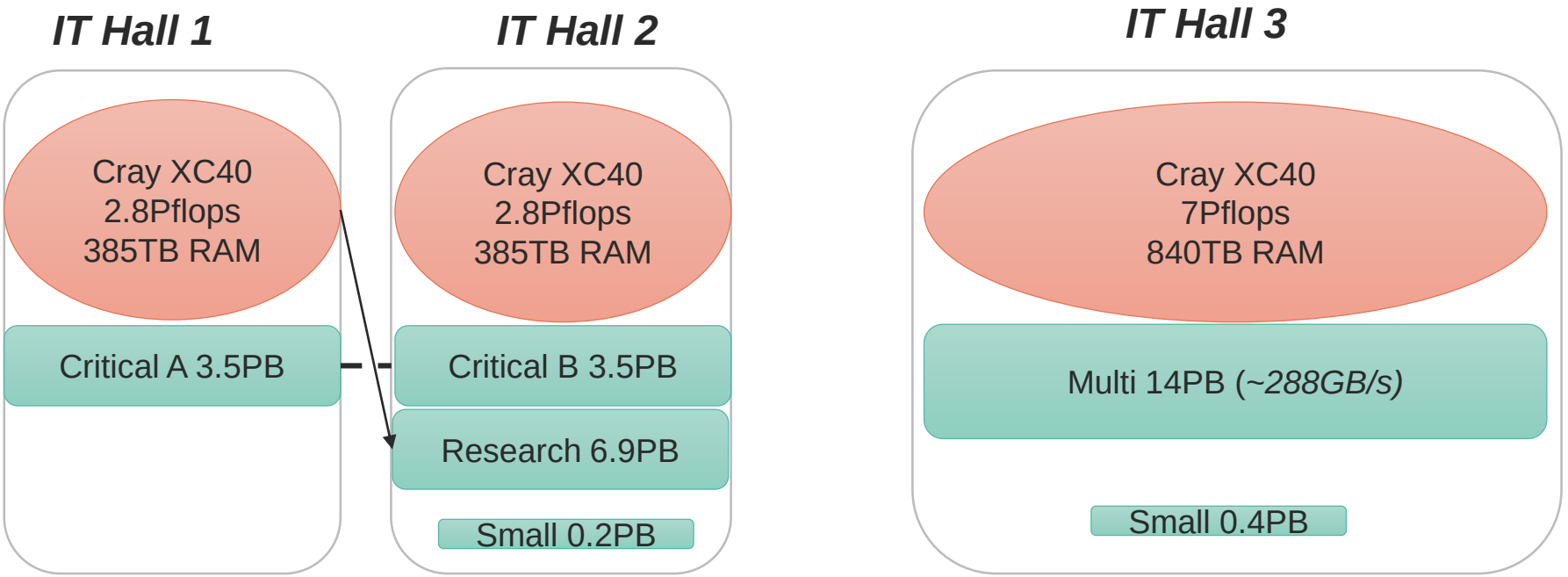
Jean-Christophe Rioual

Climate Science IT



Met Office
Hadley Centre

Met Office HPC Storage Architecture



Met Office
Hadley Centre

Met Office HPC Storage Architecture

- Hardware
  - Seagate Sonnexion Cluster
- File System
  - Lustre
  - Robinhood (scans and auto-deletion)

Met Office
Hadley Centre

Met Office HPC Storage Architecture

MASS
Long Term Tape Storage
HPSS
MOOSE Client (“ftp like”)

SPICE
Data
Processing
Platform

GPFS
3PB

Met Office
Hadley Centre

Met Office Science Workflows - IO

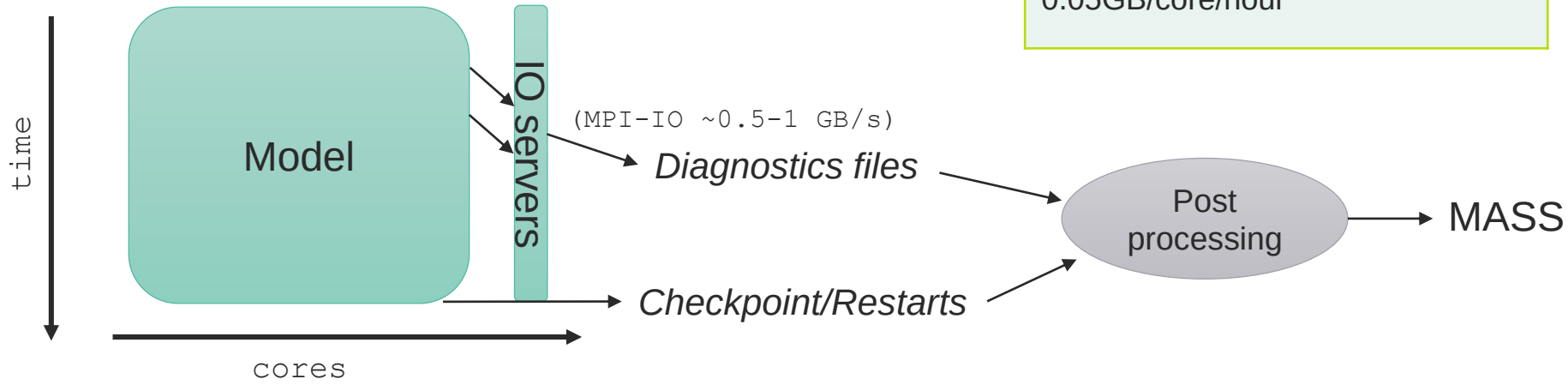
- Operational NWP
  - Models stream large files
  - Post-processing to generate products for downstream usage
    - Large number of small files.
  - Time critical capability – achieved through restricted access to the FS
  - Resiliency requirements – achieved through duplication and mirroring
- Climate Modelling
  - Long runs
  - Ensembles / bandwidth capacity workflows
  - File conversions/compressions but ~same number of files/volume
  - Data transient on file system

Met Office
Hadley Centre

Climate Models IO

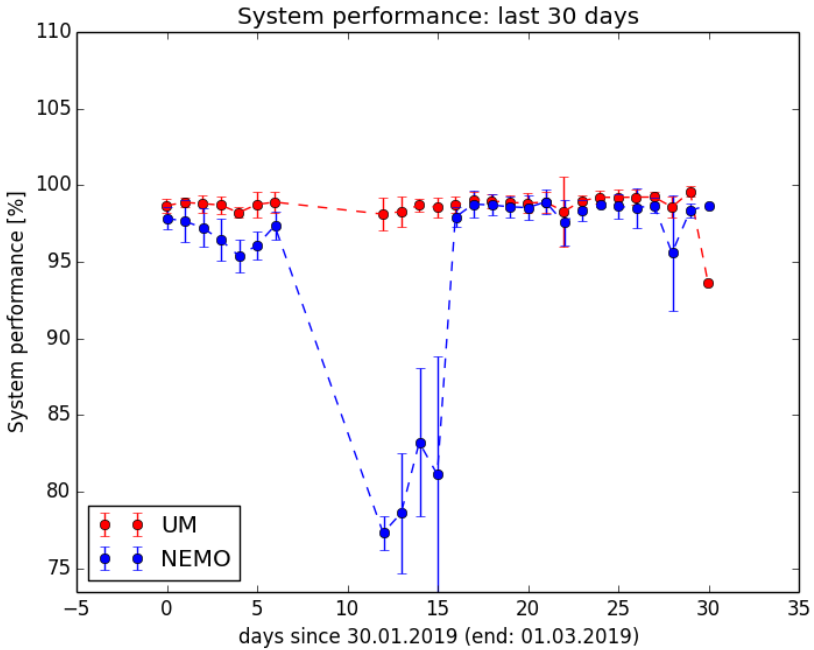
- 2 Building blocks
  - UM – Atmosphere-Land-Chemistry modelling
  - NEMO – Ocean-Ice-Biogeochemistry modelling

120 nodes simulation
10TB/day
0.05GB/core/hour



Met Office
Hadley Centre

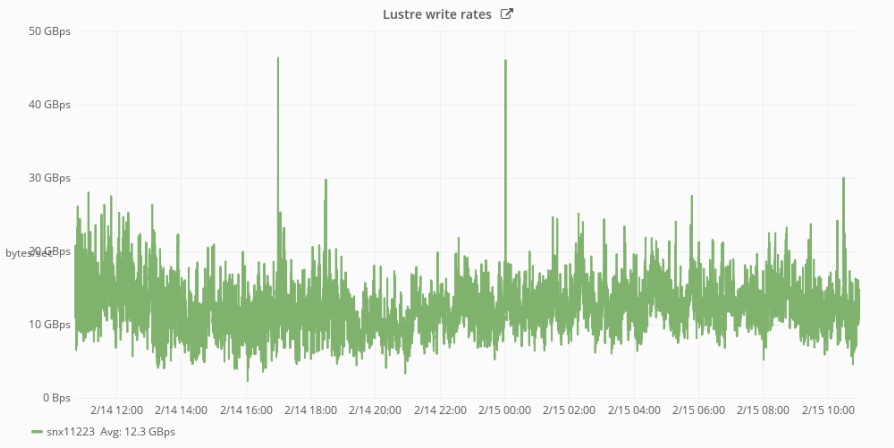
Application Performance

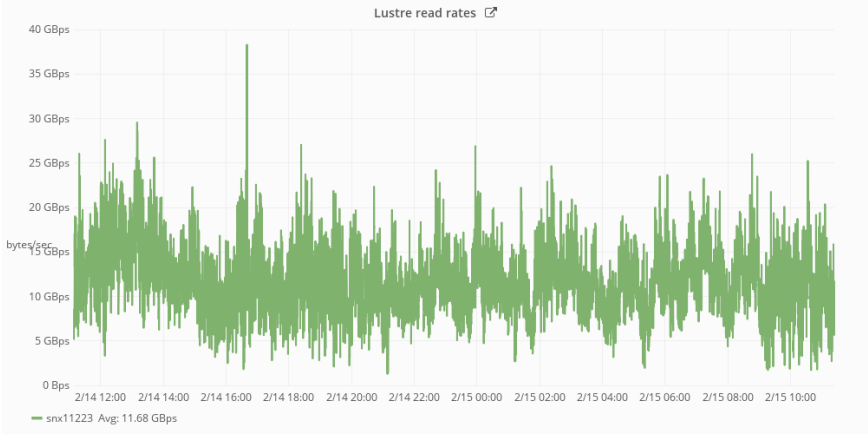


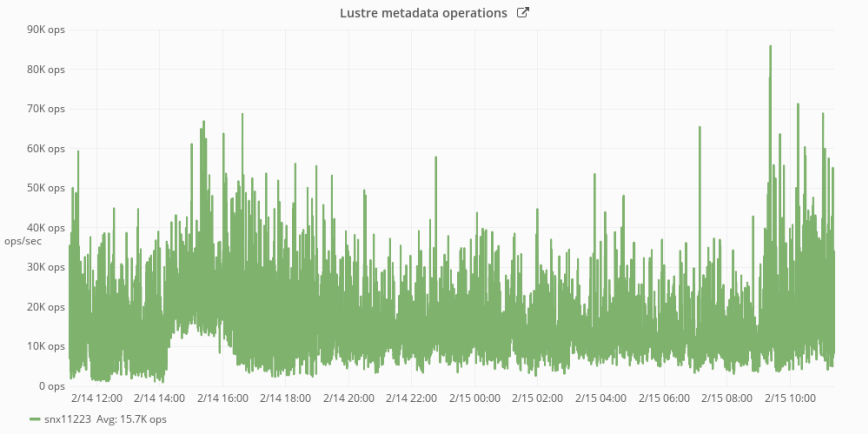
RAID checks

Met Office
Hadley Centre

Lustre Performance

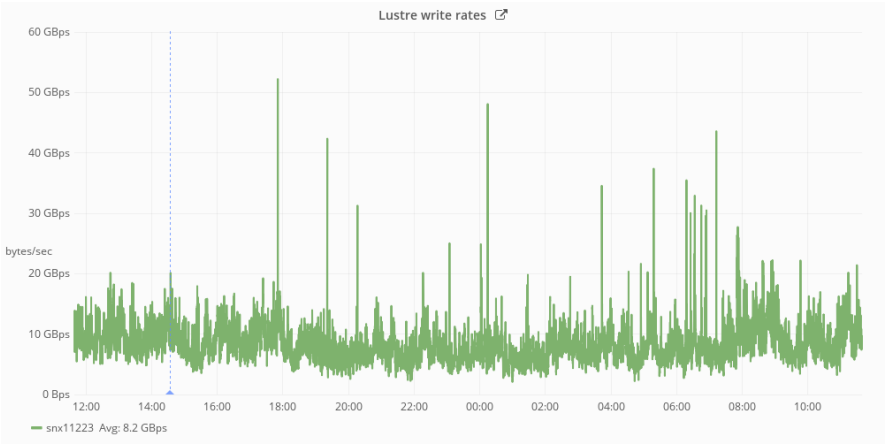


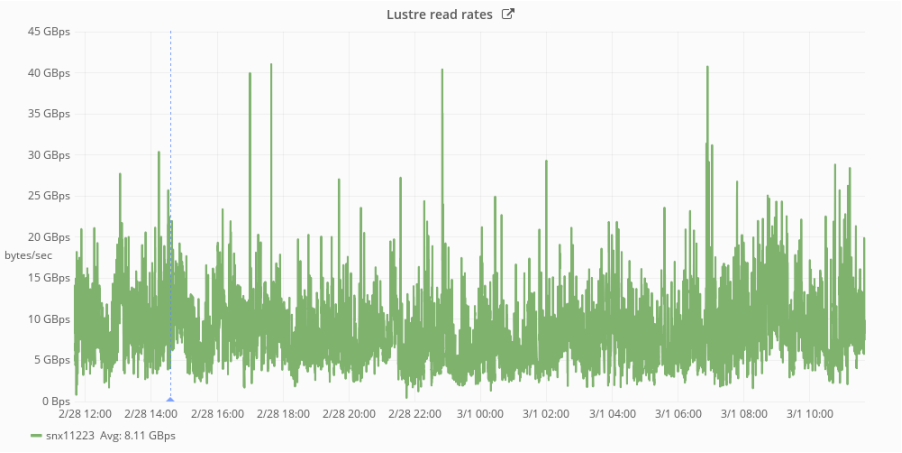


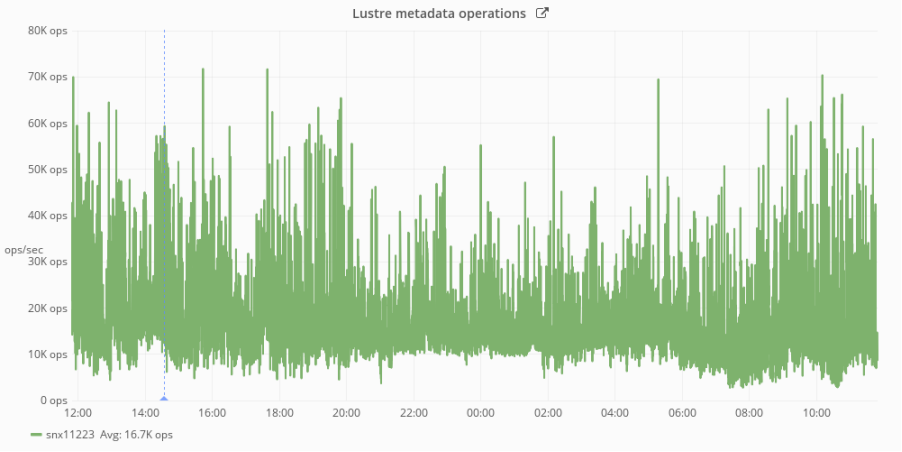


Met Office
Hadley Centre

Lustre performance







Met Office
Hadley Centre

Meta-data performance

Some workflows more stressful on meta-data
(NWP post-processing)

We expect Python workloads to increase on our HPC systems

- Performance characteristics not that well understood
- New technologies (containers, Dask, etc…)
- Flash will help ?

Met Office
Hadley Centre

Understanding the storage system

A good understanding of the strengths/weaknesses of the current system is important for writing good requirements for future systems

- End users would benefit from better tools
- System administration tools are not enough
- Linking particular applications/jobs to system events/degradation

Met Office
Hadley Centre

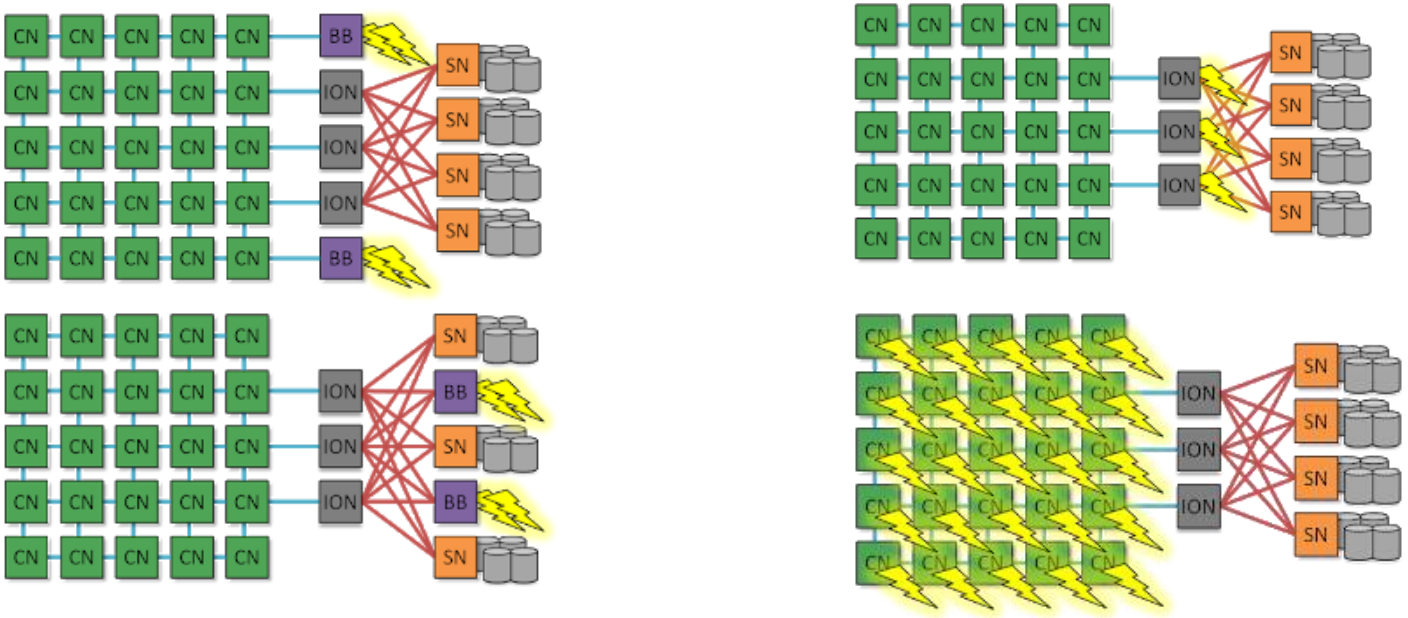
IO Benchmarking

"Classical" approach : mix of application benchmarks + IOR + MDTEST

- Application benchmarks (UM, NEMO)
  - Requirement 1 : Application Benchmarks Perf > Baseline
  - Capture capability MPI-IO bandwidth, Netcdf performance
  - Checkpoint restart speed
- IOR and MDTEST
  - Requirement 2 :
    - IOR (read/write) > N1 GB/s
    - MDTEST (file create/delete)> N2 kOPS
  - Captures capacity

Met Office
Hadley Centre

Accelerators



Glen Lockwood https://glennklockwood.blogspot.com/2017/03/reviewing-state-of-art-of-burst-buffers.html

Met Office
Hadley Centre

- In an ideal world, clients provide benchmark requirements and vendors/OEMs come back with an architecture.

- The “classical” approach (IOR/MDTEST) is not informative for
  - determining architectural choices
  - sizing (potential) acceleration layers
  - capturing some of the advantages of burst buffers
    - (Transient data in workflows)

Met Office
Hadley Centre

We (customers/users) are probably all facing similar problems

Communication between HPC centres can be complicated by commercial considerations.

Can SIGIO be a forum to discuss these issues in an open forum ?

Let’s exchange ideas !