GridPP

UK Computing for Particle Physics

Next Generation Tier 1 Storage

Shaun de Witt (STFC)
With Contributions from:
James Adams, Rob Appleyard, Ian Collier,
Brian Davies, Matthew Viljoen

HEPiX Beijing 16th October 2012

GridPP
UK Computing for Particle Physics

Introduction

- Why are we doing this?
- What were the evaluation criteria?
- Candidates
- Selections
  - And omissions
- Tests
- Status
- Timeline
  - Aim for production service for 2014 data run

Science & Technology Facilities Council

GridPP
UK Computing for Particle Physics

Why are we doing this

- CASTOR is working well for us **but:**
  - CASTOR optimised for many disk servers per pool
    - Getting harder as ‘cost optimal’ size is getting larger & we have many storage pools
  - Scheduling overhead/’hot spotting’
    - TransferManager (LSF replacement) has improved this A LOT
  - Oracle Costs!
  - Nameserver is Single Point of Failure
  - Not mountable file system
    - Limits take-up outside of WLCG/HEP(Matters as we also use Castor for storage for other STFC user groups)
  - Requires (quite a lot of) specific expertise
- Disk-only not very widely deployed now
  - EOS (CERN ATLAS+CMS), DPM (ASGC)
  - Could cause delays in support resolution
  - Reduced ‘community support’
  - Greater risk in meeting future requirements for disk-only

Science & Technology Facilities Council

Criteria

- ‘Mandatory’:
  - **Must** make more effective use of existing h/w
  - **Must not** restrict hardware purchasing
  - **Must** be able to use ANY commodity disk
  - **Must** support end-to-end checksumming
  - **Must** support ADLER32 checksums
  - **Must** support xrootd and gridFTP
  - **Must** support X509 authentication
  - **Must** support ACLs
  - **Must** be scalable to 100s Petabytes
  - **Must** scale to > 10^{12} files
  - **Must** at least match I/O performance of CASTOR

Criteria

- Desirable
  - **Should** provide NFS mountable file system
  - **Should** be resilient to hardware failure (disk/memory/node)
  - **Should** support checksum algorithms other than ADLER32
  - **Should** be independent of licensed software
  - Any required database **should** be supported by STFC DB Team
  - **Should** provide an HTTP or WebDAV interface
  - **Should** support SRM interface
  - **Should** provide a POSIX interface and *file* protocol
  - **Should** support username/password authentication
  - **Should** support kerberos authentication
  - **Should** support ‘hot file’ or ‘hot block’ replication

GridPP
UK Computing for Particle Physics

What else are we looking for?

- Draining
  - Speed up deployment/decommissioning of new hardware
- Removal
  - Great if you can get data in fast, but if deletes are slow...
- Directory entries
  - We already know experiments have large numbers of files in each directory
  - Need support for lots
- Support
  - Should have good support and wider usage
- IPv6 support
  - Not on roadmap for Castor
  - RAL & Tier 1 starting to look at IPv6

Science & Technology Facilities Council

Candidates

- HEP (ish) solutions
  - dCache, DPM, STORM+Lustre, AFS
- Parallel and ‘Scale-Out’ Solutions
  - HDFS, OpenStack, OrangeFS, MooseFS, Ceph, Tahoe-LAFS, XtreemFS, GfamrFS, GlustreFS,GPFS
- Integrated Solutions
  - IBM SONAS, Isilon, Panasas, DDN, NetApp

- Paper (twiki) based review carried out
  - plus using ‘anecdotal’/reported experience where we know sites that run things in production (i.e. reports at HEPiX etc.)

Selection

HEP

	POSIX	SRM	HTTP	NFS	WebDAV	xroot	CDMI	HW Flexibility[1]	HW Loss[2]	Distributed Metadata	Automated Replicas (Hotfiling)	End-to-end Checksumming	Notes
CASTOR	✗	✓	✗	✗	✗	✓	✗	✗	✗	✗	✗	✓	*
dCache	✓	✓	✓	✓	✓	✓	cdmi	✓	hwl	✗	✓	✓	*
DPM	✓[4]	✓	✓	✓[4]	✓	✓	cdmi	✓[3]	hwl	✗	✗	✓	*
STORM + Lustre	✓	✓	http	✓	webDAV	✓	cdmi	✓	✗	✗	ar	✓	Lustre uses dedicated metadata servers, but their load is limited because they only do pathname and permission checks.
BESTMan + POSIX FS	✓	✓	http	nfs	webDAV	xroot	cdmi	hwf	hwl	dm	ar	✗	*
AFS	✓	srm	http	✓	webDav	xroot	cdmi	hwf	hwl	dm	ar	✗	*

Notes:

1. Allows allocation of space on scales smaller than a single disk server, lacks vendor lock-in
2. Is the system resilient to loss of hardware; is there protection against data loss associated with normal hardware failure at the disk server level
3. Minimum allocatable size is 1 file system

Selections

- dCache
  - Still has SPoFs
- CEPH
  - Still under development; no lifecycle checksumming
- HDFS
  - Partial POSIX support using FUSE
- orangeFS
  - No file replication, fault tolerant version in preparation
- Lustre
  - Requires server kernel patch (although latest doesn't), no hot file replication?

GridPP
UK Computing for Particle Physics

Some rejections...

- DPM
  - Well known and widely deployed within HEP
  - No hot-file replication, SPoFs, NFS interface not yet stable
  - Some questions about scalability
- AFS
  - Massively scalable (>25k clients)
  - Not deployed in ‘high throughput production’, cumbersome to administer, security
- GPFS
  - Excellent performance
  - To get all required features, locked into IBM hardware, licensed
- EOS
  - Good performance, auto file replication
  - Limited support

Science & Technology Facilities Council

Tests

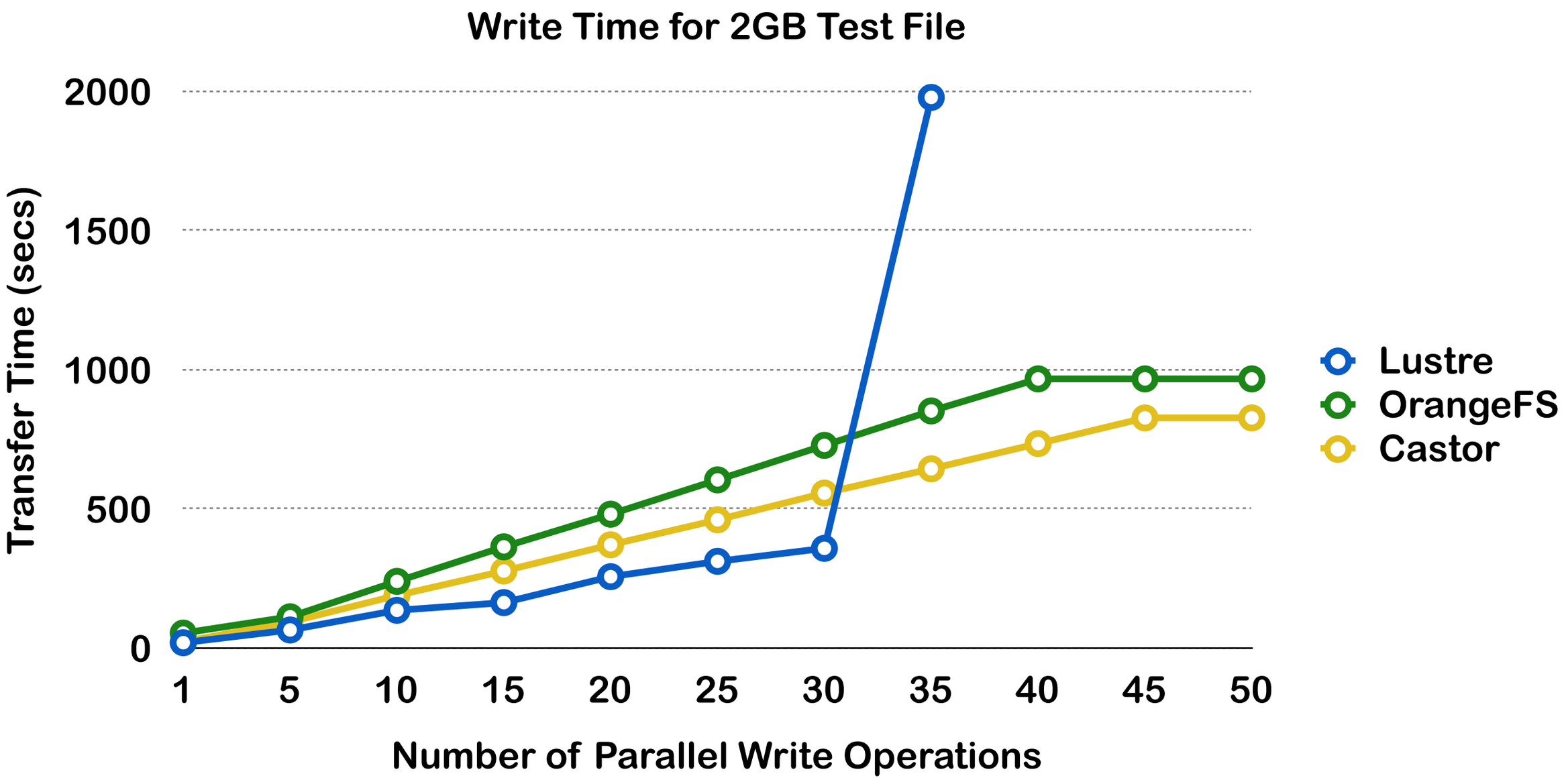
- IOZone tests
  - A-la disk server acceptance tests
- Read/Write throughput tests (sequential and parallel)
  - File/gridFTP/xroot
- Deletions
- Draining
- Fault tolerance
  - Rebuild times

Status

- dCache
  - Under deployment. Testing not yet started
- Lustre
  - Deployed
  - IOZone tests complete, functional tests ongoing
- OrangeFS
  - Deployed
  - IOZone tests complete, functional tests ongoing
- CEPH
  - RPMs built, being deployed.
- HDFS
  - Installation complete, Problems running IOZone tests, other tests ongoing

GridPP
UK Computing for Particle Physics

Results (Preliminary)



Note Lustre hitting a wall doing parallel writes

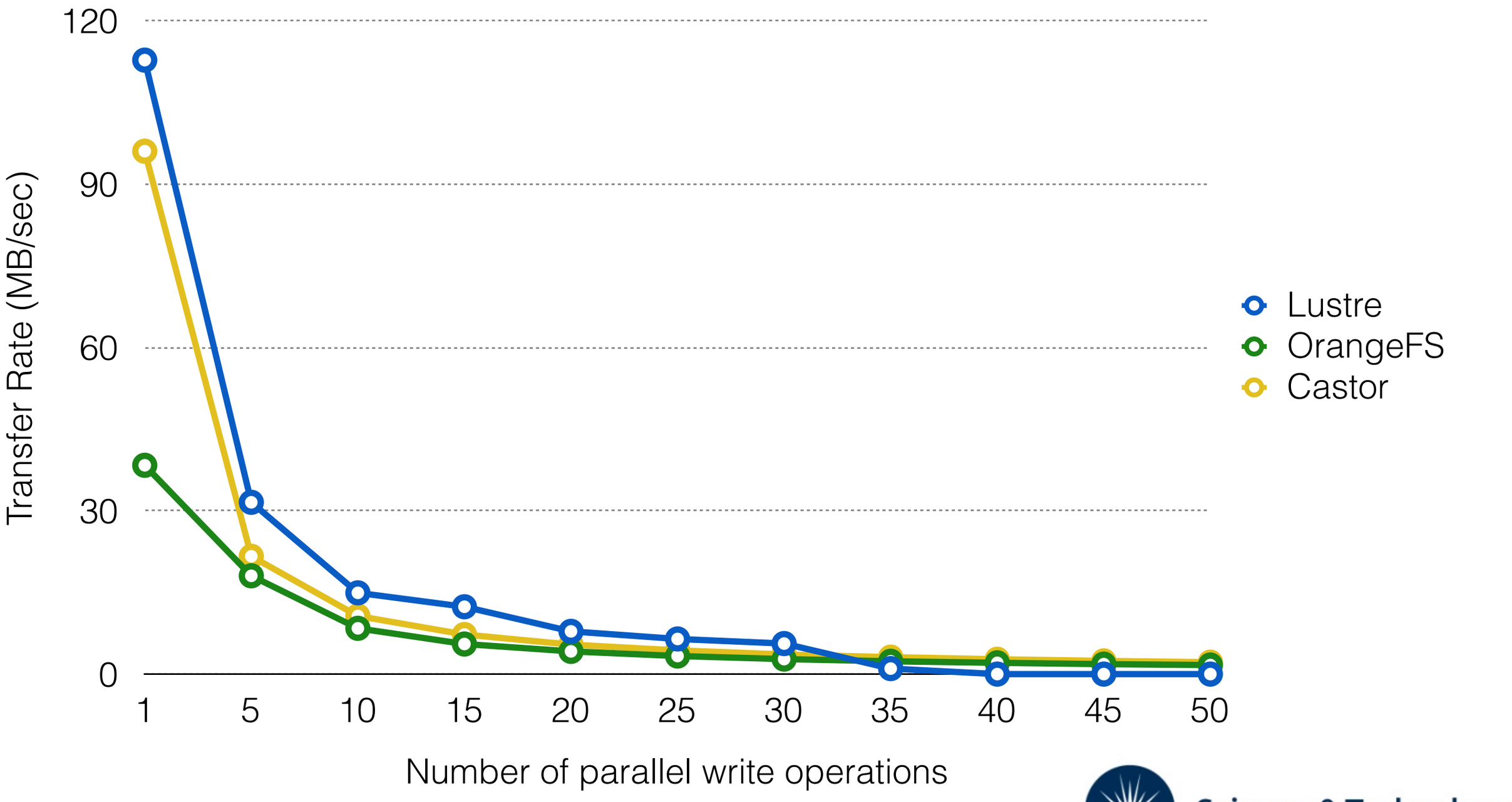
- Not entirely understood yet (have some hints from other sites
- Assume this is a setup/tuning issue
- Exactly the kind of thing that could disqualify it if we can't fix

Science & Technology Facilities Council

GridPP
UK Computing for Particle Physics

Results (Preliminary)

Write Rates for 2GB Source File

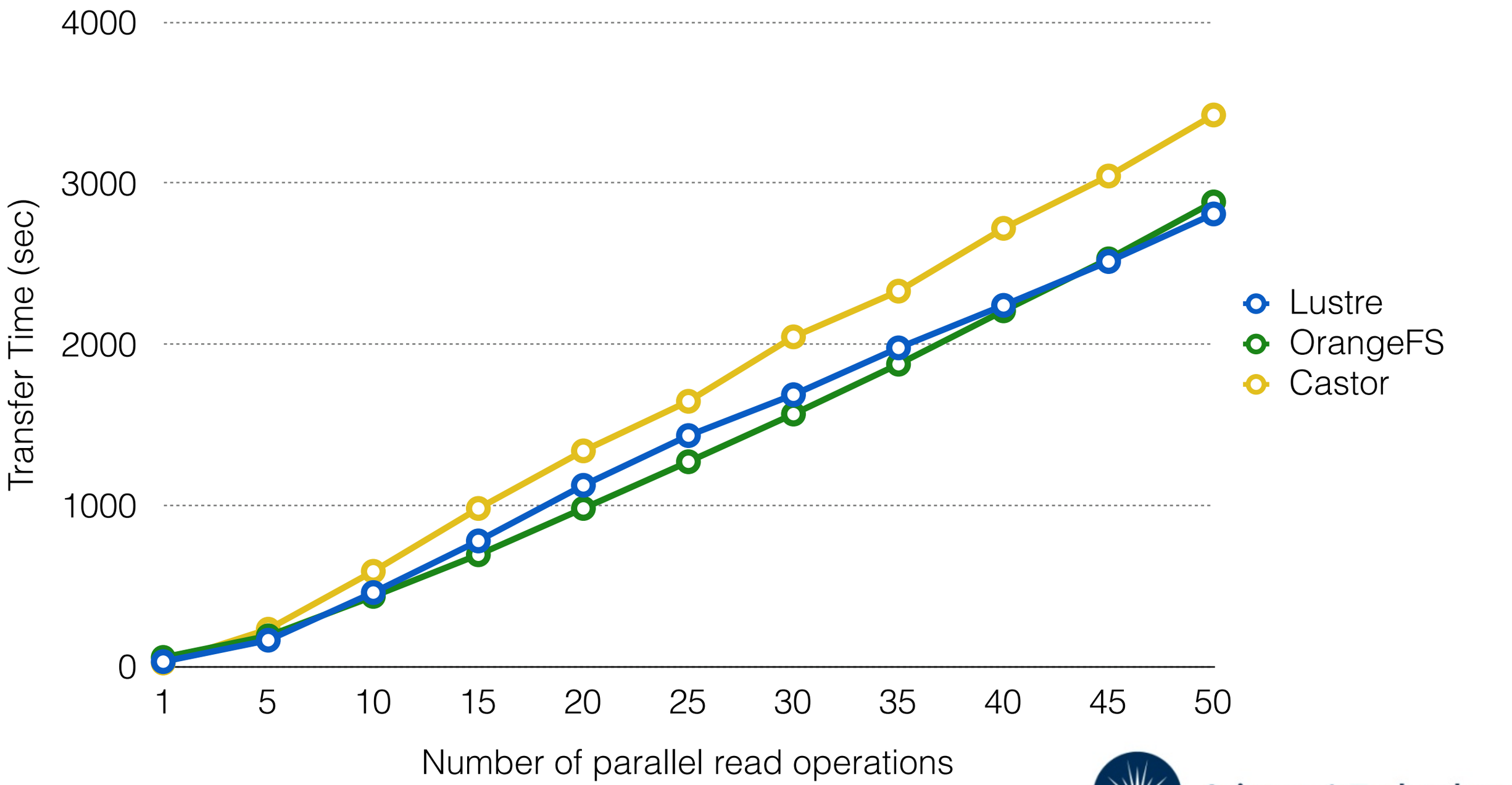


Science & Technology Facilities Council

GridPP
UK Computing for Particle Physics

Results (Preliminary)

Read Time for 2GB Source File

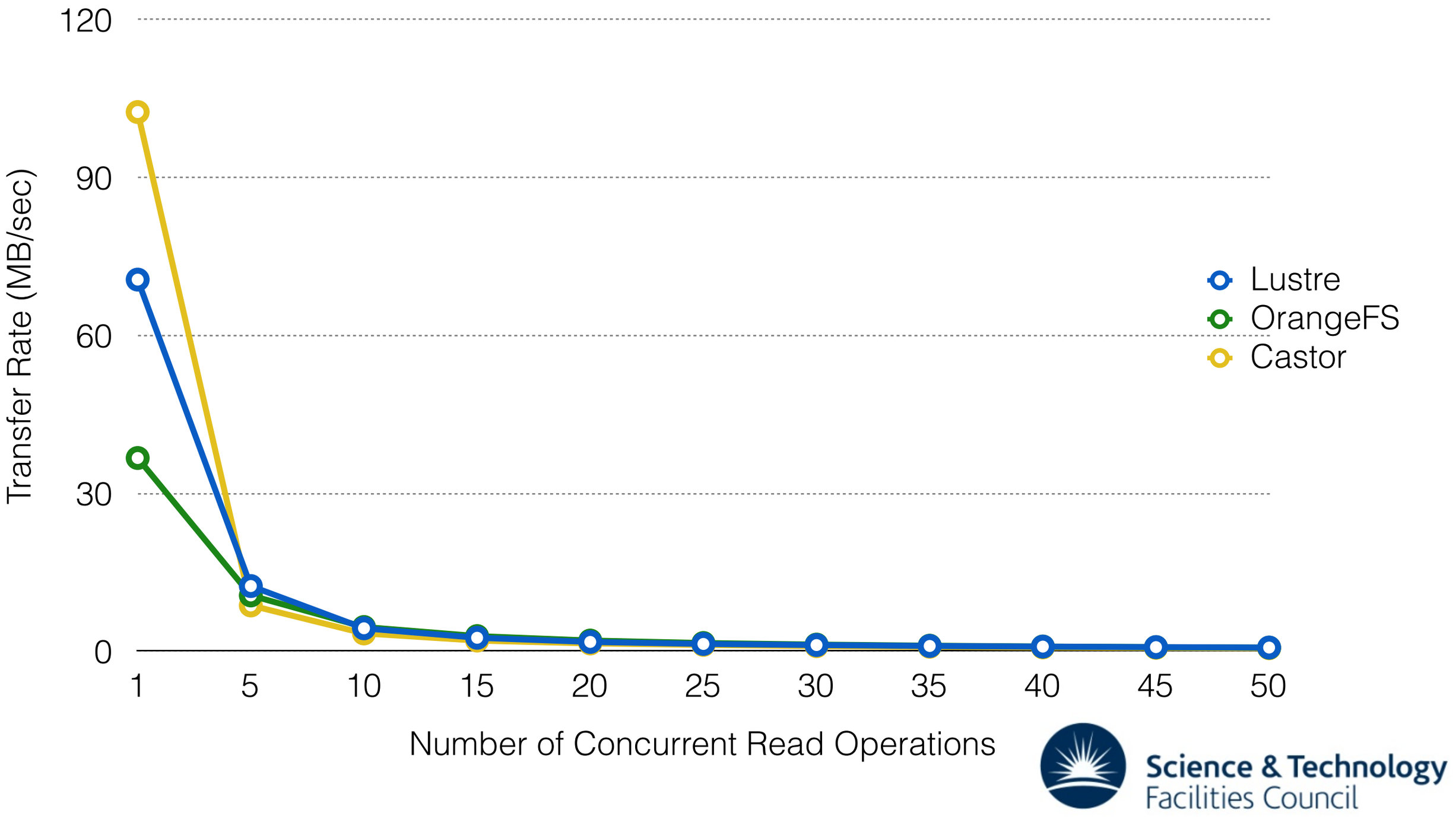


Science & Technology Facilities Council

GridPP
UK Computing for Particle Physics

Results (Preliminary)

Read Rate for 2GB Source Files



Science & Technology Facilities Council

Results

- Very Provisional so far
  - Castor rather well tuned at RAL
  - Lustre & OrangeFS hardly tuned
- Non-binding summary
  - OrangeFS & Ceph
    - look promising in long term but immature
  - dCache
    - surely could do most of what we need
    - Still file based
  - Lustre
    - Promising
    - Like that it is block based
    - Like no SPoFs
    - Stable
      - Could live with occasional downtimes for upgrades

Timeline

- Provisional
  - Dec 2012 – Final Selection (primary and reserve)
  - Mar 2013 – Pre-Production Instance available
    - WAN testing and tuning
  - Jun 2013 – Production Instance
- Depends on...
  - Hardware availability
  - Quattor Profiling - configurartion should be less complex than CASTOR
  - Results from test instances
- Open Questions:
  - One large instance or multiple smaller ones as now?
    - Large instance with ‘dynamic’ quotas has some attractions
  - Migration from CASTOR
    - Could just use lifetime of hardware