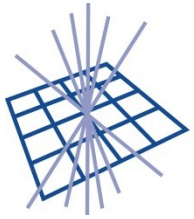
Science & Technology Facilities Council
Rutherford Appleton Laboratory



GridPP
UK Computing for Particle Physics

Cross Site Data Movement: The UK as a “Data Grid”

Jens Jensen, Mad Scientist
Scientific Computing Dept
UKRI-STFC
March 2019

Context

- GridPP – UK grid for particle physics
  - STFC-funded; infrastructure spans ~20 sites in UK
  - Connected into WLCG
- IRIS – STFC funded research
- This talk focuses on data and "plumbing"
  - Not compute
  - Nor metadata, nor information systems
  - Nor storage accounting

Note these slides will work best as powerpoint because they have some animations

Context

Google map of WLCG sites (source: WLCG, resp. Google…)
WLCG == Worldwide LHC Computing Grid
LHC == Large Hadron Collider
Clearly weighted towards northern hemisphere but a global endeavour

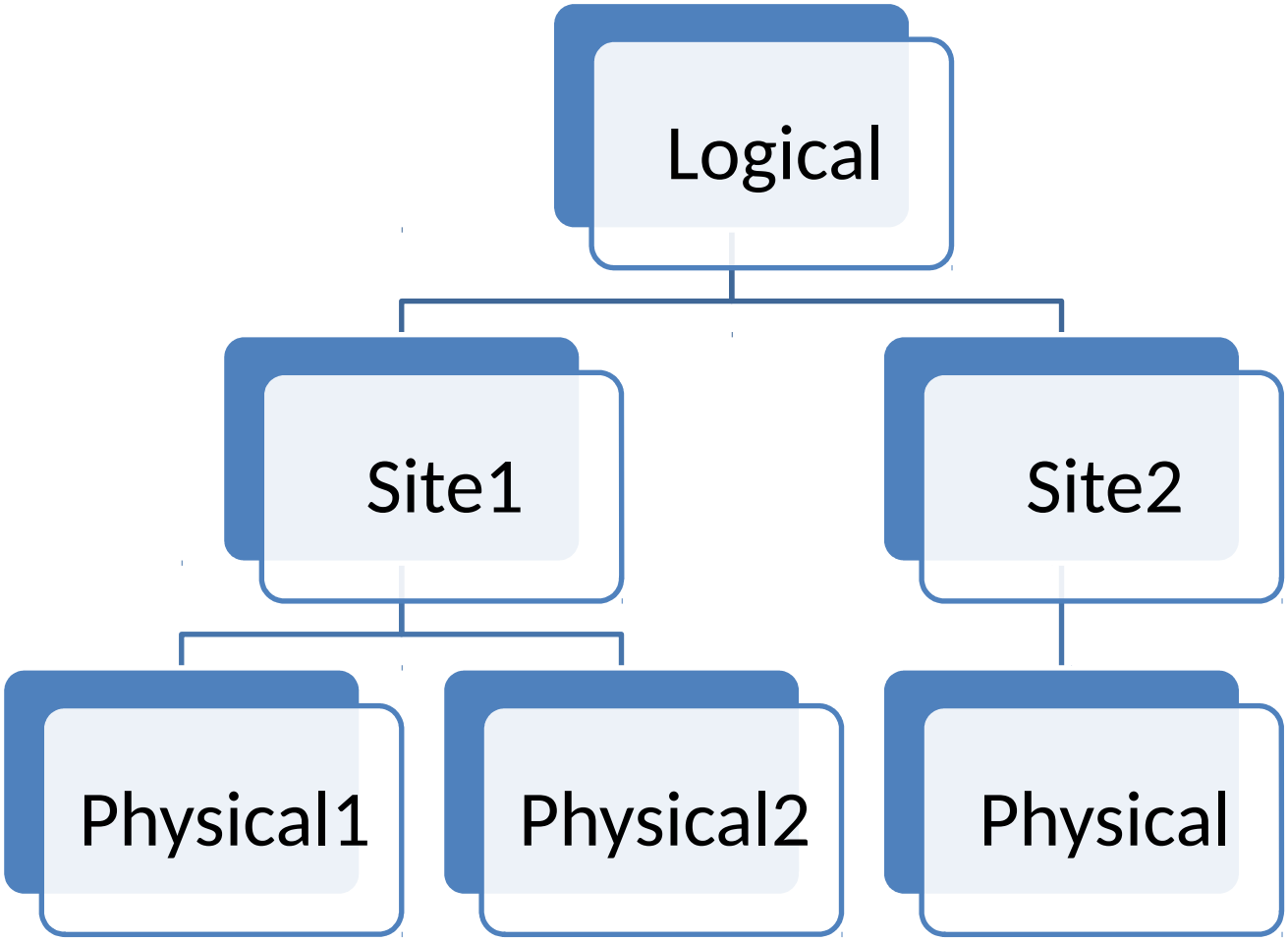
Context

Similar map of UK: GridPP
(Source: WLCG, and Google for the map…)

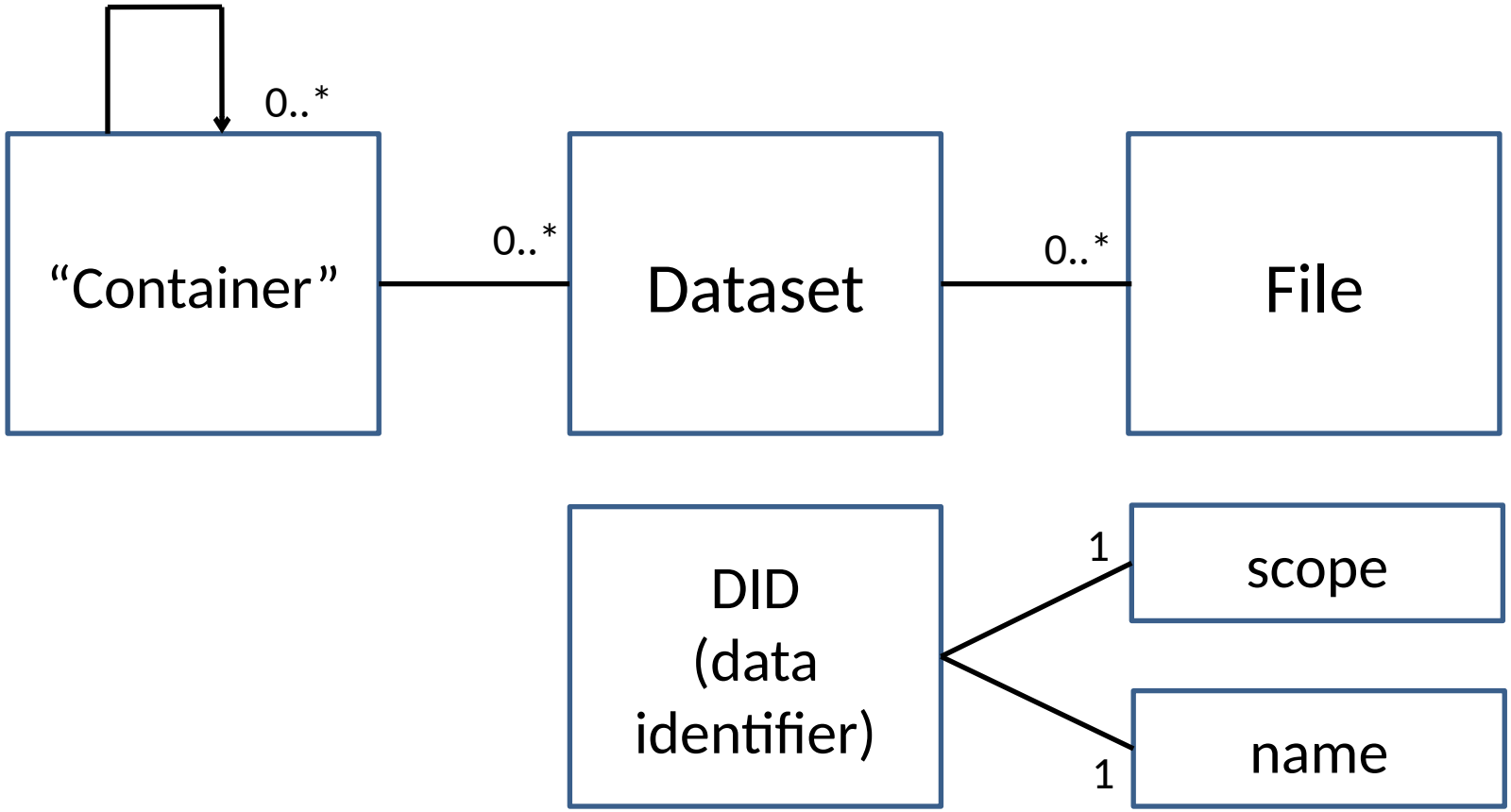
But it's not just the LHC
LIGO, LSST, SKA, DUNE, T2K, …

WLCG is mostly HTC but other users need more HPC or big memory machines

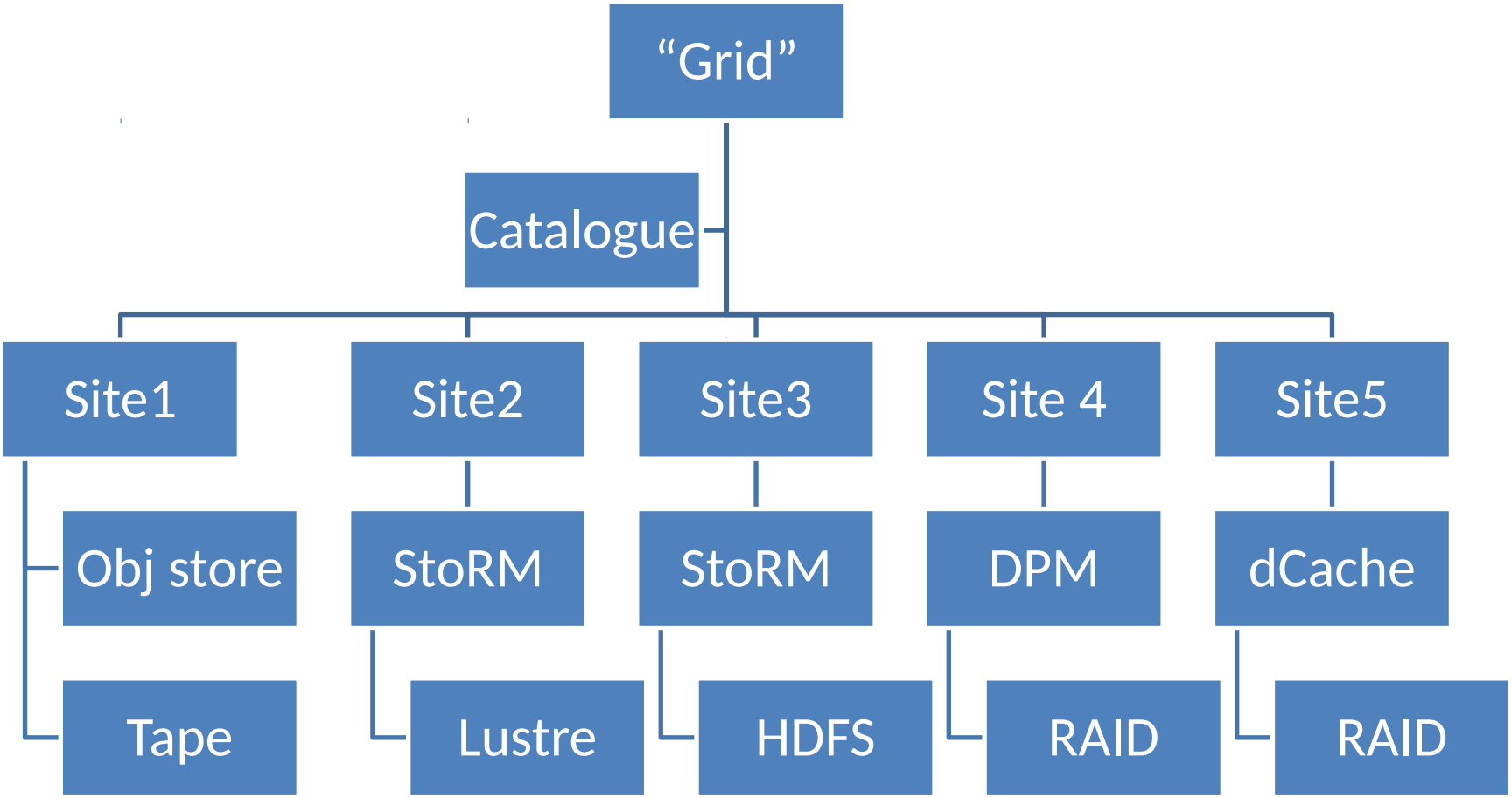
File Structure



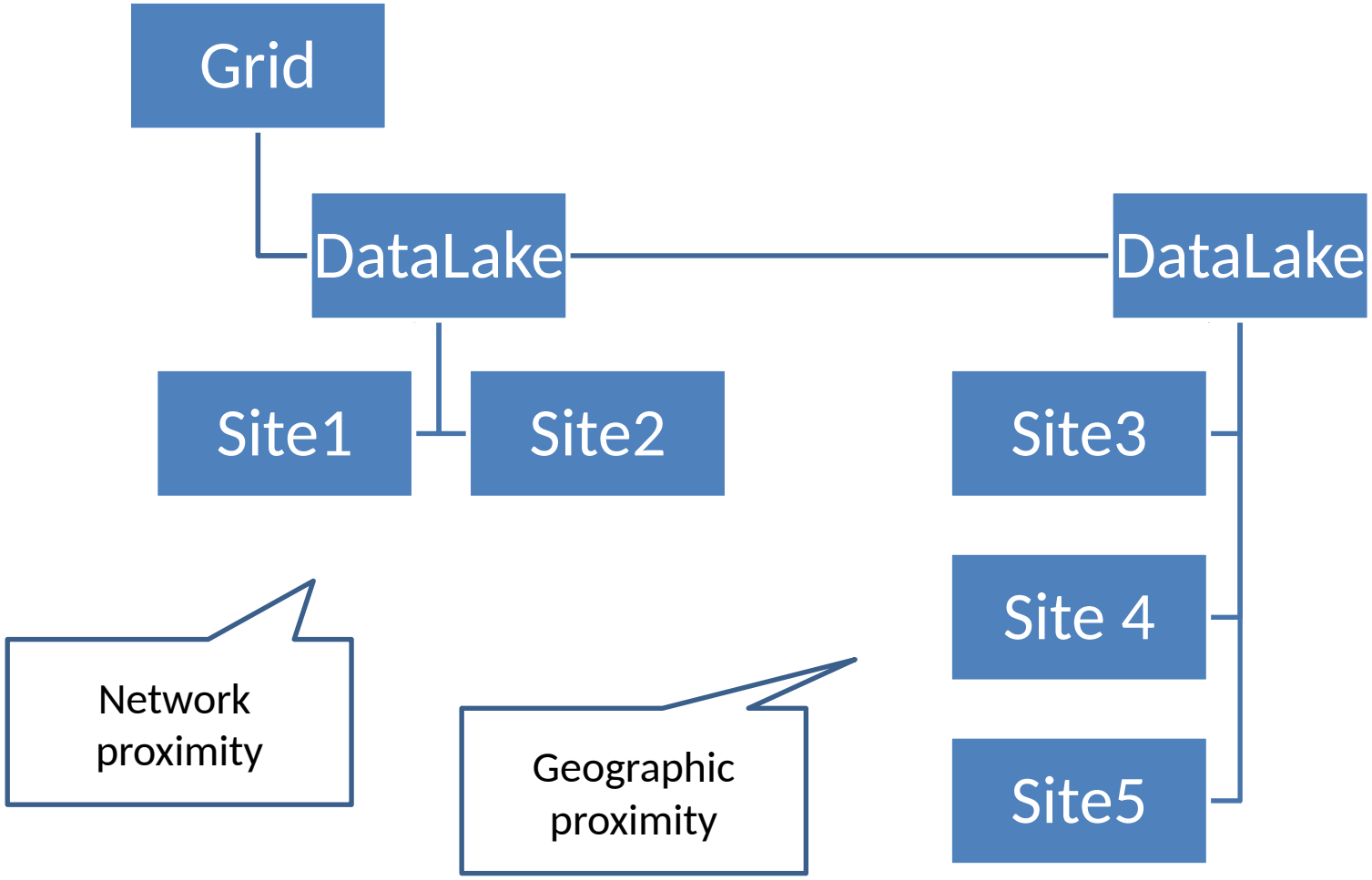
Dataset Structure (Rucio)



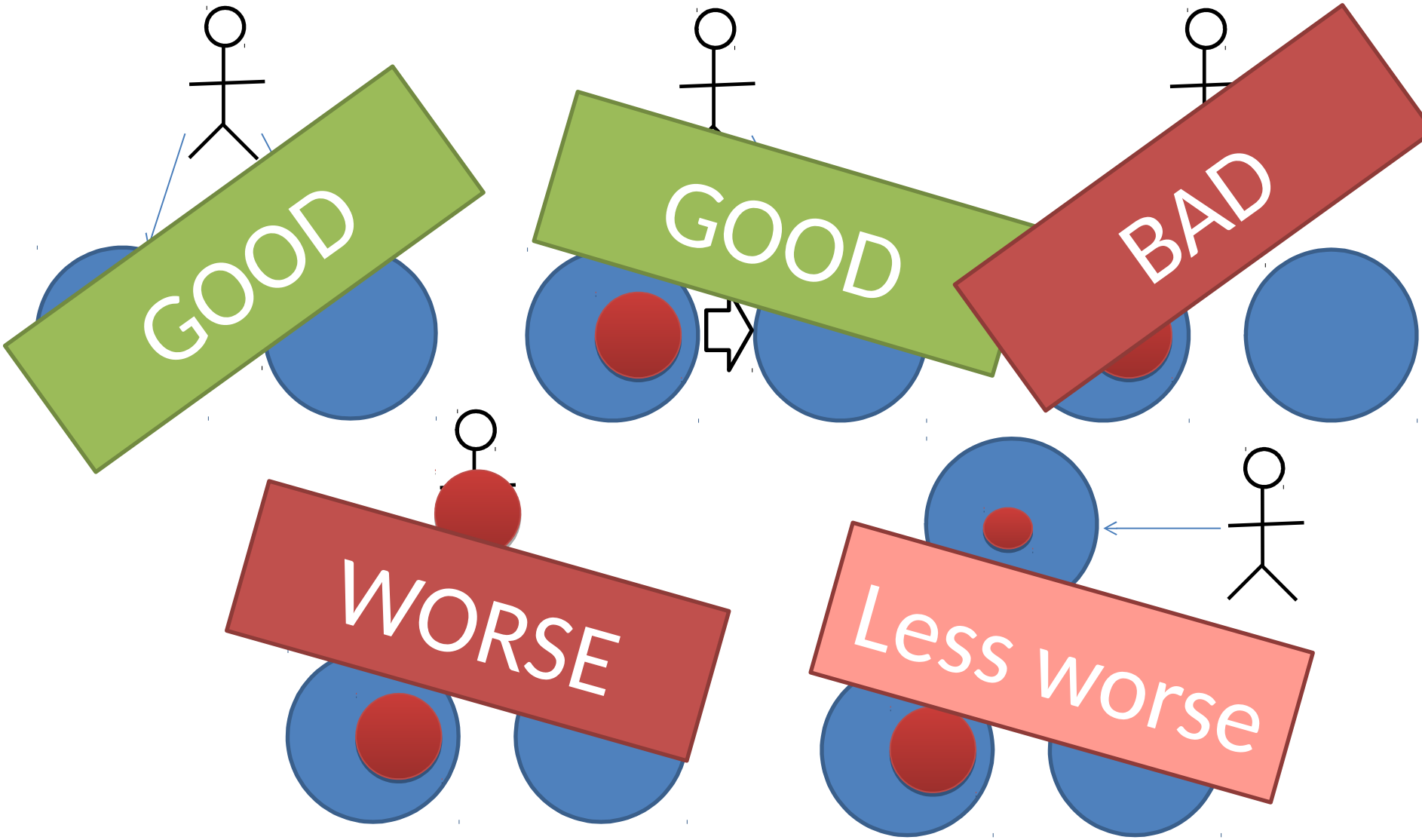
Storage Architecture



Storage Architecture



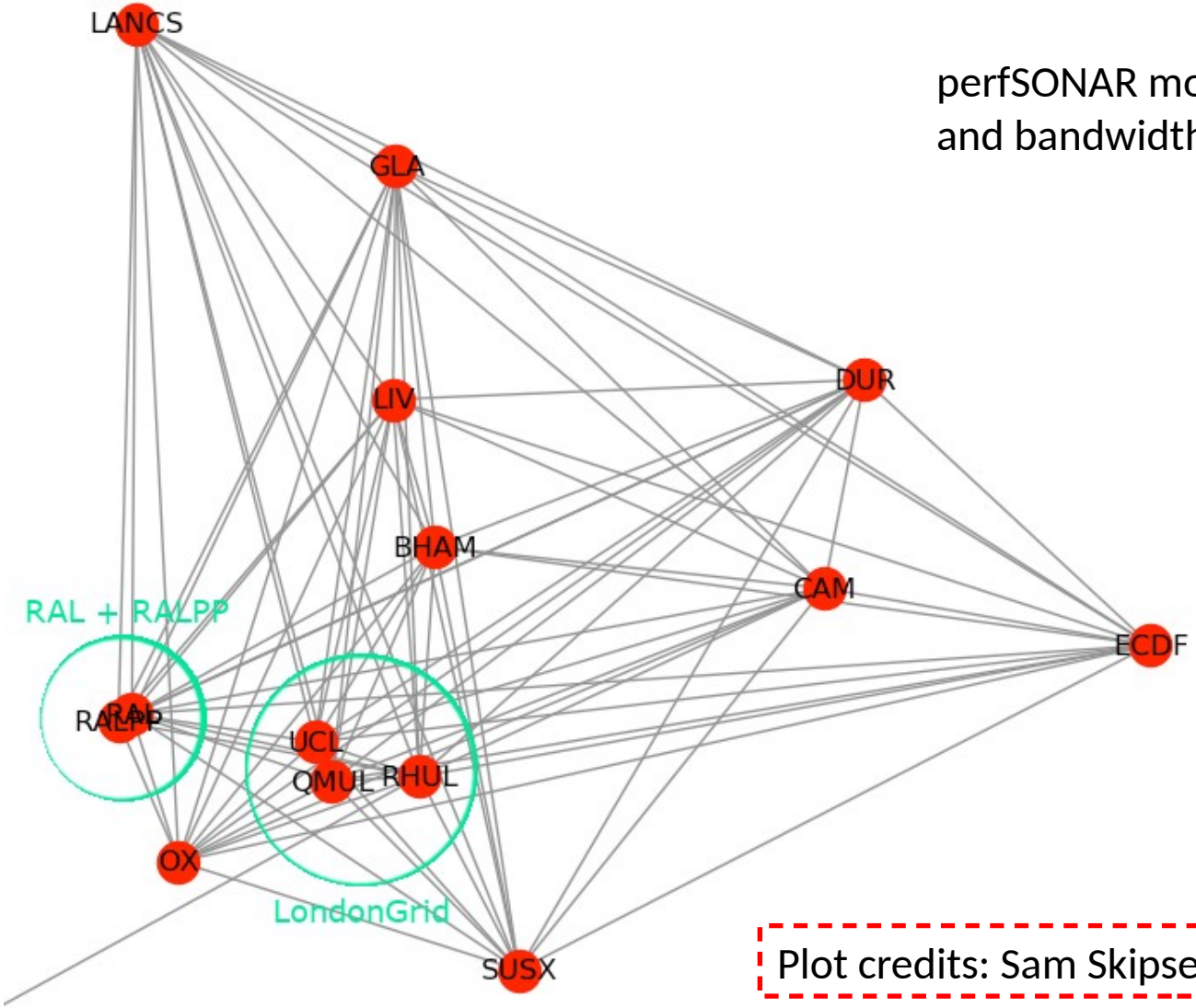
Copying Data



Transfer Protocols

- Source and destination need to share transfor protocol
- Parallel streams (e.g. GridFTP)
- Standards-based:
  - GridFTP (GFD.47)
  - HTTP/WebDAV
  - SRM (GFD.154)
- xroot

Networks

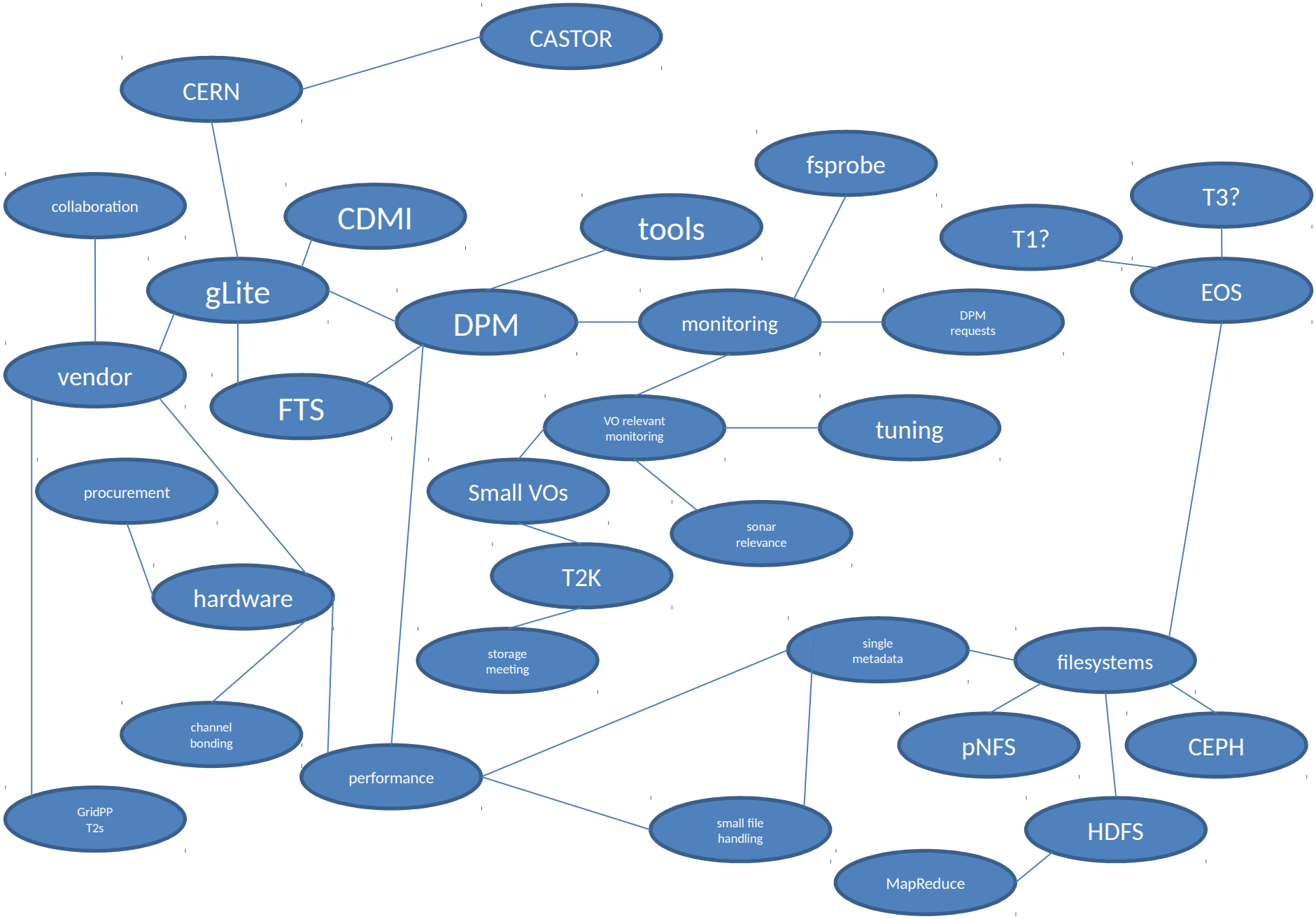


perfSONAR monitors latency and bandwidth

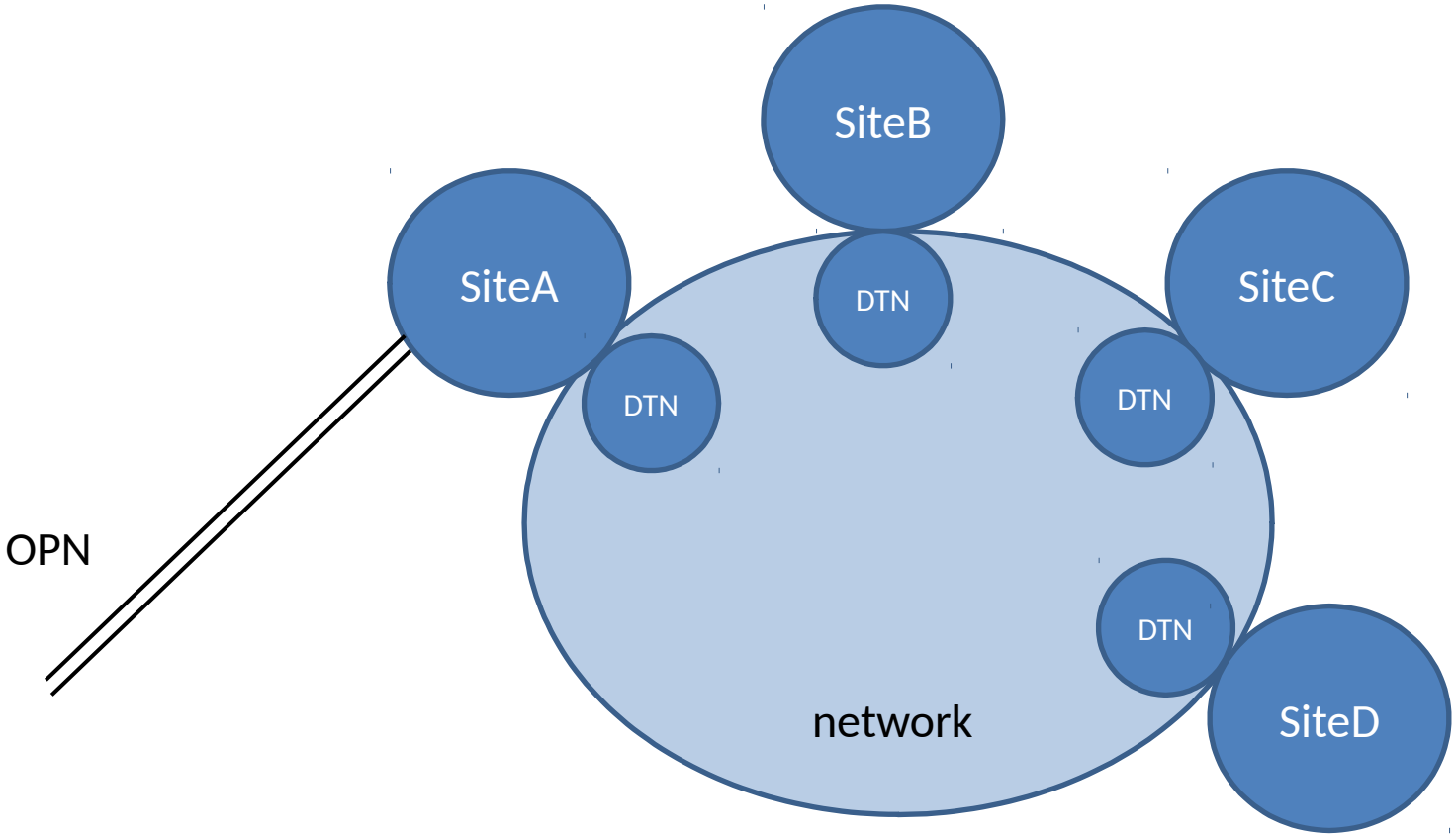
Plot credits: Sam Skipsey, GridPP, U Glasgow

Authentication, Authorisation, Delegation

- X.509 certificate authentication
  - In WLCG most users have individual certificates (IGTF)
  - Some GridPP communities generate on-the-fly (e.g. RCauth, Pathfinder)
- RBAC
  - Simple VO-defined roles through VOMS
- Token-based authorisation
  - JWT (RFC7519)

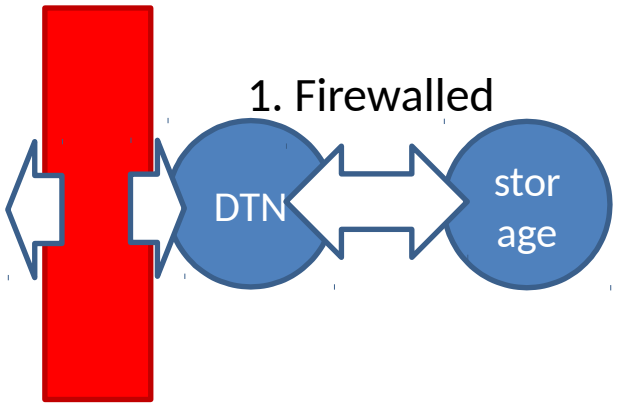

CASTOR
CERN
fsprobe
collaboration
CDMI
tools
T3?
T1?
gLite
DPM
monitoring
DPM
requests
EOS
vendor
FTS
VO relevant
monitoring
tuning
procurement
Small VOs
sonar
relevance
T2K
hardware
storage
meeting
single
metadata
filesystems
channel
bonding
pNFS
CEPH
performance
GridPP
T2s
small file
handling
HDFS
MapReduce

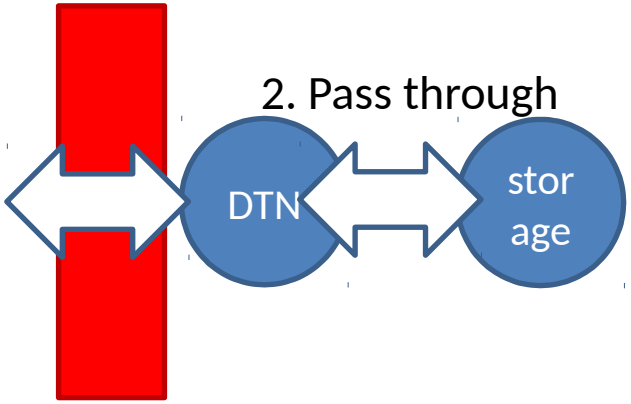
Data Transfer Zone (ESNET's "Science DMZ")

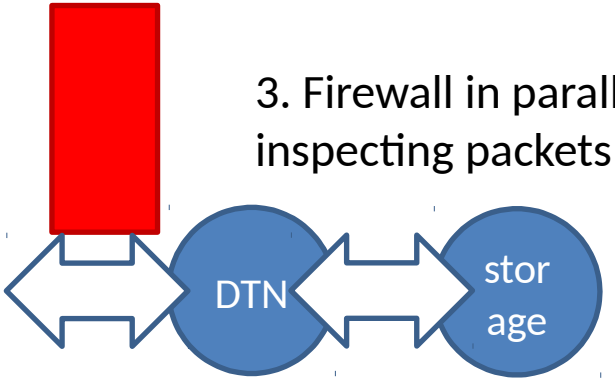


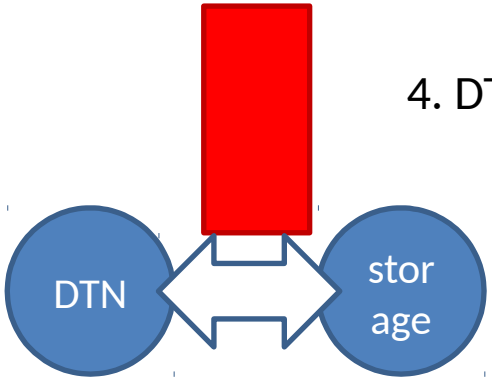
Sites run Data Transfer Nodes connecting over JANET
Nodes are secured through the IGTF PKI (incl client authentication)

Firewalling DTNs









Transfer Tools

- Globus Connect
- FTS
- Long history of low-level data toolsets and APIs
  - edg-*, lcg-*, gfal-*, globus-*
  - davix-*
- davix is CERN's WebDAV implementation

  - Supports AWS S3, Azure

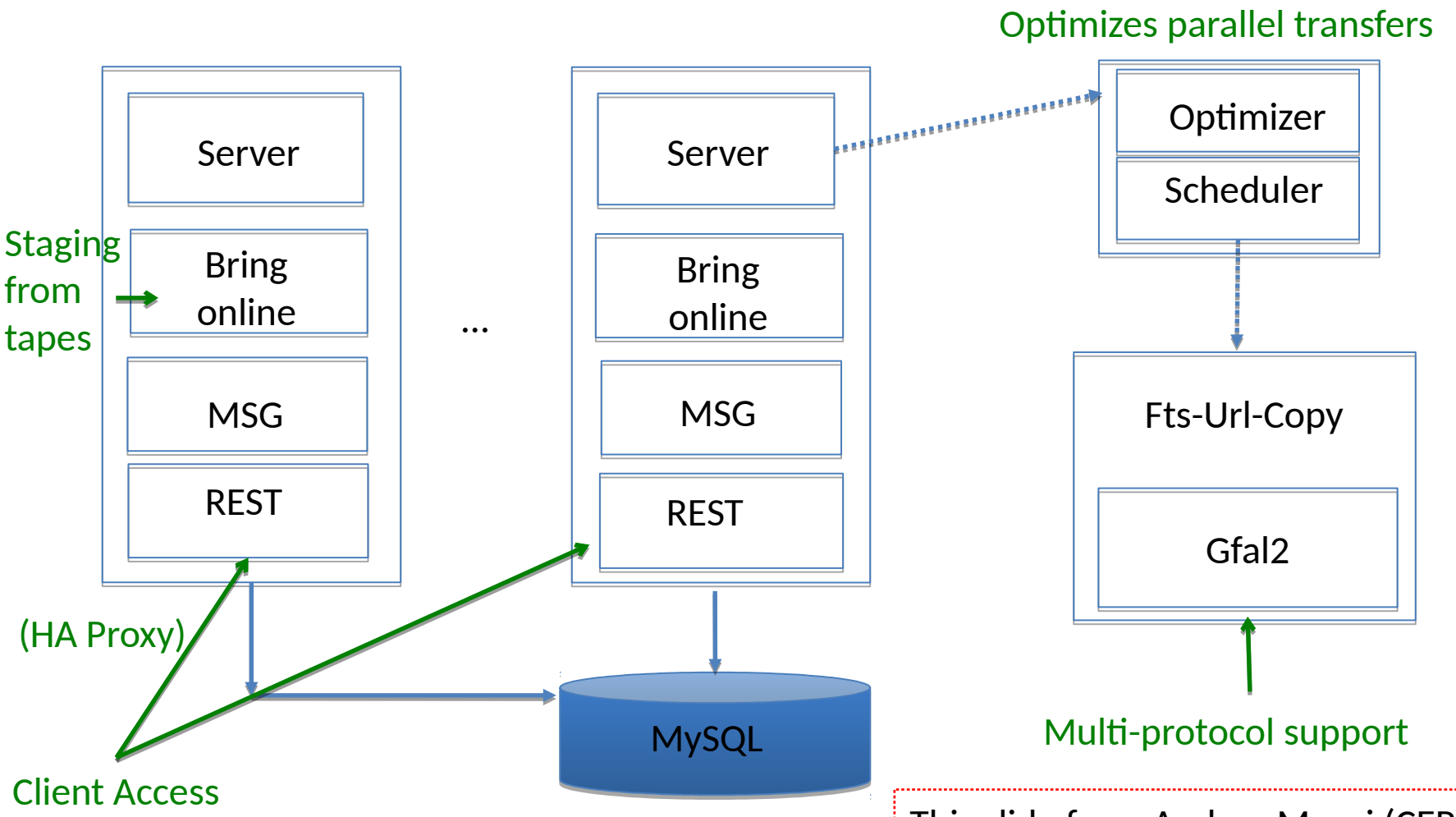
File Transfer Service - FTS

- In numbers
  - 17 instances (across WLCG) support 20 VOs
  - Move 20PB data in 26M transfers per week (~1EB/yr)
- Scheduler
  - Prioritisation
- Optimiser
  - Reorder based on throughput, success rate
  - Also optimises #parallel streams
- Automatic retries
- Small file optimisation (= conn. reuse)

Science & Technology Facilities Council
Rutherford Appleton Laboratory

GridPP
UK Computing for Particle Physics

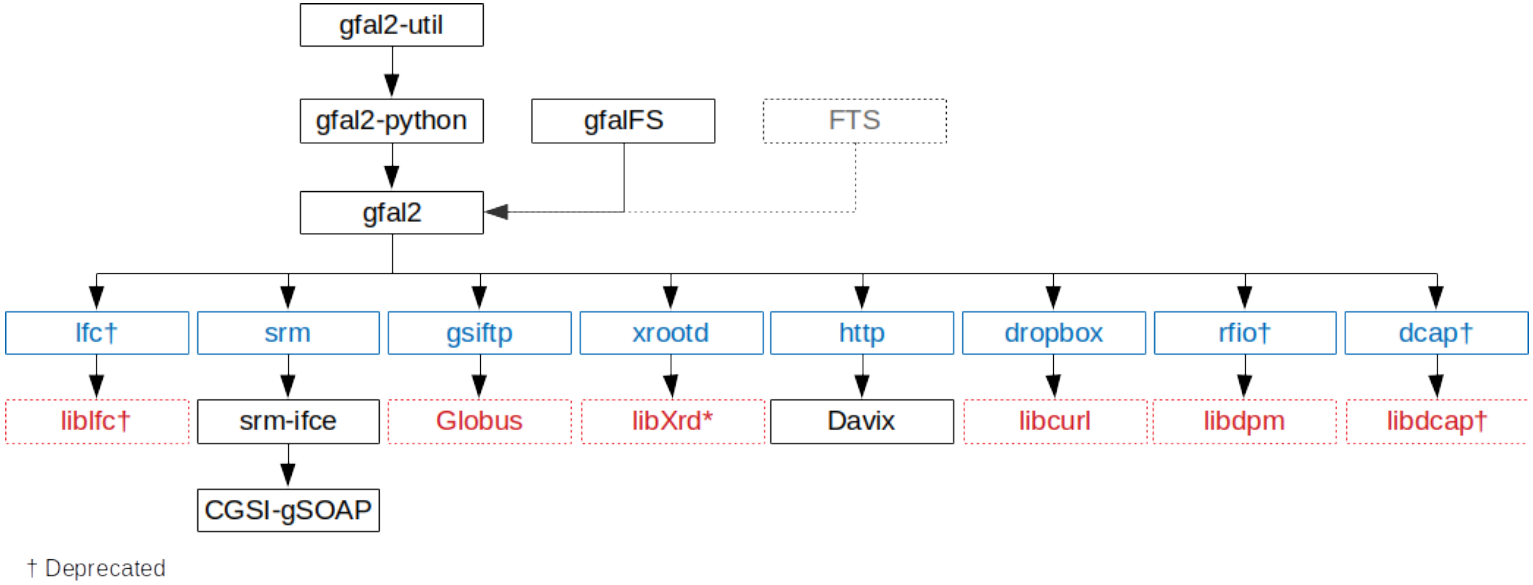
FTS
File Transfer Service



This slide from Andrea Manzi (CERN)

Multiprotocol support: gfal2

- FTP/GSIFTP, HTTP, XROOTD, SRM, S3, GCLOUD, ..
- TPC (3rdParty copy) or protocol translation (streaming)

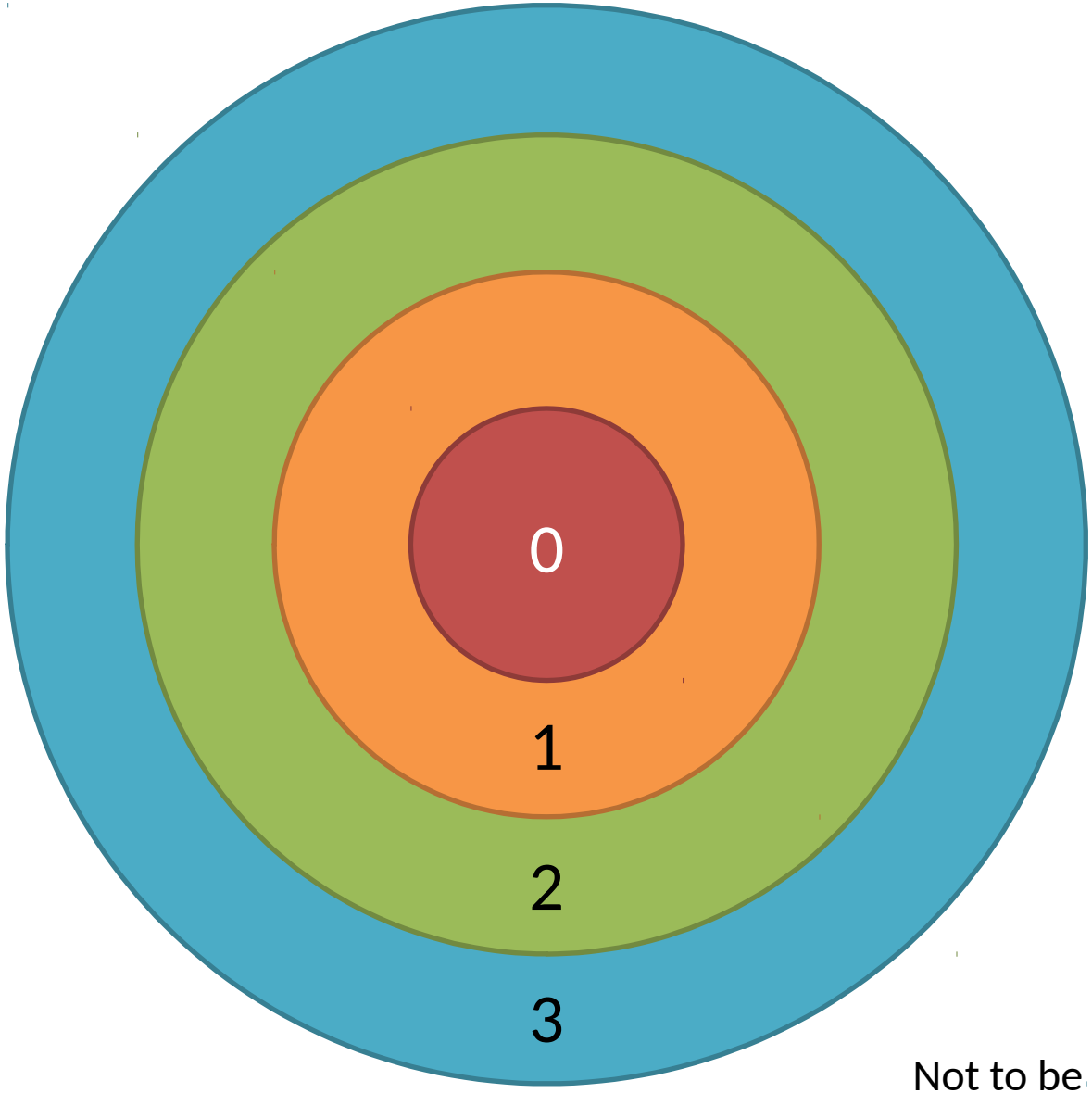


This slide from Andrea Manzi (CERN)

Turtles

- High Level data management (e.g. Rucio)
  - Replication policy
  - Deletion policy
  - User-facing APIs
- File transfer service
- Storage Elements
- Distributed File System, Object Stores, etc.
- Storage Fabric

Scaling to Exabyte



Tier0: instrument, preprocess

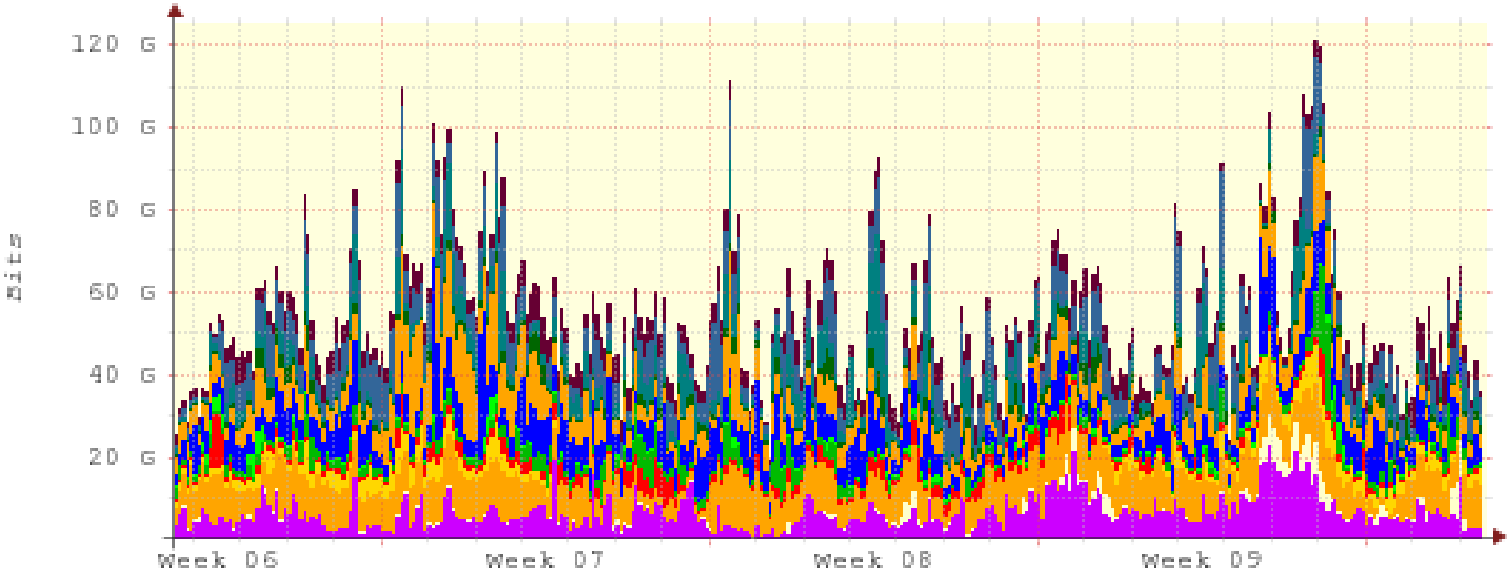
Tier1: preprocess, global replication

Tier2: user analysis, regional replication, local cache

Tier3: end user analysis

Not to be confused with data centre tiers!

LHCOPN TOTAL Traffic (CERN -> Tiers1)

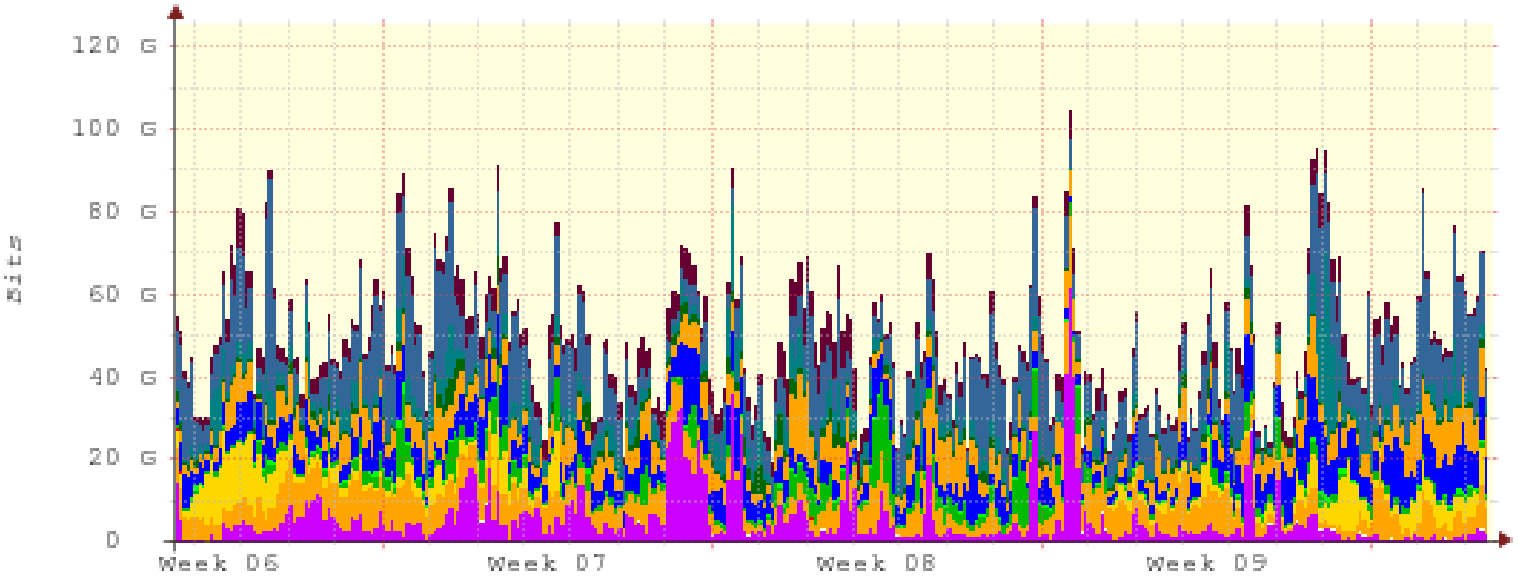


	Avg	Max	Peak	Curr
To DE-KIT	6.74G	22.87G	53.16G	3.05G
To KISTI	823.80M	8.92G	9.03G	198.58M
To RU-T1s	8.84G	16.31G	24.92G	10.28G
To IN2P3	1.80G	11.66G	26.68G	3.82G
To NDGF	1.94G	13.21G	26.41G	385.90 M
To NLT1	2.40G	21.38G	71.75G	595.17 M
To ASGC	652.18M	5.99G	8.95G	294.58M
To CNAF	6.91G	38.02G	99.06G	3.75G
To RAL	6.95G	28.23G	29.69G	3.70G
To TRIUMF	1.66G	8.95G	10.55G	1.57G
To BNL	4.69G	36.88G	50.14G	2.77G
To FNAL	6.80G	24.39G	51.03G	4.24G
To ES-PIC	4.03G	8.50G	9.85G	1.36G

Total to Tiers1 Avg: 53.99G Max:121.34G Curr: 36.01G

Last update: Tue Mar 05 2019 13:03:59

LHCOPN TOTAL Traffic Flow (Tiers1 -> CERN)



	Avg	Max	Peak	Curr
From DE-KIT	5.57G	61.58G	78.29G	1.98G
From KISTI	41.61M	499.32M	2.27G	36.68M
From RU-T1S	5.80G	19.79G	23.92G	6.64G
From IN2P3	2.63G	18.69G	24.63G	2.07G
From NDGF	6.73M	760.32M	5.28G	305.35
From NLT1	2.62G	20.92G	53.52G	460.72M
From ASGC	314.26M	2.56G	9.07G	918.00M
From CNAF	6.27G	21.14G	48.98G	8.67G
From RAL	5.72G	19.21G	26.96G	9.49G
From TRIUMF	1.48G	7.05G	8.58G	696.84M
From BNL	4.04G	40.38G	42.69G	2.34G
From FNAL	11.60G	51.41G	55.33G	8.55G
From ES-PIC	3.42G	9.73G	9.83G	94.44M

Total from Tiers1 Avg: 49.11G Max:104.28G Curr: 41.95G

Last update: Tue Mar 05 2019 13:04:00

Optimising data/compute

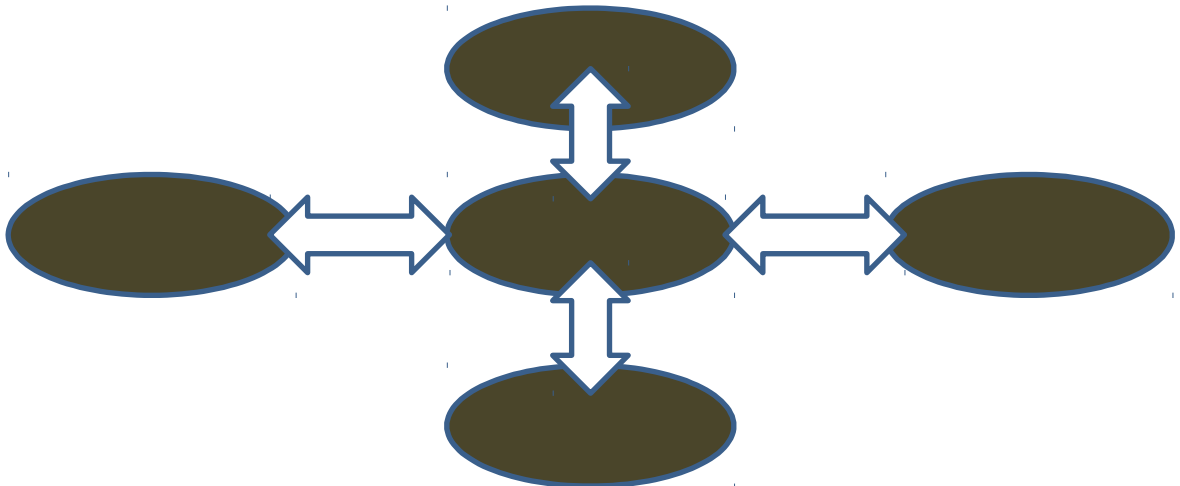
- Generally sending compute to where data is
- "Federated storage" – cross site access
  - E.g. if a replica is missing
- Pilot jobs provide late binding of workload to job slot
  - Job slot is allocated to experiment but once the job starts, it figures out what to do…

(Other) Future Directions

- Better support for non-wizard users
  - (not CLI, federated id)
- More inter-turtle communication
  - Cache-aware data layers
  - Make use of Redfish (DMTF)/Swordfish (SNIA)
- Accommodating WLCG evolutions
  - Cache only sites
  - Further increasing “federated” storage (cross site access)
- More interfacing to other infrastructures
- Supporting IRIS (STFC funded researchers)
  - Then UKRI research communities?

Conclusions

- Exascale: regimented data model
- Many-turtled approach
  - Individual turtles have been replaced over the years
  - Turtles work well, they have some independence and can talk to each other



References

- GridPP: www.gridpp.ac.uk
- WLCG: wlcg.web.cern.ch
- Rucio: rucio.cern.ch
- GridFTP: www.ogf.org/documents/GFD.47.pdf
- SRM: www.ogf.org/documents/GFD.129.pdf
- FTS: fts.web.cern.ch
- IGTF: www.igtf.net
- Redfish www.dmtf.org/standards/redfish
- Swordfish www.snia.org/tech_activities/standards/curr_standards/swordfish