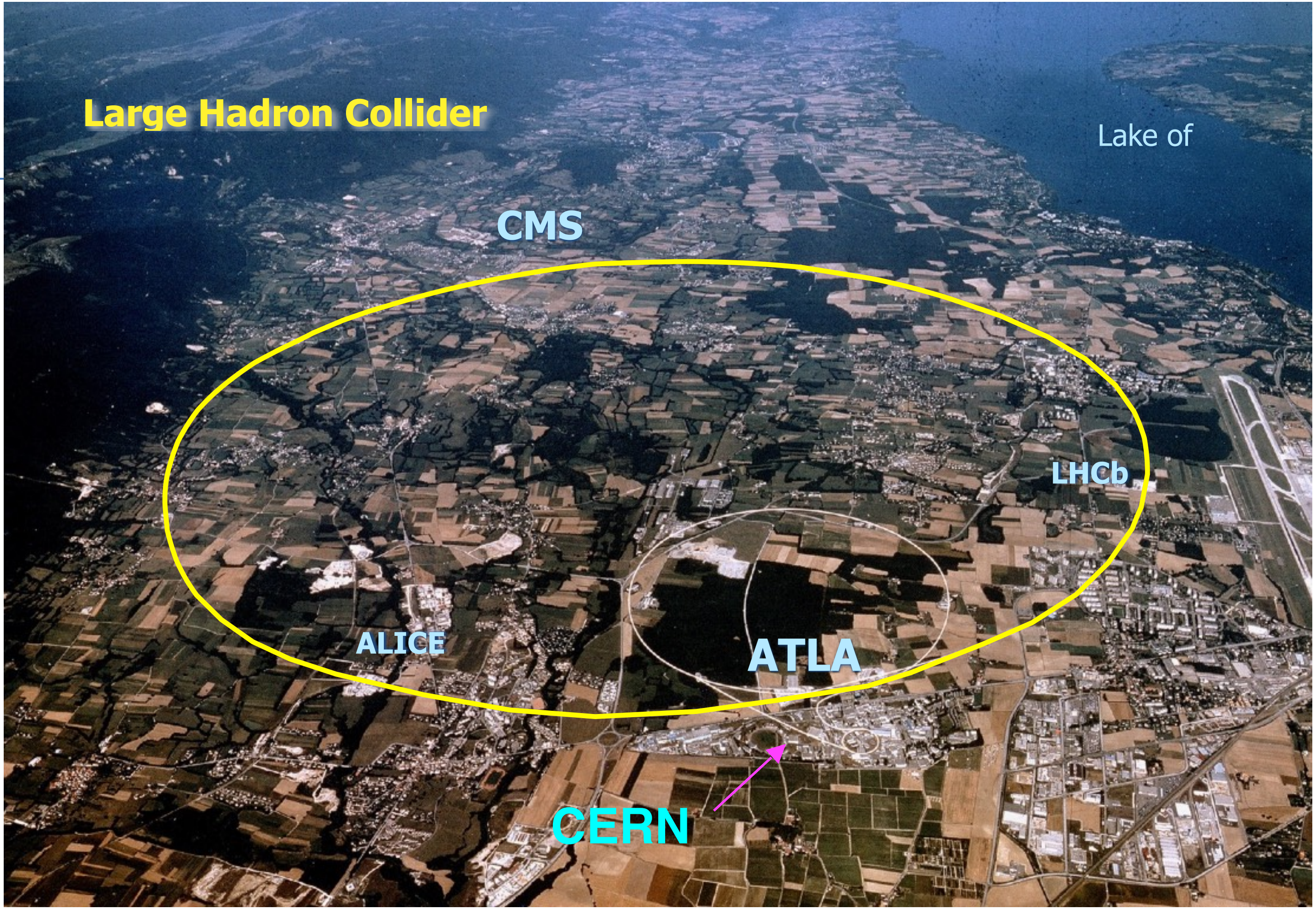
Data management and LHC

Eric Lancon

BROOKHAVEN
NATIONAL LABORATORY

U.S. DEPARTMENT OF
ENERGY

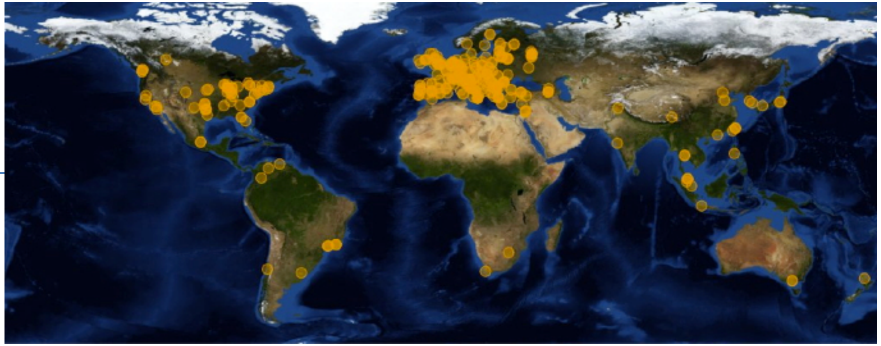
The LHC



This is ATLAS

Computing for LHC: WLCG

Worldwide LHC Computing Grid



170 Data centres

40 Countries

800'000 Cores

500 PB Disk

750 PB Tape

3 Tbps Network

Tiered Structure

Tier-0	CERN
Tier-1	Large data centres
Tier-2+3	Universities and Laboratories

Heterogeneous Computing

Data centres (partly) supported by national Funding Agencies

Centres may host and support other projects

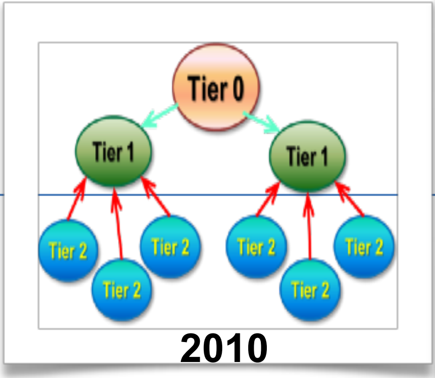
Pledged storage and compute based on an MoU

WLCG Tiers

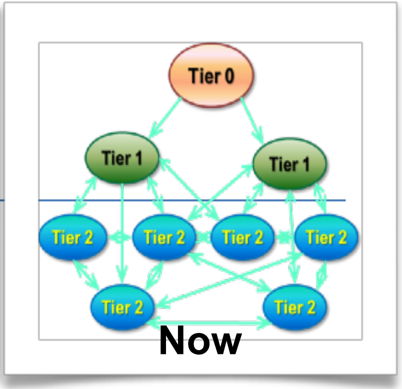
- From hierarchical to ‘democratic’ structure, data exchange between any tier, thanks to network capabilities!

- Sites are still providing resources with defined CPU/Storage ratio (No CPU only tier)

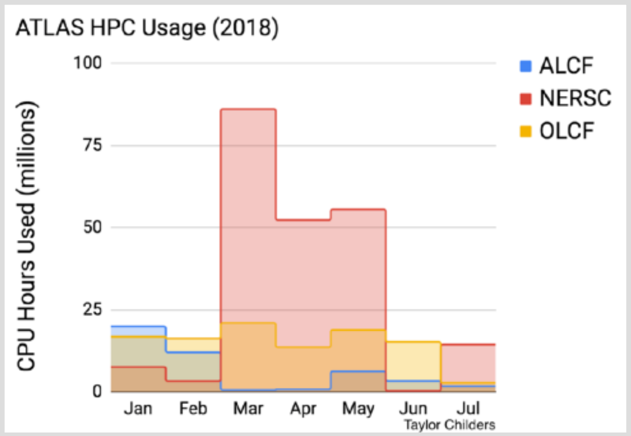
- BUT decoupling of storage and compute resources with increased usage of non-WLCG CPU provider like HPCs



2010



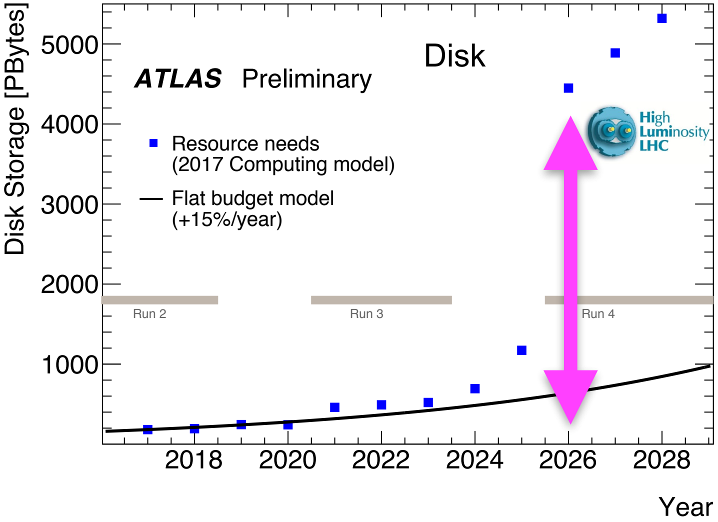
Now



Taylor Childers

Upcoming (2026) : High Luminosity-LHC

ATLAS resource requirements



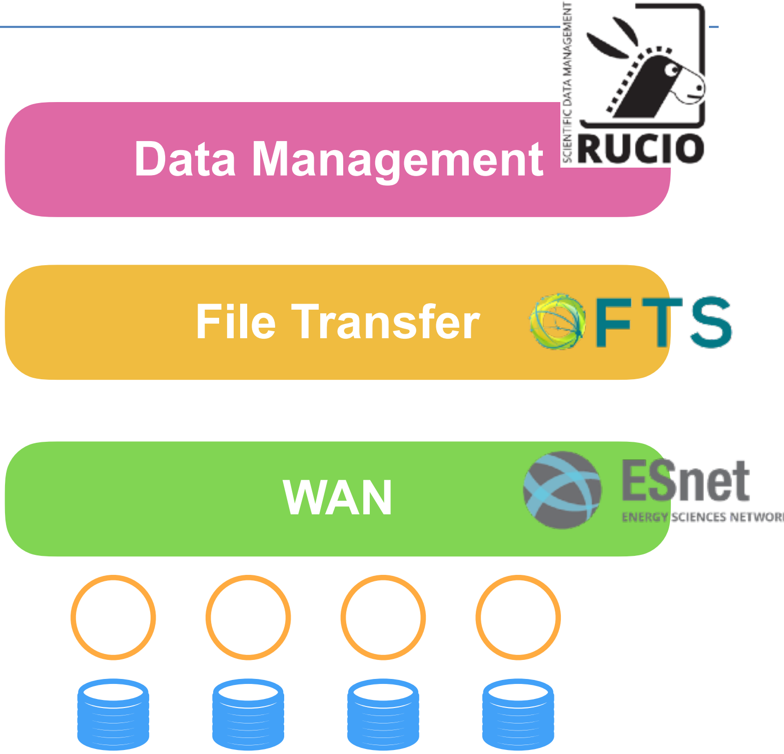
Required increased storage capacity ~10x today

(under current computing models)

Time to re-think data distribution

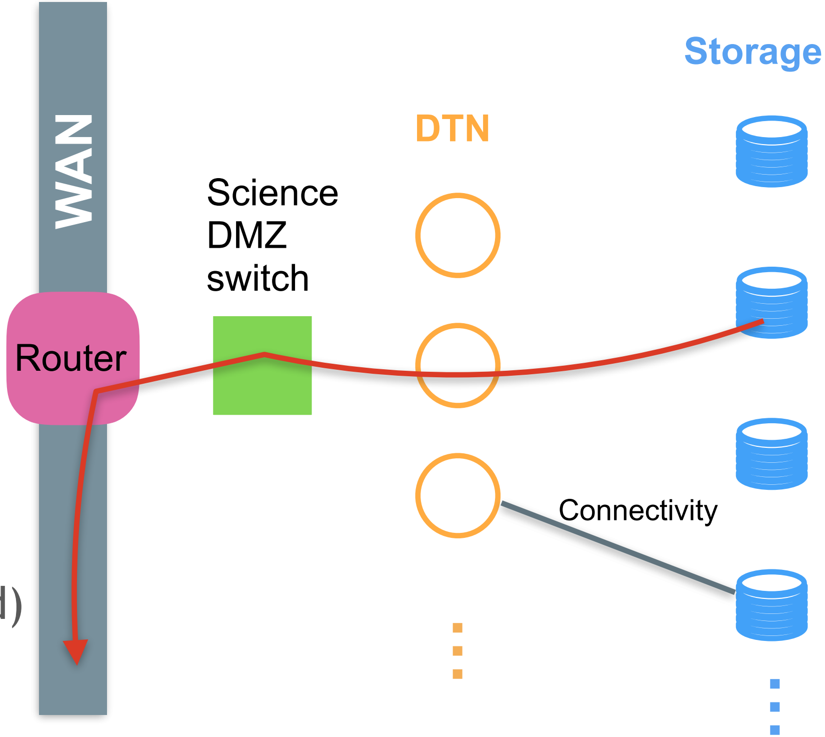
Elements of ATLAS experiment data movement

- Data Management Layer : RUCIO
- File Transfer Layer : FTS
- WAN: ESnet (+ GEANT, Signet,...)
- Storage and interfaces



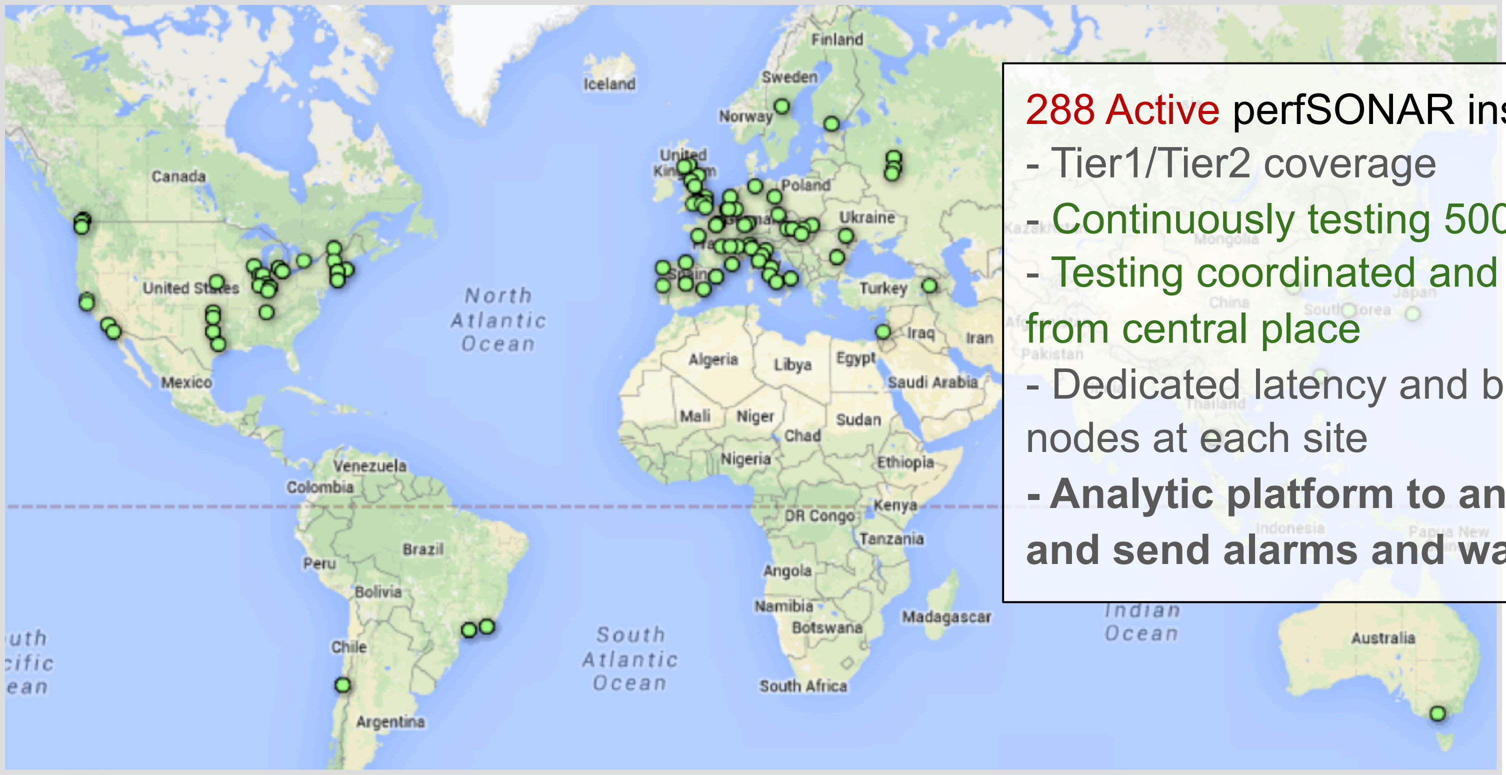
(Some) performance parameters for data movement

- Storage technology & hardware
- Connectivity
- DTNs
- Switch & Router
- WAN
- Protocols
  - Difference between filling the bandwidth and efficiently reading data
  - Currently GridFTP for transfers and
  - Xrootd for reading (caching/read-ahead)
- … (file size)



WAN

WAN: perfSONAR

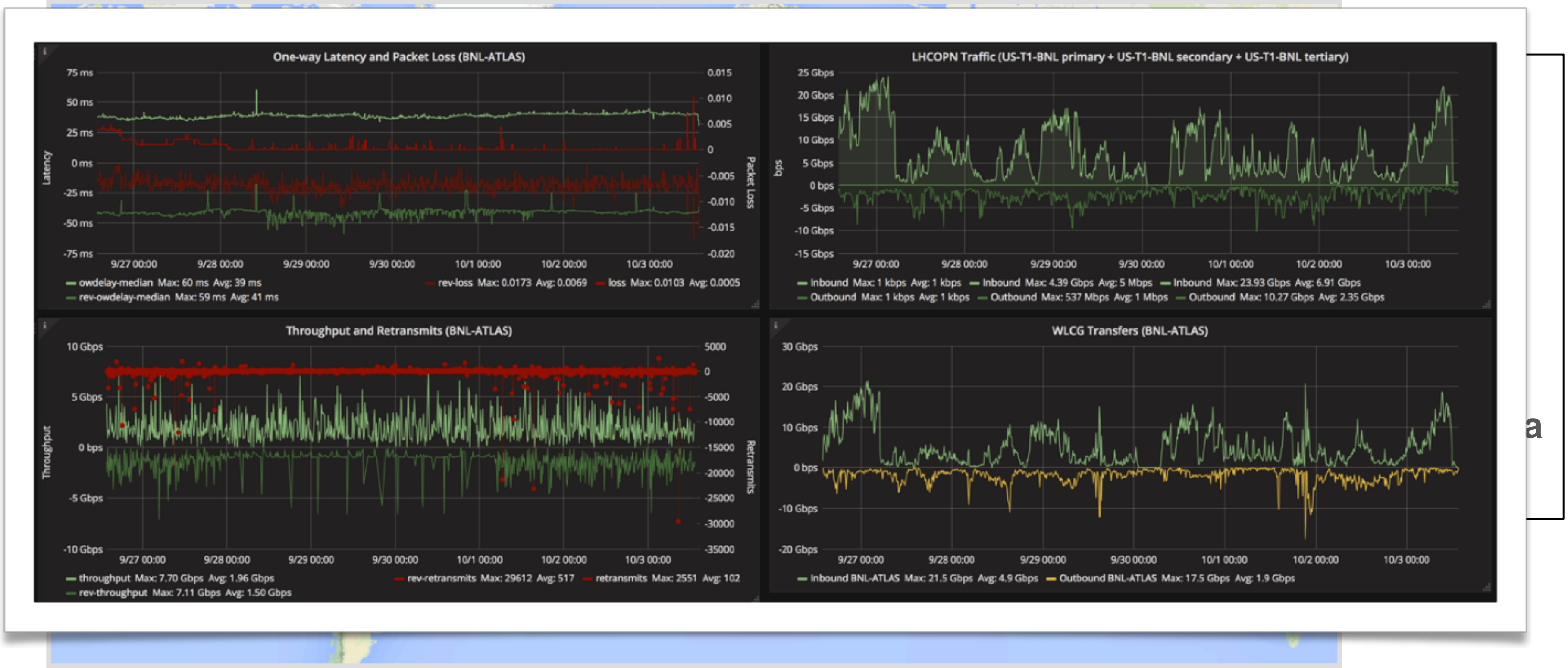


288 Active perfSONAR instances

- Tier1/Tier2 coverage
- Continuously testing 5000+ links
- Testing coordinated and managed from central place
- Dedicated latency and bandwidth nodes at each site
- **Analytic platform to analyses data and send alarms and warnings**

WAN

WAN: perfSONAR


One-way Latency and Packet Loss (BNL-ATLAS)
75 ms
50 ms
25 ms
0 ms
-25 ms
-50 ms
-75 ms
Latency
0.015
0.010
0.005
0
-0.005
-0.010
-0.015
-0.020
Packet Loss
9/27 00:00
9/28 00:00
9/29 00:00
9/30 00:00
10/1 00:00
10/2 00:00
10/3 00:00
owdelay-median Max: 60 ms Avg: 39 ms
rev-owdelay-median Max: 59 ms Avg: 41 ms
rev-loss Max: 0.0173 Avg: 0.0069
loss Max: 0.0103 Avg: 0.0005
LHCOPN Traffic (US-T1-BNL primary + US-T1-BNL secondary + US-T1-BNL tertiary)
25 Gbps
20 Gbps
15 Gbps
10 Gbps
5 Gbps
0 bps
-5 Gbps
-10 Gbps
-15 Gbps
bps
Inbound Max: 1 kbps Avg: 1 kbps
Outbound Max: 1 kbps Avg: 1 kbps
Inbound Max: 4.39 Gbps Avg: 5 Mbps
Outbound Max: 537 Mbps Avg: 1 Mbps
Inbound Max: 23.93 Gbps Avg: 6.91 Gbps
Outbound Max: 10.27 Gbps Avg: 2.35 Gbps
Throughput and Retransmits (BNL-ATLAS)
10 Gbps
5 Gbps
0 bps
-5 Gbps
-10 Gbps
Throughput
5000
0
-5000
-10000
-15000
-20000
-25000
-30000
-35000
Retransmits
throughput Max: 7.70 Gbps Avg: 1.96 Gbps
rev-throughput Max: 7.11 Gbps Avg: 1.50 Gbps
rev-retransmits Max: 29612 Avg: 517
retransmits Max: 2551 Avg: 102
WLCG Transfers (BNL-ATLAS)
30 Gbps
20 Gbps
10 Gbps
0 bps
-10 Gbps
-20 Gbps
Inbound BNL-ATLAS Max: 21.5 Gbps Avg: 4.9 Gbps
Outbound BNL-ATLAS Max: 17.5 Gbps Avg: 1.9 Gbps

ATLAS Experiment & RUCIO

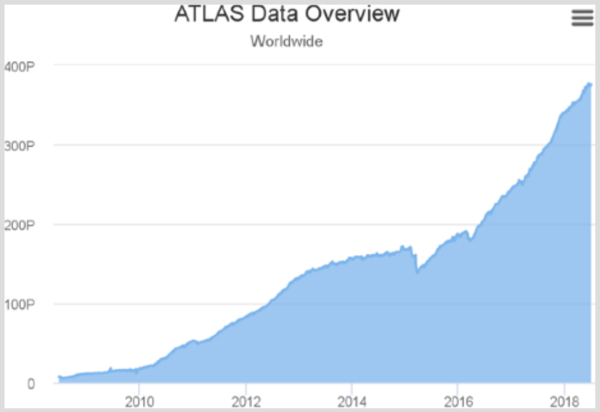
Data volume approaching 400PB

10M containers, 20M datasets, 1B files

5K accounts

1-2PB transfered/day, 3PB deletions/day

130 sites, 600 storage endpoints



Data Management

rucio.cern.ch/

Global namespace to federate across different storage systems

Control & accounting of data and users

Declarative data management with policies and rules

Transfer orchestration with priorities, shares and activities

Popularity-based replication, caching and deletion

Events & messages for synchronisation with other tools

Consistency & repair of broken and missing data

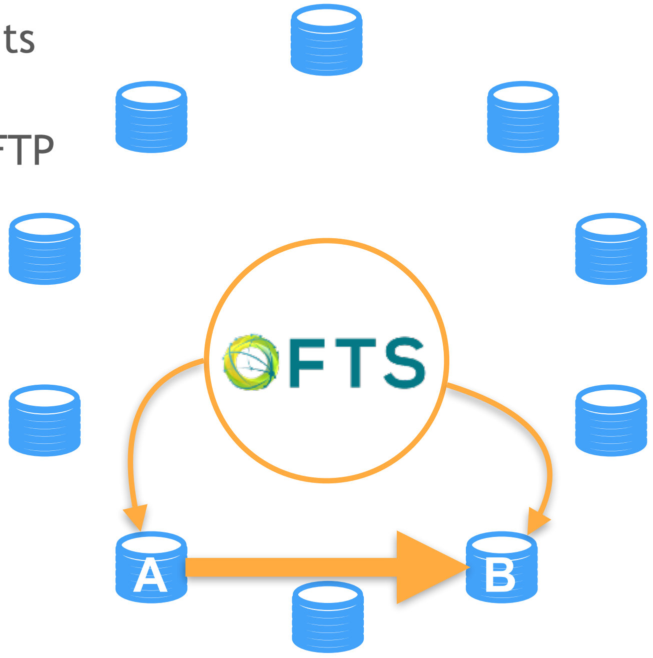
and much more …

File Transfer System

File Transfer

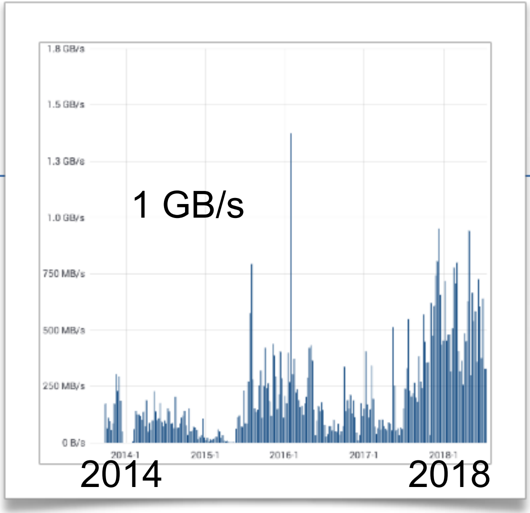
- Basic principle: 3rd party transfer between 2 end points
  - FTS should be authorized to talk to A & B
  - A & B should talk same protocol, currently GridFTP
  - Testing HTTP & XRootd

- Accept bulk requests
- Scheduler, shares
- Parallel file transfers
- Adaptive auto-tuning
- Multihop
- Session re-use
- …

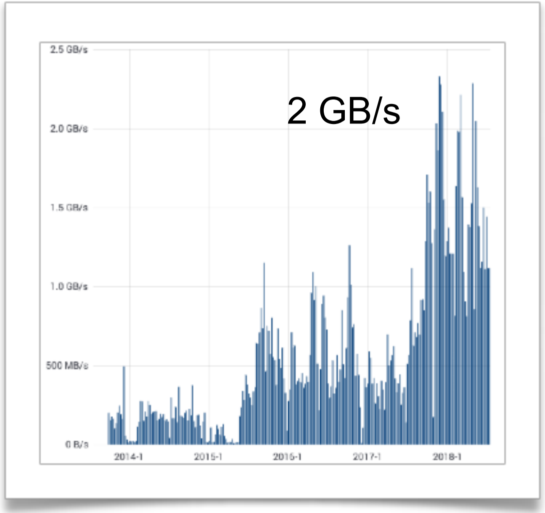


Server at BNL

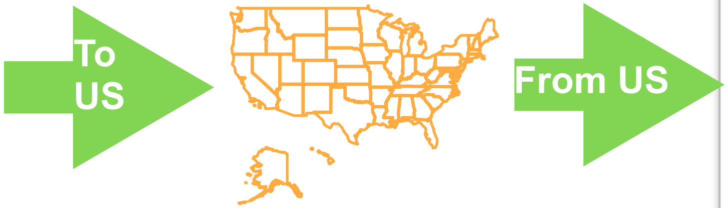
FTS transfers

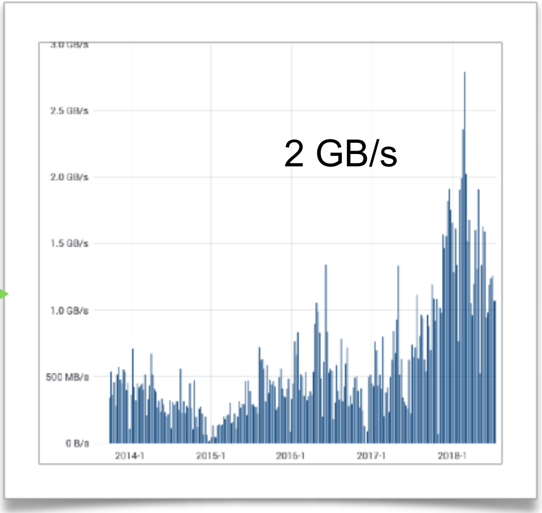


Average transfer rates
Since 2014



Within US





Heterogeneity of storage

- Age
- Latency
- Resiliency
- Size (US: 38 usable PB at 5 locations)
- Can be geographically distributed
- Overhead (raw->usable space)
  - 2-3 actual copies
  - RAID
  - Erasure

- Technology :
  - dCache
  - XRootd
  - Ceph
  - GPFS
  - …

Different funding, Different locations, History….
MoU specify availability to receive data only

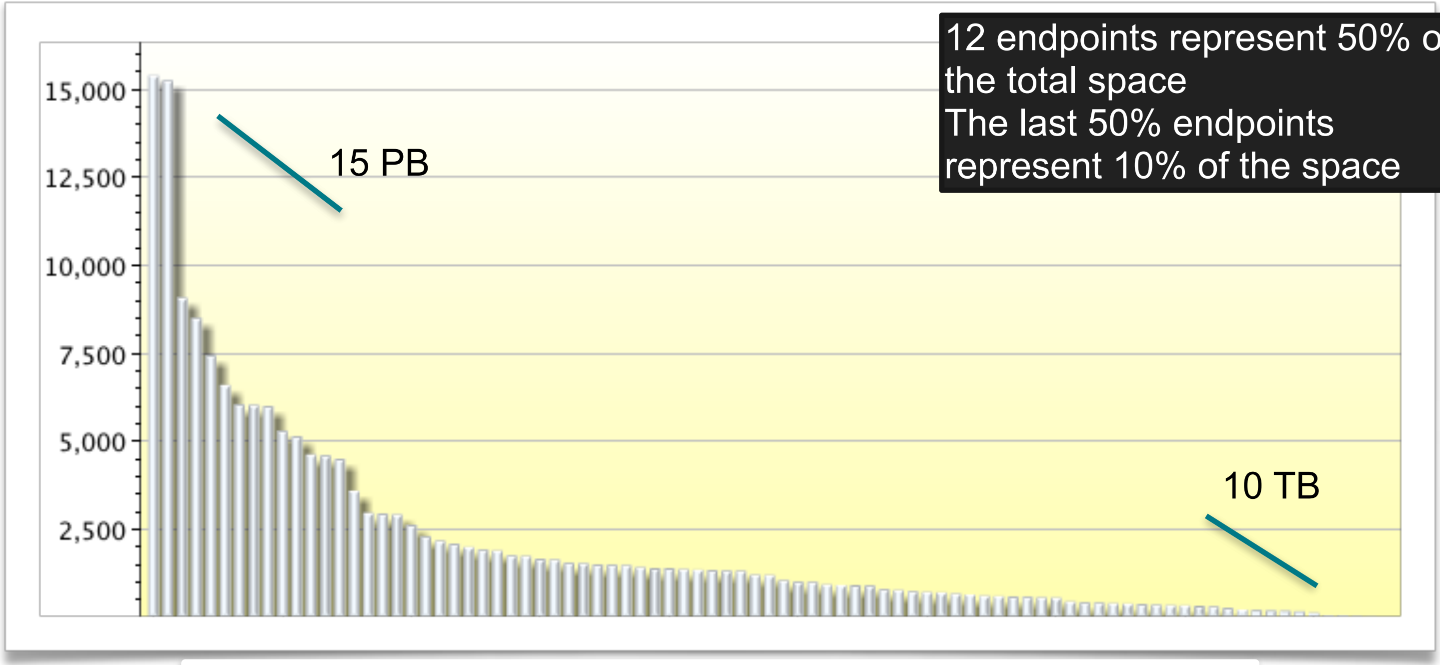
Heterogeneity of storage

- Age
- Latency
- Resiliency
- Size (US: 38 usable PB at 5 locations)
- Can be geographically distributed
- Overhead (raw->usable space)
  - 2-3 actual copies
  - RAID
  - Erasure

- Technology :
  - dCache
  - XRootd
  - Ceph
  - GPFS
  - …

Differences are (almost) not taken into account by current data placement policies

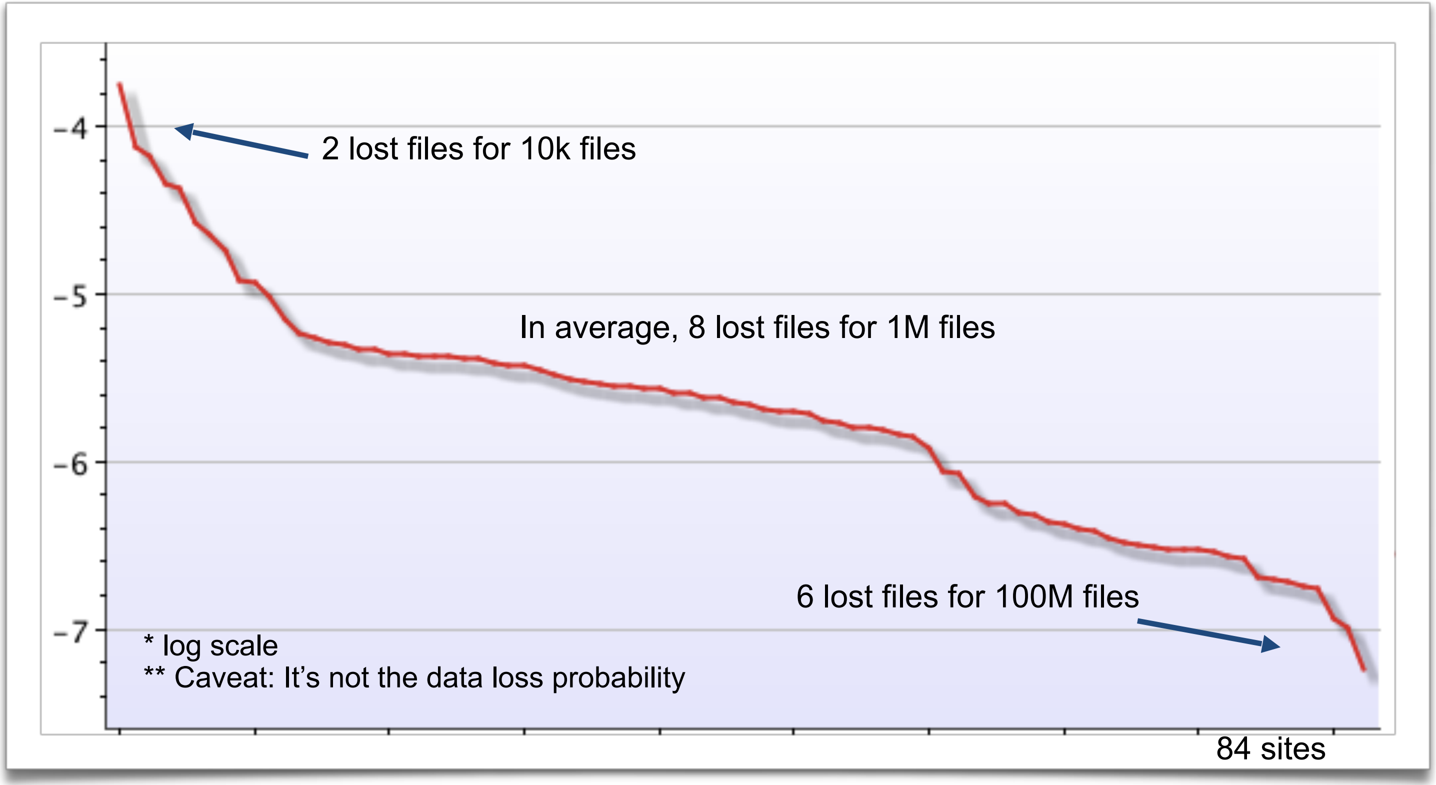
ATLAS: Storage size / end point



12 endpoints represent 50% of the total space
The last 50% endpoints represent 10% of the space

Fragmentation of disk space + operational effort

ATLAS: Lost files frequency / site



2 lost files for 10k files

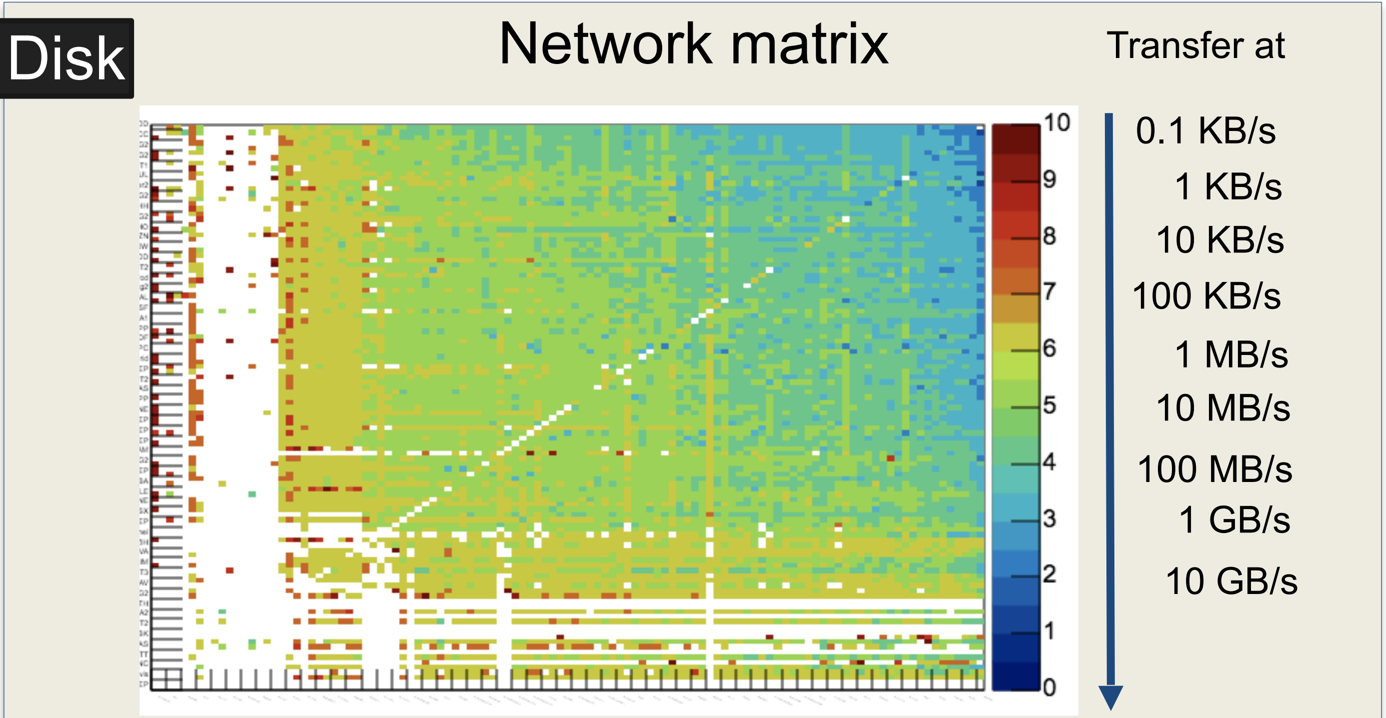
In average, 8 lost files for 1M files

6 lost files for 100M files

* log scale
** Caveat: It's not the data loss probability

ATLAS: Actual transfer rates between end points

Disk to Disk



* One month statistic: Maximum throughput during one hour
** Throughput < 100 KB/s can be due to less transfers and statistics

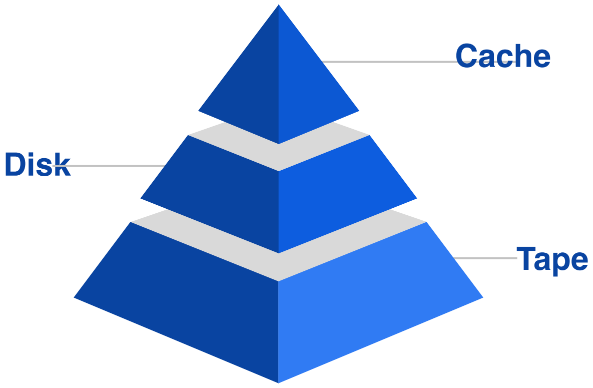
The real figure of merit!

And tapes ???

- Available at a few locations (Tier-1s)
- Decoupled from disk storage
  - Used as archive
  - Scheduled access
- Underutilized
- Reliable, cheaper than disk
- Ongoing tests of optimized tape access via ‘tape carrousel’

Storage distribution evolution

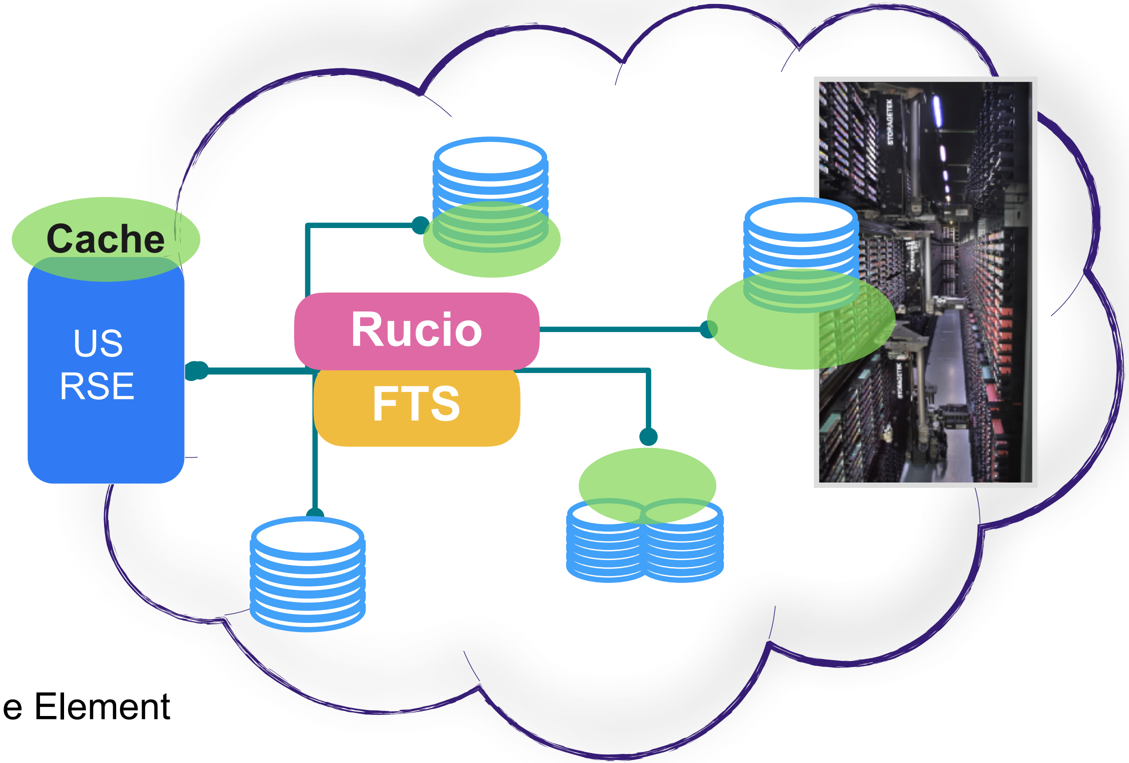
- Reduce number of end points
  - Larger storage entities
- Increase tiered hierarchy
- Introduce QoS
  - Reliability
  - Availability
  - Throughput
  - Redundancy
  - …



Proposed prototype for US ATLAS

- Expose unique storage entry point to WLCG
- Internally
  - Different QoS for various components
  - Increasing usage of tape as foundation
- Caching when needed at storage (and CPU) locations
- Rucio (and FTS)
  - To provide unified name space
  - To manage storage hierarchy and data placement and replication

Internally : Redirect, Move & Cache



RSE: RUCIO Storage Element

Thank you…

ESnet
ENERGY SCIENCES NETWORK