

Evolution in Storage Services and Data Access at Run2

Wahid Bhimji

11th November 2013

Caveats etc.

- ❖ Maybe an ATLAS bias. Maybe a site bias
- ❖ More on Storage rather than Data Management following the title.
- ❖ Not much about Tape; Not much about Security
- ❖ Starting place: [Technical Evolution Group \(TEG\)](#); (Recommendation numbers given in some places as TEG:Rx - see also backup slides)
- ❖ And [CHEP13](#); [WLCG data WG](#); [XLDB](#); [GridPP Big Data Workshop](#)
- ❖ Asked to **stimulate discussion** so some more provocative statements:
To accept blame I'll say those statements are "mostly my own"
To-give-credit I'll say they are "inspired by others"

Technology statements

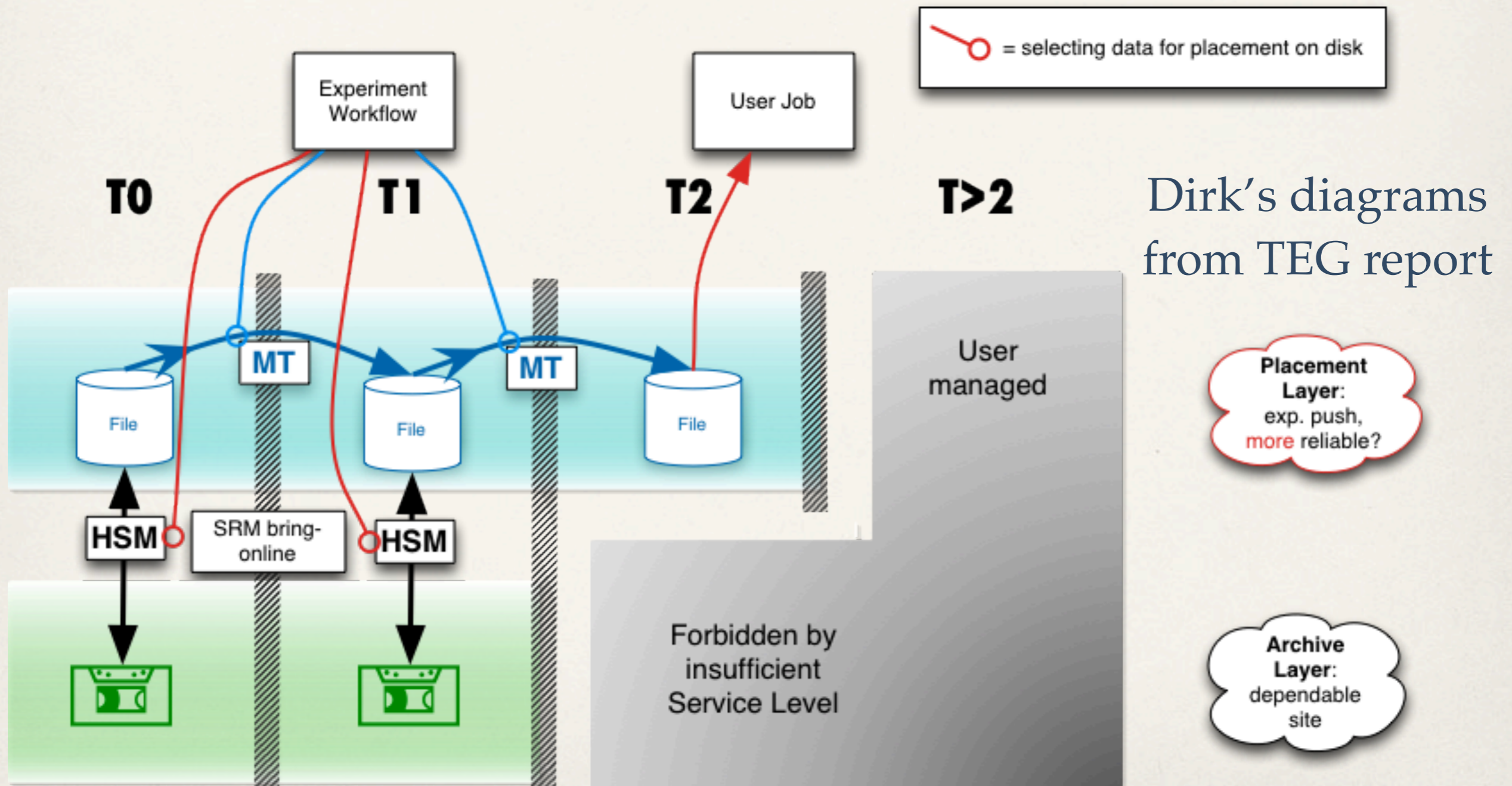
- * Run 2 technology exists now: at least the hardware and filesystems. Specific predictions are still bound to be wrong
- * Manage well, but data activities still very inefficient in many areas.
- * Need to improve: money tight; needs growing; limits on disk IO/ TB
- * “Big-data” is much bigger than before. TEG report maybe undervalued role this area and that of “Cloud” storage might have.
- * “Common” technologies; “industry” solutions will rise. We’ll benefit from being flexible enough to make use of these technologies. E.g.:
 - Thin-layer, future-looking, middleware
 - Future-looking data access interfaces
 - Interoperation with academic and other “big-data” communities

Overview

- ❖ Data Management
 - ❖ Data federations and managed transfers
- ❖ Interlude: Big Data
- ❖ Storage Management
 - ❖ Interfaces ; I/O
 - ❖ Site Perspective and Technologies

Data Management

Managed Data Placement



Dirk's diagrams from TEG report

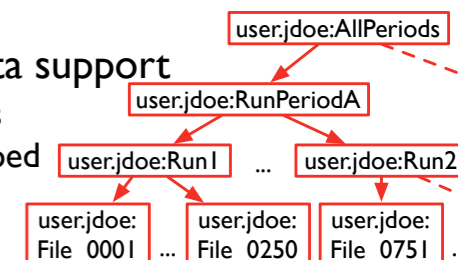
Point to Point protocols and Managed Transfer (MT)

- ❖ **FTS3** moving into production. Good interaction with experiments and experiences so far. (TEG:R7-8)
 - ❖ Flexibility to use gridFtp session reuse and xrootd or http also for managed transfers (TEG:R4-6) - not yet really used by LHC VOs.
 - ❖ Looks **well placed to satisfy needs for Run2 Managed Transfer**
- ❖ **LFC**: LHC experiments migration away **well underway** (TEG: R9) (ATLAS still use but transition to rucio progressing fast)
 - ❖ Nice to have: a good recommendation / product for small-VOs (e.g. part of rucio / alien / stick with LFC?)

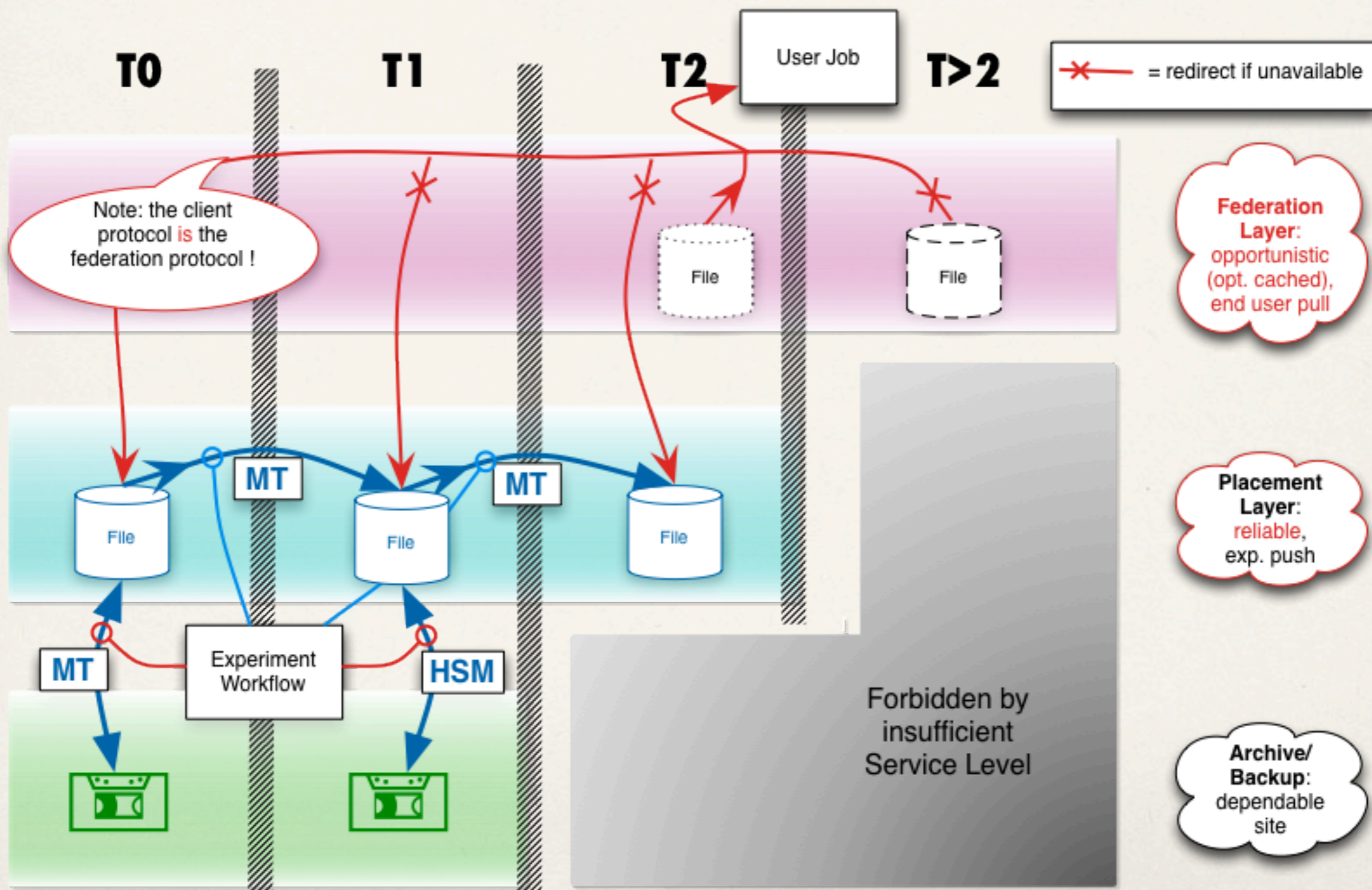
Atlas Rucio: [Vincent GARONNE](#)

Concepts - Highlights

- Better management of users, physics groups, ATLAS activities, data ownership, permission, quota, etc.
- Data hierarchy with metadata support
 - Files are grouped into datasets
 - Datasets/Containers are grouped in containers
- Concepts covering changes in middleware
 - Federations
 - Cloud storage
 - Move towards open and widely adopted protocols



Data Federations



Data Federations

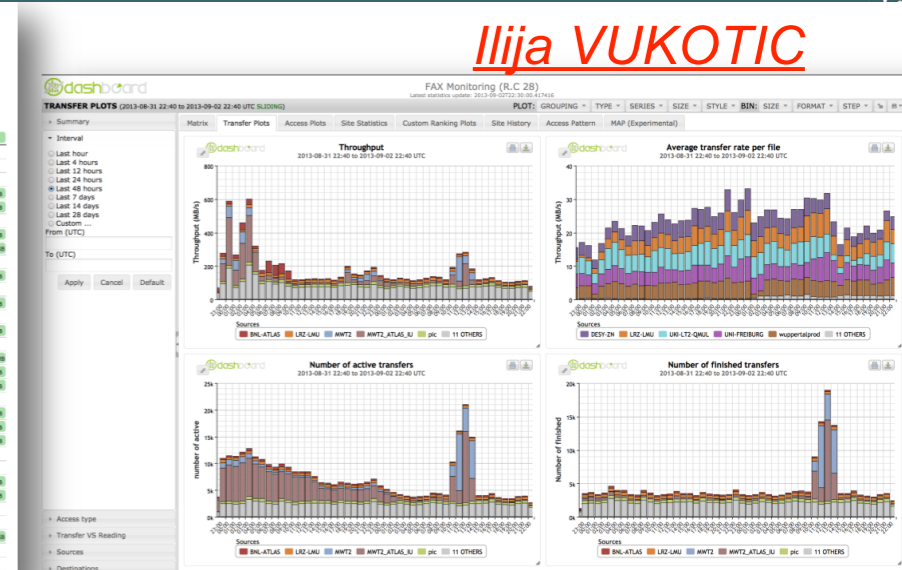
- Now **production-level service** on both ATLAS and CMS

- ALICE already do WAN fallback for some time

- Fallback; “Overflow”; Remote Data Production; Opportunistic

- Not much use of caches yet

- Lots of monitoring but do we “track the relative balance between managed and opportunistic transfers” as written in TEG : R2 ?



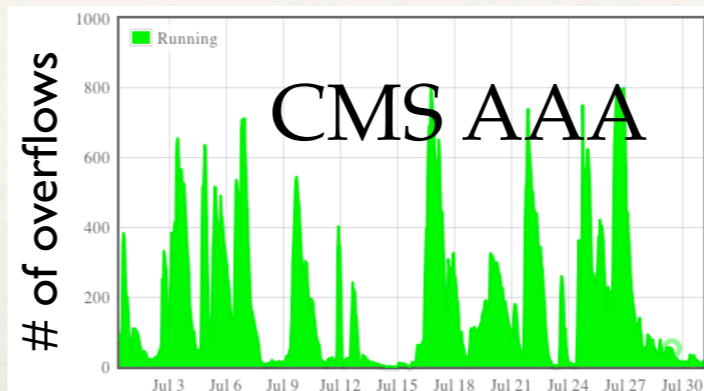
Panda Monitor
Times are in UTC

Panda report on jobs failovers to FAX over last 24 hours

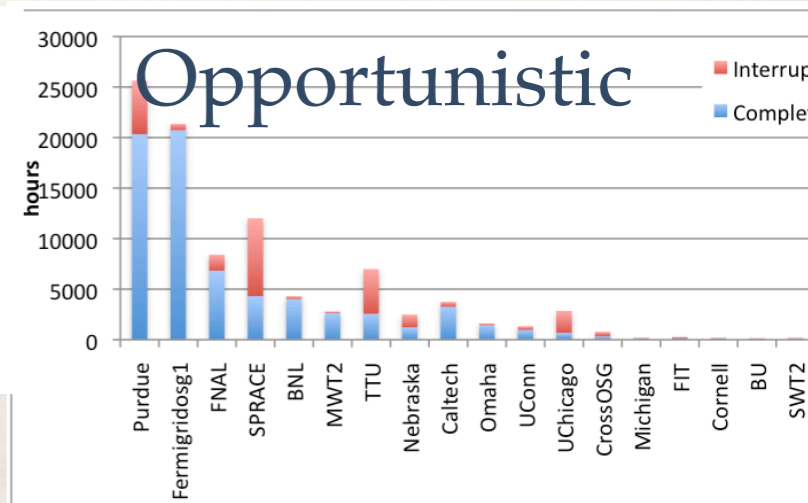
Record count: 138

Site	Jobs	WithFAX [files]	WithoutFAX [files]	WithFAX [GB]	WithoutFAX [GB]		
DE: GoeGrid	1	1	19	0.17	2.15		
FR: ANALY_LPSC	1	1	1	0.15	0.06		
PandaID	Time	WithFAX	WithoutFAX	WithFAX [GB]	WithoutFAX [GB]	Status	User
1951899183	2013-10-10 13:57:36	1	1	163463017	60679111	finished	mark hodgkinson
US: ANALY_MW2_SL6	127	136	6428	52.68	1089.72		
US: OU_OCHEP_SW2	9	9	99	5.38	38.39		

- Small number of jobs failing
- But these failures cost the most in terms of user's turn around time



Kenneth BLOOM



Data Federations

- ❖ Need to share lessons learnt across the experiment
- ❖ CMS is really going for “AAA” - total location independent
- ❖ I think we need to evaluate usage and revisit projected use (<~10% of bandwidth was written in TEG report) (TEG:R2-3)
- ❖ Also site / countries need to know how the changing requirements affects storage purchases / network provisioning etc.
- ❖ Various promising ideas with http that will be realized during Run 2 (e.g. http ecosystem ; rucio download , dmlite plugin). Http plugin to xrootd (TEG R1) in progress but not quite available yet.

How will this heterogeneous (xrd/http) landscape evolve...

BIG DATA

BBC

New Yorkers are learning to love big data

5 November 2013 Last updated at 00:03 GMT



Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

How behavioural design can overcome the dark side of big data

theguardian

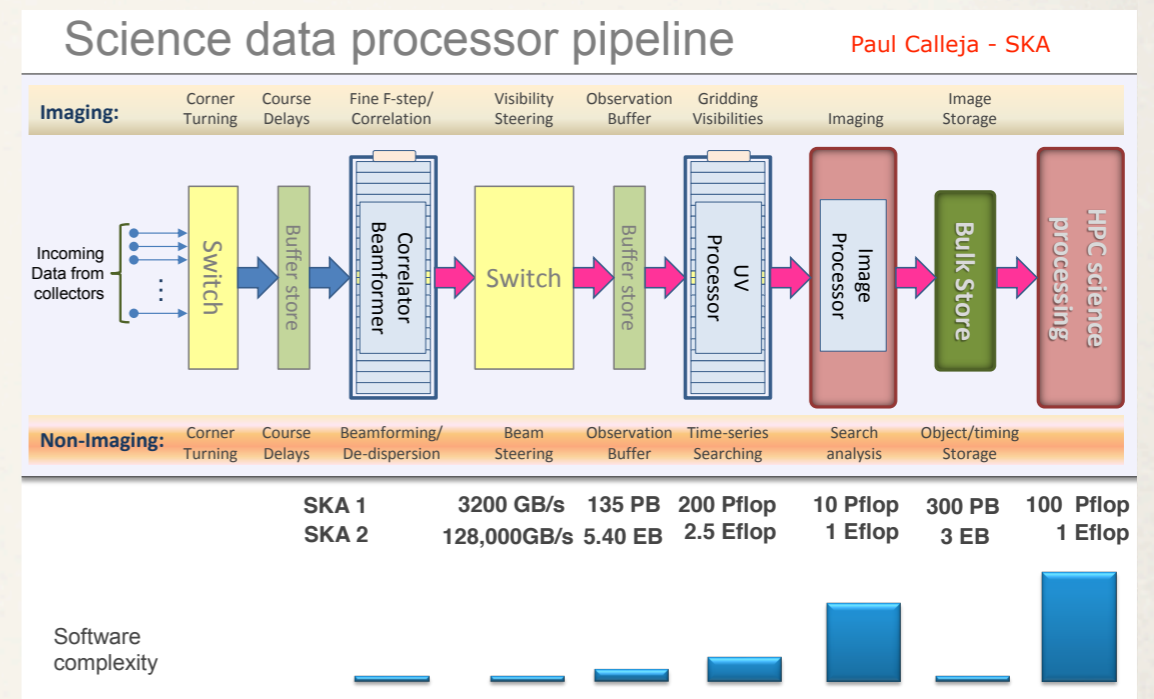
To avoid misreading the digital smoke signals, key insights from behavioural design should be a vital part of big data analysis

Christoph Burmester

[Guardian Professional](#), Friday 1 November 2013 17.40 GMT

Big Data - big crowd

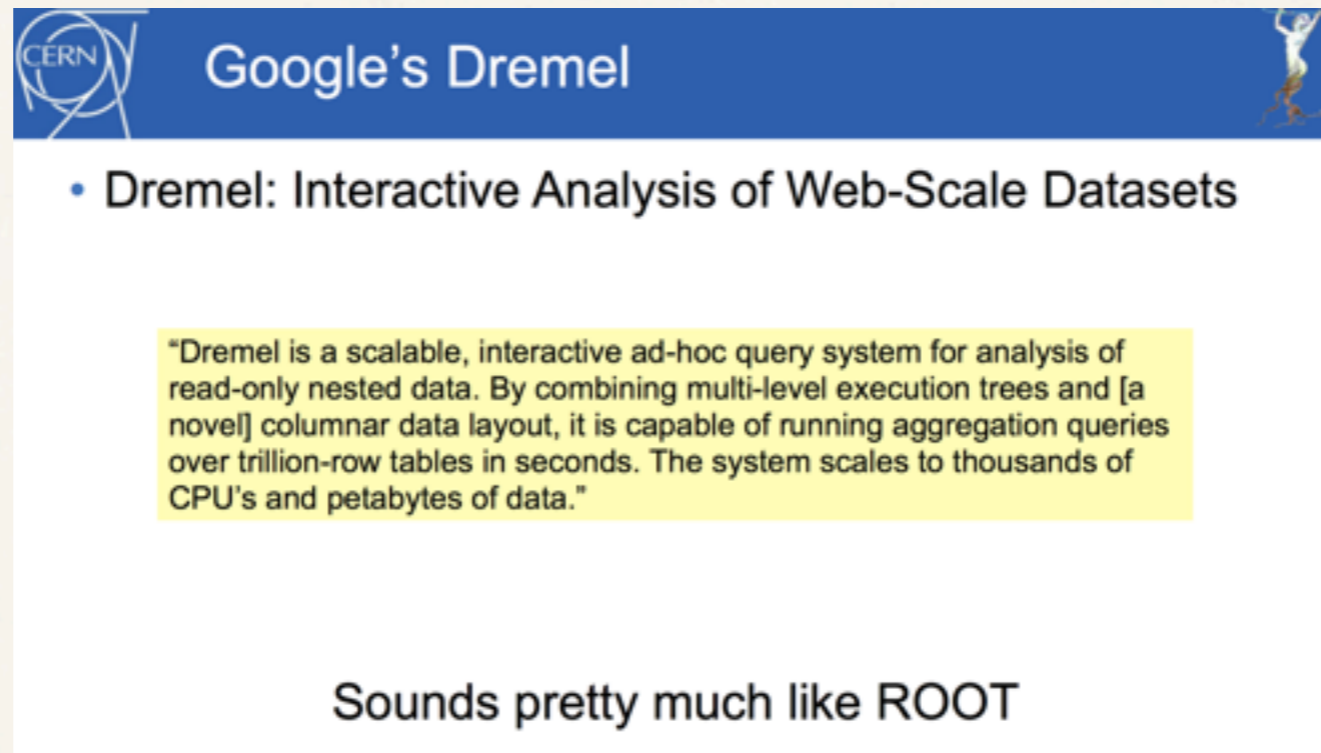
- ❖ Obviously HEP is no longer unique in data requirements.e.g. SKA->
- ❖ Other communities have different but overlapping challenges



- ❖ Opportunities to benefit from growth of “big data”
 - ❖ Impact of our work
 - ❖ sharing technologies / ideas.
- ❖ But we don't use the same tools (BD-hype in industry = Hadoop)
- ❖ And recently perhaps have not interacted enough....

Big data interactivities:

e.g. ROOT (Fons talk at GridPP Imperial Big data event):



CERN Google's Dremel

- Dremel: Interactive Analysis of Web-Scale Datasets

"Dremel is a scalable, interactive ad-hoc query system for analysis of read-only nested data. By combining multi-level execution trees and [a novel] columnar data layout, it is capable of running aggregation queries over trillion-row tables in seconds. The system scales to thousands of CPU's and petabytes of data."

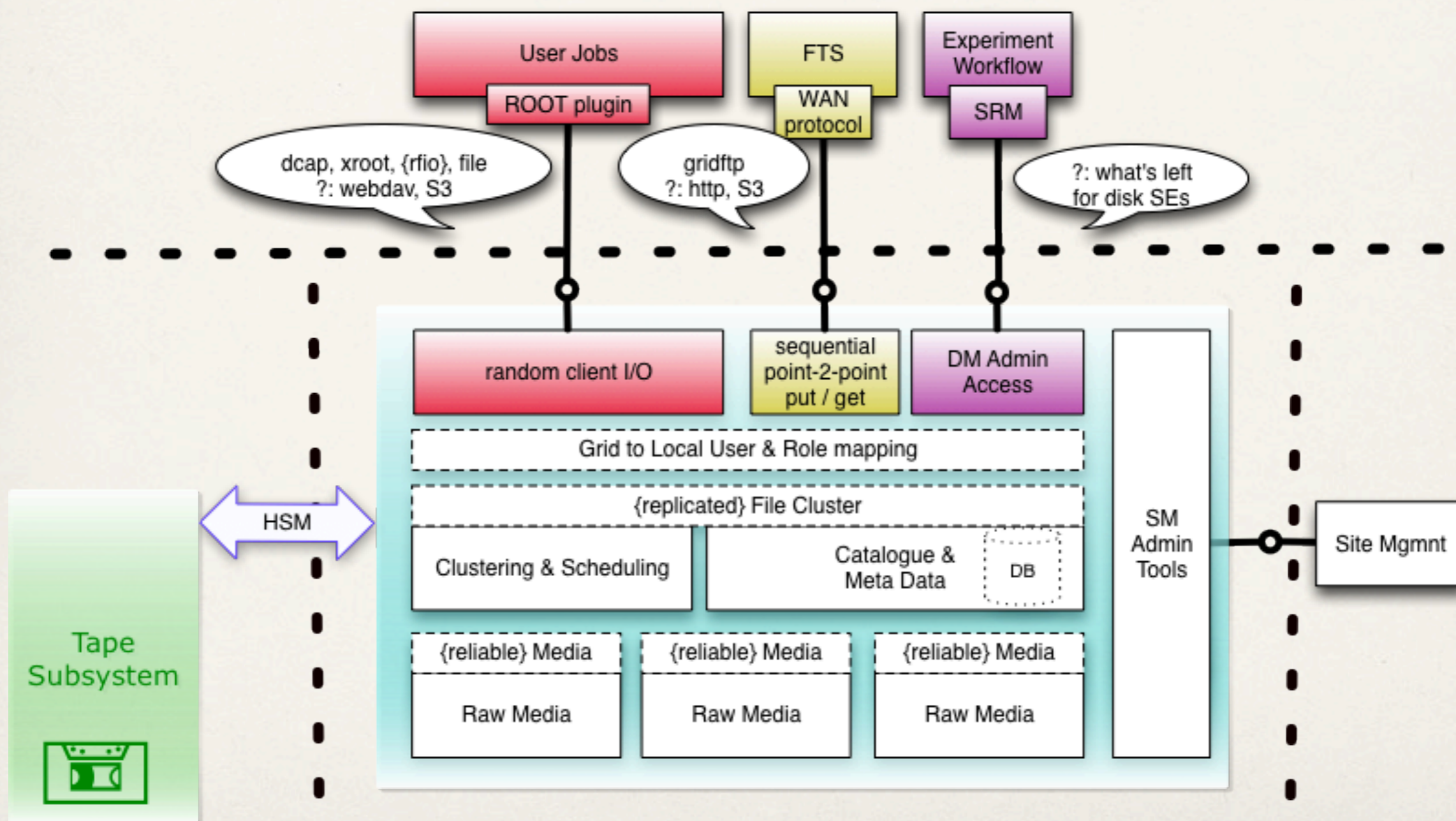
Sounds pretty much like ROOT

Dremel Paper goes on to say: "Column stores have been adopted for analyzing relational data [1] but to the best of our knowledge have not been extended to nested data models."

- ❖ This is a shame ... but, looking forward, we are building some fora for communication: e.g. Gridpp Big Data event
- ❖ This interaction will grow: should be both organisational and technical

Storage Management

Storage Interfaces

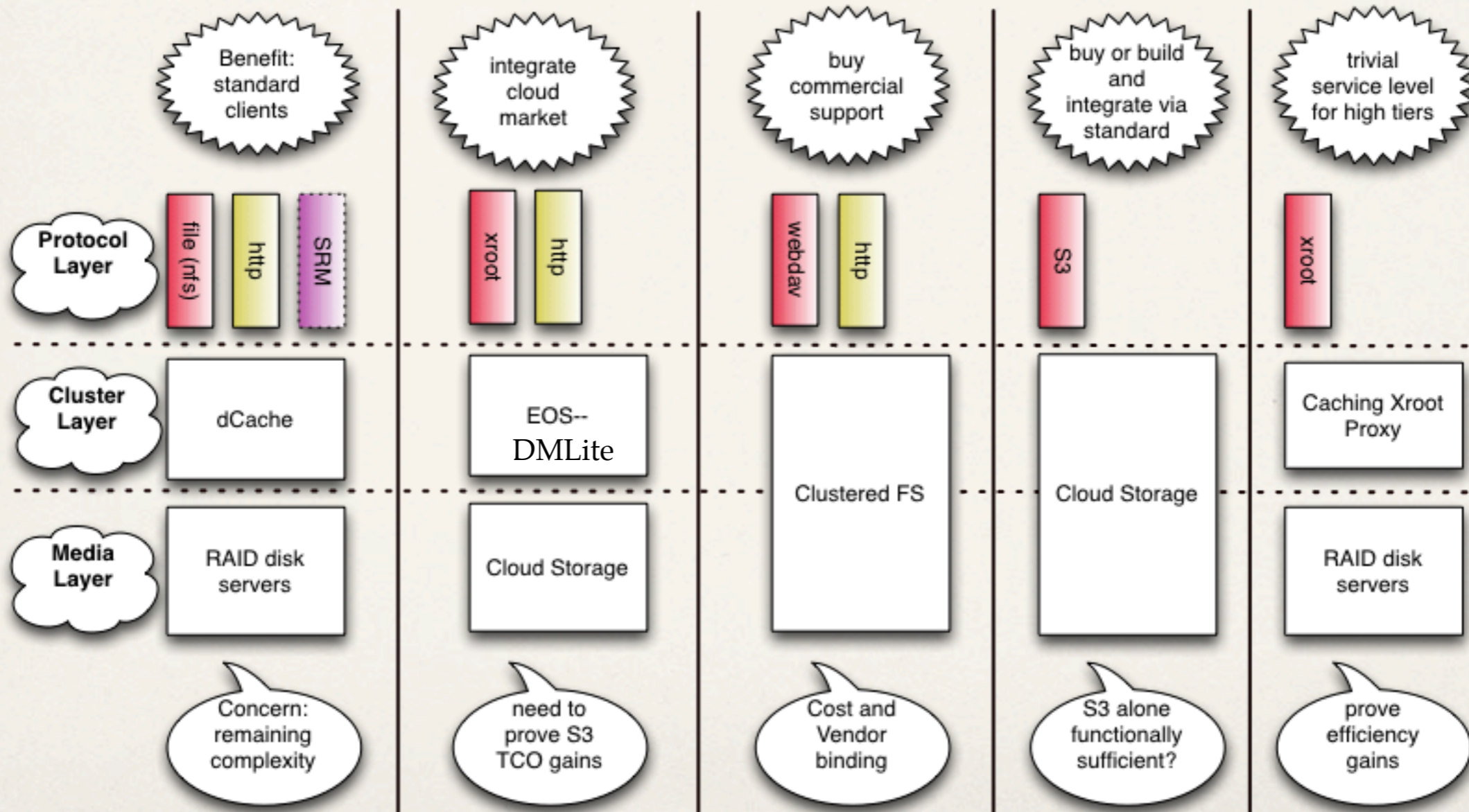


Storage Interfaces

- * TEG report discussed our Run1 storage management protocol “SRM” and current possible alternatives (e.g WebDav, CDMI):
 - * For archive sites SRM will continue to be used in Run 2 (TEG:R12)
 - * For disk-only sites non-SRM is being accommodated (TEG:R13) (gridFTP, xrootd, https for transfer; WebDav for management)
 - * Some experiments (e.g. CMS) there are SRM-free sites, others (e.g. ATLAS) still have some development - but this is progressing
 - * Being tracked in working group (TEG:R14 - see e.g GDB summary)

“Future interfaces”

TEG R14 said: “monitor and evaluate emerging developments in wider storage interfaces (e.g. Clouds) so experiments work together on long term solutions.”



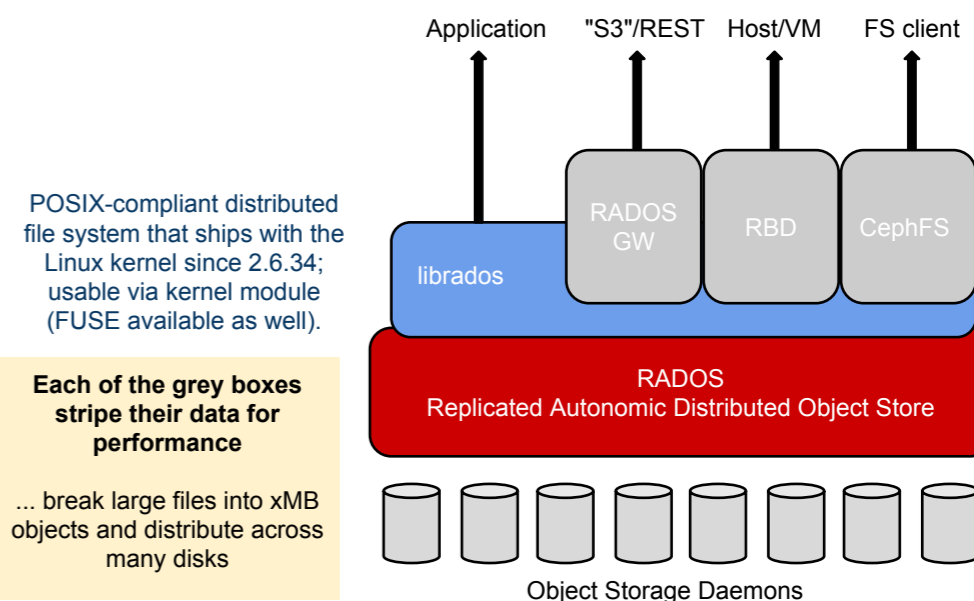
Site Storage - filesystems

- ❖ CEPH , HDFS, Lustre, GPFS
- ❖ HEP-specific storage layer will continue to be needed (in Run2)
- ❖ Some requirements relaxed: e.g. SRM
- ❖ Some not: accounting; authentication (and data access peculiarities: WAN access, random access etc.)
but why not (see later)?
- ❖ Our storage layers are evolving to support “modern” technologies

Ceph's architecture

CERN IT
Department

Daniel VAN DER STER



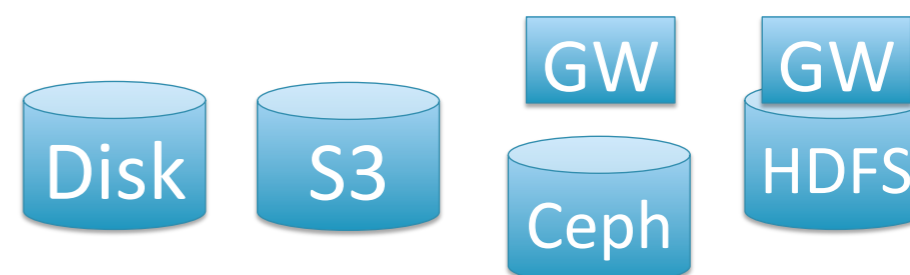
Our software: New DPM



anyDB

MD

Martin HELLMICH



Cloud Interfaces

Here I mean S3 / object stores

- ❖ Cloud will / has taken over the world (outside HEP).
- ❖ Various pieces of work on HEP use of S3 (see CHEP examples here ->)
- ❖ Can we use object stores as object stores (ATLAS probably could with rucio)

The image shows a ROOT C++ macro named `drawCloudHisto.cxx` and its output. The macro code is as follows:

```
1 void drawCloudHisto(const char* fileName)
2 {
3     // Open the remote file which contains the histogram
4     TFile* inputFile = TFile::Open(fileName);
5
6     // Load the histogram
7     TH1F* histogram = (TH1F*)inputFile->GetObjectChecked("h1gauss", "TH1F");
8
9     // Draw the histogram
10    histogram->Draw();
11 }
12
```

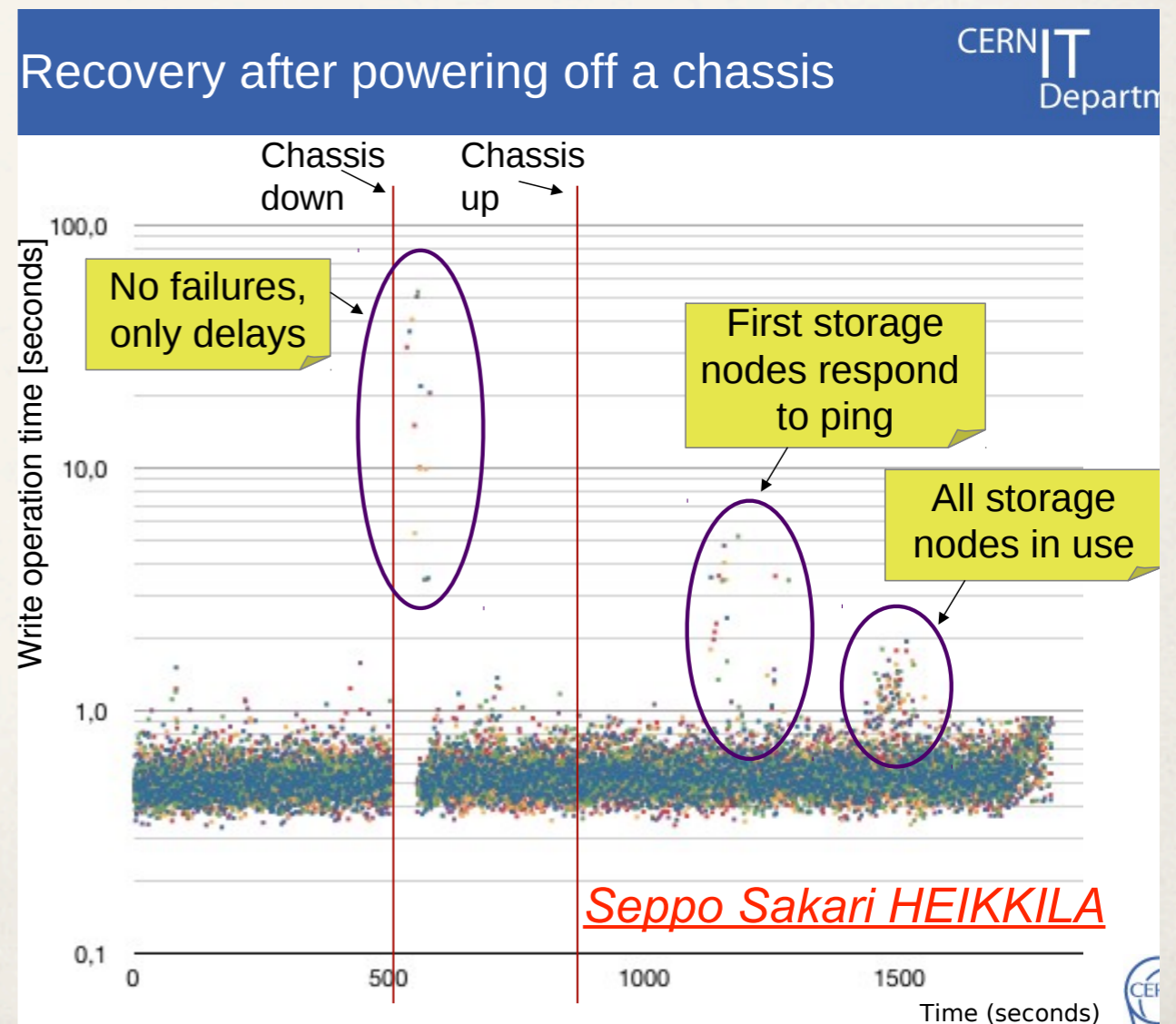
The output shows a histogram titled "Gaussian Distribution" with the following statistics:

h1gauss	
Entries	10000
Mean	0.008217
RMS	1.004

Annotations on the image:

- Backwards compatible**: Points to the `TFile::Open(fileName)` line in the macro.
- No cloud-specific code**: Points to the `TH1F* histogram = (TH1F*)inputFile->GetObjectChecked("h1gauss", "TH1F");` line.
- Load ROOT C++ macro**: Points to the `.L drawCloudHisto.cxx` command in the ROOT prompt.
- Draw the histogram contained specified in the remote Swift file**: Points to the `drawCloudHisto("swift://fsc.ihep.ac.cn:8080/root/gaussHistogram.root")` command.

With this extension, BES III can transparently use cloud storage



Really scaling

- ❖ Can we relax requirements for a certain part of our storage layer

- ❖ e.g. Object-store but also:

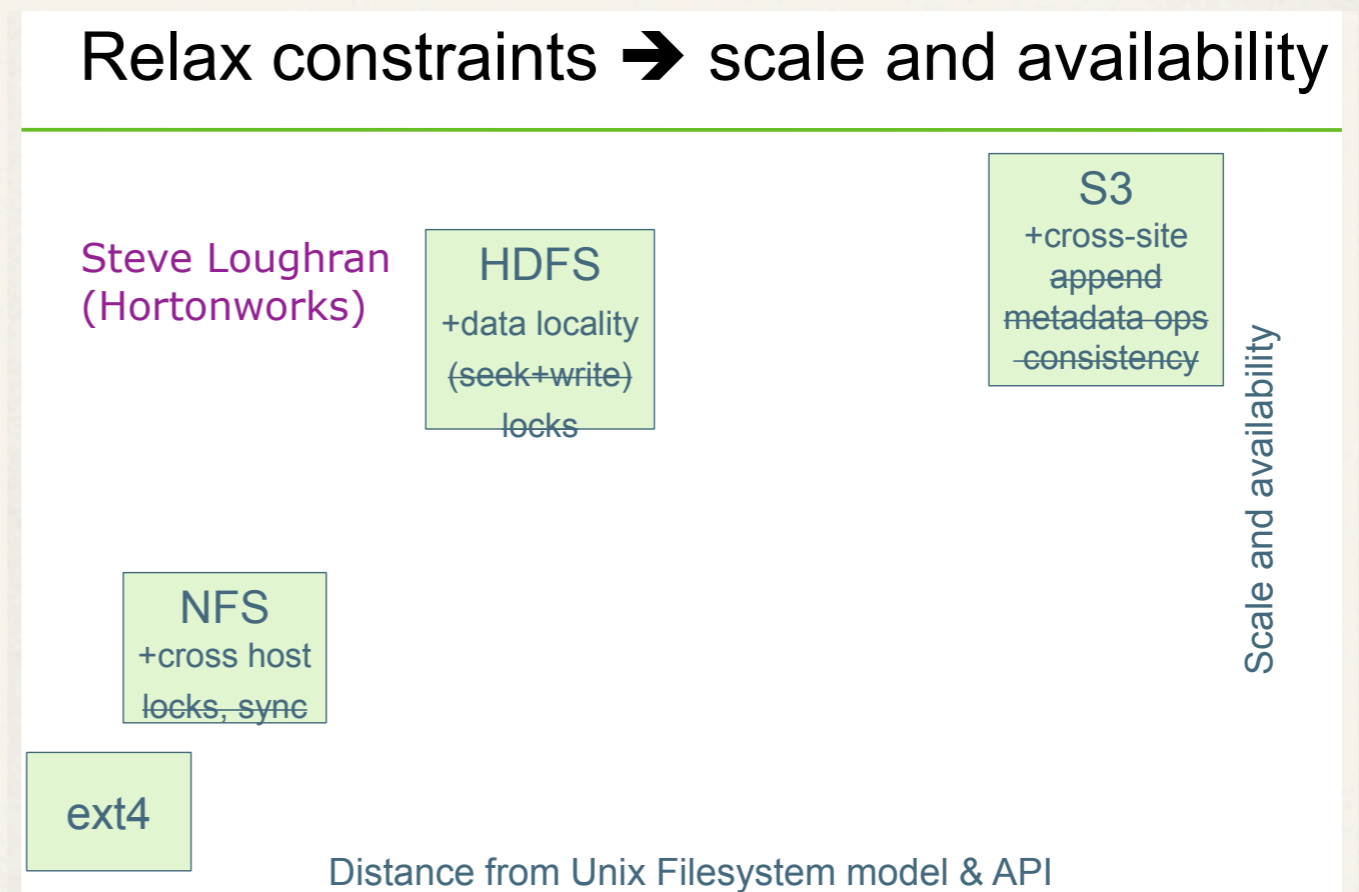
- ❖ Read-only

- ❖ Security (world readable)

- ❖ Relaxed consistency

- ❖ Needed also for caches

- ❖ And in practice many of our current spaces are read-only (write-once), world readable



I/O and benchmarking

TEG: R15-17,22

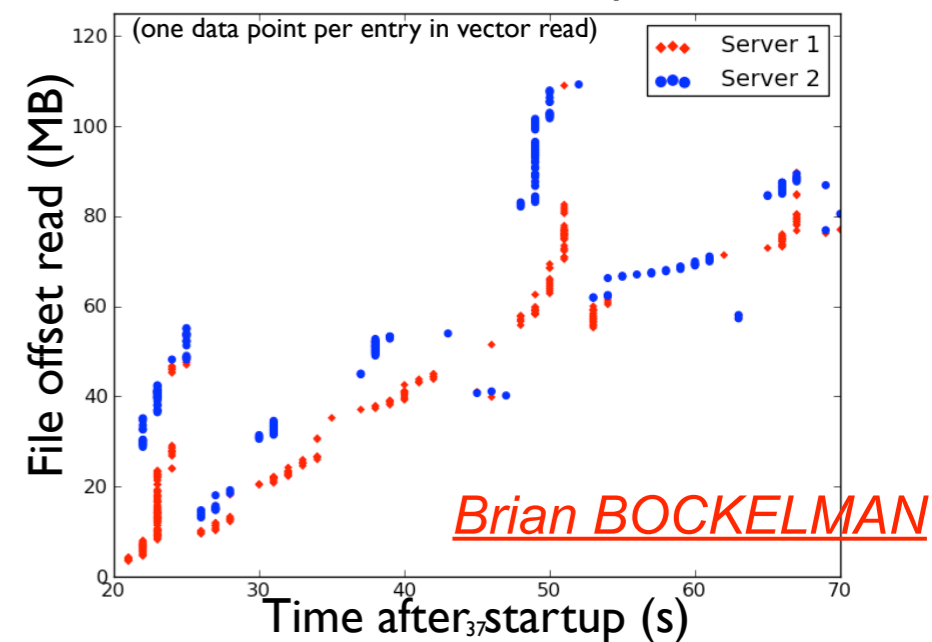
- ❖ Big improvements in experiments ROOT IO
- ❖ But our IO requirements need to be understood, stated more clearly, measured

Have rich data sets from xrootd monitoring

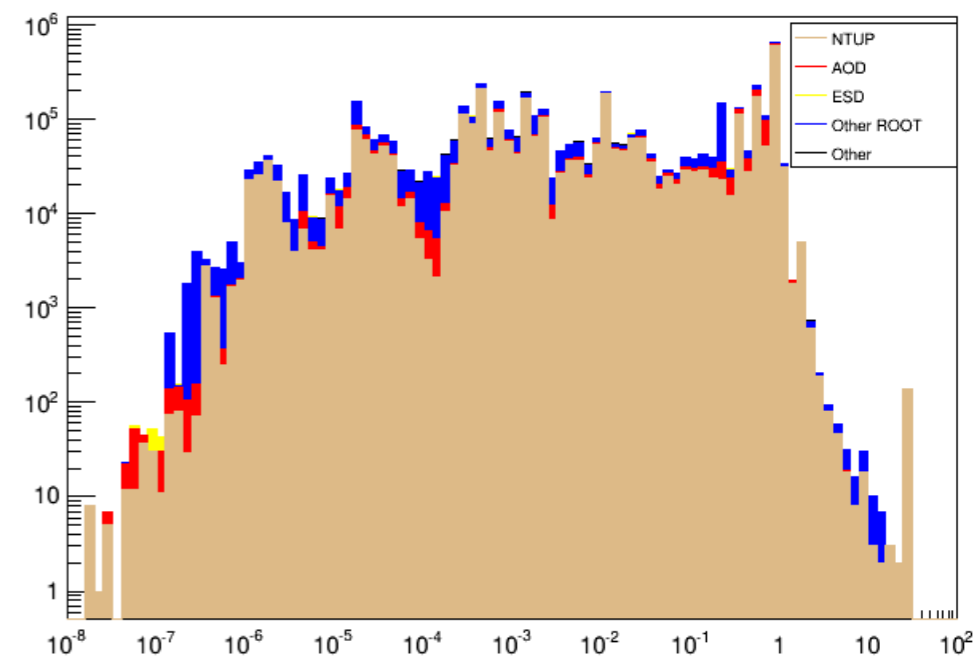
- ❖ Must further optimise: can also learn from other Big-data communities TEG: R19
- ❖ Organisational issues can have a big impact
 - ❖ E.g. New Atlas Analysis model needs to stay sane and not be circumvented

Multisource Illustration

Read offset versus time, per source



Fraction of file read (non xrdcp)



Site / Experiment interaction

- ❖ Currently the most effective sites are where there is someone involved in the experiment there. Possibly we rely on this too much.
- ❖ This is particularly true on the storage side:
 - ❖ Even if compute is outsourced to Cloud providers storage will probably still need knowledgeable “sites” (in Run2)
 - ❖ Though they could just run a front end with S3 etc. on backed
- ❖ WLCG provides a site/experiment interaction layer. This is not just in operations. Much to learn from sites expertise (TEG: R20, 21)

E.g. Have wider Data WG meetings with more site participation

Storage Hardware

Very brief here- would be valuable to have:

TEG: R18: Storage technology review

Incorporating vendors; spreading information between sites.

- ❖ Vendors are moving servers to satisfy “big data” market; disk supply influenced by e.g. changes in laptop market
- ❖ 5 year timescale - nothing that radical:
 - ❖ Larger drive capacities 4TB now
 - ❖ Also options for cheaper less reliable drives
 - ❖ SSDs / hybrid offerings
- ❖ Can we make use of popularity for site level tiered placement - keep available info for the site to do this (TEG: R22)

Summary

- * Many areas of **evolution**: filesystems, storage systems, data management services, federations.
 - * New, and established solutions, are being built and developed incorporating standards and flexibility to change - **Good news !**
- * **Performance is good** ... is it **good enough** for I/O challenges to come:
 - * How can we learn from and impact on “Big Data” communities
 - * Can we better understand (and relax) our requirements to scale
 - * I/O, Tiered storage; Object stores, Read-only caches, etc.

Backup: *A reminder of TEG
recommendations (v. brief version)
and a couple of other things*

Federations

- Current option is only xrootd
 - Activity in http that should be supported (e.g. [DPM](#))
 - (NFS 4.1 possible but not near happening for this)
 - R1: HTTP plugin to xrootd
- Activity in ALICE ; CMS; ATLAS
- All anticipate < 10 % of traffic this way
 - R2: Monitoring of federation network bandwidth
- Breakdown of what features experiments expect.
 - R3: Topical working groups on open questions

Point-to-point (WAN) Protocols

- GridFTP is ubiquitous and must be supported in medium term
 - R4: gridFTP: use recent versions; exploit session reuse.
- Xrootd is currently used alternative:
 - R5: Ensure xrootd well supported on all systems
- HTTP again a serious option (DPM \leftrightarrow dCache tests)
 - R6: HTTP: continue tests; explore at scale

Managed Transfer (FTS)

- FTS is the only tool and used for more than transfer
 - Though experiments will go their own way if need be
- R7: Update FTS3 workplan to include all requirements in report e.g. use of replicas; http transfers; staging from archive
- R8: Cross-experiment test of FTS3 features

Management of Catalogues and Namespaces

- R9: LFC not needed (by LHC) in med-term:
- Could be repurposed and useful tools (e.g. for consistency checking) should work with other catalogues
- Also : Storage system quotas not needed (handled by experiment)

Separation of archives and disk pools/caches

- All experiments will split archive (tape) and cache (disk pools):
 - Atlas; LHCb; Alice already: CMS plan for this year
 - R10: HSM still to be supported for managing disk buffer.
- A large separate disk pool managed through transfer offers advantages:
 - Performance: Lots of spindles.
 - Practicality: Need not be at same site.
 - R11: FTS should support staging (see R7); Experiment workflows should support this transfer model

Storage Interfaces: SRM and Clouds

SRM:

- Ubiquitous;
- Needed in short-term buried in exp. frameworks;
- Practical advantages from common layer

BUT:

- Not all functions needed/implemented;
- Performance concerns;
- Industry not using (and developing alternatives);
- Experiment frameworks adapting for alternatives.

SRM: Looked at each functional component:

Which used: (see big table in report for details)

Functional Group	Usage Observation
Storage Capacity Management	For Space Management: Only space querying used (LHCb; ATLAS) (not dynamic reservation, moving between spaces etc.)
File Locality Management	For Service Classes: on medium term, spacetokens could be replaced by namespace endpoints (no orthogonality required)
	For Archives: bringOnline (and pinning) needed – no replacement.
Transfer protocol negotiation	Data access interface (get tURL from SURL): needed by LHCb: Alternatives exist: e.g algorithms or rule-based lookup
	Load balancing and backpressure: Needed but alternatives exist (and backpressure not imp. in SRM)
Transfer and Namespace	FTS and lcg-utils at least should support alternatives

Looked at alternatives:

Some used by WLCG currently ([GridFTP](#) ; [xrootd](#))

Some in industry ([S3](#); [WebDav](#); [CDMI](#))

[Mapped to functions](#): (see big table in report for details)

Storage Interfaces: Recommendations

R12: Archive sites: maintain SRM as there's no replacement

- Non-archive no alternative yet for everything:
 - But experiments already looking at integrating

R13: Working group should evaluate suitability targeting subset of used functions identified in report

- Ensuring alternatives are scalable and supportable
- must be supported by FTS and lcg_utils for interoperability

R14: Working group should monitor and evaluate emerging developments in wider storage interfaces (e.g. Clouds) so experiments work together on long term solutions.

Storage Performance:

(Experiment I/O usage, LAN protocols, evolution of storage)

R15: Benchmarking and I/O requirement gathering

Develop benchmarks; Experiments forecast bandwidth IOPS and bandwidth needs; storage supports measurement of these.

R16: Protocol support and evolution

Experiments can use anything ROOT supports

But move towards fewer protocols and direct access supported.

ROOT; http direct access; and NFS4.1 should be developed

R17: I/O error management and resilience

Explicitly determine storage error types and ensure application handling

R18: Storage technology review

Incorporating vendors; spreading information between sites.

R19: High-throughput computing research

Not restricted to current data formats (ROOT);

Hadoop style processing or NextBigThing™

Storage Operations:

Site-run services: monitoring; accounting etc

R20: Site involvement in protocol and requirement evolution:

ie. site representatives on storage Interface working group to ensure proposals are manageable by them

R21: Expectations on data availability. Handling of data losses

Experiments should state data loss expectations (in MoU) and reduce dependence on “cache” data.

Common site policies for data handling (examples in report)

R22: Improved activity monitoring:

Both popularity and access patterns

R23: Storage accounting

Support [StAR accounting record](#)

POOL Persistency

- Recently LHCb moved so now ATLAS specific sw.
- Atlas also plan a move so:

R24: POOL development not required in medium term

Security

Separate [document with Security TEG](#).

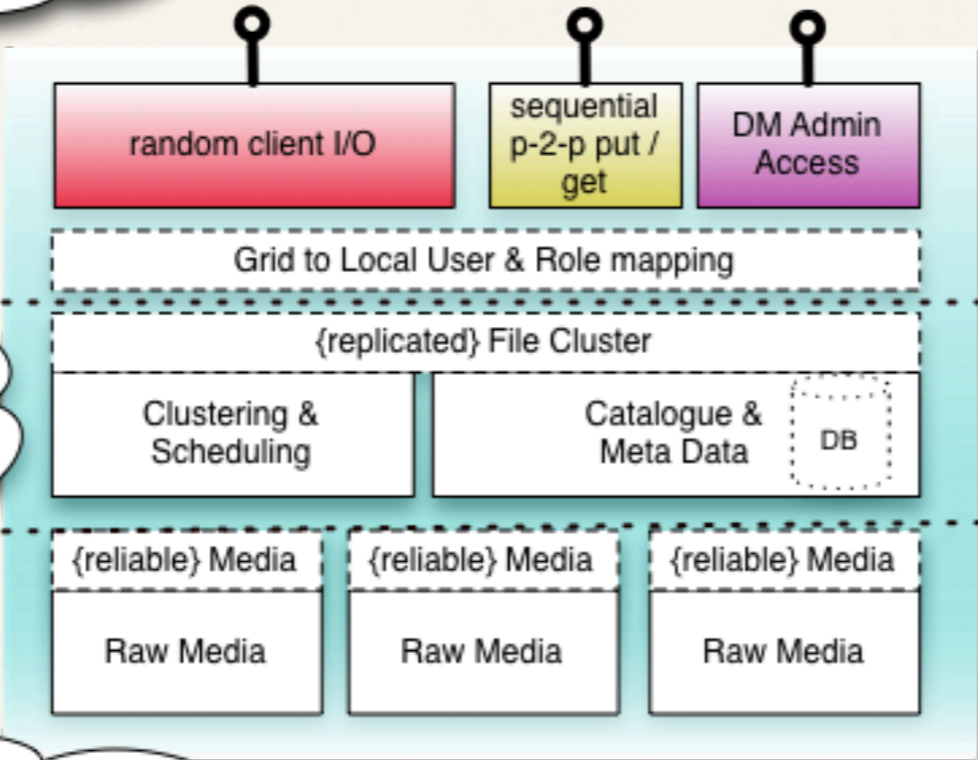
Areas that need attention in the near term:

R25: Removal of backdoors from CASTOR

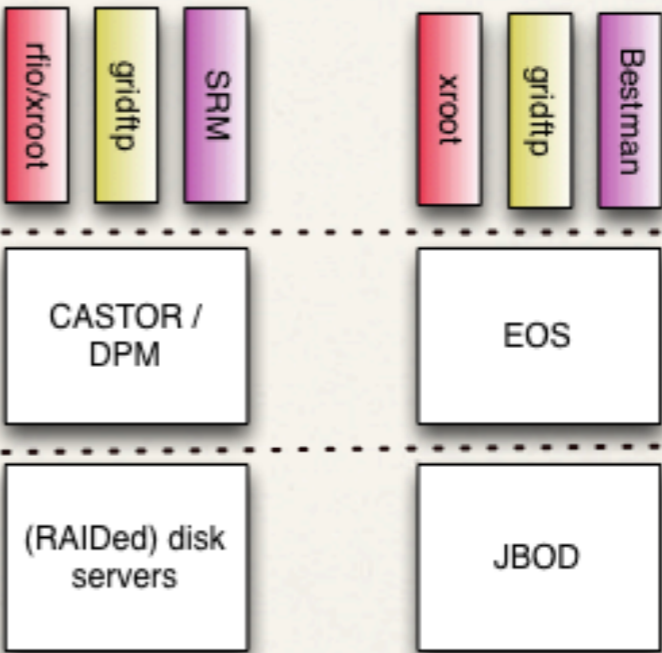
R26: Checks of the actual permissions implemented by Storage Elements.

R27: Tackling the issues with data ownership listed in document (e.g. ex. VO members; files owned by VO rather than individual)

User Protocol Layer
 local & WAN efficiency,
 federation support, identity
 & role mapping



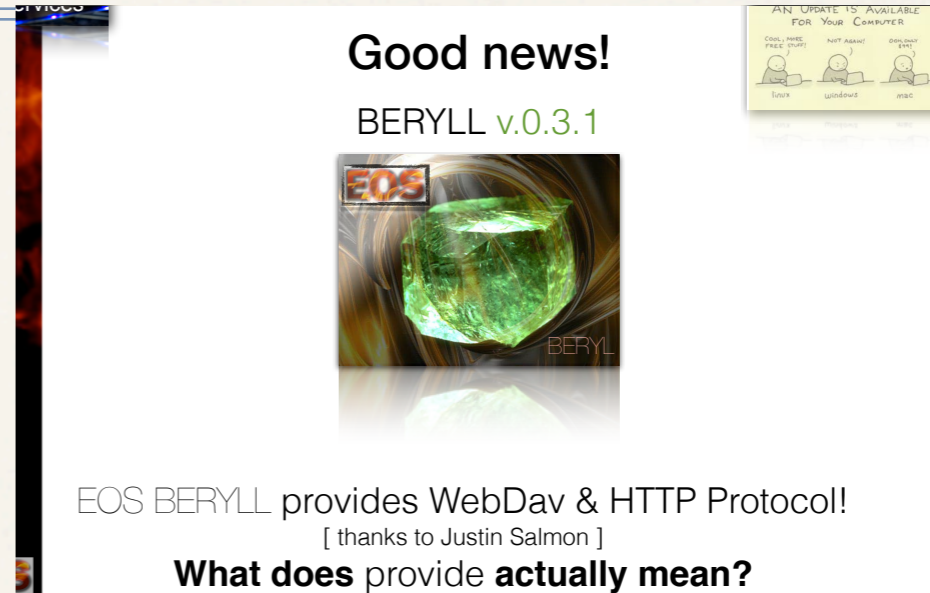
Cluster Layer
 scaling for large
 numbers of
 concurrent clients




Media Layer
 Stable manageable storage,
 scaling in volume per \$
 (including ops effort)

Interfaces: WebDav

- ❖ HTTP(s) / Dav - interfaces now in EOS , dCache , DPM , Storm .
- ❖ Used by atlas for renaming and plans for download etc.
- ❖ Davix allows for performant IO



Good news!
BERYLL v.0.3.1



EOS BERYLL provides WebDav & HTTP Protocol!
[thanks to Justin Salmon]
What does provide actually mean?

AN UPDATE IS AVAILABLE FOR YOUR COMPUTER

COOL, HAVE FREE STUFF! NET AGAIN! DOWNLOAD THAT!

YOUR WINDOWS MISC

