Data Curation in the Grid

Jamie Shiers, Information Technology Department, CERN, Geneva, Switzerland

Agenda

- Classify the problem(s) – the "Use Cases"
  1. Re-analysis of data from a previous facility, e.g. LEP
  2. Use of scientific data in education & outreach
  3. **Data Curation for a running machine, e.g. LHC**

- What are the specific issues related to, or benefits of "the Grid"
  - Briefly define "grid computing";
  - Differentiate between grid and Grid:
  - ➢ **What is our current experience with data & storage management in grid and Grid?**

- Outlook

What makes up data curation?

Data Curation comprises:

- Data management
- Adding value to data
- Data sharing for re-use
- Data preservation for later re-use

Data Curation Vision Statement:

- Data curation is not an end, but rather a means to collect, organize, validate and preserve data to address the grand research challenges that face society. Successful data curation will require strategic infrastructure building efforts that encompass hardware, software, and human resource development.

Conclusions – UNESCO Debate

- As long as advances in storage capacity continue there are **no significant issues** related to the **volume** of scientific data that must be kept [**experience later**]
- Periodic migration between different types of storage media must be foreseen [**more later…**]
- Specific storage formats must also be catered for – this can require much more **significant** (time consuming and expensive) migrations [**watch for paradigm shifts**]

- By far the biggest problem concerns **understanding** the data – there is currently **no** clear solution in this domain

How much data is involved?

- In 1998, the following estimates were made regarding the data from LEP (1989 – 2000) that should be kept

Experiment	Analysis dataset	Reconstructable dataset
ALEPH	250GB	1-2TB
DELPHI	2-6TB	
L3	500GB	5TB
OPAL	300GB	1-2TB

- **By today's standards, these data volumes are trivial**

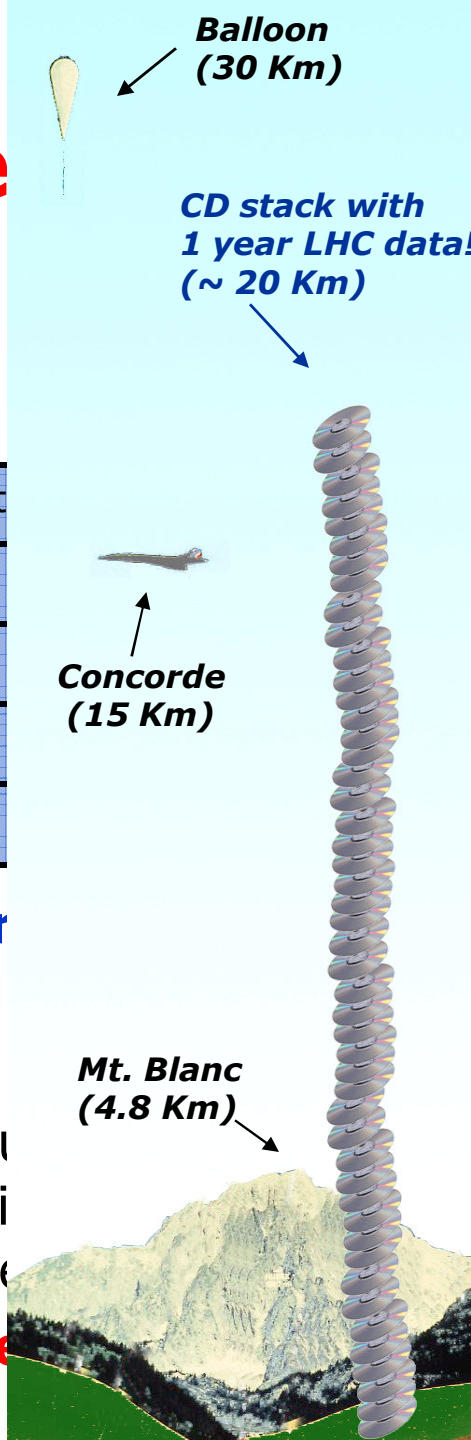
  - A 2TB storage device – with built in RAID – costs a couple of hundred CHF at MediaMarkt!
  - Even though the total volume of data at the LHC is much much higher, the data that must be kept beyond the life of the machine (2007 to ~2020) will be easily handled by then
- **The LHC will generate some 15PB of data per year!**

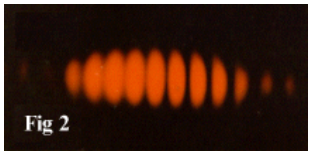
How much data is involve

- In 1998, the following estimates were made regarding data from LEP (1989 – 2000) that should be kept

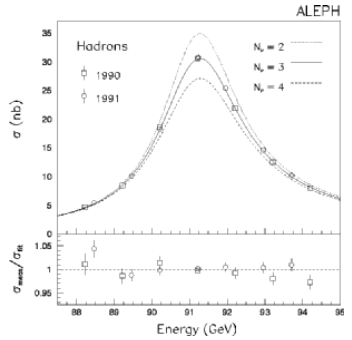
Experiment	Analysis dataset	Reconst
ALEPH	250GB	1-2TB
DELPHI	2-6TB	
L3	500GB	5TB
OPAL	300GB	1-2TB

- **By today's standards, these data volumes are tr**
  - A 2TB storage device – with built in RAID – costs of hundred CHF at MediaMarkt!
- Even though the total volume of data at the LHC is mu much higher, the data that must be kept beyond the li machine (2007 to ~2020) will be easily handled by the
- **The LHC will generate some 15PB of data per ye**





Use Cases Revisited



1. **Re-analysis of data – e.g. from LEP:**
   - Data volume: a few TB today;
   - Duration: a few years has stretched to **>1 decade**.
   - ➢ **Where will the analysis be done? [Not on "museum system"]**

2. **Use of data for education:**
   - e.g. perform fit on # neutrino families – a result that was widely publicized in the early days of LEP;
   - Duration: **100 years**? cf "Young's fringes" experiment;
   - 💣 **Nothing** is "standard" on the timescale of 100 years: multiple minor (e.g. "Excel version") and regular major data & storage migrations);
   - ➢ **Where will the analysis be done? [In "the cloud"?]**

3. **Data & meta-data curation for a running experiment:**
   - **Data volume: extends to many PB;**
   - **Duration: decades?**

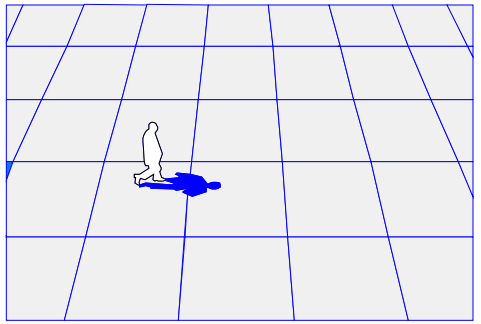
➢ **We will need to solve this last case for the LHC! A solution for the other Use Cases?**

What is Grid Computing?

- Today there are many definitions of ***Grid computing***:
- The definition provided by [1] Ian Foster in his article:

"What is the Grid? A Three Point Checklist" [2] is:

1. Computing resources are not administered centrally;
2. Open standards are used;
3. Non-trivial quality of service is achieved.



Why Grid Computing?

- Grid computing addresses two important issues:

1. The significant political issue of **funding**: it allows countries / funding agencies to spend money on computing & storage resources locally;

2. Scientific ***and*** socio-economic **benefits**: it allows labs and Universities to play a significant role in data processing and analysis – this reduces one of the oft raised criticisms of HEP

- It has also been **demonstrated** through a series of "challenges" to satisfy the needs of the LHC experiments – and now production data taking!

The Worldwide LHC Computing Grid

- Tier0 (CERN), ~10 Tier1s, ~100 Tier2s;
- Sum of resources at each tier approximately constant
- Specific roles assigned to each tier
  - Variations in computing models by experiment

- Tier1 sites must provide "custodial storage" for a significant fraction of the data! **[fortunately geo-plexed]**

- Storage management is much more than just "storage";
  - e.g. many Tier2s provisioned and configured for capacity – not access
- Data management is much more than storage management – involves multiple meta-data systems, databases (also required for storage management), file transfer and aggregation systems etc.

➢ **Both involve multiple complex hardware and software systems – all of which can and do fail! Regularly!**

The Worldwide LHC Computing Grid

- **Tier0** is at CERN. It receives the raw and other data from the Experiments' online computing farms and records them on **permanent** mass storage. It also performs a first-pass reconstruction of the data. The raw and reconstructed data are distributed to the Tier1 Centres.
- **Tier1** Centres provide a distributed **permanent** back-up of the raw data, **permanent** storage and management of data needed during the analysis process, and offer a grid-enabled data service. They also perform data-intensive analysis and re-processing, and may undertake national or regional support tasks, as well as contribute to Grid Operations Services.
- **Tier2** Centres provide well-managed, grid-enabled disk storage and concentrate on tasks such as simulation, end-user analysis and high-performance parallel analysis.
- In addition, CERN provides an **Analysis Facility** that has the functionality of a combined Tier1 and Tier2 Centre, except that it does not offer permanent storage of back-up copies of raw data.

Jim Gray's Advice

On one of his visits to CERN, Jim recommended we:

1. **Geo-plex our Data**
2. **Scrub it continually for errors**

- By "geo-plexing" he meant store multiple copies in different locations – perhaps in different formats
  - e.g. to suit specific access patterns
- By "following" his advice, we have recovered from data loss affecting ~100K files (several times…)
- But its not an inherent part of our global data management strategy… (Even if built in to the experiments' models.)

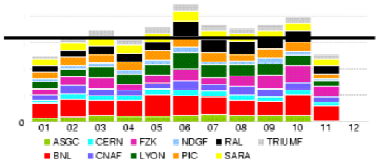
WLCG & Data Movement

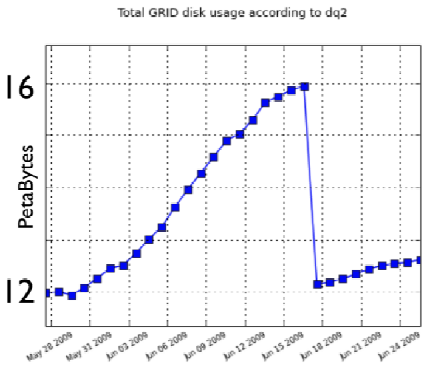
- Data movement is an intrinsic part of WLCG:
  - Pit to Tier0; Tier0 to Tier1s; Tier1s to Tier2s (and other Tier1s); Tier2s to Tier1s etc.
- CMS PhEDEx can "source" data from multiple sites – not just site having "custodial responsibility"

Data Distribution Results

- 4PB of data distributed
  - Large files! Large files!
- T1s all passed, some small problems in T1-T1 distribution identified and cured
- 54/67 T2 sites also made the metric (ranged from 100% to 5% share of data)
  - 13 fell short from slightly (99.7% complete) to catastrophically (25.9% complete)
  - Problems numerous: transfer service misconfiguration, SE instability, out of space, network bottlenecks

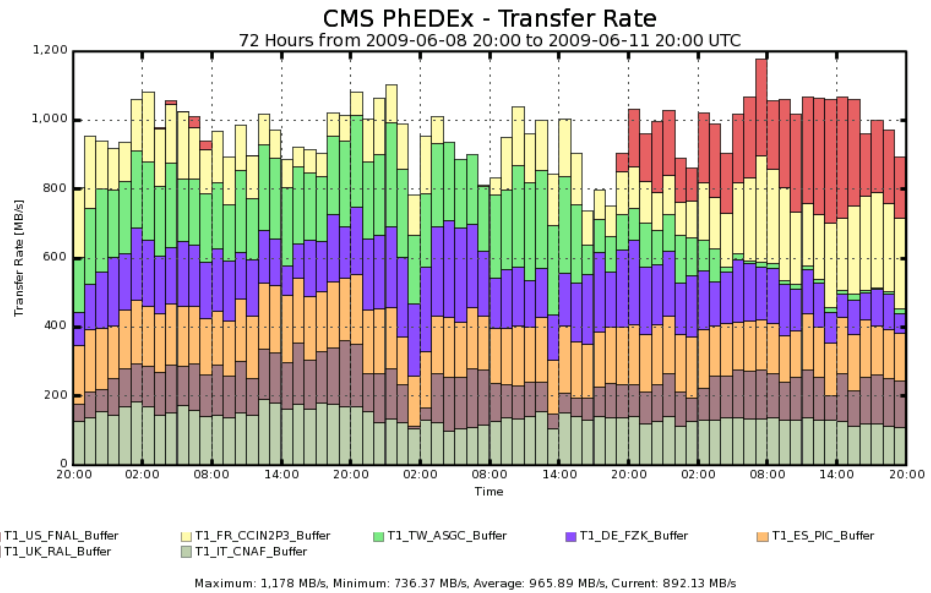
3GB/s





		Transfers
Cloud	Efficiency	Throughput
ASGC	99%	397 MB/s
BNL	84%	1128 MB/s
CERN	100%	334 MB/s
CNAF	98%	561 MB/s
FZK	85%	556 MB/s
LYON	96%	620 MB/s
NDGF	84%	137 MB/s
PIC	93%	429 MB/s
RAL	99%	838 MB/s
SARA	53%	262 MB/s
TRIUMF	100%	297 MB/s

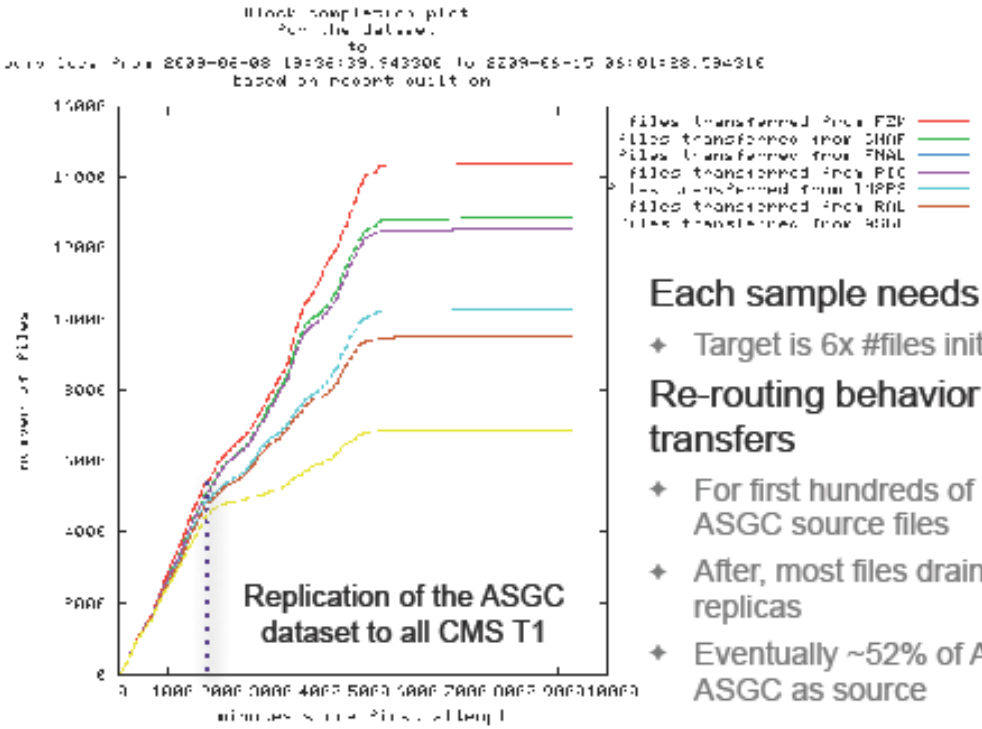
Peaks of 5.5GB/s



Transfer re-routing / load sharing (T1→T1)

As files are exported from source T1, PhEDEx starts to select other new replicas from other T1s

- Results in files not being routed from original T1 but rather redistributed within the full set of T1 sites



Each sample needs to be replicated at 6 other T1s

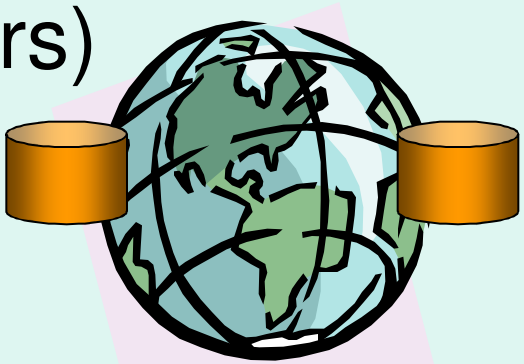
- Target is 6x #files initially at ASGC

Re-routing behavior clearly visible in STEP transfers

- For first hundreds of mins, most transfers start from ASGC source files
- After, most files drain to destination from off-ASGC source replicas
- Eventually ~52% of ASGC files were not started from ASGC as source

It's Hard to Archive a Petabyte
It takes a LONG time to restore it.

- **At 1GBps it takes 12 days!**
- Store it in two (or more) places online (on disk?).
  A *geo-plex*
- Scrub it continuously (look for errors)
- On failure,
  - use other copy until failure repaired,
  - refresh lost copy from safe copy.
- Can organize the two copies differently
  (e.g.: one by time, one by space)



http://research.microsoft.com/~gray/talks/cern_2001.

The Grid: Part of the Solution or Part of the Problem

- In the Grid we talk **interfaces** and not **implementation**
- Storage is a good example: SRM is the interface – there are multiple (partial) implementations and the full range of back-ends
  - e.g. dCache + HPSS or ENSTORE or TSM or DMF or
  - Storage devices and configurations vary significantly too!
- This has – on at least one occasion – saved us when silent data corruption only affected one family of storage **[recovery typically by experiments]**
- But this huge degree of **complexity** and the absence of a consistent high-level data management **vision** are probably not maintainable in the long term…

Current Data Management vs Database Strategies

Data Management

- Specify only interface (e.g. SRM) and allow sites to chose implementation (both of SRM and backend s/w & h/w mass storage system)

Databases

- Agree on a single technology (for specific purposes) and agree on detailed implementation and deployment details

WLCG experience from both areas shows that you need to have very detailed control down to the lowest levels to get the required performance and scalability.
How can this be achieved through today's (or tomorrow's) Cloud interfaces?
Are we just dumb???

The Way Forward…

- The minimum that we require is an integrated data & storage management **service** – even if implemented on top of independent (and separately managed) components (both site & VO)

- There is a large **opportunity** to provide a consistent data management strategy – building on what we have learned in 10 years of grid computing and taking today's technology into account

💣The current situation – with both data loss and / or corruption – is not **sustainable**

Just Startup Woes?

- Twenty years ago – in the early days of LEP – were things really much better?
- Some of the key data and storage management components were still being written or not fully deployed
- Major changes were around the corner: e.g. mainframe to distributed computing shift –
  "from supercomputing to super-market computing"

- The fact that we have **repeatedly** moved 1PB of data grid-wide a day and have achieved production status across a **world-wide** grid is a huge achievement!

Conclusions and outlook

- ATLAS has a robust Software & Computing system that can stand the impact of the first LHC collision data
- Nevertheless software development is never finished, as code optimization and improvements are always needed
- The Grid infrastructure, if used carefully and through official tools, can benefit all members of the Collaboration by providing computing power and data storage independently of the geographical location of the collaborators
- Operating this system needs a considerable amount of manpower but, thanks to its distributed nature, operations can be run for most tasks from home institutes

We are eagerly waiting for the first LHC data!

Summary

- Storage: solved in theory – still very (manpower) expensive in practice
- Data management: a major rethink of data management for grid & cloud environments is required – it will come because we need it
- Data access: an ignored problem
- Metadata: still in its infancy. When we can approach the level of a musical score we can claim progress but not success…

Conclusions

- We marvel how recent generations performed "cultural atrocities" – e.g. removing marble from the pyramids
- ***How will posterity consider us for failing to preserve our scientific legacy and their heritage?***
- Preserving knowledge in a way that it can be used by future generations might not be cheap but does this alone remove the **obligation** to make all efforts?
- There are many technical and cultural issues to be solved – e.g. "freedom" of data access, consistent use of digital metadata – these would also benefit current work
- **And the archives will only live as long as they are actively (and financially) supported**

The End