Data Storage Accounting at RAL

Rob Appleyard (GridPP/STFC),
Jens Jensen (GridPP/STFC)

Introduction

- First, a little about what who I am and what I do:
  - I manage the data storage at RAL - the UK's LHC Tier 1 site
  - CASTOR storage of data for WLCG* and local facilities
    - **C**ERN **A**dvanced **Stor**age manager
    - Disk & Tape
    - My responsibility
  - RAL is also developing a Ceph-based object storage system known as 'ECHO'
- This talk: Recent developments in our information system

Science & Technology Facilities Council

*Worldwide LHC Computing Grid

What's an 'Information System'?

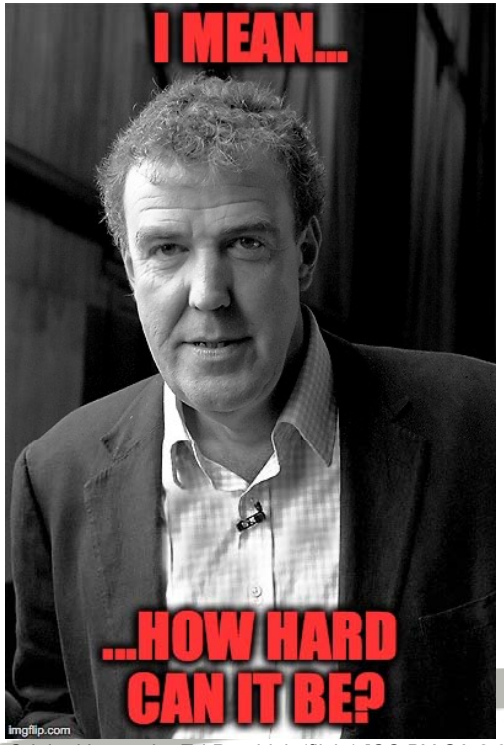
- It's a system that provides accounting information…
- What does an information system help people do?
  - Resource discovery
  - Resource accounting
- What information is provided?
  - Accounting metadata (space used/total)
  - Other metadata, like site location, contacts, etc.
- Who are the users?
  - Individual VO members ("Where should my data go?")
  - VO computing admins ("Do we have enough storage?")
  - WLCG management ("Are sites meeting their pledges?")

WLCG Information Systems

- Historically (from late ‘90s) used Globus Monitoring & Discovery System (MDS)
  - CERN implemented the BDII (Berkeley Database Information Index)
    - LDAP (Lightweight Directory Access Protocol) based, mostly using GLUE 1 (Grid Laboratory Uniform Environment) information schema
    - Actual file format is LDIF (LDAP Data Interchange Format)
- Hierarchical – resource (such as SE) -> site -> top level)

Isn’t This a Trivial Problem?

- So we should just publish the amount of data and capacity we have in whatever format is required, right?
- But it’s a bit more complicated than that…
  - Everyone has a different idea of what we should ‘obviously’ publish...
- In 2009, a WLCG document was published that tried to standardise all this.
  - We know it as “Installed Capacity”*
  - Doesn’t map well to modern requirements
  - Requires GLUE1 format
  - No longer viewed as authoritative



Original image by Ed Perchick (flickr) [CC BY-SA 2.0 (http://creativecommons.org/licenses/by-sa/2.0)], via Wikimedia Commons, modifi by author using imgflip.com

Science & Technology Facilities Council

*https://twiki.cern.ch/twiki/pub/LCG/WLCGCommonComputingReadinessChallenges/WLCG_GlueSchemaUsage-1.8.pdf

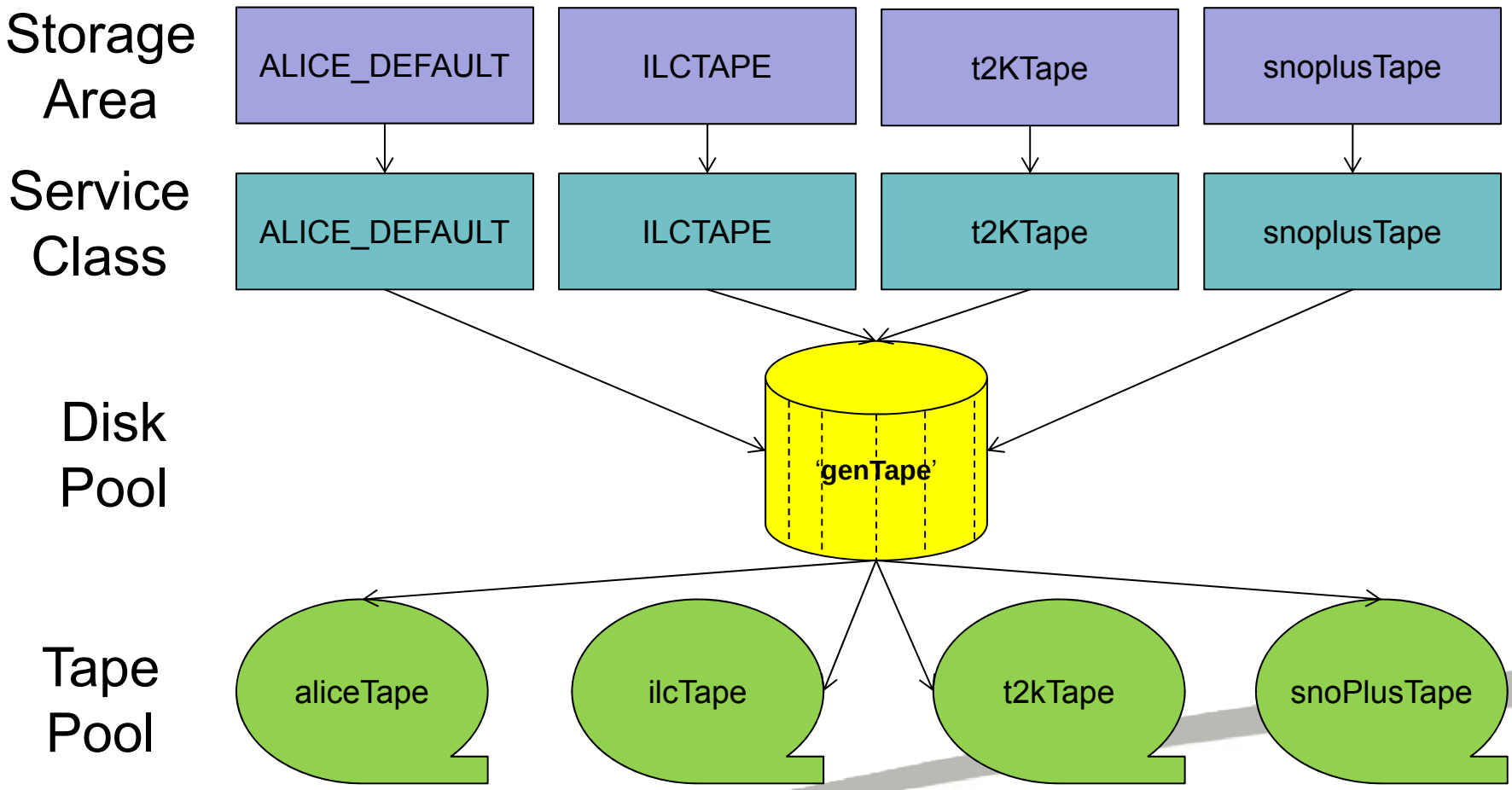
History Time

- RAL's existing accounting system:
  - 'CASTOR Information Provider' – "The CIP"
  - Written to be compliant with Installed Capacity
  - Output is in GLUE1-complient LDIF format for LDAP
- Now dated
  - We should have moved to GLUE 2 by 2012 ☹
  - Output data also viewed as 'inaccurate' by users due to mismatch between requirements of Installed Capacity and actual use cases.
- So: Project to build a replacement from scratch

Why is This Complicated?

Science & Technology
Facilities Council

Complication 1: Shared Spaces



Science & Technology Facilities Council

Complication 1: Shared Spaces

- Let's say we want to account for the disk cache element, which uses automatic garbage collection
  - 4 VOs share space, using u_1, u_2, u_3, u_4 out of total T
  - We also report a total T_n to each VO.
  - How to define T_n?
    - If we just say T_n = T for all VOs, we are giving the most accurate information to the users (it's a cache!)
    - But WLCG is not happy because we are quadruple-accounting.
    - If we say T_n = T/4 for all VOs, then we are reporting the real number (although WLCG still suspicious)
    - But then we are lying to users!

Complication 1: Shared Spaces

- Defining ‘Free’ is even worse…
  - 4 VOs share space, using u1, u2, u3, u4 out of total T
  - If Free for VO1 = F1: which is true?
    - ‘All the space that nobody is using’
      F1 = Total – (u1 + u2 + u3 + u4)
      Result: Double-accounting
    - ‘An equal share of the space that nobody is using’
      F1 = (Total – (u1 + u2 + u3 + u4)) / 4
      Result: Lying to user
    - ‘Divide free space in proportion to how much you are currently using’
      F1 = (Total – (u1 + u2 + u3 + u4)) * u1 / (u1 + u2 + u3 + u4)
      Result: Confused user
    - Something else?

Operation: Gordian Knot

- New implementation:
  - Just ignore the caches
    - They are largely transparent to users
    - Data throughput rates from users of shared tape resource rarely high enough to cause trouble.
  - We still have disk storage in CASTOR, but it is not shared.

Image from http://www.maa.org/external_archive/devlin/devlin_9_01.html, marked for noncommercial resue

Science & Technology Facilities Council

Complication 2: Tape

- Deleted files stay as “gaps” on tape until repacked
- Compression:
  - Data compresses on tape, so tape total capacity is “variable” according to how well data compresses…
  - Pledge is volume pre-compression
    - However users pay for the tapes ☺
  - Old solution: Previously based guess on average compression ratio
    - Assumed free multiplied by same factor
    - Which average – arithmetic or geometric mean?

Complication 2: Tape

- Tape allocated is “infinite”
  - Tape is in pool and allocated on demand
  - How does one publish ∞ in GLUE…?
- ‘Deletion’ is also a slippery concept here
  - Deleting a file on tape doesn’t immediately result in free space
  - Gain in space is not realised until we ‘repack’ the tape
  - Should we account for this?

Solving the Tape Problem

- Report used space from the VO perspective
  - Compression is not their problem
  - Saves VO getting worried because their catalogue says they are x bytes but we report a smaller number
- We now use our pledged data capacity as our ‘total’ capacity
- Repack is also not a VO problem, so don’t worry about the deletion issue
  - We can free tapes by repacking as/when required

Complication 3: Broken hardware

- What do you do when a storage node needs repair?
  - Assume data will be recovered…
  - But should we reduce our capacity while the node is unavailable?
  - Yes?
    - VO worried by ‘disappearing’ data
  - No?
    - VO has less usable capacity than we report
    - May hit trouble if storage full
- Eventual solution dicated by other considerations...

Complication 4: Multiple copies of data

- Replication within the SE (but in same storage tech/layer)
  - For durability (e.g. 0.999999s)
  - For availability (multiple copies of “hot data”)
- E.g. writing 1MB creates 3 copies
  - Has the user used 1MB? Or 3MBs?
  - Used goes up by 1MB, free goes down by 3MB?
  - Or is free pre-divided by 3? (like “usable space” for RAID)
  - But what if not all users/files have three copies, eg for dynamic replication?
- Solution: publish usable space as free; dynamically created copies (for availability, garbage-collectable) do not count

Complication 5: ECHO

- Ceph-based object store – new paradigm
  - One big instance, sliced up between users
  - 8+3 erasure coding – raw data is 37.5% bigger than nominal size
  - Capacity is defined by allocation, not underlying hardware
  - What do we do when the hardware fails

The Solutions

Science & Technology Facilities Council

CASTOR Disk Spaces

OLD	Production	Draining	Disabled	Readonly
Used	Used	-	-	Used
Free	Free	-	-	-
Total	Total	-	-	Total
Reserved	Total	Total	Total	Total

NEW	Production	Draining	Disabled	Readonly
Used	Used	Used	Used	Used
Free	Free	-	-	Free
Total	Total	-	-	Total
Reserved	Total	Total	Total	Total

CASTOR Tape Spaces

OLD	Full	Part-Full	Archived	Disabled
Used	Used	Used	Used	Used
Free	Free	Free	-	-
Total	Total	-	-	Total
Reserved	Total	Total	Total	Total

NEW: Ignore state of tape

- Used = Σ(Sizes of all user's files on tape)
- Free = Pledge - Σ(Sizes of all user's files on tape)
- Total = Pledge
- Reserved = Pledge

ECHO Accounting (Proposed)

- **Disclaimer: The ECHO development team haven't finalised their requirements; this is a proposal…**
- ECHO is analogous to tape from an accounting perspective
  - We propose to handle erasure coding like tape compression
  - Neither users nor WLCG care about underlying capacity, they care what can be used.
- Deal with hardware failure by hardware overcommit

ECHO Accounting (Proposed)

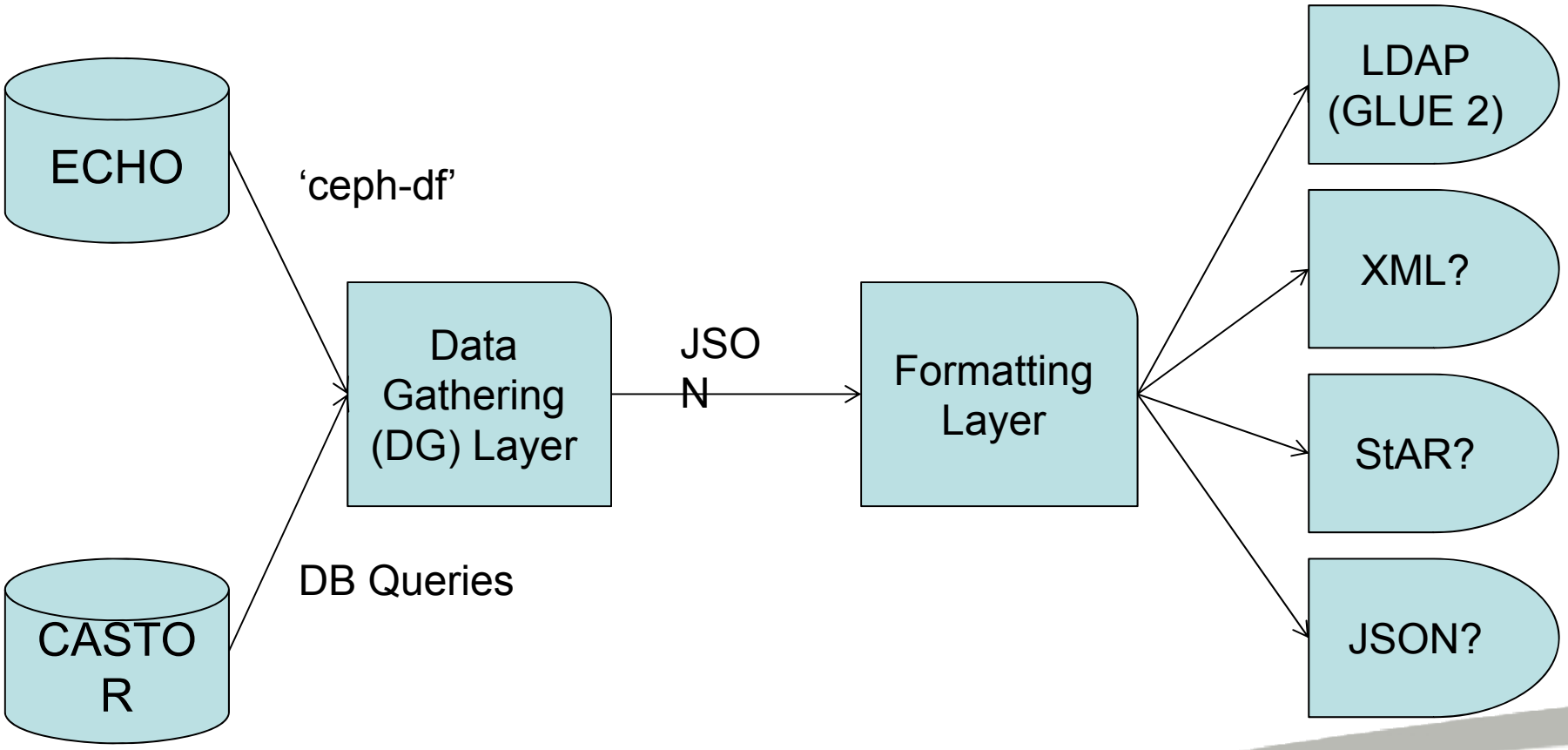
Handle much like tape!

- Used = Σ(All object sizes)/Erasure Code ratio (1.375)
- Free = Pledge – Used
- Total = Pledge
- Pledge = Pledge

Data Gathering

- Most of CASTOR data gathered from CASTOR's 'name server' DB
  - Big Oracle DB holding contents of namespace tree
  - RAL DBA Andrey Smirnov wrote an SQL query to return size of all data below given point in namespace
- Have to ask another DB ('Stager') for value of 'Free' and 'Total for disk.
- ECHO very easy
  - All necessary information can be trivially gathered using Ceph command line utilities
  - They can output JSON! :D

Accounting System Architecture



The Formatting Layer

- Converts JSON input to output suitable for external use
- The formatting layer's primary target is GLUE 2-formatted LDIF for LDAP for a GIP.
  - GIP: Generic Information Provider; a BDII plugin
- This satisfies WLCG requirements

Data Publishing (WLCG)

- Other accounting formats exist...
- Path not taken: If the output from the DG layer was in XML rather than JSON, StAR could be created trivially with using XSLT (XML -> XML conversion language)
  - StAR-formatted data can be published via SSM, a daemon which reads the StAR from a file and publishes to APEL
- https://wiki.egi.eu/wiki/APEL/SSM
- StAR is designed for cloud, and uses timestamps to show publishing “freshness”

Data Publishing (VOs)

- Big VOs often have idiosyncratic accounting requirements
  - ATLAS simply require JSON file uploaded *into the SE* in a known location…
  - Other VOs have been ignoring the old information system entirely, due to perceived mismatch between Installed Capacity and user requirements
- Users see content with proposed new implementation
  - We are hopeful for a good take-up of new system

Acknowledgements

- Funded by GridPP and STFC's scientific computing department
- John Gordon (STFC) for background information, especially on StAR
- Andrey Smirnov for the Oracle query
- Bruno Canning & RAL ECHO development team
- Maria Alandes Pradillo (CERN) and Guenter Grein (KIT) who have patiently waited nearly 18 months for us to solve the GGUS ticket that prompted this work!

Science & Technology Facilities Council

Any Questions?

Image by NASA, ESA, and the Hubble Heritage Team (STScI/AURA) [Public domain], via Wikimedia Commons

Science & Technology Facilities Council