Nordic Data Glaciers and Fjords

NelC NT1 Manager
Mattias Wadenstein
<maswan@ndgf.org>

2018-06-05
WLCG DOMA kick-off
CERN
Overview

• What are we doing
• Main challenges
• Future outlook
What are we doing

• NDGF is a single dCache instance with common namespace etc with disk and tape distributed over many sites
  - Central administration team
  - Storage pools bought and run by local sites

• Tier-1 pledged storage plus some integrated tier-2 storage
  - NDGF pledges for ATLAS and ALICE tier-1 storage
  - ATLAS storage part of Swedish and Slovenian tier-2 pledges
What are we doing

- admin
- GridFTP door
- xrootd door
- SRM door

ALICE VOBox:
- xrootd entry-point

NAGIOS

Pledged in 2018:
- 75 CPU (KHS06)
- 9 Disk (PB)
- 12 Tape (PB)
What we are doing

• Site operations for disk pools
  – Maintain hardware, networking, and OS
  – Provide a non-privileged user account with ssh login to central admin team
  – Takes 5-10% FTE, compared to the 20-50% FTE of a small Tier2 SE
  – Only providing office hours coverage (distributed setup means loss of a site is only a “service degradation <20%”)

• For tape pools
  – Some more complications in configuring the HSM connector
  – In addition to running tape library etc
ARC in a data lake

• With only a few large distributed storages computing will not rely on just data stored at local storage elements
• For batch CPU efficiency, some kind of caching mechanism seems to be needed
• ARC already supports “storage-less” computing for WLCG with local cache
ARC in a Data Lake

- ARC is location agnostic
- No problem getting some data from other lakes
- By staging data before running jobs, good CPU efficiency
ARC for caching

- ARC with 100TB of cache can support full ATLAS data flow on 5-10k cores
  - A size where IOPS-friendly storage is not horribly expensive
  - Reliability only needs to be OK, impact is crashed jobs not lost data
  - Makes it possible to buy the bulk lake storage optimized for size, aggregate bandwidth, and reliability
- Good cache reuse (>80% hit rates)
  - Size of cache impacts cache hit rate over time
- Protocol agnostic
  - Easy to add new protocols to inputs, no application support needed
Main challenges

• Overcoming resistance
  – Which sites should be folded into which?
  – Some sites have non-WLCG users of their storage elements
  – Different site policies on how to provide storage

• Visibility of tier-2 storage vs pledges
  – SIGNET-T2 in Slovenia has 0PB installed and 3.5PB pledged
  – NDGF-T1 has 3.5PB more disk installed than pledged
  – If we want this to be an attractive model, this needs some kind of fix on WLCG level
  – This issue is agnostic to technical implementation
Main challenges

• The building blocks need to be reasonable hardware
  – TB vs IOPS and MB/s (from disk and to network)
  – Reliability

• And have reasonable network policy
  – Direct access to pools from the world or only headnodes?
  – Firewalls in the way?

• Running out of internal bandwidth is painful
  – And with a distributed funding model remediation can be slow

• Large transition costs
Future outlook

• Large consolidated namespaces could open up for more classes in QoS
  - For some possible win in either more PB/€ or fewer copies needed
  - Only makes sense when there is sufficient scale for each class

• Standard protocols (i.e. https) for future-proofing and attractiveness to other communities

• More and more reliable storage for our users, the LHC experiments
Future outlook in details

• We could certainly integrate a “few” more sites
  – We are confident of being able to grow a factor 5-10 in terms of disk space, even when distributed

• Other sites with could also do this
  – If you can choose which sites should be folded into which
  – Switching from the “reading directly from local storage element” paradigm is probably the biggest shift
  – Using the AGLT2 approach of caching disk on the “nearby” disk is an option for handling this – or switch to use a caching CE
Nordic Data Glaciers and Fjords

Photo: Erlend Bjørtvedt (CC-BY-SA)
Questions?