Creating Federated Data Stores For The LHC

Summary

Andrew Hanushevsky
SLAC National Accelerator Laboratory

September 13-14, 2012
IN2P3, Lyon, France

Status

- FAX (ATLAS)
  - Increase analysis opportunities (WAN, failover, etc)
    - Ensures only "good" sites join the federation
  - *Adopting regional federation topology*
    - Vector clients to the closest currently available data source
  - Looking at WAN access vs local caching
    - Caching seems usually better go but hit-rate and storage issues
  - *LFC look-up is the major stumbling block*
    - Will this be the last time such DM will be developed?
  - Type of access in user jobs is also a challenge
  - Uses a rather detailed site certification process
  - Goal is >90% of available data by the end of the year

Status

- AAA (CMS)
  - *Increase analysis opportunities* (WAN, failover, etc)
    - Ensures only "good" sites join the federation
  - Covered the US sites and now working on worldwide sites
  - WAN access is supportable and works well
    - Caching seems usually better go but hit-rate and storage issues
  - Next years tasks include
    - Hardening
    - Public cloud usage (depends on pricing)
    - Data aware job management.
    - Caching proxy
  - Client changes take a long time!

Workload Management

- Panda
  - Stage in data from the federation as fallback
    - Could be used as a tape avoidance system
  - Direct access later (1st for input, later for output data)
    - Jobs shipped to available sites with lowest data access cost

WAN Access

- Need to optimize data access for high latency
  - TTreeCache is key here (chachinh, async pre-reads, etc)
- Caching also helps reduce WAN load
- Monitoring is crucial to track WAN access
  - Identify badly behaving applications

Recent Development

- New xroot client available in September
- EOS being extended for federation & monitoring
  - Pretty much done, minor additions needed
  - LFC translation is ready to go for ATLAS
- DPM (rewritten)
  - *Multi-VO data access* that can federate using xroot
    - Implemented as plug-ins to basic xrootd front-end
  - Waiting for the final 3.2.x version of xrootd going into EPEL
    - Will then be available in EMI repository
- dCache is adding N2N plugins equivalent to the xrootd ones.

Monitoring

- *UCSD Monitoring is in a mature state*
  - ActiveMQ feed almost done (needed for world-wide view)
- Monitoring useful for measuring data popularity
  - Rich set of data for mining to see n-order effects
  - Countless ways to render the data to gain usage insights
- Information is being rolled into dashboards
- There is 3-5 months more work to what we need
  - *But how long before people deploy the new stuff*?
- Seems like this will continue to be very active
  - At least for the next year
- We need to start thinking about monitoring for multi-VO sites.

Public Clouds

- Federation on cloud storage
  - Example single data point: 11-16MB/sec/stream on EC2.
  - Storage: 1TB $100-125/month
  - Data flow in cloud 2-3x better than reading from outside
- There is no real cloud standard
  - This makes moving from cloud to cloud difficult
  - EC2 is the biggest player and most popular
    - Google & Microsoft are showing promise as competitors
- Using federated storage (in & out) is key
  - Leaving data in cloud is expensive

WLCG & Federations

- Publicize the concept of federated storage
  - Working group to explore concepts (ends at end of year)
- *Separate "federation" from the apps that use it*
  - Allows better exploration of fail-over, self-healing, caching ...
- Biggest issue is the historical lfn->sfn translation
  - Complicates creating an efficient global name space
- Federated storage is rapidly progressing
  - Need still to understand security & monitoring issues
- Working group still active

EUDAT Federation

- In year 1 or of 3 year project
  - Provide storage infrastructure for sharing data
- Very diverse group of people
  - Much like a multi-VO system on steriods
- Decided to use iRODS
  - Provides federation glued via a database (similar to LFC)
    - Works well within the confines of itself but has scalability issues.
    - Does not seem to integrate very naturally with other systems.
- Looking for other systems (Xrootd, HTTP) to externalize access

NorduGrid Federation

- dCache based storage federation with caching
  - Data caching is significant activity for worker-node jobs
    - Competing I/O – caching vs jobs
    - Cache should be fast for random I/O but managed store need not
      - Cache also solves a number of operational issues.
      - Have a rule-of-thumb sizing caches for WLCG sites: 100TB cache per 1PB store.
- Looking at new ways of federating the caches
  - Either xrootd or ARC based HTTP
- Lesson highlights
  - Have product developers on your own staff
  - Availability is an upper bound for user's happiness
  - One system for all is the (unobtainable) holy grail

HTTP Federations

- EMI funded project revolving around DPM
  - Current system based on Apache + lcgdm_day + dmlite
    - Plus arbitrary plugins
  - Development is ongoing to handle edge cases
    - Endpoint changes (e.g. life, content)
- Project work to align xroot & http approaches
  - http plug-in for xrootd signed off
  - Other possibilities being explored

HTTP Federations

- EMI funded project revolving around DPM
  - Current system based on Apache + lcgdm_day + dmlite
    - Plus arbitrary plugins
  - Development is ongoing to handle edge cacses
    - Endpoint changes (e.g. life, content)
- Project work to align xroot & http approaches
  - http plug-in for xrootd signed off
  - Other possibilities being explored

Goals

- Driving forces for federation
  - Create more opportunities for data access.
  - This seems to be strong enough to foster several efforts.
- Outline broad technical solutions
  - We outlined several monitoring solutions for cohesiveness
  - Protocol alignment to allow more inter-play
- Establish framework for technical co-operation
  - We have this meeting and WLCG framework
    - As witnessed by the protocol alignment project
- Revisit our definition…..

Definition: Storage Federation

- Collection of disparate storage resources managed by co-operating but independent administrative domains transparently accessible via a common name space.
- Maybe we don't change the definition, but differentiate the things unique to the work discussed here:
  - Single protocols? Maybe not
  - From any entry point, access to all the data in the system.
  - Direct end-user access to files from source?

Next Meeting

- So, should we have another meeting?
  - If yes, Jeff Templon offered to host at NIKHEF
- Half-day virtual meeting @ pre-GDB meeting in April.
- Dates (green preferred):

M	T	W	R	F	S	S
18	19	20	21	22	23	24
25	26	27	28	29	30	

November 2013

Thank You

- IN2P3
  - For hosting this meeting
- Jean-Yves Nief
  - For local organization & meeting web page
- Stephane Duray & administration team
  - For excellent logistical support