Users experience in data access (Life sciences)

Geneva, September 28, 2006

Johan Montagnat

Requirements

- **Both files and associated metadata**
  - Distributed files
  - Distributed metadata
  - Complex metadata schemas (keyed to files or not)
- **Access control**
  - Fine grain, role based (ACLs including groups/roles)
- **Third party protection**
  - Encryption (protection from malicious users and resource admins)
- **Confidentiality**
  - Data anonimization (application level)
  - Hiding owners identity
  - Yet, re-identification is sometimes needed (pseudonimization)

Data Management SystemS

- EDG RLS (Replica Location Service)
  - Used in the biomed VO until spring 2006.
  - Security problems. Now deprecated.
- LFC (LCG File Catalog) + GFAL (File Access Library)
  - Production file catalog used by default
  - Some stability problems encountered after activating it (Spring 2005)
    - LFC failures, periodical restarts
    - Database connection re-initialization
  - Hopefully solved in latest versions
  - Currently limited in terms of security (VO wide access to SURLs)

Data Management SystemS

- **GliteIO (DMS client/server) + FiReMAN (File Catalog) + Hydra (Encryption + Key Sharing Service)**
  - High level security (ACLs, encryption...)
  - Still partly supported but phased out
  - Client available in gLite 3.0, servers to be installed by users (in particular gLiteIO servers need to be installed for each target SE)
  - Plans to integrate Hydra + extended security in LFC as a replacement
- **Perroquet + Encfile**
  - Encfile: encryption service (similar to Hydra)
  - Perroquet: PAROT based file access wrapper
  - used on LCG production infrastructure

SEs limitations in terms of security

- **SRM v1 security limited**
  - no access control to SURLs
- **SRM v2 will provide SE-side ACLs**
  - leading to an ACL coherency problem (different replicas of a same file on different SEs have different ACLs)
- **The alternative is gLiteIO server**
  - The user is not directly authorized to access storage. The gLiteIO service has access instead. It enforces access controls checking.

Drug Discovery (DC) experience

- **Avian Flue Data Challenge**
  - Spring 2006, using LFC/GFAL
  - Thousands of jobs submitted daily
  - A wrapper script attempts to recover in case of any middleware error to improve the system robustness
- **Regular DMS errors**
  - Up to 5% failures
  - Both for data retrieval (lcg-cp) and output data replication (lcg-rep)
- **A globus-url-copy based back-up procedure used in case of failure**
- **Problems with file systems limitations (number of files)**

Medical Data Manager experience

- **Very high security requirements**
  - Medical image
  - Direct interface to clinical storage
- **The MDM is coupled to gLiteIO + Fireman + Hydra**
  - All files (images) are registered in Fireman
  - A gLiteIO server controls access through ACLs
- **Still on-going development**
  - Bug fixes
  - Problems related to certificate chains and services authorization
  - Changes in default security procedures
- **Two independent file catalogs have to be maintained**
  - Two different clients on the application side

Interfacing to medical storage

- **Mostly an application level problem**
  - MDM is relying on the DICOM (image format + commication protocole standard)
  - Application level anonimization of data
- **But middleware could help**
  - Hooks to application-specific modules for security, data decoding, etc.
- **Much more to be done on in the medical area**
  - Various Picture Archiving and Communication Systems (PACS)
  - Non-DICOM based patient information (Radiological Information System, Hospital Information System, HL7 standard...)

Metadata

- **Currently using AMGA grid-interface to DBMS**
  - ACL based access control
  - Encrypted communications through SSL
- **Satisfactory situation on the middleware layer**
  - Functionality and performance OK
  - Much more to do in the application-specific area
- **Major challenge: distributed metadata federation**
  - Single view
  - Single query interface
  - Metadata storage distribution

Coupling WMS with data

- **We once were in the DataGrid project**
  - Exploit massive data parallelism: Single Processing, Multiple Data
  - Batch-oriented system efficient in this respect
- **Multiple-data jobs**
  - Independent description of processing and input data to process (not possible with JDLs)
  - Single submission for a job to be repeated over as many input data as desired
- **Multiple-data workflows**
  - The same when considering workflows: decoupling workflow description and data to process (not possible with DAGMan)

Wrap-up

- **Complex requrirements related to data management**
- **Different file catalogs**
  - LFC / FiReMAN
  - non interoperable
- **Different tools**
  - GFAL / gLiteIO, different security levels
  - Hydra / EncFile
- **Instability problems still experienced recently**
- **Some progress at an application area level to develop specific services**

Perspectives

- **Secured SEs (SRM v2 with ACLs)**
- **Integration of LFC/GFAL and Hydra**
- **More application level data management**
  - Anonymization
  - Metadata schema
  - Study more complex clinical use cases related to authorization
- **Databases federation**