POOL & ARDA
a few first thoughts
(not much discussed inside POOL yet..)

Dirk Düllmann,
IT-DB & LCG-POOL

ARDA workshop, 21 January 2004

POOL interest in ARDA

- POOL will be one client of ARDA services among others
  - ARDA will be a framework for running POOL based (analysis) tasks
    - At the moment POOL is mainly used in production
    - POOL will need to address analysis area this year
      - Request for integration of POOL ref and collections with ROOT as a shell is the next step
  - POOL needs to be able to integrate with ARDA provided services: mainly catalog & collections
- POOL needs to be well aligned with the main ARDA concepts
  - Eg Collection, Dataset, File

POOL Collections

- **Role of event level collections and meta data in a grid environment needs clarification and prototyping**
    - Expect active collaboration with ARDA to come up with a model for deploying collections in production and analysis environments
    - Integration with experiment frameworks just starting
    - Still many open questions about requirements
        - Is there a Collection Catalog (like the File Catalog)?
            - Re-use of catalog implementations?
            - Or are collections/datasets just more general files?
        - Consistent catalog & meta data distribution – how ?
        - What collection meta level data needs to be kept?
        - How do POOL collections tie in with ARDA and with grid middleware?
- **Several implementations exist in POOL and are being picked up for prototyping**
- **Collection implementation in POOL is only a first step**
    - But the real issue is not the implementation but the conceptual model
    - Need active experiment involvement (and some agreement) in this area

ARDA interest in POOL

- POOL provides deployment models, interfaces and implementations interesting for ARDA
  - Provide a model for cross population between implementation using catalog meta data queries
- POOL file catalog may be seen as an implementation of file (and later dataset) catalogs
  - See first successful use of POOL file catalogs in production activities
- POOL collections should become an implementation of the (more generic) ARDA datasets concept
- More importantly:
  - by now widely accepted model of cascading catalog implementations should be considered for ARDA
    - Central services alone will not fulfil all use cases
- Extraction and publishing mechanism should be designed into all catalogs and associated meta data systems

A common deployment model?

- Several implementations of distributed analysis frameworks exist but..
  - There are too(?) many parallel streams
  - Need to factorise out the commonalities
- Need a minimal but commonly shared model first
  - Minimal set of end user services/concepts
    - Eg Collection, Dataset, Task, Job, Application, Run …
    - Ideally user requirements should be expressible/expressed only using those concepts
  - Minimal set of lower level (implementation) concepts
    - Eg Replica, GUID, File, FileCatalog, ..
    - Many system admin / production manager type requirements can be expressed using these concepts.
  - Which level "owns" (defines and implements) it?
  - What does each service/concept provide semantically?
  - Can we define a shared minimal interface..
    - …and multiple implementations which may extend the basic interface

- End-to-end services in a layered system
  - May involve several concepts/services
  - Each of which may "live" on a different layer
- Example: POOL Collections
  - POOL (Event) Collection - POOL end user concept
    - Access to collections of POOL objects
    - Integrated with PROOF analysis back end
  - Dataset – ARDA layer concept
    - Access to abstract data elements from a set of closely related files
  - File – generic GRID middleware concept
    - Access to a unstructured set of bytes in file somewhere on the grid

Summary

- POOL will be a major ARDA client
  - Provides end user concept of persistent object collections
  - Needs to stay aligned with ARDA concepts and services
    - In particular in the File Catalog and Collection area
- Propose a common work package between POOL and ARDA in particular in the Collections/Dataset area
  - If POOL interface does not fulfil experiments needs we should change it – rather than having eg different catalogs
  - If POOL implementations don't fulfil experiments needs we'll improve them or change to agreed ARDA ones
- POOL provides deployment models, service interfaces and several implementations to ARDA
  - Would like to participate in the more detailed definition of a common ARDA deployment model