Data Catalogue.

ICAT at DESY: Challenges in Integration and Deployment

Jürgen Starek
DESY Scientific Computing Dept.
Hamburg, January 2013

Data Sources

- DESY Photon Science sources
  - Petra III
  - CFEL
  - Flash
- Currently approximately 1.8 PB
  - 300 TB on disk, remainder on tape
  - dedicated dCache instance (see www.dcache.org)
  - Many tar'ed small files, little to no Nexus containers
- 67 TB already available in internal catalogue (Gamma Portal)
- Expected data input from Petra III:
  - O(500) TB per year
  - O(3000) beamtimes per year

Current data flow management work with Petra III

> Data access for participants

- **Gamma-Portal**
- dCache, Oracle, APEX
- internal catalogue
- staging as well as downloading

> Data access for the public

- **ICAT**
- dCache, Oracle, J2EE
- public downloads
- federated catalogue

> Cross-workflow problems

- Authentication and authorisation of guest researchers
- Mapping guest researchers in DOOR and Registry (internal WUO systems)
- ACLs and POSIX-rights in the backend file system
- Access rights to analysis machines

> Supporting the migration to Nexus/HDF5

Talking to the users

- No clear use case
- Hard to agree on metadata sets (Nexus “application definitions”)
- No manual work for beamline staff or researchers at data ingestion time
- Worries about rights management
- BUT: Hopes for
  - searchability
  - ease of data transfer
  - speeding up users' visits to the beamlines

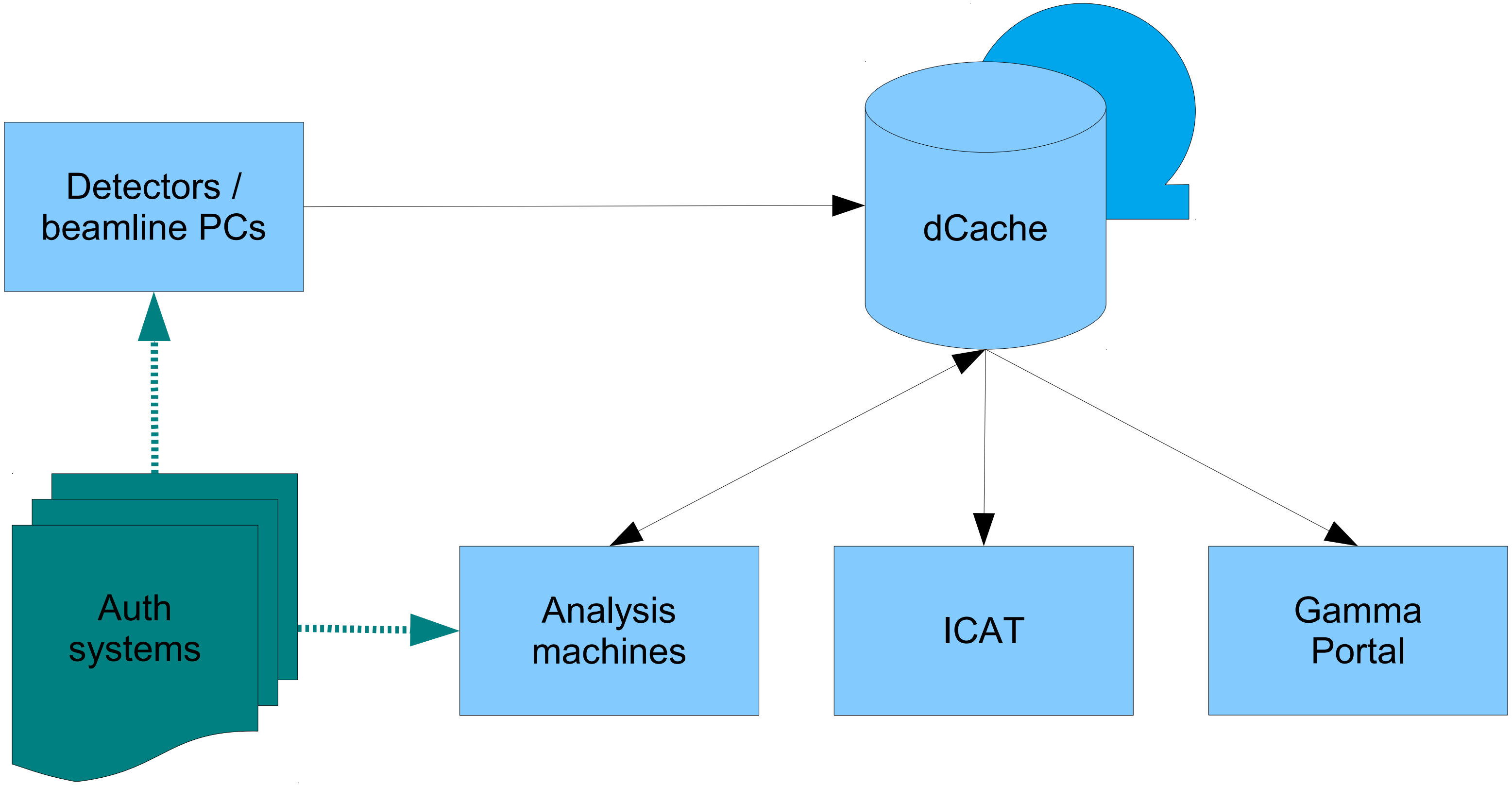
Lessons from talking to the users

- No clear use case
  - Provide service along with sample data to judge interest
- Hard to agree on metadata sets (Nexus “application definitions”)
  - Start with data from beamlines with known, little-changing experiments
- No manual work for beamline staff or researchers at data ingestion time
  - Experiment metadata get written into Nexus files by Tango server at beamline
  - Utility program to add experimenter-defined metadata like comments
  - Nexus file is ingested into catalogue by scripts, metadata gets extracted in the process
- Worries about rights management
  - Coordination work with different departments
  - Introduction of ACLs into dCache
  - Extension of user management systems

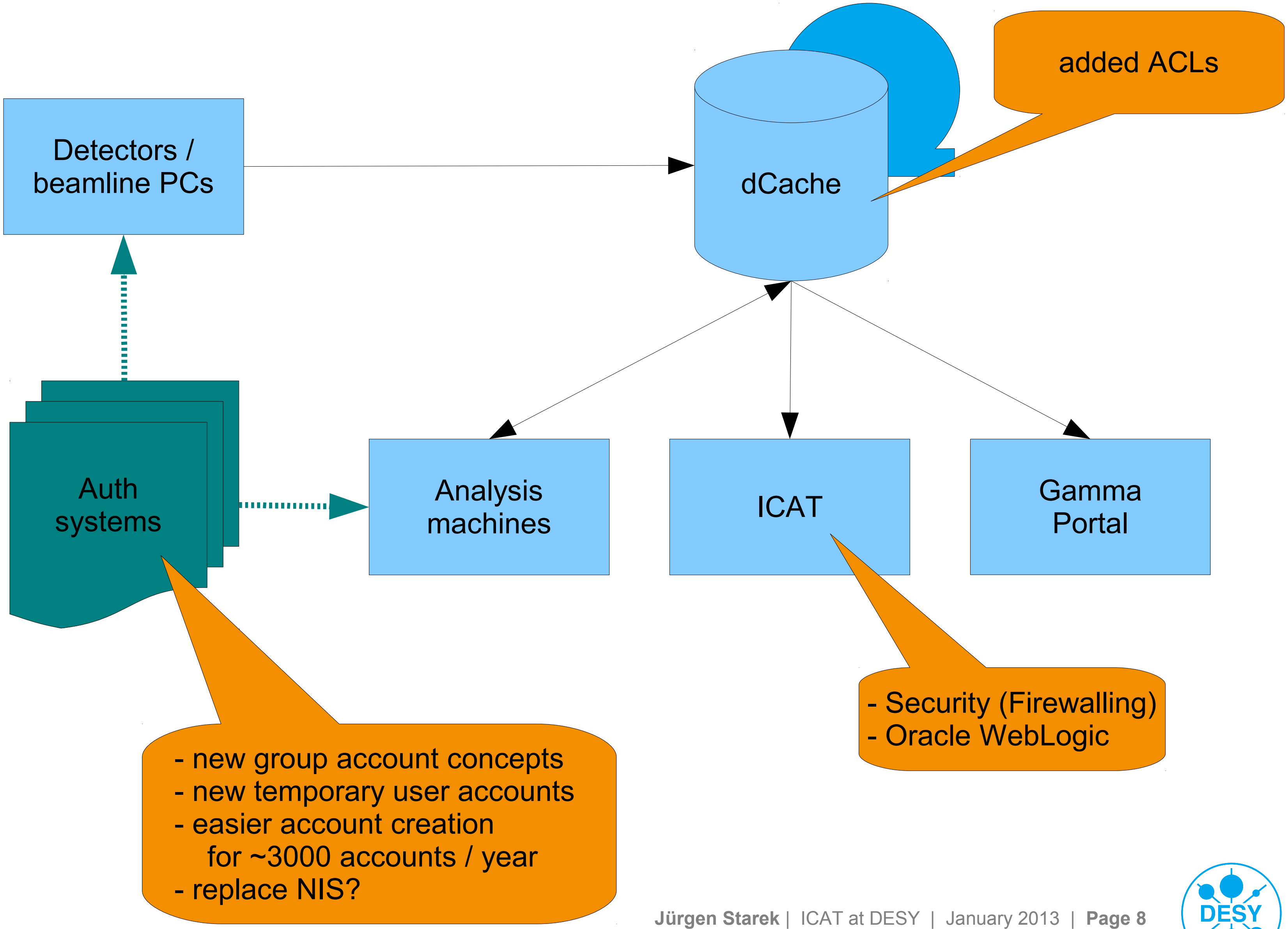
Open Access and rights management

- Drafts for data policy are being discussed
- No institute-wide opinion on DOIs yet
- Very strong data protection requirements from commercial beamline users
  - Data may not be visible to any other users in any stage of the archival process

Resulting changes in the data pipeline



Resulting changes in the data pipeline



Detectors / beamline PCs

dCache

added ACLs

Auth systems

Analysis machines

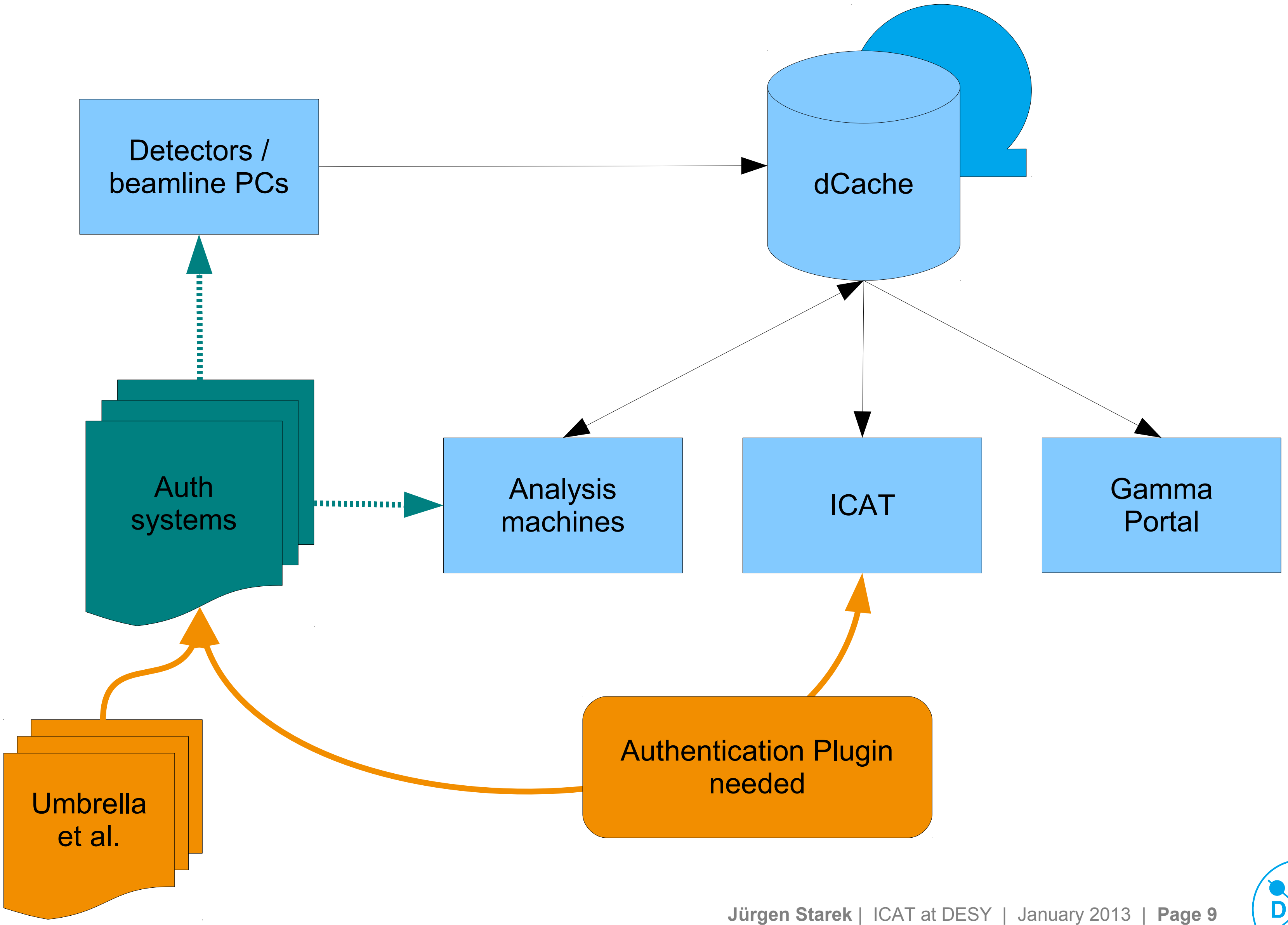
ICAT

Gamma Portal

- Security (Firewalling)
- Oracle WebLogic

- new group account concepts
- new temporary user accounts
- easier account creation for ~3000 accounts / year
- replace NIS?

Resulting changes in the data pipeline



Rights management in the backend

- dCache offers both ACLs and POSIX rights
- All files belong to admin user
- Gamma Portal uses this admin user
- ICAT data server will get ACL-based rights
- Normal users are not expected to write to dCache directly
  - Read access is controlled by ACLs

Outlook

- Unified data management and analysis pipelines for most photon science applications
- ICAT to serve as public web-facing catalogue
- Federation of identity information via Umbrella
- Federation of data via ICAT
- Research into Object Store backends based on dCache