Storage management in biomed

Franck Michel, Tristan Glatard
for the biomed VO

LFC and DPM Synchronization, May 31th 2012

biomed virtual organization

- **Users**
  - 280 users from ~20 different countries
  - Two SMEs (non commercial activities)
  - Application fields: Bioinformatics, Drug discovery, Medical Imaging
- **Large infrastructure, loosely controlled**
  - 238 CEs (batch queues) from 129 sites ; 111 Storage Elements
  - No formal agreement with sites
  - VO support and management from user groups on a voluntary basis
- **Heterogeneous application environments**
  - Heterogeneous tooling: portals, file catalogs, workflow engines, pilot-job systems
  - No central control point

Issue summary

- **SE cleanup is required**
  - When user leaves the VO
  - When SE is full
  - When SE is decommissioned
- **SE cleanup procedure**
  - List VO files and DNs on SE from LFC entries (LFCBrowseSE)
  - Notify users, and assist file migration or cleanup
- **File listing is unwieldy and not reliable**
  - Listing can take days
  - Zombie files (a.k.a dark data): stored on SE, no entry in LFC
  - Ghost files: entry in LFC, SURL does not exist
- **Cleanup is extremely cumbersome**
  - Zombie files are only known to sites
  - Some file owners left the VO (permission issues, lcg-del fails silently)

SE implementations in biomed

SE implementation		Number of SEs		Total space (GB)		Used space (GB)	
DPM		87	(78%)	3738399	(77%)	2039549	(99%)
	1.7.4	10		27881		13251	
	1.7.3	2		59251		42676	
	1.8.1	2		18110		9556	
	1.8.0	33		2358414		1387648	
	1.7.2	7		482902		66132	
	1.8.2	33		791841		520286	
dCache		13	(11%)	89927	(1%)	10401	(0%)
	1.9.10	1		268		76	
	1.9.12	6		73144		8093	
	1.9.1	1		1100		123	
	1.9.5	5		15415		2109	
StoRM		11	(9%)	994153	(20%)	1333	(0%)
	1.5.6	2		526		21	
	1.5.0	1		988956		61	
	1.6.2	2		95		75	
	1.8.0	1		1000		274	
	1.8.1	1		1001		393	
	1.8.2	3		2075		343	
	1.8.11	1		500		166	

- **For cleanup, technical teams will focus on DPM.**
- **Evolution of other implementations?**
- **File catalogs: LFC used in all known cases, except 1**

LFC-DPM synchronization

- **Expected use cases**
  - Facilitate data cleanup
    - Only LFNs have to be removed
    - Simplifies permission management
  - Simplify storage management
    - Prevent creation of zombie/ghost files
    - Curate existing ones?
- **Warnings**
  - Don't automatically clean zombie files
  - Some users (1 group) don't use LFC
- **Questions**
  - How to test?
  - What is the deployment plan?
  - Will VOs still have to clean zombie/ghost files when synchronization is deployed?