Light weight Disk Pool Manager status and plans

Jean-Philippe Baud, IT-GD, CERN

April 2005

Disk Pool Manager aims

- Provide a solution for the small Tier-2s in LCG-2
  - This implies 1 to 10 Terabytes in 2005
- Focus on manageability
  - Easy to install
  - Easy to configure
  - Low effort for ongoing maintenance
  - Easy to add/remove resources
- Support for multiple physical partitions
  - On one or more disk server nodes
- Support for different space types – volatile and permanent
- Support for multiple replicas of a file within the disk pools

Manageability

- Few daemons to install
  - Disk Pool Manager
  - Name Server
  - SRM
- No central configuration files
  - Disk nodes request to add themselves to the DPM
- Easy to remove disks and partitions
  - Allows simple reconfiguration of the Disk Pools
  - Administrator can temporarily remove file systems from the DPM if a disk has crashed and is being repaired
  - DPM automatically configures a file system as “unavailable” when it is not contactable

Features

- DPM access via different interfaces
  - Direct Socket interface
  - SRM v1
  - SRM v2 Basic
  - Also offer a large part of SRM v2 Advanced
    - Global Space Reservation (next version)
    - Namespace operations
    - Permissions
    - Copy and Remote Get/Put (next version)
- Data Access
  - Gridftp, rfio (ROOTD, XROOTD could be easily added)
- DPM Catalog shares same code as LCG File Catalog
  - Possibility to act as a "Local Replica Catalog" in a distributed catalog

DPM Details

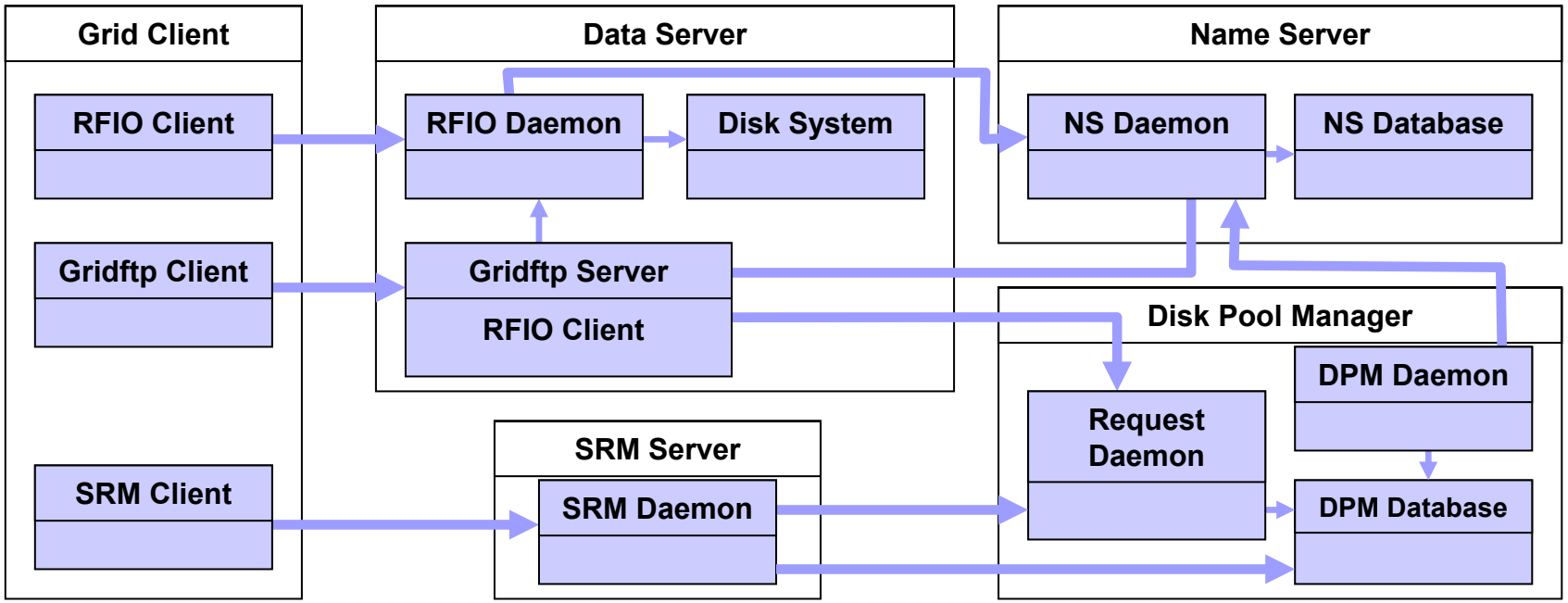
Features (1/2)

- Namespace operations
  - All names are in a hierarchical namespace
  - mkdir(), opendir(), etc…
- Security – GSI Authentication and Authorization
  - Mapping done from Client DN to uid/gid pair
  - Authorization done in terms of uid/gid
  - VOMS will be integrated
    - VOMS roles appear as a list of gids
  - Ownership of files is stored in catalog, while the physical files on disk are owned by the DPM
  - Permissions implemented
    - Unix (user, group, other) permissions
    - POSIX ACLs (group and users)

Features (2/2)

- Retries and timeouts
  - Make client resilient to temporary outage of server

Architecture



Policies

- The policies are pool attributes
- There are currently 4 types of policies:
  - File system selection for storing a new file
    - The default policy is Round Robin as long as there is enough free space
  - Garbage collector
    - The default policy is to remove the least recently used files which are not pinned
  - Request selection
    - The default policy is FIFO
  - Migration policy: to automatically migrate durable files from Tier2 to Tier1 when space is needed for example
- All policies can be replaced online (shared library) and do not require code recompilation nor daemon restarts (next version)

Pool selection

- The pools have 2 attributes for this: space type (Volatile or Permanent) and restriction to certain VOs
- However a given pool might have no restriction: the pool is shared by all users and for any type of space
- We recommend to have different pools for Volatile and Permanent space: reliable hardware for permanent storage.
- Disks on CPU servers (worker nodes) can be used for Volatile space
- Hot files can be replicated to several disks and the DPM selects the best replica (less used or closest to the CPU)

Disk Pool Manager APIs

- There are 2 categories of APIs:
  - Administrative: disk pool configuration
    - dpm_addfs (char *, char *, char *, int);
    - dpm_addpool (struct dpm_pool *);
    - dpm_getpoolfs (char *, int *, struct dpm_fs **);
    - dpm_getpools (int *, struct dpm_pool **);
    - dpm_modifyfs (char *, char *, int);
    - dpm_modifypool (struct dpm_pool *);
    - dpm_replicate (char *);
    - dpm_rmfs (char *, char *);
    - dpm_rmpool (char *);
  - User: these map pretty well to the SRM v2.1 calls

Status

- DPNS, DPM, SRM v1 and SRM v2 (without Copy nor global space reservation) have been tested for 4 months
- The secure version has been tested for 6 weeks
- GsiFTP has been modified to interface to the DPM
- RFIO interface is in final stage of development

Plans for Service Challenge 3

- Provide a possible solution for the small Tier-2s
  - This implies 1 to 10 Terabytes in 2005
- Focus on manageability
  - Easy to install
  - Easy to configure
  - Low effort for ongoing maintenance
  - Easy to add/remove resources
- Support for multiple physical partitions
  - On one or more disk server nodes
- Replacement of 'Classic SE'
  - Only metadata operations needed (the data dos not need to be copied)