LHCb plans for SC3

A.Tsaregorodtsev,
CPPM, Marseille

SC3 Workshop, 14 June 2005, CERN

LHCb SC3 goals

- Phase 1
  - Demonstrate Data Management to meet the requirements of the Computing Model
- Phase 2
  - Demonstrate the full data processing sequence in real time
  - Demonstrate full integration of the Data and Workload Management subsystems

General approach

- Maximum use of centralized components
  - LHCb is a “small” experiment
  - Do not have 24/7 support by LHCb experts on sites
    - No dedicated LHCb sites
  - Minimize synchronization problems
  - Add extra components (mirrors) as a matter of load balancing as need would be
- Keep a fallback solution for all the components
  - Catalogs, data moving tools, monitoring, etc

Phase 1: Data Moving

Phase 1 goals

a) **Moving of 8 TB of digitised data from CERN/Tier-0 to LHCb participating Tier1 centers in a 2-week period.**

   - **The necessary amount of data is already accumulated at CERN**
   - **The data are moved to Tier1 centres in parallel.**
   - **The goal is to demonstrate automatic tools for data moving and bookkeeping and to achieve a reasonable performance of the transfer operations.**

b) **Removal of replicas (via LFN) from all Tier-1 centres**

Phase 1 goals (cont’d)

c) **Moving data from Tier1 centre(s) to Tier0 and to other participating Tier1 centers.**

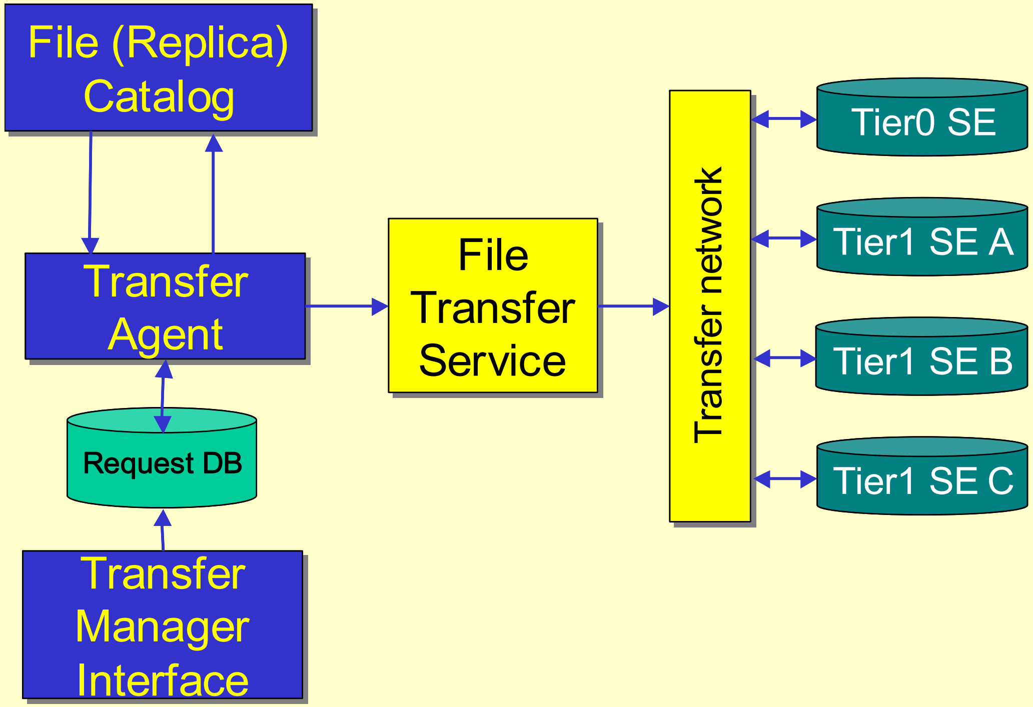
- **The goal is to demonstrate that the data can be redistributed in real time in order to meet the stripping processing.**

d) **Moving stripped DST data from CERN to all Tier1’s**

- **The goal is demonstrate the tools with files of different sizes**
  - **Necessary precursor activity to eventual distributed analysis**

File Transfer with FTS

- Start with central Data Movement
  - FTS+TransferAgent+ RequestDB
- Explore using local instances of the service at Tier1's
  - Load balancing
  - Reliability
  - Flexibility
- TransferAgent+ReqDB are to be developed
  - Requires access to FTS service now



Transfer Agent

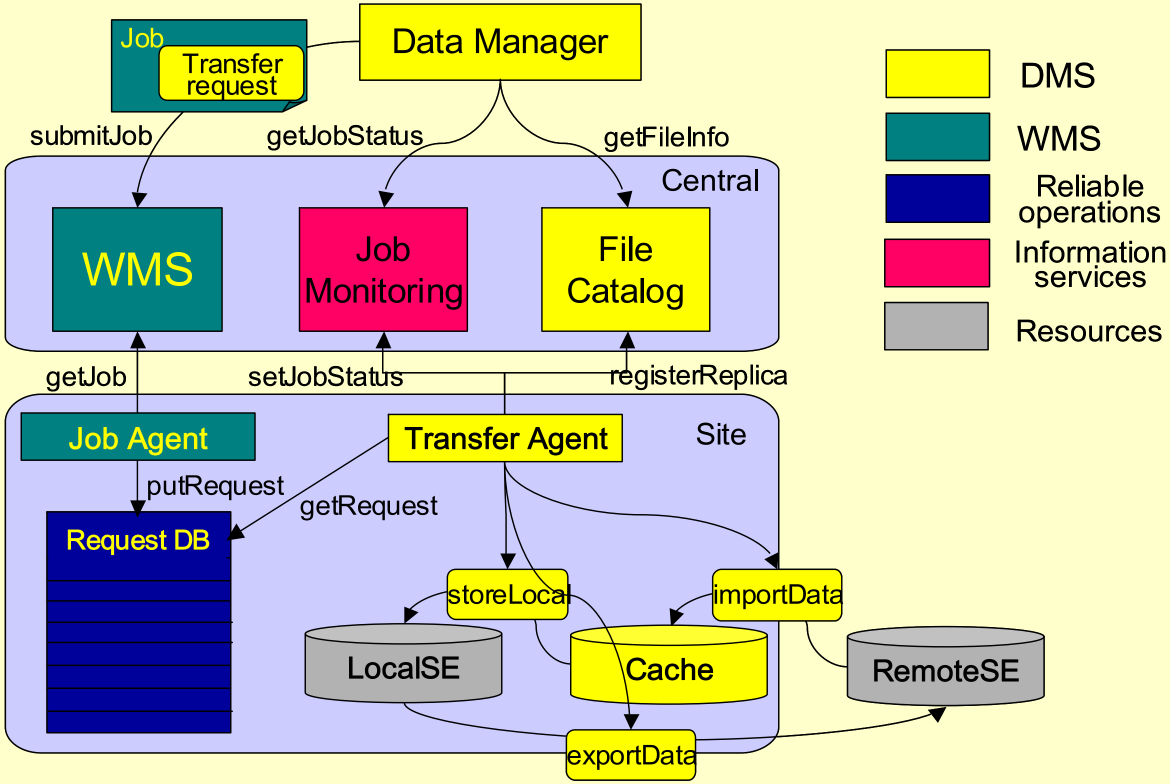
- Gets transfer requests from Transfer Manager ;
- Maintains the pending transfer queue ;
- Optimizes transfers in terms of:
  - Number of simultaneous transfers for a given channel (end point source/destination);
  - Optimal source replica for a given destination
- Validates transfer requests ;
- Submits transfers to the FTS ;
- Follows the transfers execution, resubmits if necessary ;
- Updates the replica information in the File Catalog ;
- Accounts for the transfer characteristics:
  - Start/execution time;
  - Effective bandwidth.

FTS requirements

- Handles transfer requests
- Provides transfer accounting information
  - Transfer start time
  - Transfer execution time
  - Effective bandwidth, percentage of the total available bandwidth
- Notifications of the transfer state changes:
  - States: received, ready, running, done
  - Otherwise keep polling

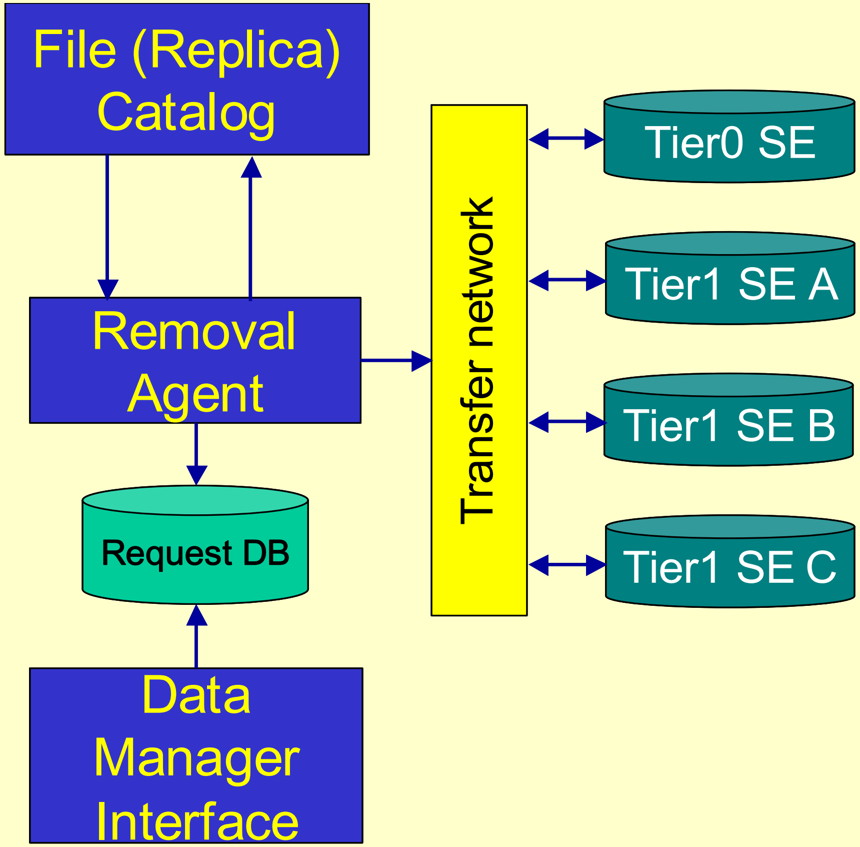
Existing File Transfer framework

- Keep existing tools as a fallback solution
- Using both gridftp or FTS file transport for data import/export
- Might merge eventually in a single system



File removal

- Should be fast to allow efficient storage management
- Central Removal Agent
  - Might be delegated to local agents
- Removing all the remote replicas with eventual retries of failures
- Update of the File Catalog

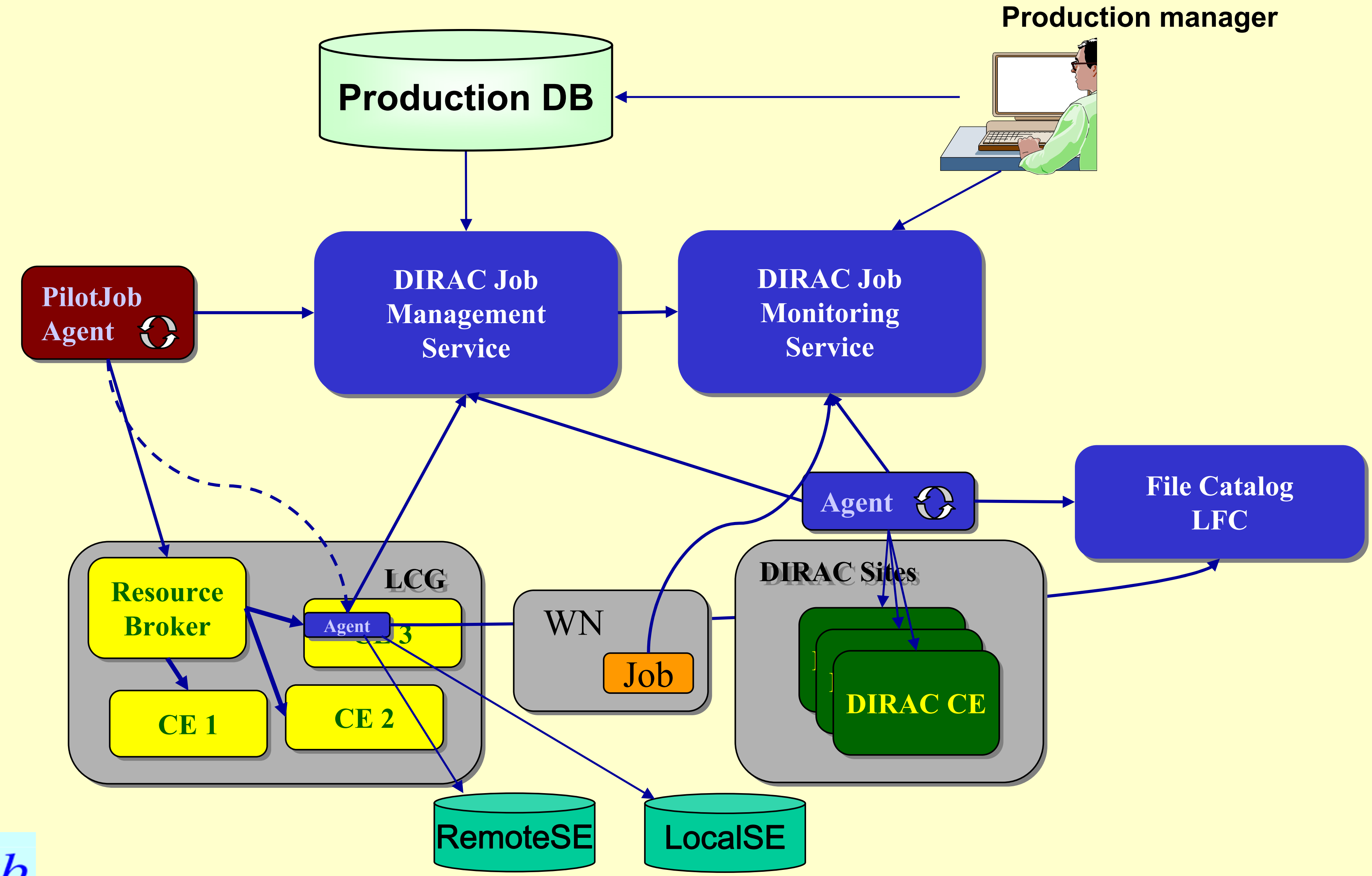


Phase 2 : Full Data Processing chain

Phase 2 goals

- MC production in Tier2 and Tier1 centers with DST data collected in Tier1 centers in real time followed by Stripping in Tier1 centers
  - MC events will be produced and reconstructed. These data will be stripped as they become available
- Data analysis of the stripped data in Tier1 centers.

Data production on the grid

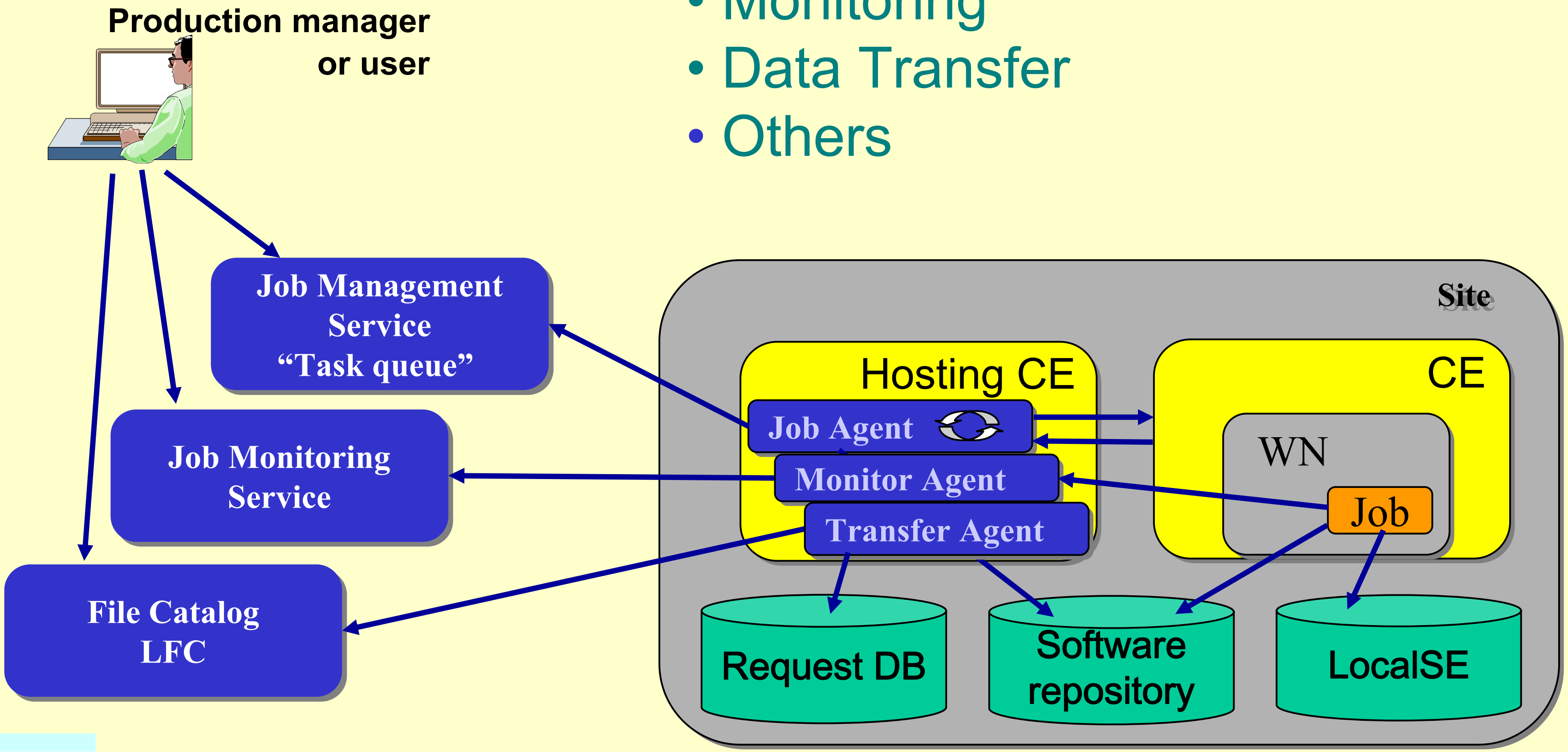


DIRAC overlay network

- The DIRAC overlay network paradigm is first of all there to abstract heterogeneous resources and present them as single pool to a user :
  - LCG or DIRAC sites or individual PC's
  - Single central Task Queue is foreseen both for production and user analysis jobs
- The overlay network is dynamically established
  - No user workload is sent until the verified LHCb environment is in place

On-site LHCb agents

- Host CE runs LHCb specific agents :
  - WMS agents
  - Monitoring
  - Data Transfer
  - Others



VO specific agents

- Dedicated VO box is an attractive solution
- LHCb offers to explore another solution - Hosting CE
  - Might be more acceptable on (smaller) sites.
- Agents submitted as jobs
  - Through jobManager-fork queue
- Agents credentials:
  - User certificates
    - Need MyProxy service available
  - Host certificate ?
- Running fully under responsibility of the VO
  - Site managers might want to examine the start-up scripts and software to be executed
- Need access to managed local storage
  - Software installation
  - Request “Database”

VO specific agents: MonitorAgent

- Jobs are sending monitoring information through job wrappers:
  - Application status
  - Environment parameters
- MonitorAgent
  - Buffers the monitoring information for reliable transfer for Job Monitoring service

VO specific agents: TransferAgent

- TransferAgent:
  - Collects data transfer requests from successful jobs
    - Maintains data transfer/registration requests database
      - Files, sqlite, MySQL
  - Initiates transfer:
    - Direct gridftp
    - Through FTS
  - Monitors the transfers
    - retries transfers in case of failures
  - Registers the newly created replicas to the File Catalog
    - Retries registration in case of immediate FC unavailability.

VO specific agents: Other

- Other services can be also considered
  - MonALISA, xrootd – possibly shared with others
  - JobAgent
    - Can be added when gLite CE will become available
    - Getting jobs from DIRAC Task Queue
    - Installing the necessary software
    - Submitting to local CE
  - …

SC3 services needed by LHCb

- Resources
  - CE service
  - SE service
    - SRM v1.1 interface to MSS
    - gridftp accessible
- Grid Catalogs
  - Dedicated LFC central catalog
    - Read-only mirrors on Tier1 sites
  - Dedicated FiReMan central catalog
  - Dynamically generated Pool XML slices to connect to applications
- Data transfer
  - FTS
    - Central FTS engine at CERN
    - FTS clients in Tier1(2) centers
  - gridftp access to SE’s should be still available

CE service

- Provide necessary information for taking a scheduling decision:
  - VO waiting/running jobs
  - Total waiting/running jobs if resources are shared with other VO’s
- Job manipulation/information interface
  - submit(),kill()
  - getJobStatus()
- More advanced features eventually
  - getTimeLeft()
  - reserveScratchSpace()
  - …
- Stays to see if gLite CE will provide this functionality

SE service

- SE level v1.1 is foreseen for SC3
- This level is quite limited and chosen as a temporary compromise
  - LHCb was advocating v2.0 level
  - Need for file pinning
  - Need for storage name space management
  - Need for storage browsing
- LHCb will be willing to participate in early tests of v2.0
- Physical file name space management
  - The same structure as for LFN name space
  - Facilitate problems debugging, integrity checks, etc
  - Simplifies data access tools

File Catalog use

- We will start with central world-readable LFC full catalog
  - Used both as Storage Index and Replica Catalog
  - Stress test the centralized solution
    - ~10M entries, ~100M replicas, ~100Hz queries rate
- Add read-only redundant mirrors on Tier1 sites as a load-balancing optimization
  - All the updates are still through the central master catalog
  - Mirror updates “as soon as feasible”

Things to be done

For Phase 1 to start in September we have to develop:

- Data Transfer Agents
  - Using FTS as transport
  - Need FTS service and client tools now
- LHCb agents (applications) orchestrating the data processing chain in real time
  - They are using all the required services
  - We need access to these services now to start the development
- Dedicated manpower is foreseen