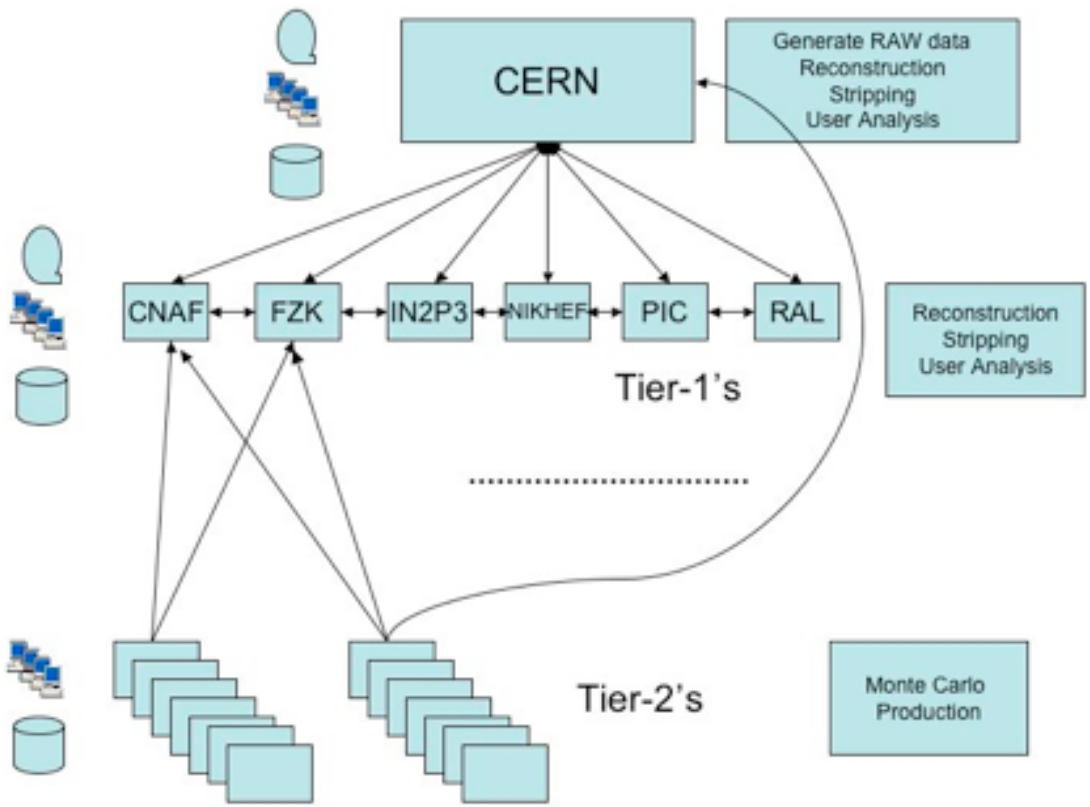
# Data Production Management

*Andrew C. Smith (CERN)*
*On behalf of the LHCb DIRAC team*

# LHCb Computing Model

- The LHCb Computing Model
  - Describes the dataflow model for all stages of processing
    - Real and simulated events
  - Defines role of LHCb Tier1 (+ CERN) and Tier2 computing centres



# DIRAC

- DIRAC is the LHCb Workload and Data Management System *
  - Initial incarnation as the LHCb production system
  - Since evolved into a generic community Grid solution
    - Either stand-alone environment or
    - Community Overlay Grid System (COGS)
    - VO specific utilities can be tailored as required
  - Demonstrated 10k concurrently running jobs
  - Management of O(10M) Data files and replicas

  - Supports the implementation of LHCb's Computing Model
    - Generic to support other VO requirements
  - Provides data driven production and data management

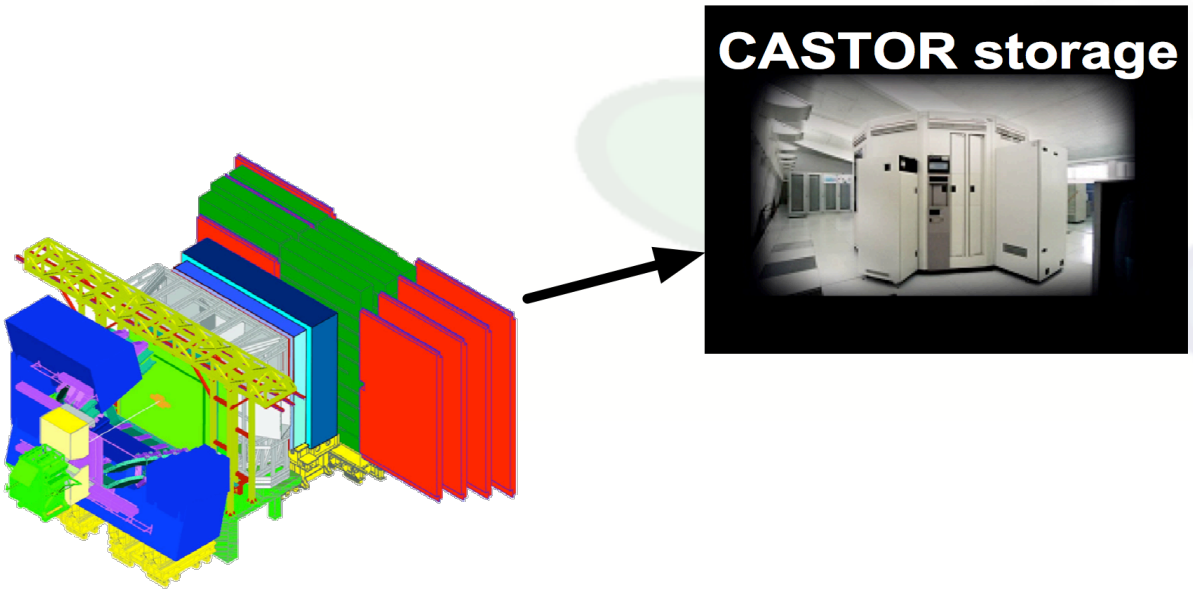
* A. Tsaregorotsev et al. DIRAC: A community grid solution, CHEP 2007 [189]

# Dressed Rehearsal Exercise

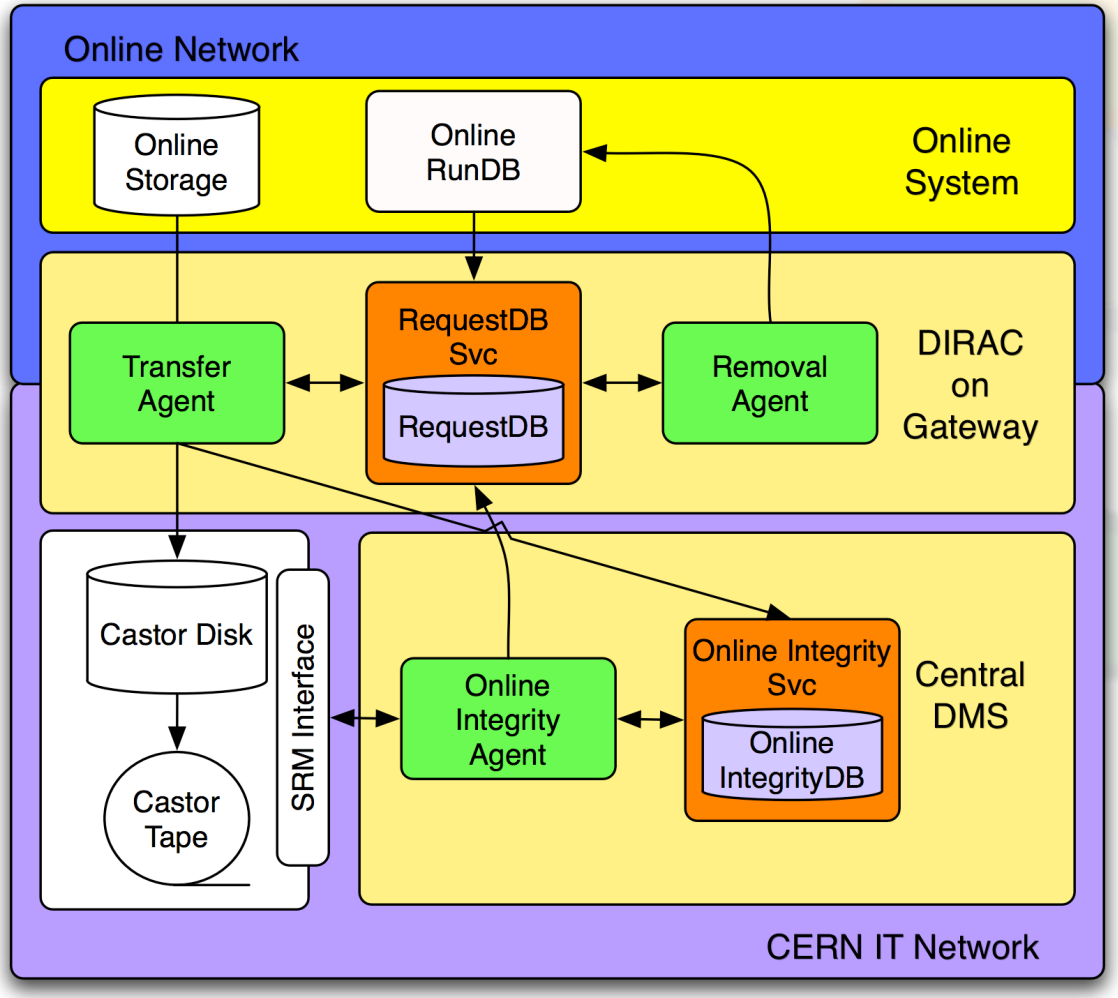
- WLCG 'Full Dressed Rehearsal' (FDR) exercise
  - Allow LHC experiments to deploy Computing Model full chain
    - Making use of WLCG services and resources
  - Preparation for real data taking
- LHCb's activity
  - Simulated RAW physics data (matching properties of eventual real data)
    - Uploaded from Online storage to Grid
    - Data driven replication to LHCb Tier1s
    - Data driven processing (reconstructed + stripped)
  - Processing produces user analysis data (DSTs)
    - Distributed to all LHCb Tier1 sites.

# LHCb to the Grid

- RAW physics data from LHCb written to Online storage
  - 40 MB/s online data acquisition rate
- Replicate RAW data to Castor at CERN
  - Dedicated 1Gb link per host
- Remove data Online when ‘safe’
  - Correctly migrated to tape
    - Checksum compared against Online file

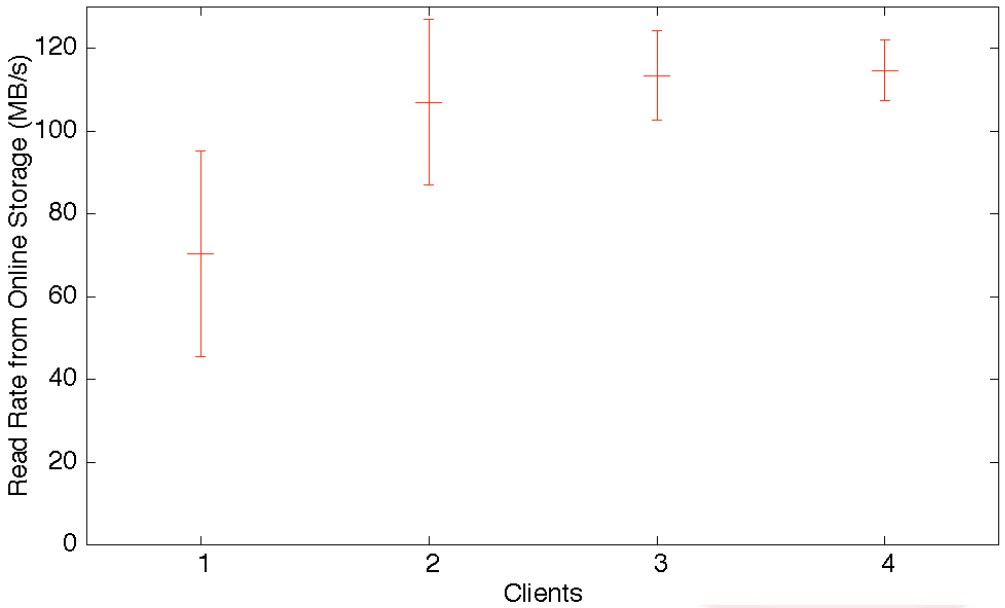


- DIRAC install at Gateway
  - Replicate RAW data to Castor pools
  - Register data to Online Integrity DB




- Persist files Online till 'safe'
- File checksum calculated online at write time
- File checksum calculated by Castor on migration to tape
- Online Integrity Agent interrogates Castor for file checksum
- If 'safe': place removal request in DIRAC RequestDB at Gateway
- Pass removal request to Online system

# LHCb to the Grid

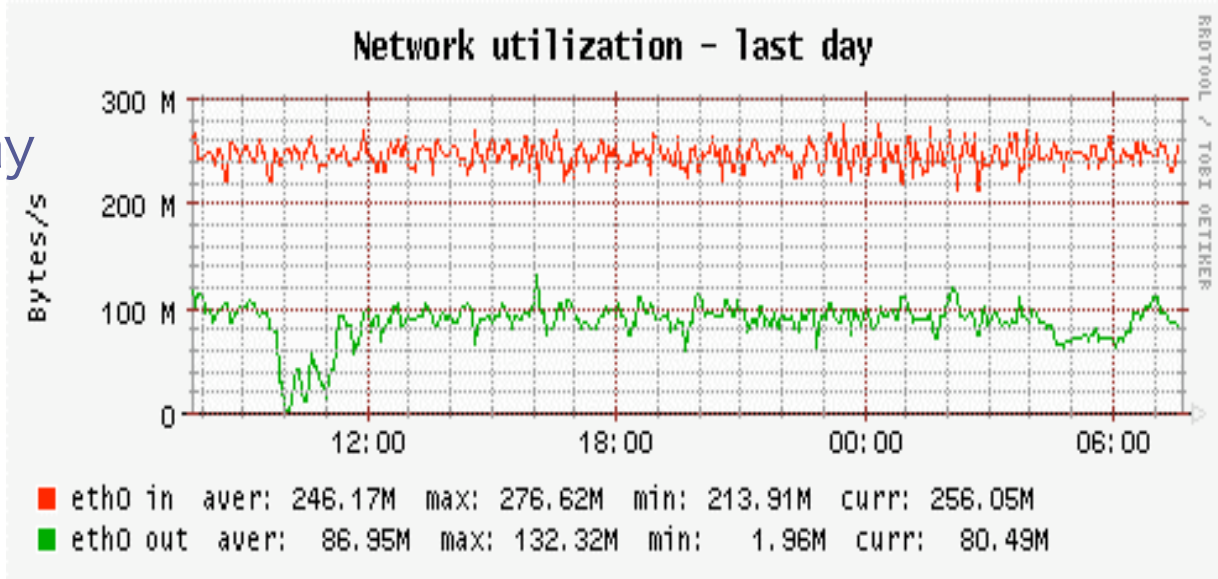



## Single host at Gateway

- Varying number of 'rfcp' clients
- Asses the read rate from Online storage

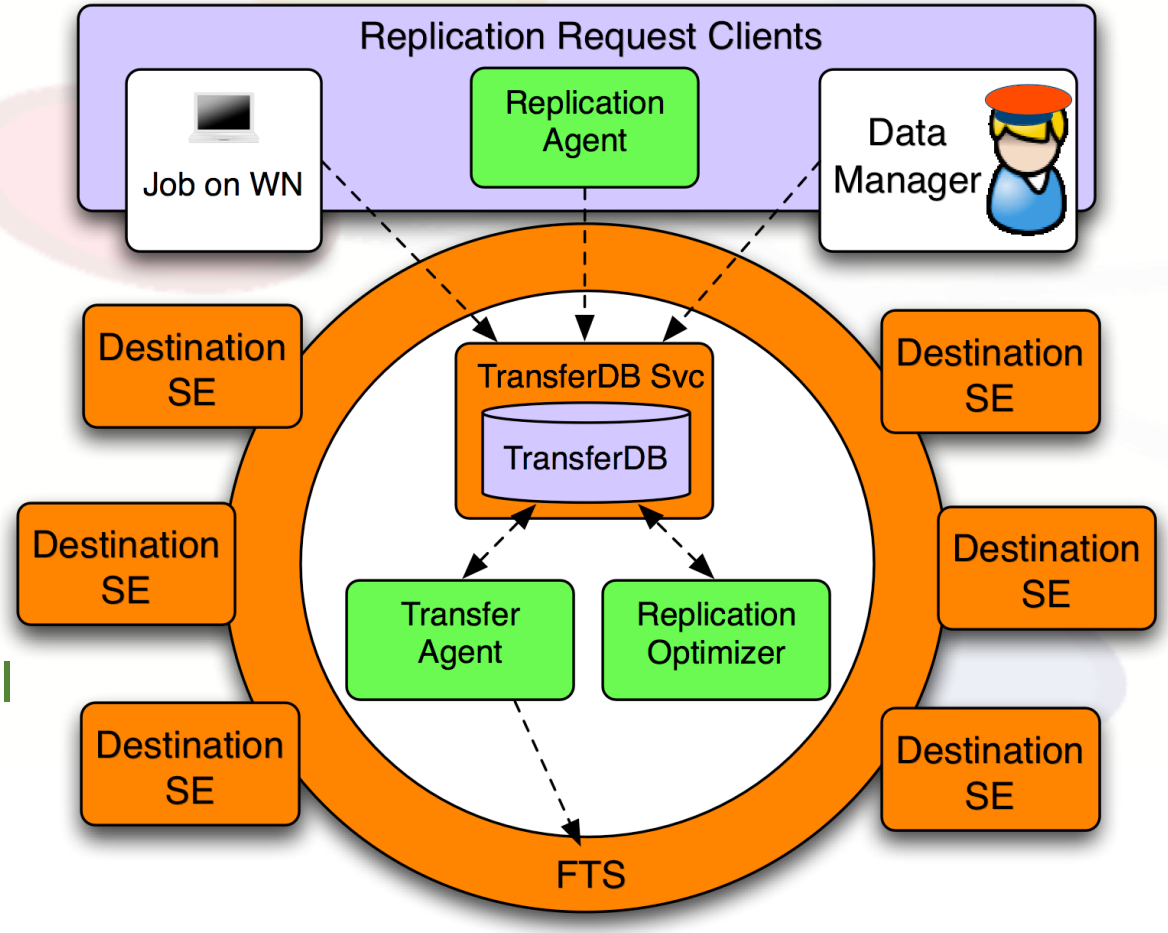
## Two hosts at Gateway

- Each running 4 clients
- Able to maintain ~250MB/s
- Meets 40MB/s requirement



# Bulk Data Replication

- Transfer requests centrally managed
  - Maintained in TransferDB
- Create bulk transfers by aggregating requests
- Transfer Agent polls TransferDB
  - Obtains bulk transfer requests
  - Submits and monitors transfers through FTS CLI
  - Requests retried till success



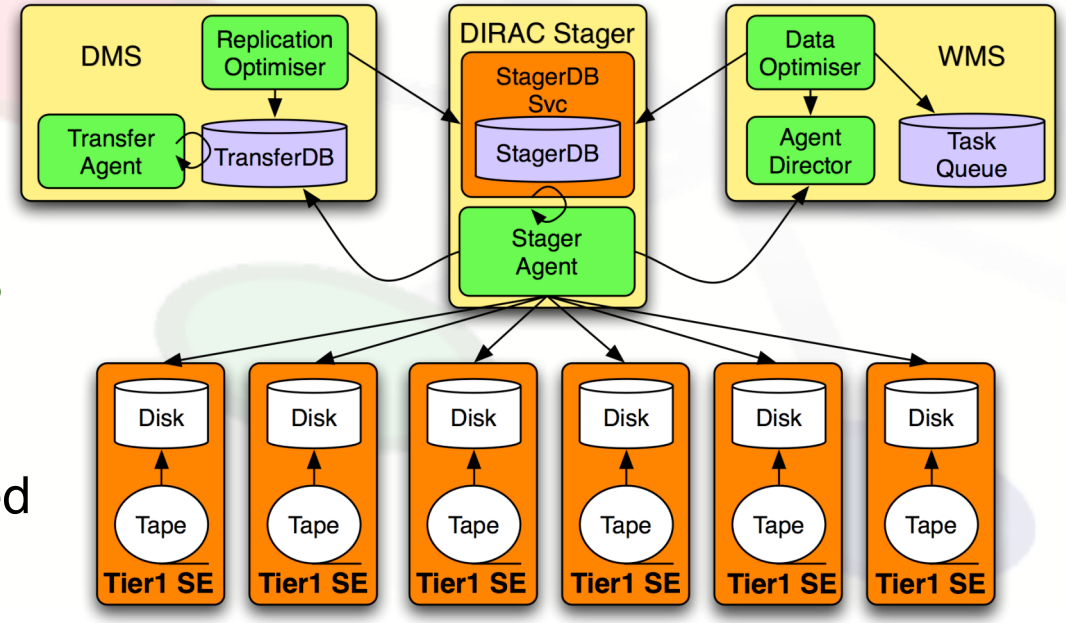
# Bulk Data Replication

- Iteration to this solution
  - Previous solution: Distributed transfer management
    - Transfers executed from distributed VO Boxes
    - This provided persistency of transfer requests
    - But, this architecture proved unmanageable
      - Distributed Transfer Agents throttle their requests
        - No knowledge of other agent requests
      - Load placed on SEs unpredictable
  - Current: Centralized TransferDB with distributed failover
    - Transfer requests placed in centralized DB
      - If this fails these requests are persisted on VO Boxes *
    - Allows resource usage optimization and intelligent scheduling

* A. C. Smith et al. DIRAC Reliable Data Management for LHCb, CHEP07 [195]

# Bulk Data Replication

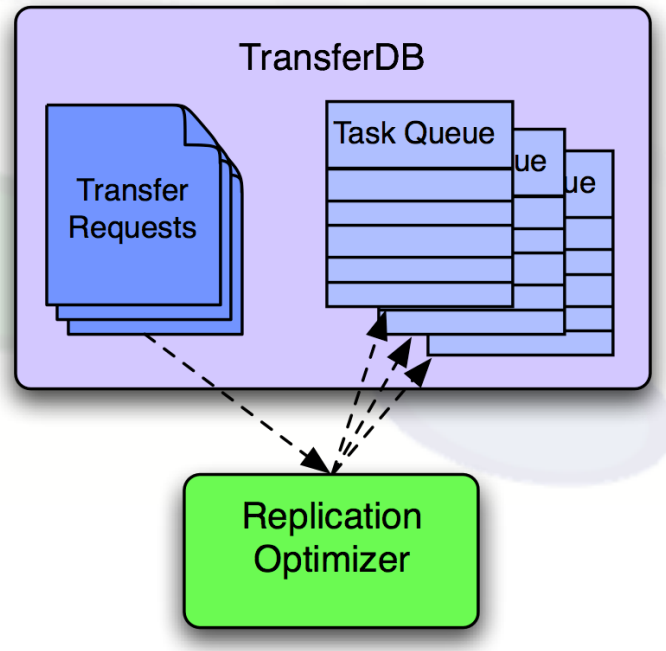
- Resource usage optimization
  - FTS transfer slots can be wasted if files not on disk
    - Results in severe throughput penalty
  - Ensure file availability on cache with DIRAC Stager *
    - Single DIRAC Stager instance used by WMS and DMSs
  - Replication Optimizer determines SURLs from LFC
    - Places stage request if file in tape storage
  - Stager Agent submits stage requests to SE SRMs
  - Stager Agent updates TransferDB once is staged
    - Replication scheduled




- * A. C. Smith et al. DIRAC Reliable Data Management for LHCb, CHEP07 [195]
- S. Paterson et al. DIRAC Optimized Workload Management, CHEP07 [174]

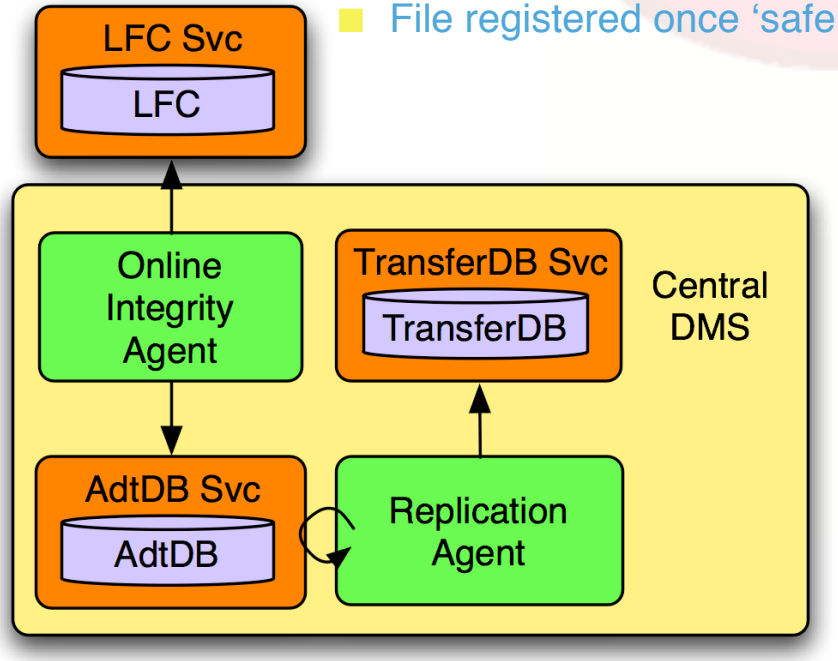
# Bulk Data Replication

- Intelligent Scheduling
  - Each FTS channel reflected in TransferDB as a task queue
    - Replication Optimizer assigns files to task queues
  - Dynamic, real-time scheduling
    - Aim to minimize latency
    - Assign files to task queues based on
      - Number of files in task queue
      - Observed recent bandwidth
  - Possible to extend architecture
    - Support pluggable static strategies
      - Multi-unicast from source
      - Seed sites for further replication
      - …many possible



# Data Driven RAW Replication

- Dataflow for RAW files
  - Master copy at CERN
  - Replicated at Tier1 site
    - 40MB/s aggregated out of CERN
  - Reconstruction based on pledged resources
    - Tier1s and CERN
- Data driven replication using AutoDataTransferDB as hook
  - File registered once 'safely' migrated




- Replication Agent splits files according to site share defined in configuration
- Places transfer requests into TransferDB
  - Handled together with other types of transfer request

# Data Driven RAW Replication

- AutoDataTransferDB (AdtDB) contains simple file catalogue
  - Exposes standard DIRAC file catalogue API
  - Each DM operation specifications contained in DB
    - Source and target SEs
    - File selection (based on LFN namespace)
    - Threshold number of files
- Replication Agent uses AdtDB API
  - Checks for files to be processed
  - Applies selection for desired files
  - Checks file location
  - Files selected if matching operation specification
  - Request created when threshold number of files found
- Replication Agent logic generalized to support multiple operation types
  - Single source to many destinations
  - Splitting files based on configurable share

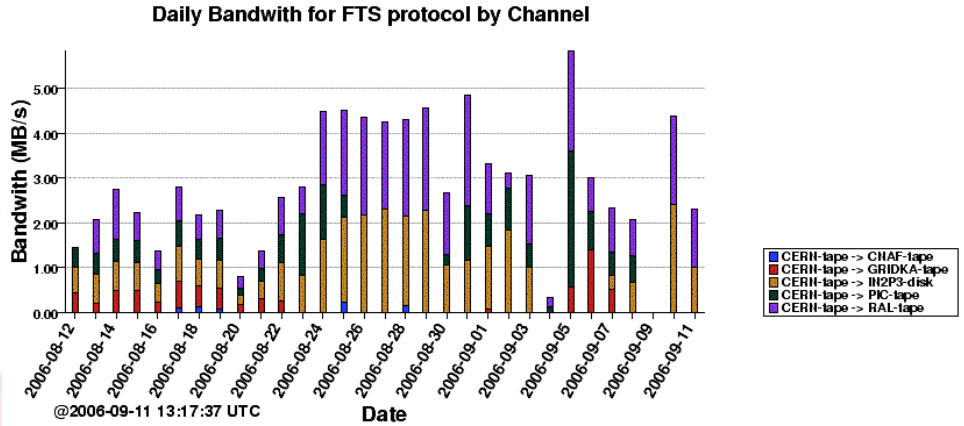
# Data Driven Reconstruction/Stripping

- Similar components exist in the Workload Management System for job creation and submission*
  - ProcessingDB
    - Analogous to AdtDB
  - Transformation Agent
    - Analogous to Replication Agent
- Files registered in ProcessingDB may be selected for processing
  - Transformations define specific processing activities
    - Reconstruction/stripping/re-processing
    - Based on file properties
  - Transformation Agent groups files for processing
    - Subsequently submits jobs
- Registration of RAW data in AdtDB initializes the chain of replication and data processing

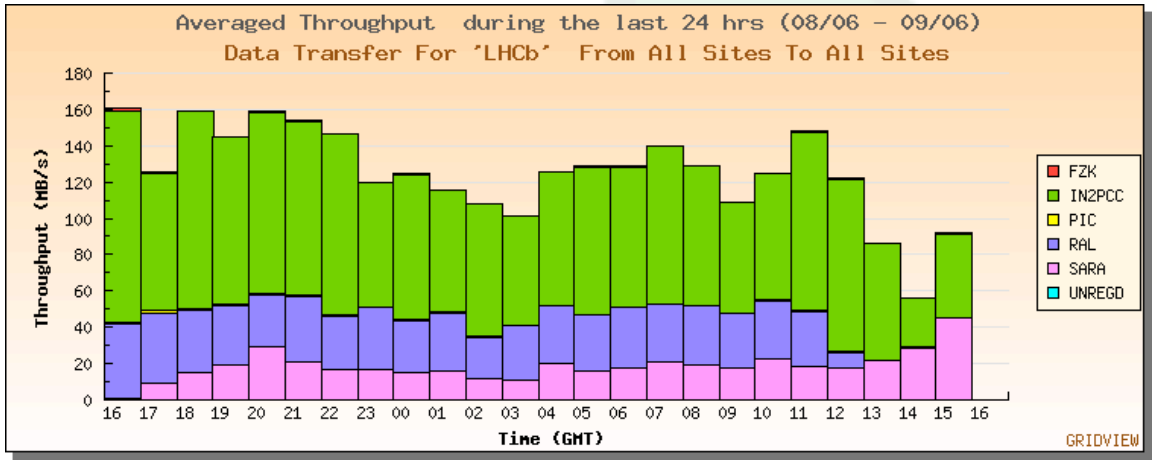
* A. Tsaregorotsev et al. DIRAC: A community grid solution, CHEP 2007 [189]

# Data Driven RAW Replication

- Functionality tests during DC06 shows machinery works
  - Concurrent transfers spawned to 5 Tier1 sites




- Export exercises have shown the ability to surpass Computing Model requirements (40MB/s)

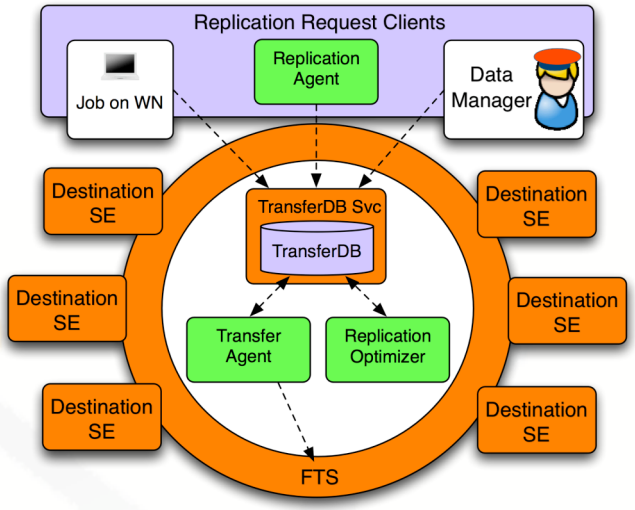


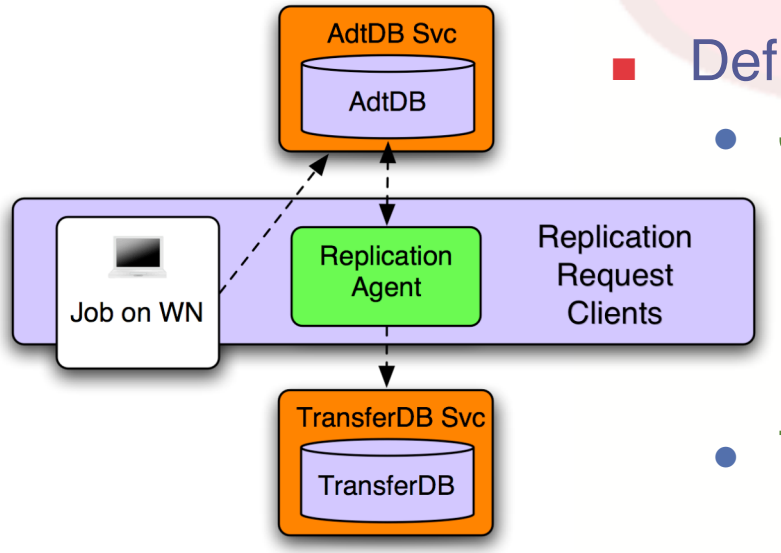
# Optimized DST Replication

- Processing activities produce user analysis files (DST)
  - DSTs from real data must be replicated to all Tier1s
    - Average ~11MB/s in and out
    - Shared 10Gb network available using FTS
  - DSTs from MC data present on a selection of Tier1s
- Two mechanisms are possible for distributing DSTs
  - Output sites explicitly specified in processing workflow
    - Job has this knowledge on the worker node
  - Output sites specified in the AdtDB
    - For each processing activity

# Optimized DST Replication

- Define output SEs in production workflow
  - Job uploads output to SE associated Tier1
  - Places replication requests to TransferDB
    - Replication by standard tools
  - Sufficient where output SEs are known




- Define output SEs in AdtDB per production
  - Job uploads output to SE associated Tier1
    - File registered to AdtDB
  - Replication Agent determines output SEs
    - Defined in configuration per production
    - Places replication request to TransferDB
  - This allows load balancing for MC DSTs

- Full throughput tests to be performed during 'FDR'
  - Determine which mechanism to use

# Outlook and Conclusions

- Full chain Computing Model supported in DIRAC
  - All components tested individually
  - FDR will bring these together for full chain exercise
- Dressed rehearsal timetable
  - First dressed rehearsal in the context of SRM2.2 testing
    - Beginning post-CHEP
  - Full Dressed rehearsal in spring