The LHCb Data Management System

Philippe Charpentier

CERN

On behalf of the LHCb Collaboration

LHCb Computing Model in a nutshell

- RAW files (3GB) are transferred to Tier0
  - Verify migration and checksum
- Transfer to Tier1s
  - Each file replicated once (whole run at a single site)

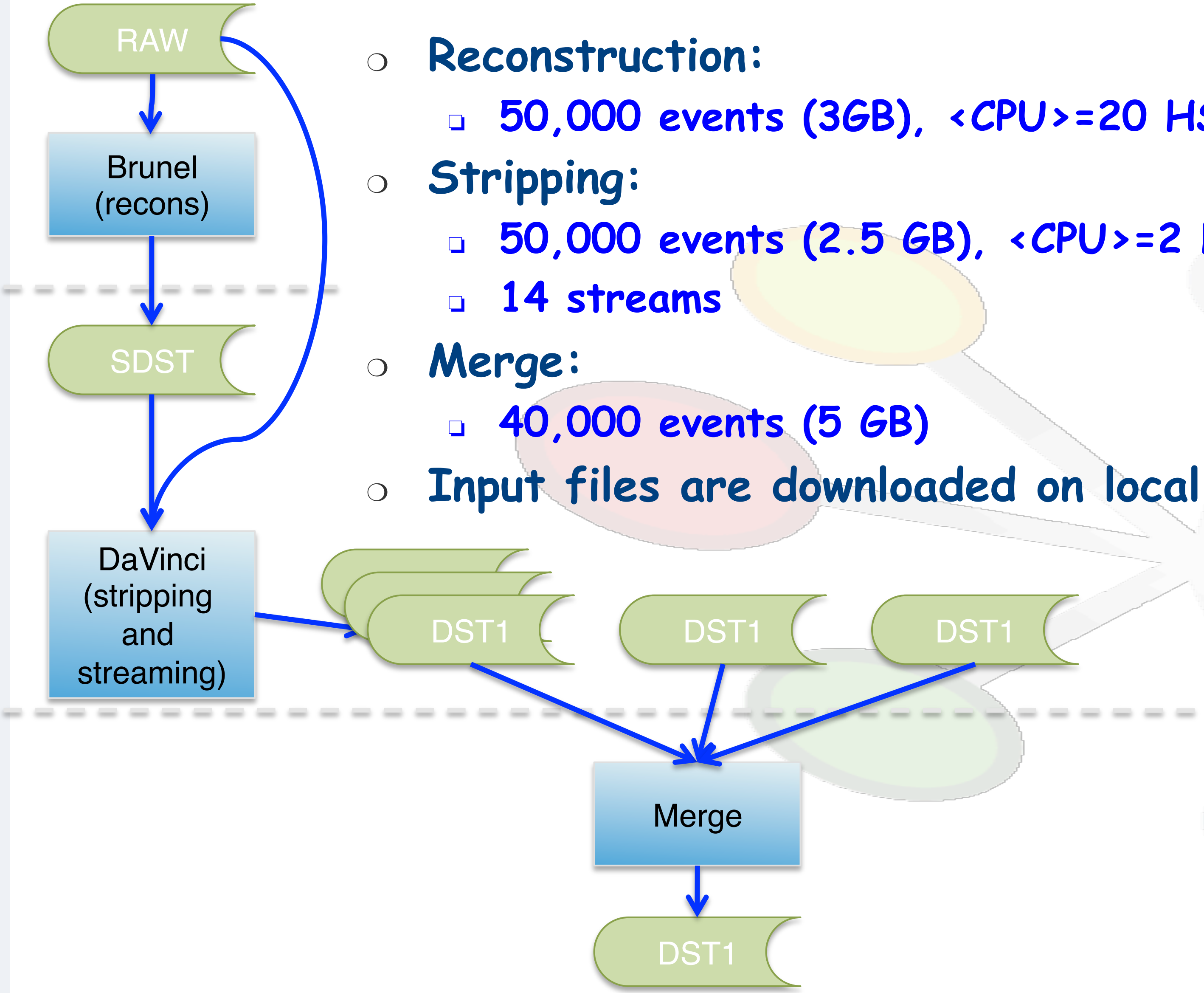
Poster #272, S.Roiser

- Reconstruction
  - At CERN and Tier1s (up to 300 HS06.hours)
    - If needed Tier2s can also be used as “Tier1 co-processor”
- Stripping
  - On Tier1 where SDST is available
- MC simulation
  - Complex workflow: simulation, digitization, reconstruction, trigger, filtering
  - Running everywhere with low priority
    - Tier2, Tier1, CERN and unpledged resources (some non-Grid)
- User analysis
  - Running at Tier1s for data access, anywhere for MC studies
- Grid activities under control of LHCbDirac
  - LHCb extensions of the DIRAC framework (cf Poster
  - Many presentations at CHEP (again this time, talks and posters)

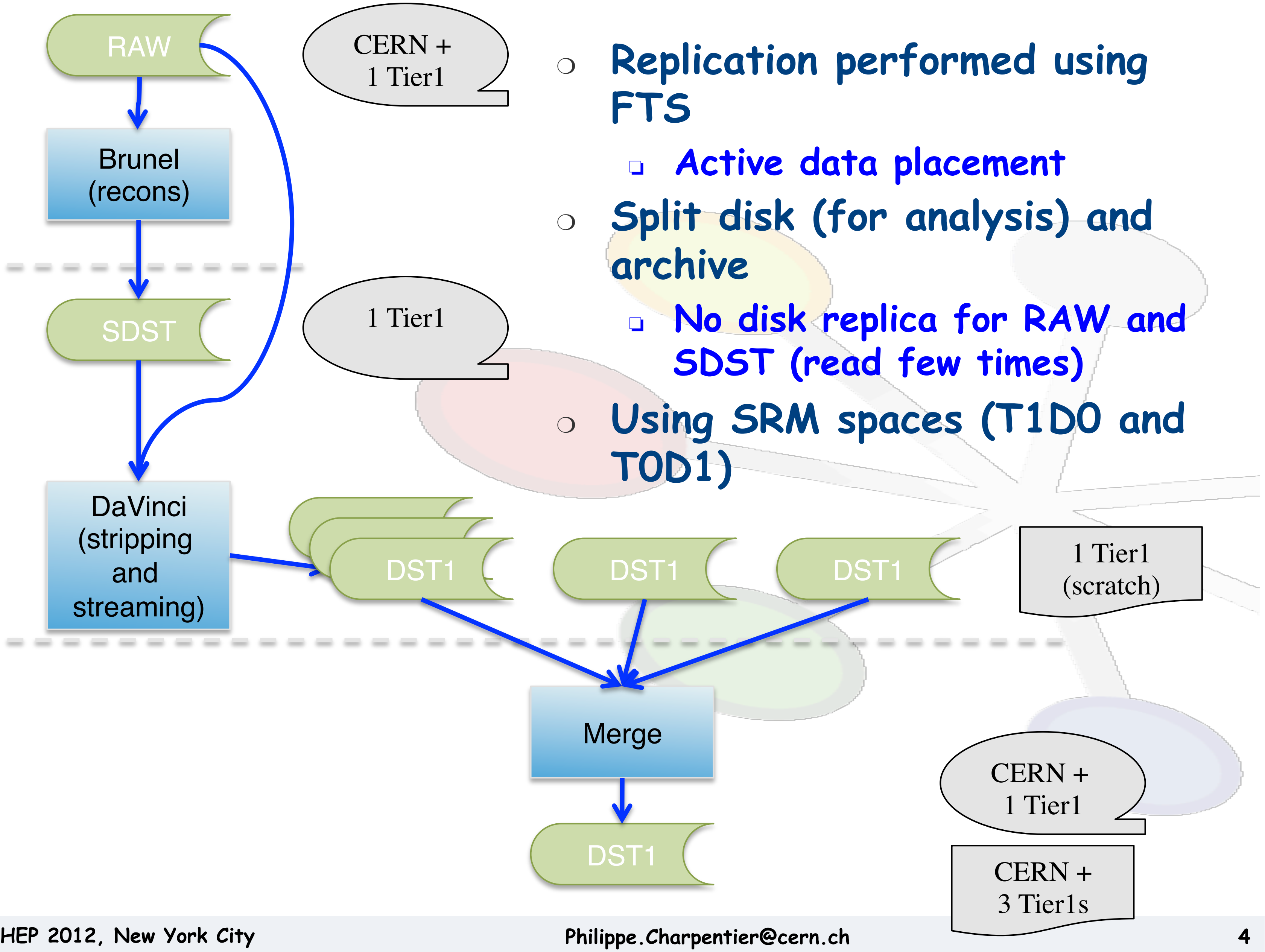
Poster #145, F.Stagni

Real data applications

- Reconstruction:
  - 50,000 events (3GB), <CPU>=20 HS06.s/evt
- Stripping:
  - 50,000 events (2.5 GB), <CPU>=2 HS06.s/evt
  - 14 streams
- Merge:
  - 40,000 events (5 GB)
- Input files are downloaded on local disk



Data replication



- Replication performed using FTS
  - Active data placement
- Split disk (for analysis) and archive
  - No disk replica for RAW and SDST (read few times)
- Using SRM spaces (T1D0 and T0D1)

Datasets

- Granularity at the file level
  - Data Management operations (replicate, remove replica, delete file)
  - Workload Management: input/output files of jobs
- LHCbDirac perspective
  - DMS and WMS use LFNs to reference files
  - LFN namespace refers to the origin of the file
    - Constructed by the jobs (uses production and job number)
    - Hierarchical namespace for convenience
    - Used to define file class (tape-sets) for RAW, SDST, DST
    - GUID used for internal navigation between files (Gaudi)
- User perspective
  - File is part of a dataset (consistent for physics analysis)
  - Dataset: specific conditions of data, processing version and processing level
    - Files in a dataset should be exclusive and consistent in quality and content

Replica Catalog (1)

- Logical namespace
  - Reflects somewhat the origin of the file (run number for RAW, production number for output files of jobs)
  - File type also explicit in the directory tree
- Storage Elements
  - Essential component in the DIRAC DMS
  - Logical SEs: several DIRAC SEs can physically use the same hardware SE (same instance, same SRM space)
  - Described in the DIRAC configuration
    - Protocol, endpoint, port, SAPath, Web Service URL
    - Allows autonomous construction of the SURL
    - `SURL = srm:<endPoint>:<port><WSUrl><SAPath><LFN>`
  - SRM spaces at Tier1s
    - Used to have as many SRM spaces as DIRAC SEs, now only 3
    - LHCb-Tape (T1D0) custodial storage
    - LHCb-Disk (T0D1) fast disk access
    - LHCb-User (T0D1) fast disk access for user data

Replica Catalog (2)

- Currently using the LFC
  - Master write service at CERN
  - Replication using Oracle streams to Tier1s
  - Read-only instances at CERN and Tier1s
    - Mostly for redundancy, no need for scaling
- LFC information:
  - Metadata of the file
  - Replicas
    - Use "host name" field for the DIRAC SE name
    - Store SURL of creation for convenience (not used)
      - Allows lcg-util commands to work
  - Quality flag
    - One character comment used to set temporarily a replica as unavailable
- Testing scalability of the DIRAC file catalog
  - Built-in storage usage capabilities (per directory)

Poster: A.Tsaregorodtsev

Bookkeeping Catalog (1)

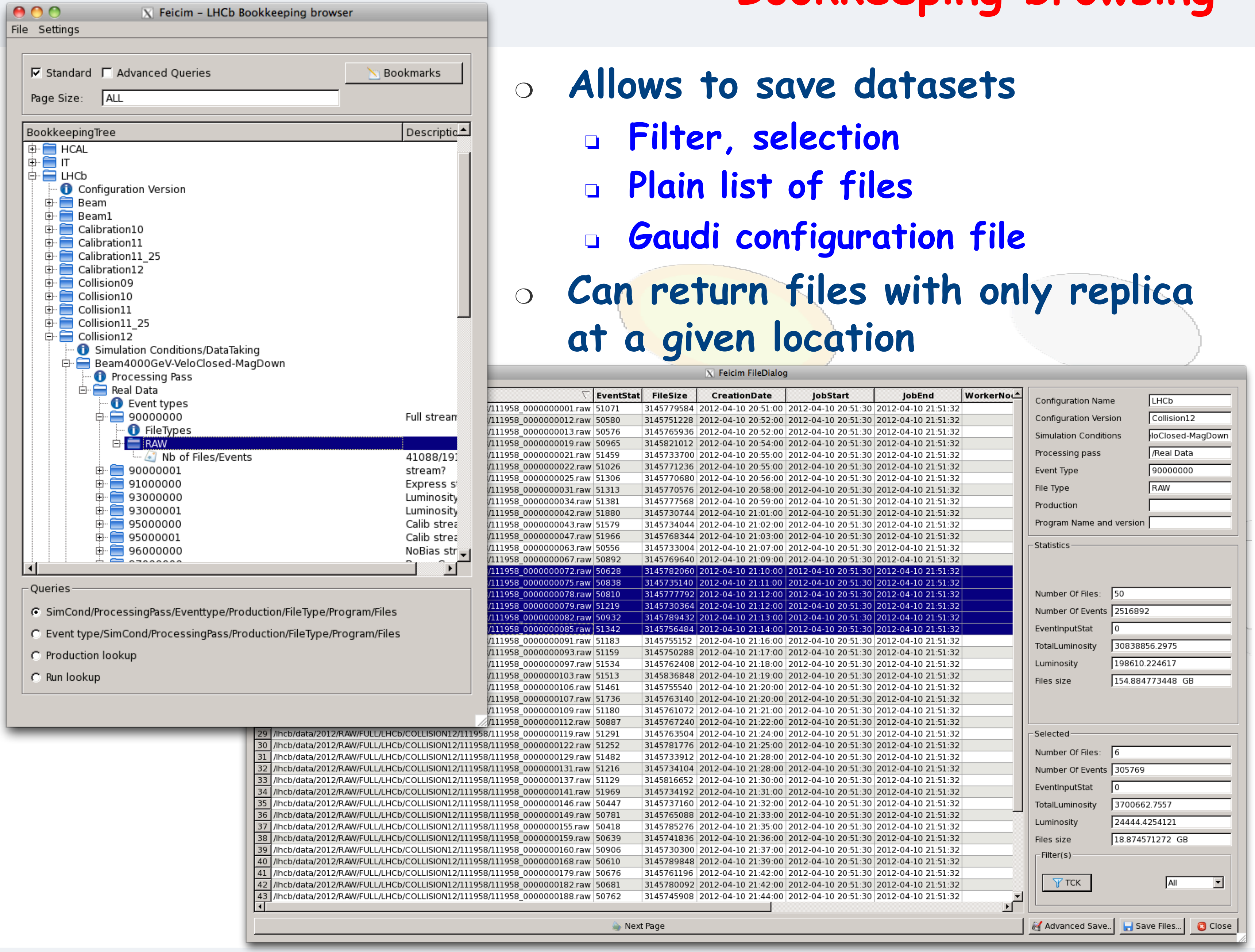
- User selection criteria
  - Origin of the data (real or MC, year of reference)
    - LHCb/Collision12
  - Conditions for data taking of simulation (energy, magnetic field, detector configuration…
    - Beam4000GeV-VeloClosed-MagDown
  - Processing Pass is the level of processing (reconstruction, stripping…) including compatibility version
    - Reco13/Stripping19
  - Event Type is mostly useful for simulation, single value for real data
    - 8 digit numeric code (12345678, 90000000)
  - File Type defines which type of output files the user wants to get for a given processing pass (e.g. which stream)
    - RAW, SDST, BHADRON.DST (for a streamed file)
- Bookkeeping search
  - Using a path
    - `/<origin>/<conditions>/<processing pass>/<event type>/<file type>`

Bookkeeping Catalog (2)

- Much more than a dataset catalog!
- Full provenance of files and jobs
  - Files are input of processing steps ("jobs") that produce files
  - All files ever created are recorded, each processing step as well
    - Full information on the "job" (location, CPU, wall clock time…)
- BK relational database
  - Two main tables: "files" and "jobs"
  - Jobs belong to a "production"
  - "Productions" belong to a "processing pass", with a given "origin" and "condition"
  - Highly optimized search for files, as well as summaries
- Quality flags
  - Files are immutable, but can have a mutable quality flag
  - Files have a flag indicating whether they have a replica or not

Bookkeeping browsing

- Allows to save datasets
  - Filter, selection
  - Plain list of files
  - Gaudi configuration file
- Can return files with only replica at a given location



Feicim – LHCb Bookkeeping browser

File Settings

☑ Standard ☐ Advanced Queries Bookmarks

Page Size: ALL

BookkeepingTree	Descriptio
HCAL	
IT	
LHCb	
Configuration Version	
Beam	
Beam1	
Calibration10	
Calibration11	
Calibration11_25	
Calibration12	
Collision09	
Collision10	
Collision11	
Collision11_25	
Collision12	
Simulation Conditions/DataTaking	
Beam4000GeV-VeloClosed-MagDown	
Processing Pass	
Real Data	
Event types	
90000000	Full strean
FileTypes	
RAW	
Nb of Files/Events	41088/191
90000001	stream?
91000000	Express s
93000000	Luminosity
93000001	Luminosity
95000000	Calib strea
95000001	Calib strea
96000000	NoBias str

Queries

- ● SimCond/ProcessingPass/Eventtype/Production/FileType/Program/Files
- ○ Event type/SimCond/ProcessingPass/Production/FileType/Program/Files
- ○ Production lookup
- ○ Run lookup

Feicim FileDialog

		EventStat	FileSize	CreationDate	JobStart	JobEnd	WorkerNo
	/111958_0000000001.raw	51071	3145779584	2012-04-10 20:51:00	2012-04-10 20:51:30	2012-04-10 21:51:32	
	/111958_0000000012.raw	50580	3145751228	2012-04-10 20:52:00	2012-04-10 20:51:30	2012-04-10 21:51:32	
	/111958_0000000013.raw	50576	3145765936	2012-04-10 20:52:00	2012-04-10 20:51:30	2012-04-10 21:51:32	
	/111958_0000000019.raw	50965	3145821012	2012-04-10 20:54:00	2012-04-10 20:51:30	2012-04-10 21:51:32	
	/111958_0000000021.raw	51459	3145733700	2012-04-10 20:55:00	2012-04-10 20:51:30	2012-04-10 21:51:32	
	/111958_0000000022.raw	51026	3145771236	2012-04-10 20:55:00	2012-04-10 20:51:30	2012-04-10 21:51:32	
	/111958_0000000025.raw	51306	3145770680	2012-04-10 20:56:00	2012-04-10 20:51:30	2012-04-10 21:51:32	
	/111958_0000000031.raw	51313	3145770576	2012-04-10 20:58:00	2012-04-10 20:51:30	2012-04-10 21:51:32	
	/111958_0000000034.raw	51381	3145777568	2012-04-10 20:59:00	2012-04-10 20:51:30	2012-04-10 21:51:32	
	/111958_0000000042.raw	51880	3145730744	2012-04-10 21:01:00	2012-04-10 20:51:30	2012-04-10 21:51:32	
	/111958_0000000043.raw	51579	3145734044	2012-04-10 21:02:00	2012-04-10 20:51:30	2012-04-10 21:51:32	
	/111958_0000000047.raw	51966	3145768344	2012-04-10 21:03:00	2012-04-10 20:51:30	2012-04-10 21:51:32	
	/111958_0000000063.raw	50556	3145733004	2012-04-10 21:07:00	2012-04-10 20:51:30	2012-04-10 21:51:32	
	/111958_0000000067.raw	50892	3145769640	2012-04-10 21:09:00	2012-04-10 20:51:30	2012-04-10 21:51:32	
	/111958_0000000072.raw	50628	3145782060	2012-04-10 21:10:00	2012-04-10 20:51:30	2012-04-10 21:51:32	
	/111958_0000000075.raw	50838	3145735140	2012-04-10 21:11:00	2012-04-10 20:51:30	2012-04-10 21:51:32	
	/111958_0000000078.raw	50810	3145777792	2012-04-10 21:12:00	2012-04-10 20:51:30	2012-04-10 21:51:32	
	/111958_0000000079.raw	51219	3145730364	2012-04-10 21:12:00	2012-04-10 20:51:30	2012-04-10 21:51:32	
	/111958_0000000082.raw	50932	3145789432	2012-04-10 21:13:00	2012-04-10 20:51:30	2012-04-10 21:51:32	
	/111958_0000000085.raw	51342	3145756484	2012-04-10 21:14:00	2012-04-10 20:51:30	2012-04-10 21:51:32	
	/111958_0000000091.raw	51183	3145755152	2012-04-10 21:16:00	2012-04-10 20:51:30	2012-04-10 21:51:32	
	/111958_0000000093.raw	51159	3145750288	2012-04-10 21:17:00	2012-04-10 20:51:30	2012-04-10 21:51:32	
	/111958_0000000097.raw	51534	3145762408	2012-04-10 21:18:00	2012-04-10 20:51:30	2012-04-10 21:51:32	
	/111958_0000000103.raw	51513	3145836848	2012-04-10 21:19:00	2012-04-10 20:51:30	2012-04-10 21:51:32	
	/111958_0000000106.raw	51461	3145755540	2012-04-10 21:20:00	2012-04-10 20:51:30	2012-04-10 21:51:32	
	/111958_0000000107.raw	51736	3145763140	2012-04-10 21:20:00	2012-04-10 20:51:30	2012-04-10 21:51:32	
	/111958_0000000109.raw	51180	3145761072	2012-04-10 21:21:00	2012-04-10 20:51:30	2012-04-10 21:51:32	
	/111958_0000000112.raw	50887	3145767240	2012-04-10 21:22:00	2012-04-10 20:51:30	2012-04-10 21:51:32	
29	/lhcb/data/2012/RAW/FULL/LHCb/COLLISION12/111958/111958_0000000119.raw	51291	3145763504	2012-04-10 21:24:00	2012-04-10 20:51:30	2012-04-10 21:51:32	
30	/lhcb/data/2012/RAW/FULL/LHCb/COLLISION12/111958/111958_0000000122.raw	51252	3145781776	2012-04-10 21:25:00	2012-04-10 20:51:30	2012-04-10 21:51:32	
31	/lhcb/data/2012/RAW/FULL/LHCb/COLLISION12/111958/111958_0000000129.raw	51482	3145733912	2012-04-10 21:28:00	2012-04-10 20:51:30	2012-04-10 21:51:32	
32	/lhcb/data/2012/RAW/FULL/LHCb/COLLISION12/111958/111958_0000000131.raw	51216	3145734104	2012-04-10 21:28:00	2012-04-10 20:51:30	2012-04-10 21:51:32	
33	/lhcb/data/2012/RAW/FULL/LHCb/COLLISION12/111958/111958_0000000137.raw	51129	3145816652	2012-04-10 21:30:00	2012-04-10 20:51:30	2012-04-10 21:51:32	
34	/lhcb/data/2012/RAW/FULL/LHCb/COLLISION12/111958/111958_0000000141.raw	51969	3145734192	2012-04-10 21:31:00	2012-04-10 20:51:30	2012-04-10 21:51:32	
35	/lhcb/data/2012/RAW/FULL/LHCb/COLLISION12/111958/111958_0000000146.raw	50447	3145737160	2012-04-10 21:32:00	2012-04-10 20:51:30	2012-04-10 21:51:32	
36	/lhcb/data/2012/RAW/FULL/LHCb/COLLISION12/111958/111958_0000000149.raw	50781	3145765088	2012-04-10 21:33:00	2012-04-10 20:51:30	2012-04-10 21:51:32	
37	/lhcb/data/2012/RAW/FULL/LHCb/COLLISION12/111958/111958_0000000155.raw	50418	3145785276	2012-04-10 21:35:00	2012-04-10 20:51:30	2012-04-10 21:51:32	
38	/lhcb/data/2012/RAW/FULL/LHCb/COLLISION12/111958/111958_0000000159.raw	50639	3145741836	2012-04-10 21:36:00	2012-04-10 20:51:30	2012-04-10 21:51:32	
39	/lhcb/data/2012/RAW/FULL/LHCb/COLLISION12/111958/111958_0000000160.raw	50906	3145730300	2012-04-10 21:37:00	2012-04-10 20:51:30	2012-04-10 21:51:32	
40	/lhcb/data/2012/RAW/FULL/LHCb/COLLISION12/111958/111958_0000000168.raw	50610	3145789848	2012-04-10 21:39:00	2012-04-10 20:51:30	2012-04-10 21:51:32	
41	/lhcb/data/2012/RAW/FULL/LHCb/COLLISION12/111958/111958_0000000179.raw	50676	3145761196	2012-04-10 21:42:00	2012-04-10 20:51:30	2012-04-10 21:51:32	
42	/lhcb/data/2012/RAW/FULL/LHCb/COLLISION12/111958/111958_0000000182.raw	50681	3145780092	2012-04-10 21:42:00	2012-04-10 20:51:30	2012-04-10 21:51:32	
43	/lhcb/data/2012/RAW/FULL/LHCb/COLLISION12/111958/111958_0000000188.raw	50762	3145745908	2012-04-10 21:44:00	2012-04-10 20:51:30	2012-04-10 21:51:32	

Next Page

Configuration Name	LHCb
Configuration Version	Collision12
Simulation Conditions	loClosed-MagDown
Processing pass	/Real Data
Event Type	90000000
File Type	RAW
Production	
Program Name and version	

Statistics

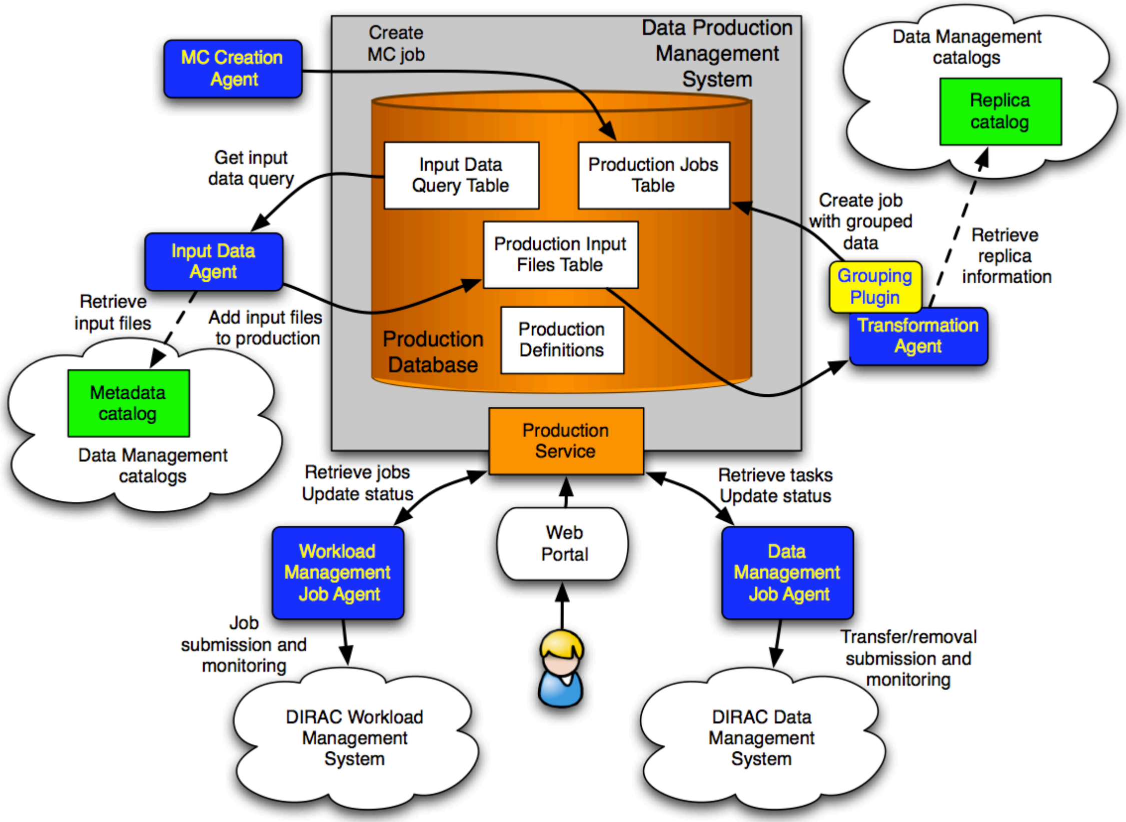
Number Of Files:	50
Number Of Events	2516892
EventInputStat	0
TotalLuminosity	30838856.2975
Luminosity	198610.224617
Files size	154.884773448 GB

Selected

Number Of Files:	6
Number Of Events	305769
EventInputStat	0
TotalLuminosity	3700662.7557
Luminosity	24444.4254121
Files size	18.874571272 GB

Filter(s)

Dataset based transformation


MC Creation Agent
Create MC job
Data Production Management System
Data Management catalogs
Replica catalog
Input Data Query Table
Production Jobs Table
Get input data query
Input Data Agent
Production Input Files Table
Create job with grouped data
Retrieve replica information
Grouping Plugin
Transformation Agent
Retrieve input files
Add input files to production
Production Database
Production Definitions
Metadata catalog
Data Management catalogs
Production Service
Retrieve jobs Update status
Retrieve tasks Update status
Web Portal
Workload Management Job Agent
Data Management Job Agent
Job submission and monitoring
Transfer/removal submission and monitoring
DIRAC Workload Management System
DIRAC Data Management System

- Same mechanism used for jobs (workload management tasks) and data management
- Input datasets based on a bookkeeping query
  - Tasks are data driven, used by both WMS and DMS

Data management transformations

- Replication transformations
  - Uses a policy implementing the Computing Model
  - Creates transfer tasks depending on the original location and space availability
  - Replication using a tree (not all from the initial source)
    - Optimized w.r.t. channel recent bandwidth and SE locality
  - Transfers are whenever possible performed using FTS
    - If not possible, 3rd party gridftp transfer
      - If no FTS channel or for user files (special credentials)
  - Replicas are automatically registered in the LFC
- Removal transformations
  - Used for retiring datasets or reducing the number of replicas
  - Used exceptionally to completely remove files (tests, bugs…)
  - Replica removal protected against last replica removal!

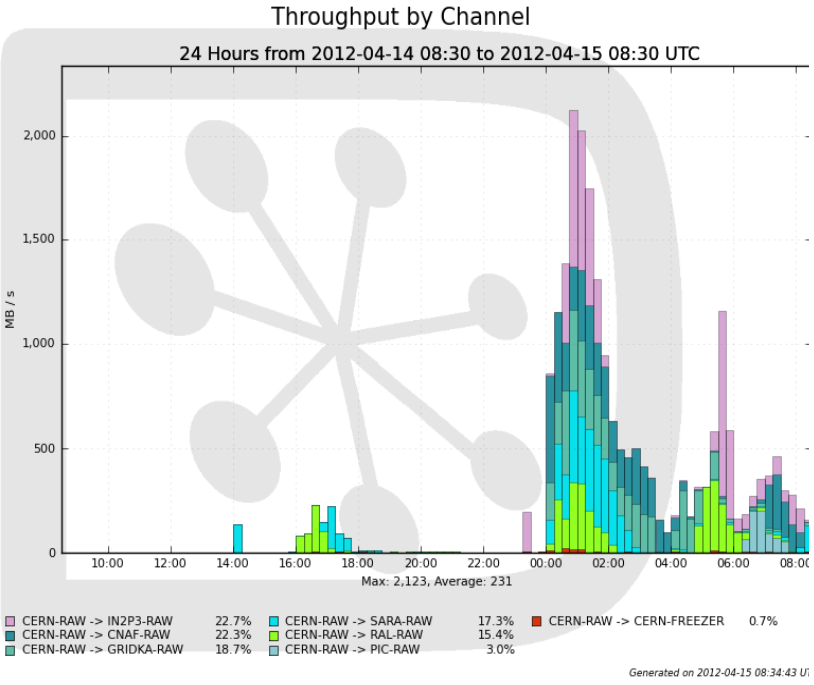
  - Transfers and removal use the Data Manager credentials

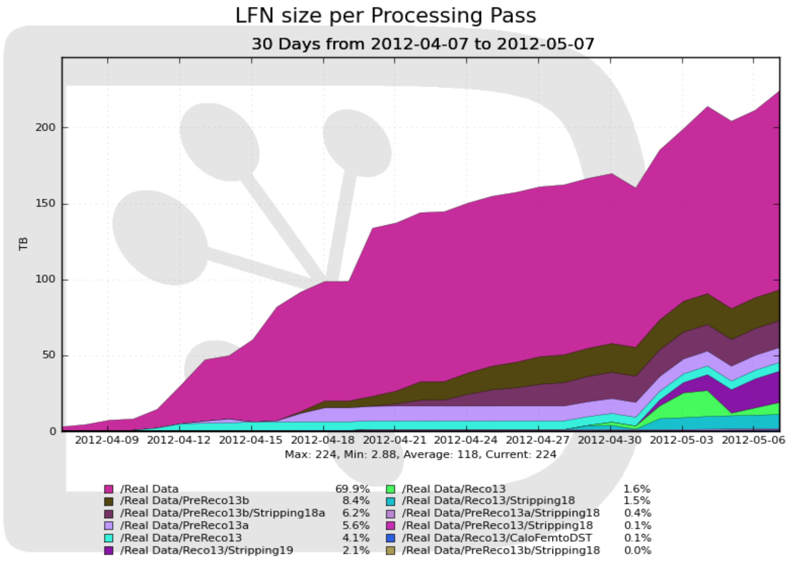
Staging: using files from tape

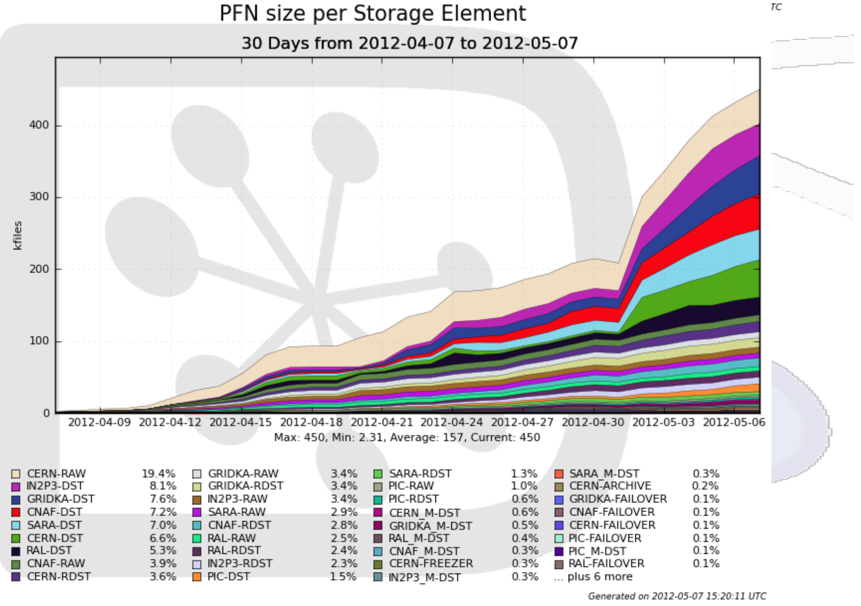
- If jobs use files that are not online (on disk)
  - Before submitting the job
  - Stage the file from tape, and pin it on cache
- Stager agent
  - Performs also cache management
  - Throttle staging requests depending on the cache size and the amount of pinned data
  - Requires fine tuning (pinning and cache size)
    - Caching architecture highly site dependent
    - No publication of cache sizes (except Castor and StoRM)
- Jobs using staged files
  - Check first the file is still staged
    - If not reschedule the job
  - Copies the file locally on the WN whenever possible
    - Space is released faster
    - More reliable access for very long jobs (reconstruction) or jobs using many files (merging)

Data Management Operational experience

- Transfer accounting
  - Per site, channel, user…
- Storage accounting
  - Per dataset, SE, user…
  - User quotas
    - Not strictly enforced…







Outlook

- Improvements on staging
  - Improve the tuning of cache settings
    - Depends on how caches are used by sites
  - Pinning/unpinning
    - Difficult if files are used by more than one job
- Popularity
  - Record dataset usage
    - Reported by jobs: number of files used in a given dataset
    - Account number of files used per dataset per day/week
  - Assess dataset popularity
    - Relate usage to dataset size
  - Take decisions on the number of online replicas
    - Taking into account available space
    - Taking into account expected need in the coming weeks
  - First rely on Data Manager receiving an advice
    - Possibly move to automated dynamic management

Conclusions

- LHCb uses two views for data management:
  - Dataset view as seen by users and productions
  - Files view as seen by Data Management tools (replication, removal)
- Datasets are handled by the LHCb Bookkeeping system (part of LHCbDirac)
- File view is generic and handled by the DIRAC DMS
- The LHCb DMS gives full flexibility for managing data on the Grid
- In the future LHCb expect to use popularity criteria for deciding on the number of replicas for each dataset
  - *Should give more flexibility to the Computing Model*