DIRAC Data Management System

A. Tsaregorodtsev,
CPPM-IN2P3-CNRS

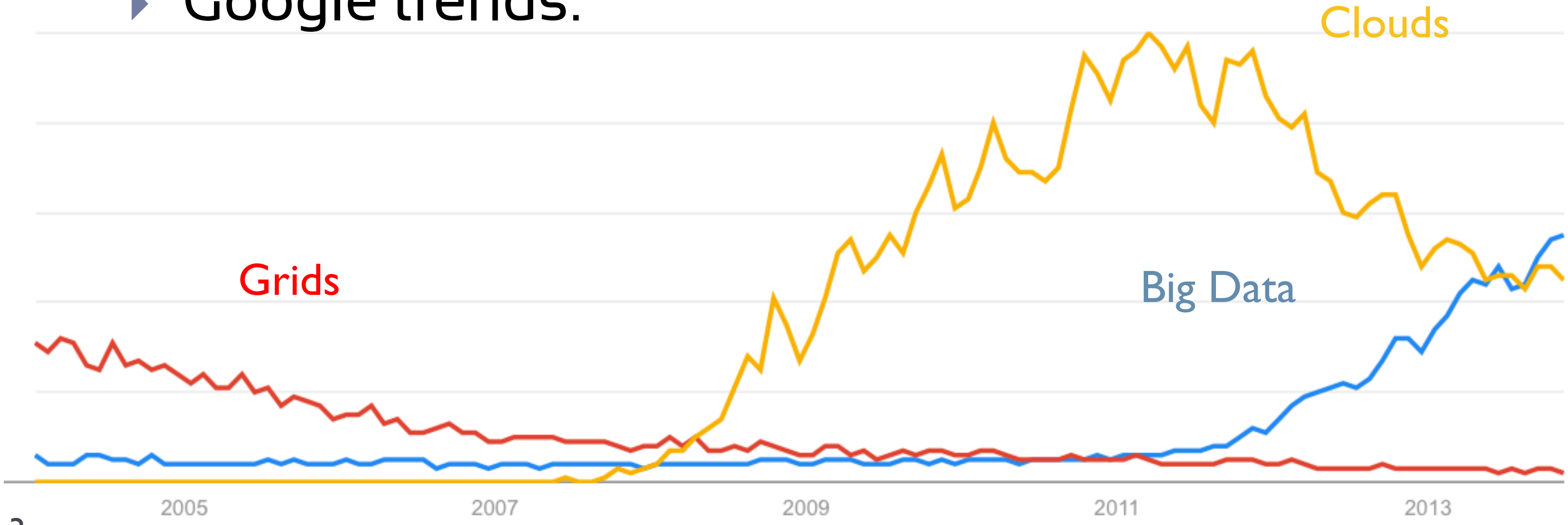
Kurchatov Institute, 26 November, 2015

Plan

- DIRAC Project brief overview
- Data Management System problem
- DIRAC Data Management Model
- DMS Basic Components
- Managing Large Data Flows
- Extending DIRAC DMS
- Conclusions

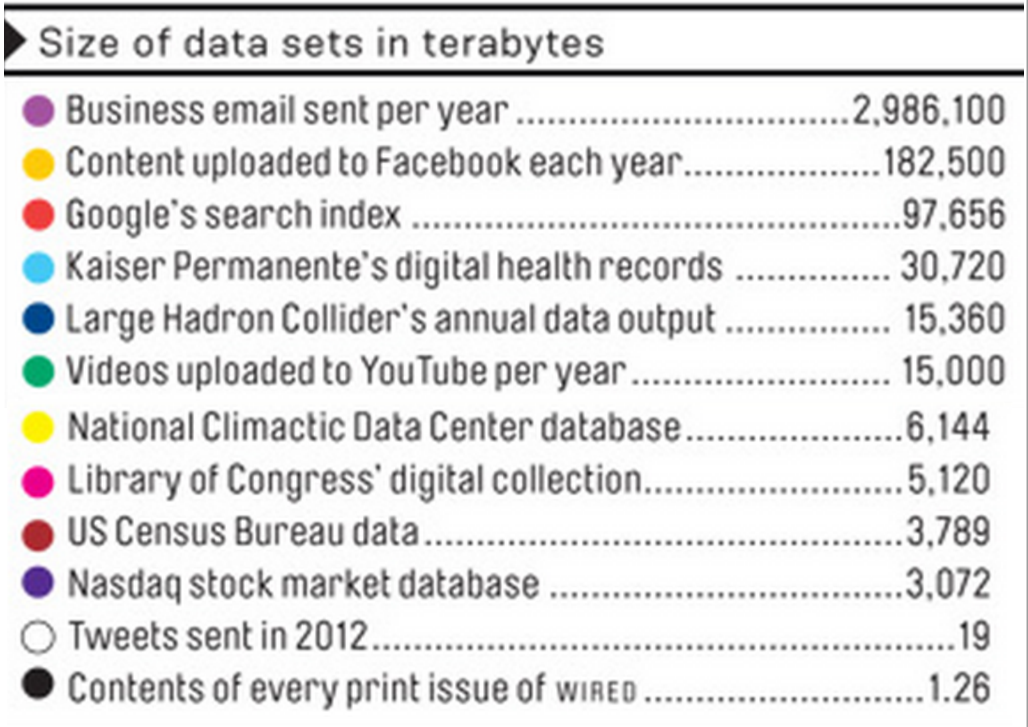
Big Data

- Data that exceeds the boundaries and sizes of normal processing capabilities, forcing you to take a non-traditional approach to its analysis
  - Challenges include analysis, capture, data curation, search, sharing, storage, transfer, visualization, and information privacy
- Telecommunications/Internet traffic increase:
  - ~350 PB in 1990 ➔ ~1ZB (zettabyte = 10^6 PB) in 2015
- Google trends:



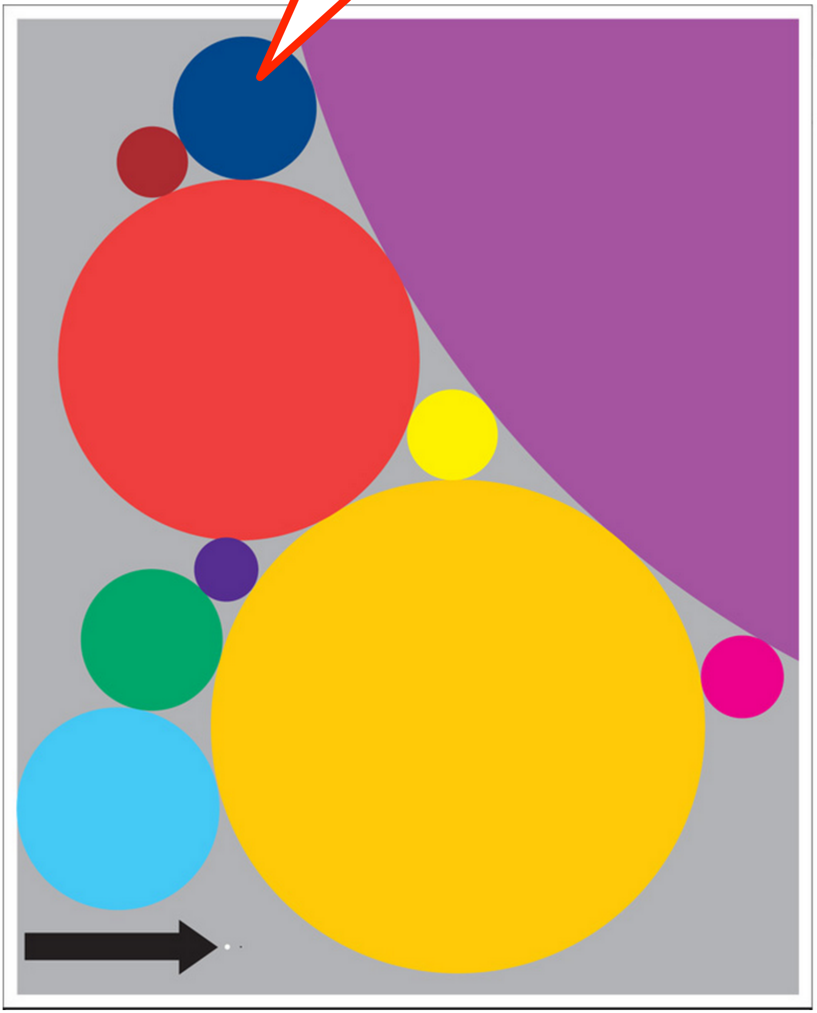
Data

- Already in 2010 the LHC experiments produced 13 PB of data
  - That rate outstripped any other scientific effort going on



Size of data sets in terabytes

●	Business email sent per year	2,986,100
●	Content uploaded to Facebook each year	182,500
●	Google's search index	97,656
●	Kaiser Permanente's digital health records	30,720
●	Large Hadron Collider's annual data output	15,360
●	Videos uploaded to YouTube per year	15,000
●	National Climactic Data Center database	6,144
●	Library of Congress' digital collection	5,120
●	US Census Bureau data	3,789
●	Nasdaq stock market database	3,072
○	Tweets sent in 2012	19
●	Contents of every print issue of WIRED	1.26

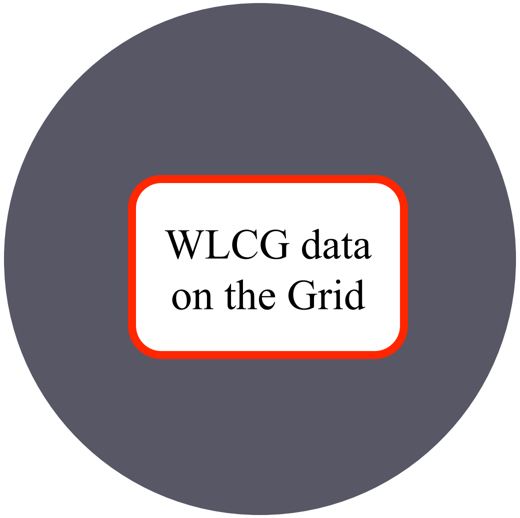


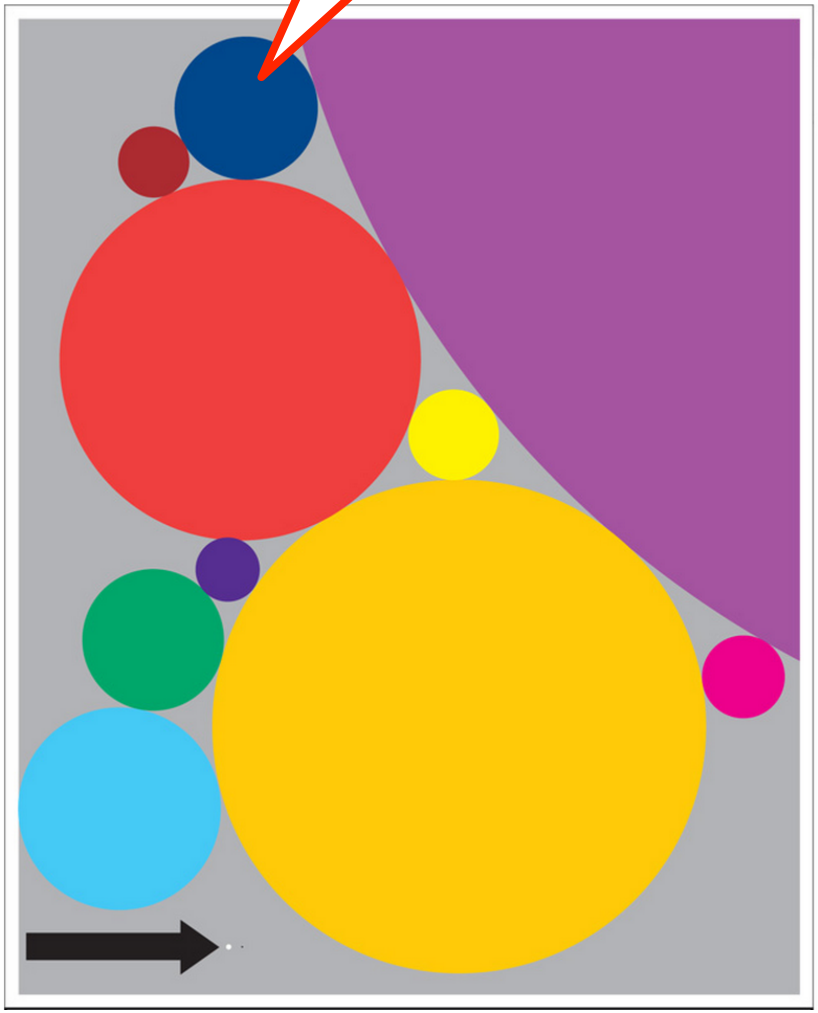
LHC RAW data per year

http://www.wired.com/magazine/2013/04/bigdata

Data

- LHC RAW data volumes are inflated by storing derived data products, replication for safety and efficient access, and by the need for storing even more simulated data than the RAW data:
- ~130 PB overall


WLCG data on the Grid


LHC RAW data per year

http://www.wired.com/magazine/2013/04/bigdata

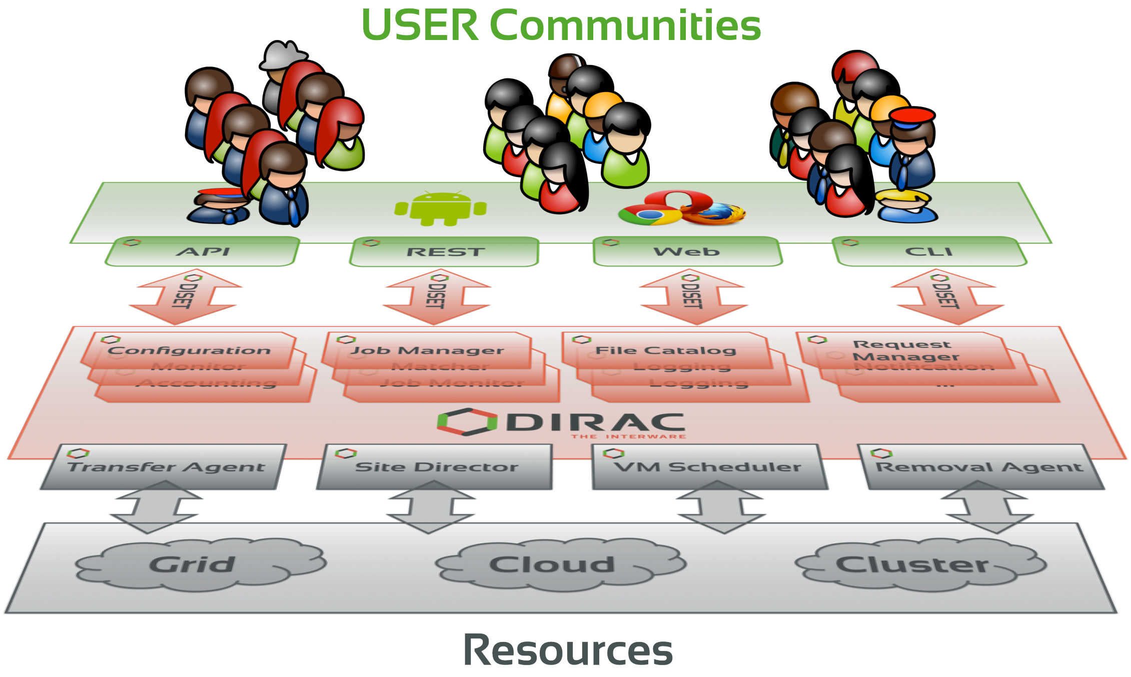
DIRAC Grid Solution

- LHC experiments, all developed their own middleware to address the above problems
  - PanDA, AliEn, glideIn WMS, PhEDEx, …
- DIRAC is developed originally for the LHCb experiment
- The experience collected with a production grid system of a large HEP experiment is very valuable
  - Several new experiments expressed interest in using this software relying on its proven in practice utility
- In 2009 the core DIRAC development team decided to generalize the software to make it suitable for any user community.
  - Consortium to develop, maintain and promote the DIRAC software
    - CERN, CNRS, University of Barcelona, University of Montpellier, IHEP
- The results of this work allow to offer DIRAC as a general purpose distributed computing framework

DIRAC
THE INTERWARE

Interware

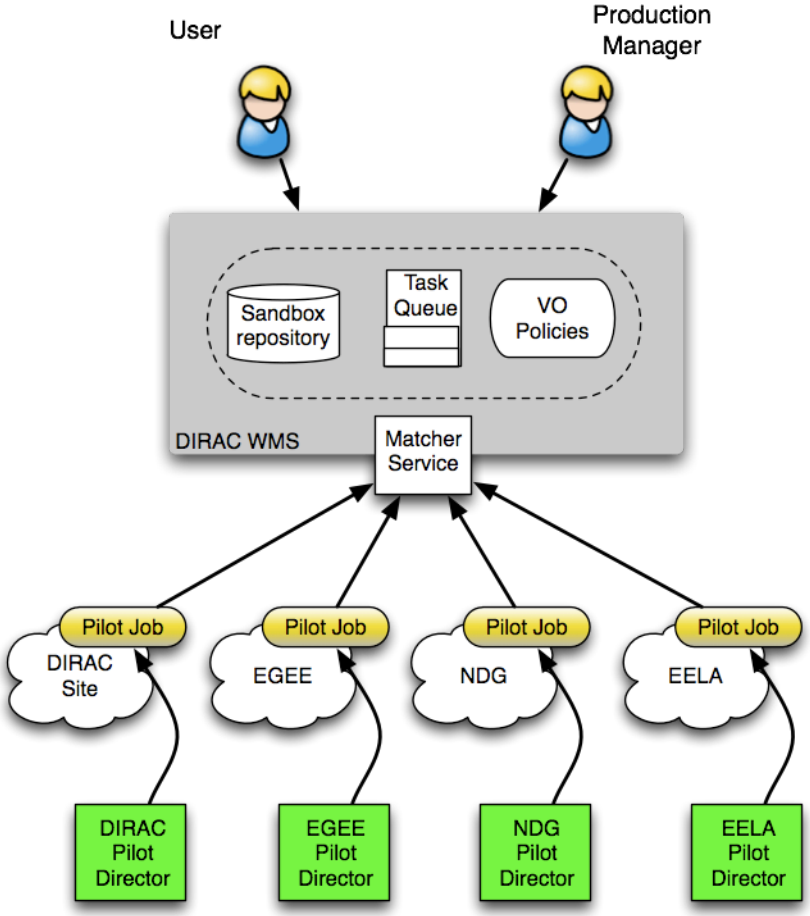
- DIRAC provides all the necessary components to build ad-hoc grid infrastructures **interconnecting** computing resources of different types, allowing **interoperability** and simplifying **interfaces**. This allows to speak about the DIRAC ***interware***.



DIRAC THE INTERWARE

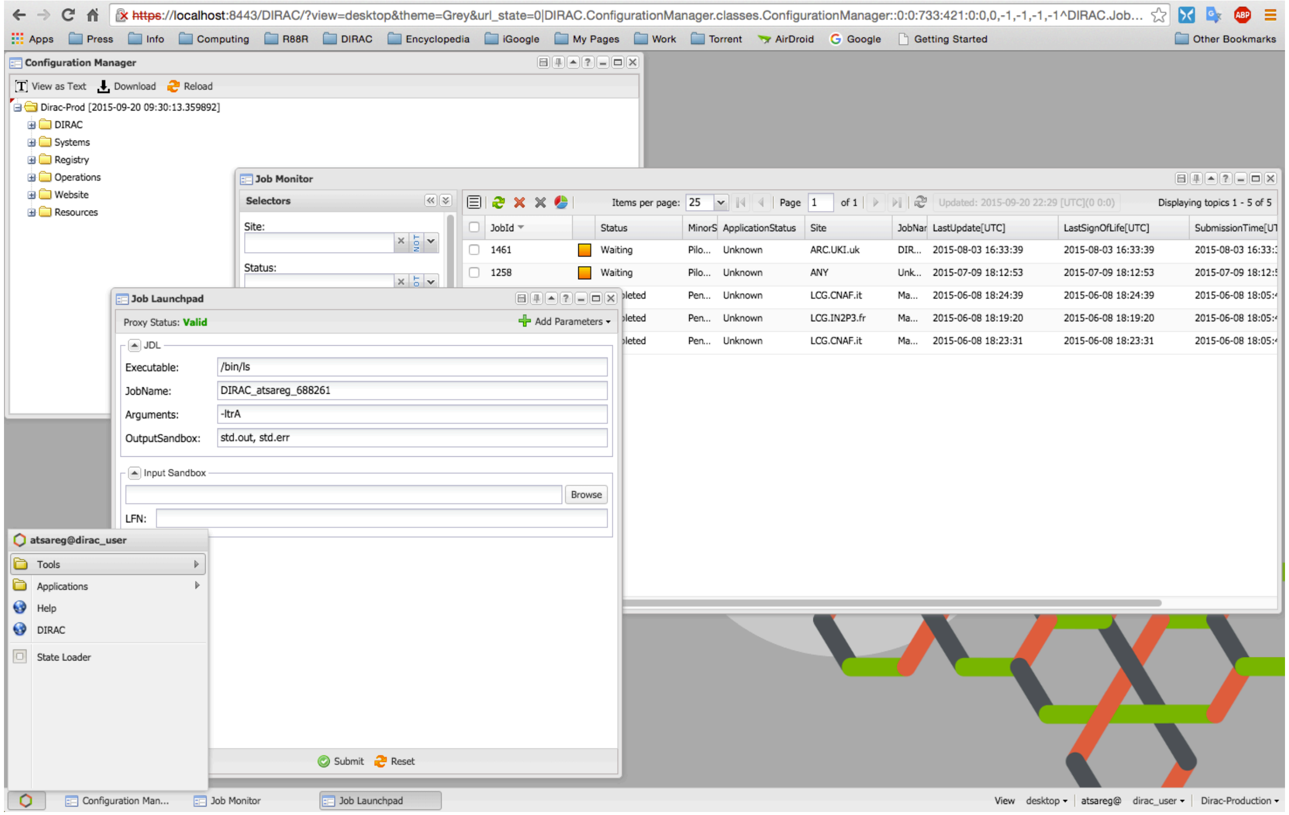
Interware: WMS

- Pilot based Workload Management
  - High user job efficiency
  - Suitable for usage with heterogeneous resources
  - Allowing application of community policies



Distributed computer

- DIRAC forms an abstraction of a simple computer which has the power of thousands of CPUs and petabytes of storage behind the scene coming from various sources and various technologies (grids, clouds, etc)



Dirac-Prod [2015-09-20 09:30:13.359892]

- DIRAC
- Systems
- Registry
- Operations
- Website
- Resources

Site:

Status:

Items per page: 25 Page 1 of 1 Updated: 2015-09-20 22:29 [UTC](0 0:0) Displaying topics 1 - 5 of 5

JobId		Status	MinorS	ApplicationStatus	Site	JobNar	LastUpdate[UTC]	LastSignOfLife[UTC]	SubmissionTime[UT
1461	🟧	Waiting	Pilo...	Unknown	ARC.UKI.uk	DIR...	2015-08-03 16:33:39	2015-08-03 16:33:39	2015-08-03 16:33:
1258	🟧	Waiting	Pilo...	Unknown	ANY	Unk...	2015-07-09 18:12:53	2015-07-09 18:12:53	2015-07-09 18:12:
		pleted	Pen...	Unknown	LCG.CNAF.it	Ma...	2015-06-08 18:24:39	2015-06-08 18:24:39	2015-06-08 18:05:
		pleted	Pen...	Unknown	LCG.IN2P3.fr	Ma...	2015-06-08 18:19:20	2015-06-08 18:19:20	2015-06-08 18:05:
		pleted	Pen...	Unknown	LCG.CNAF.it	Ma...	2015-06-08 18:23:31	2015-06-08 18:23:31	2015-06-08 18:05:

Proxy Status: **Valid**

Executable:	/bin/ls
JobName:	DIRAC_atsareg_688261
Arguments:	-ltrA
OutputSandbox:	std.out, std.err
LFN:	

- DIRAC Web Portal is following the computer desktop paradigm
  - Natural for a non-expert user

DMS Problem

- Data is partitioned in files
  - 0 to 10-20 GB
- File replicas are distributed over a number of Storage Elements world wide
  - Distribution according to a Computing Model of the experiment
- Files are registered in a single logical name space
  - With metadata and ACL info
- For most of applications the file access as simple as in a File System
  - With additional upload/download if necessary

DMS objects

- Files
- Replicas
- Datasets

- Storage Elements
  - Access protocols
- Transfer channels
  - Transfer services

- Catalogs
- File Systems

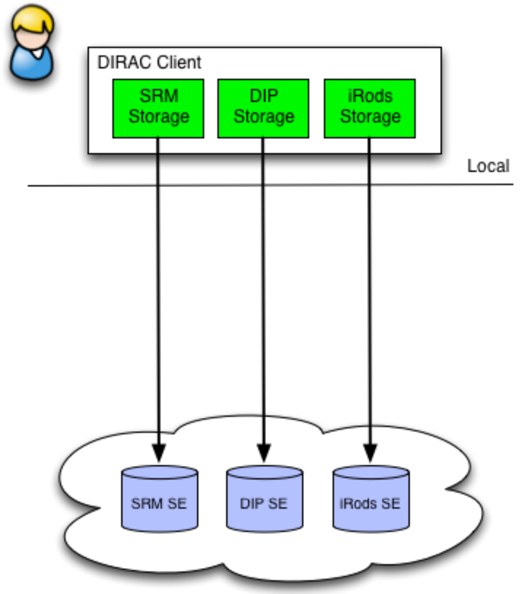
DMS tasks

- Initial File upload
- Catalog registration
- File replication
- File access/download
- Integrity checking
- File removal

- Usually we are working with 10K to 1M files at a time
  - Make sure that ALL the elementary operations are accomplished

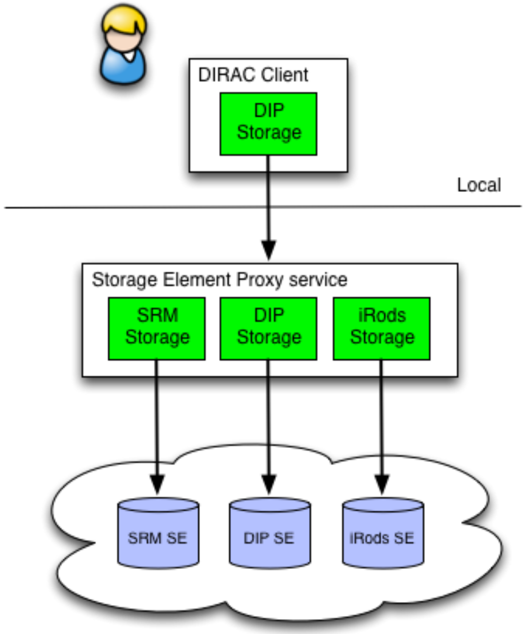
Data Management Model

- Storage element abstraction with a client implementation for each access protocol (SRM, XROOTD, gfal2 based, etc)
- Each SE is seen by the clients as a logical entity
  - With some specific operational properties
  - New SE technologies, e.g. Federated Cloud, EOS are available after the proper configuration
  - SE's can be configured with multiple protocols



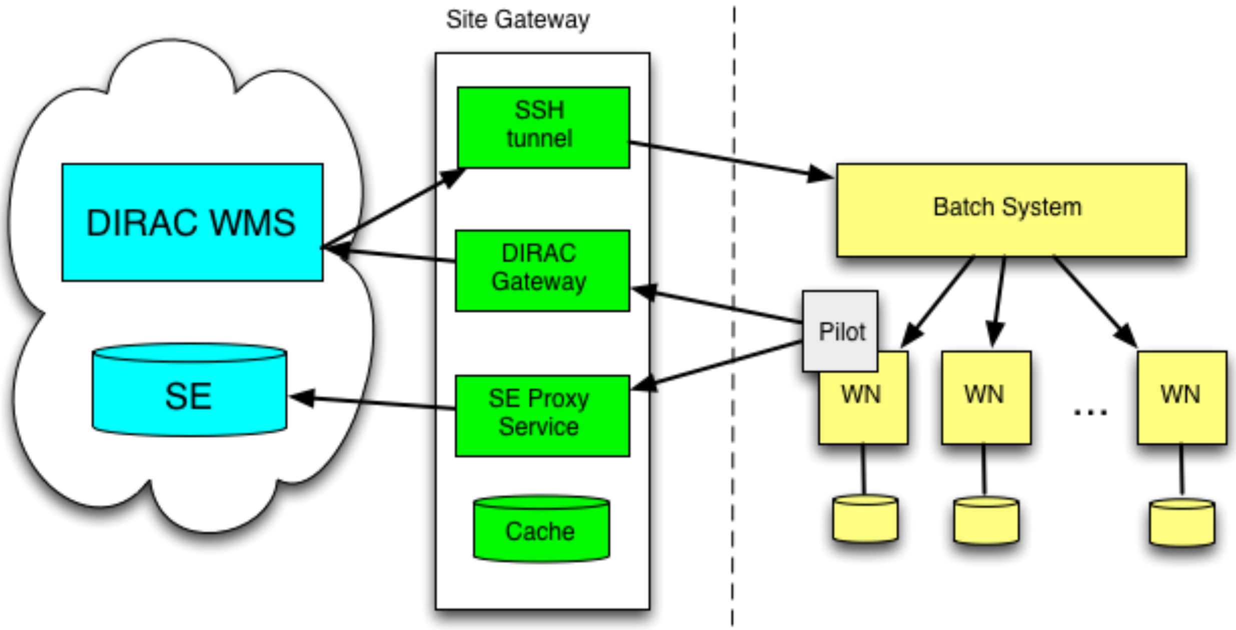
Storage Element Proxy

- SE Proxy Service translates the DIRAC data transfer protocol to a particular storage protocol
  - Using DIRAC authentication
  - Using credentials specific to the target storage system
- SE Proxy Service allows access to storages not having access libraries on a given client machine
- Allows third party like transfers between incompatible storages



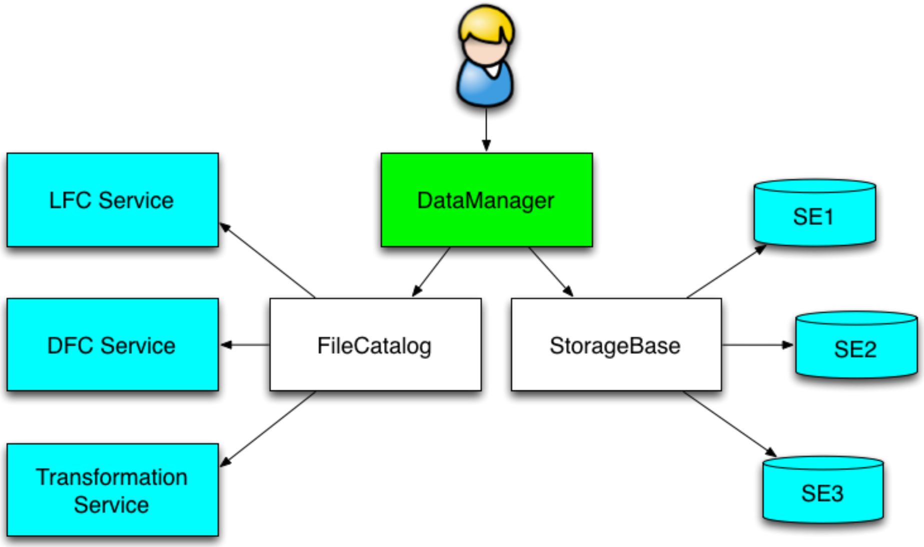
HPC example

- Pilot submitted to the batch system through the SSH tunnel
- Pilot communicates with the DIRAC service through the Gateway proxy service
- Output upload to the target SE through the SE proxy



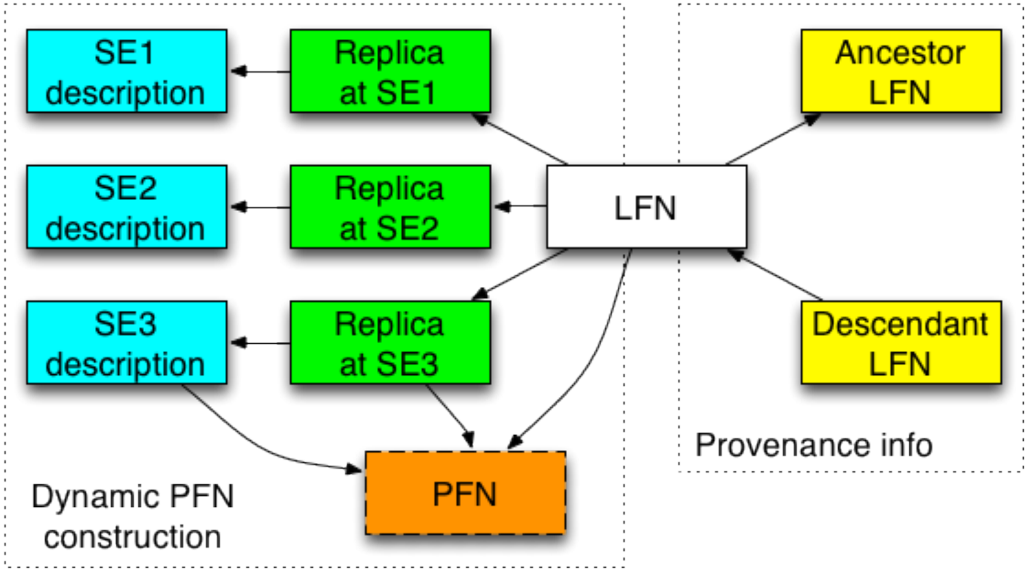
File Catalog

- Central File Catalog (DFC, LFC, …) is maintaining a single global logical name space
- Several catalogs can be used together
  - The mechanism is used to send messages to "pseudocatalog" services
- DataManager is a single client interface for logical data operations



Replica Catalog

- File standard metadata
  - Size, ownership, time stamps, ACL, checksum
- Standard Replica Catalog functionality
  - Optimized for bulk queries
- On the fly PFN construction
  - Small database footprint
  - Full PFN can be stored if necessary

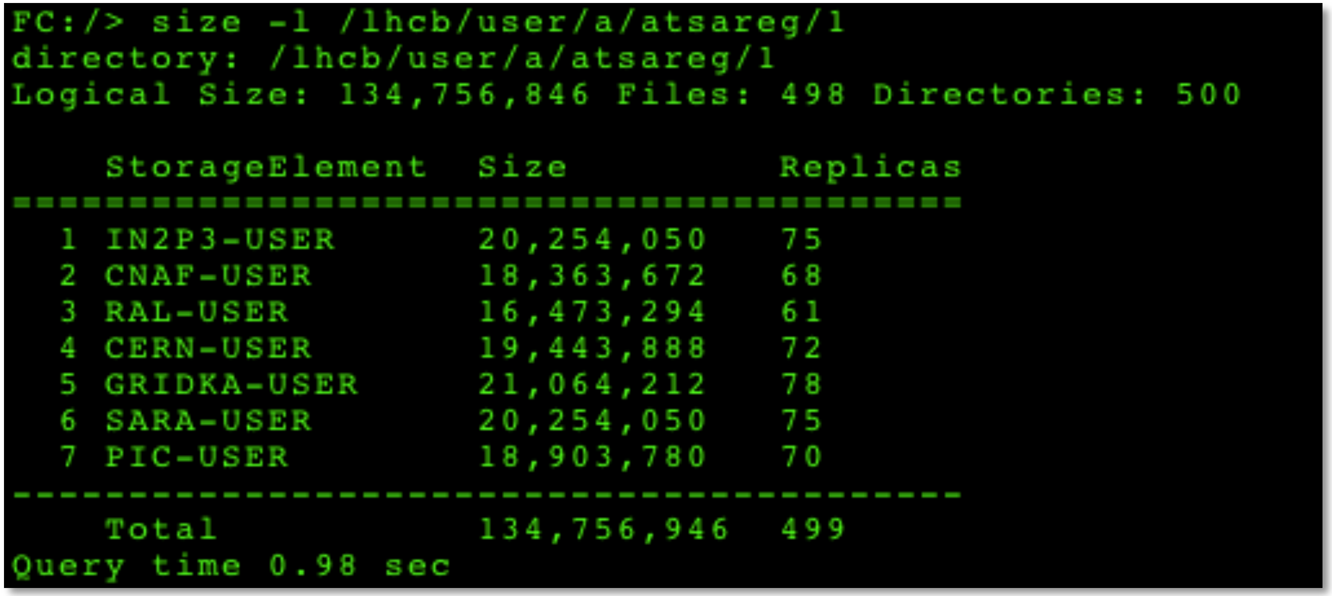


- Ancestor-descendent relations
  - Basic provenance information
  - Possibility to select ancestors in a given generation

Storage Usage

- Efficient Storage Usage reports
  - Necessary for
    - quota policy management
    - storage management
- Using special prefilled tables
  - Updated at each new file or replica insertion
    - More efficient with bulk insertion
  - Instant reports for any directory
  - Possibility of instant "*du*" command

Storage Usage



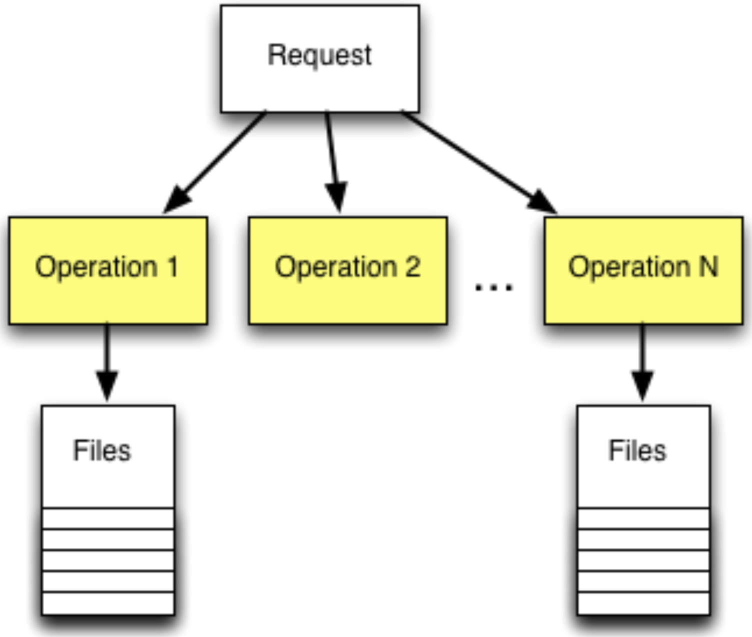
```
FC:/> size -l /lhcb/user/a/atsareg/l
directory: /lhcb/user/a/atsareg/l
Logical Size: 134,756,846 Files: 498 Directories: 500

    StorageElement   Size            Replicas
=========================================
  1 IN2P3-USER       20,254,050    75
  2 CNAF-USER        18,363,672    68
  3 RAL-USER         16,473,294    61
  4 CERN-USER        19,443,888    72
  5 GRIDKA-USER      21,064,212    78
  6 SARA-USER        20,254,050    75
  7 PIC-USER         18,903,780    70
-----------------------------------------
    Total            134,756,946   499
Query time 0.98 sec
```

- Report of storage usage for any directory
  - Whole community data
  - Per user data
  - “Logical” storage
    - LFNs, sum of the LFN sizes
  - “Physical” storage
    - Physical replicas, total volume per Storage Element

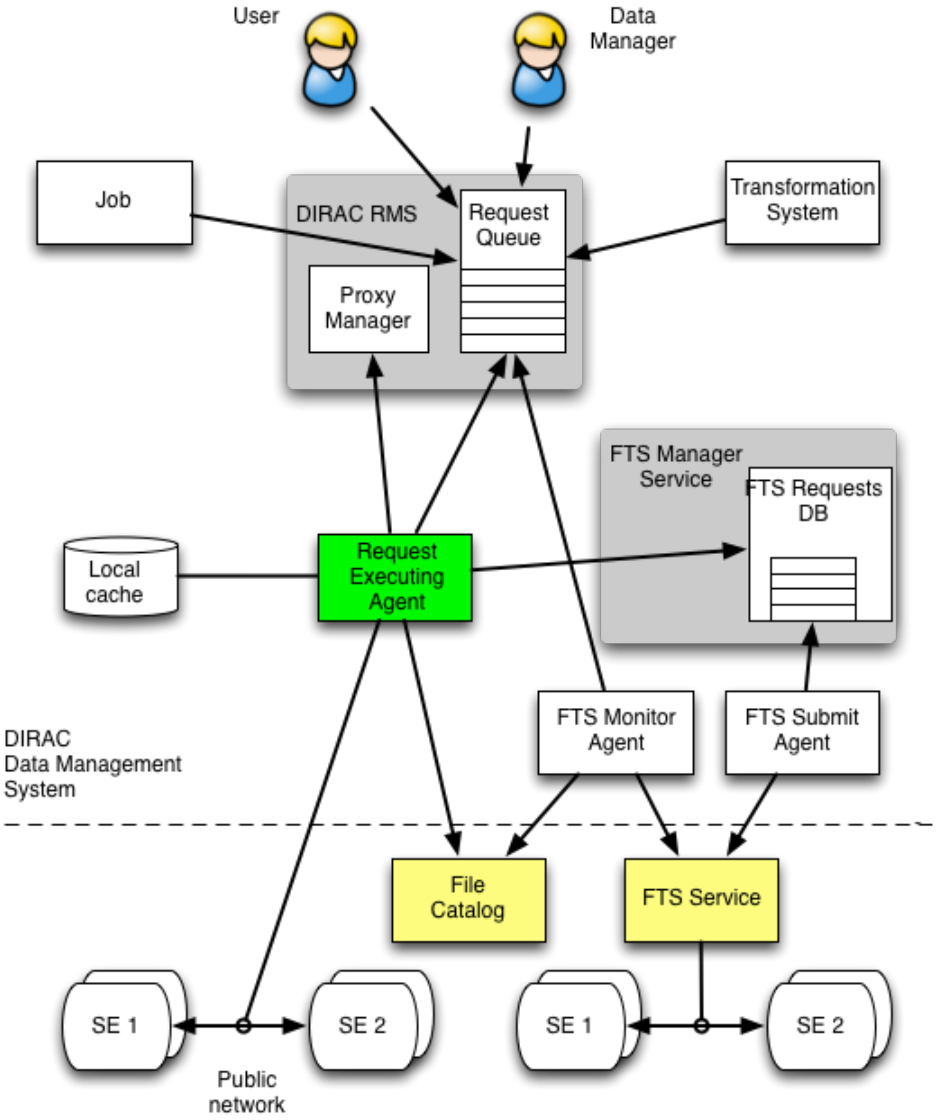
Asynchronous operations

- Request Management System (RMS)
  - Keeps the database of Requests
- Request is a sequence of Operations executed in a certain order
  - Operations can have associated Files
- Each Operation type has a dedicated Executor
  - Execution is done with the credentials of the Owner of the Request
    - E.g. user defined operations
  - Examples: ForwardDISET, ReplicateFile, RemoveFIle
- Executors are invoked by an agent running in a background
  - Retry logic in case of failures



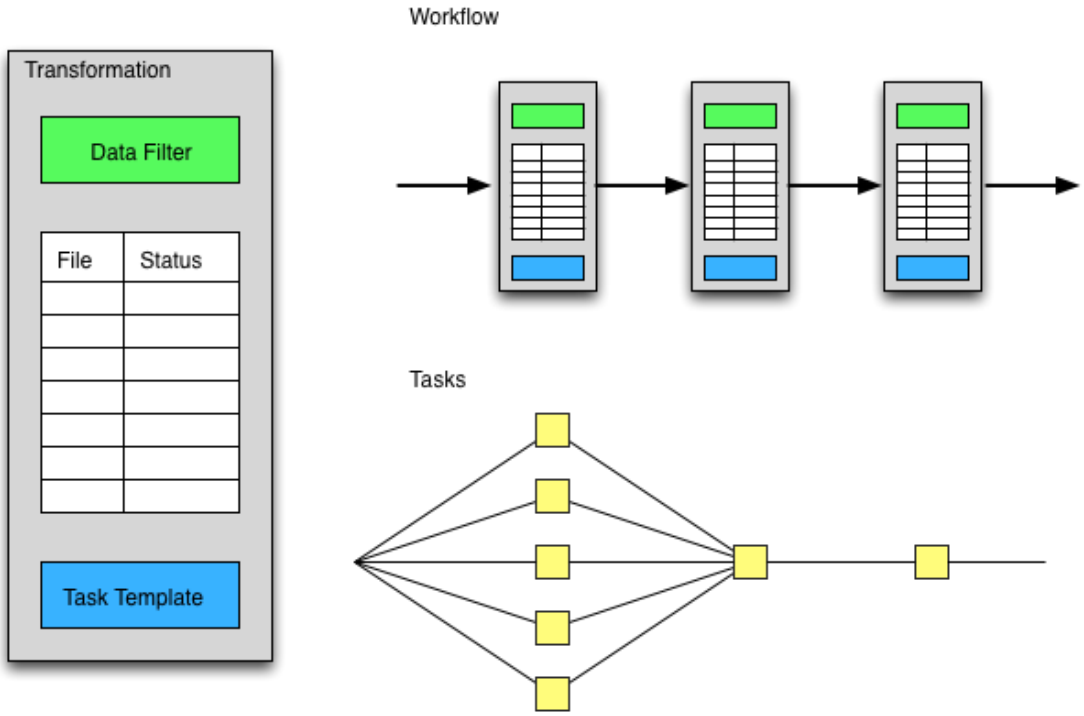
Bulk transfers

- Replication/Removal Requests with multiple files are stored in the RMS
  - By users, data managers, Transformation System
- The Replication Operation executor
  - Performs the replication itself or
  - Delegates replication to an external service
    - E.g. FTS
  - A dedicated FTSManager service keeps track of the submitted FTS requests
  - FTSMonitor Agent monitors the request progress, updates the FileCatalog with the new replicas



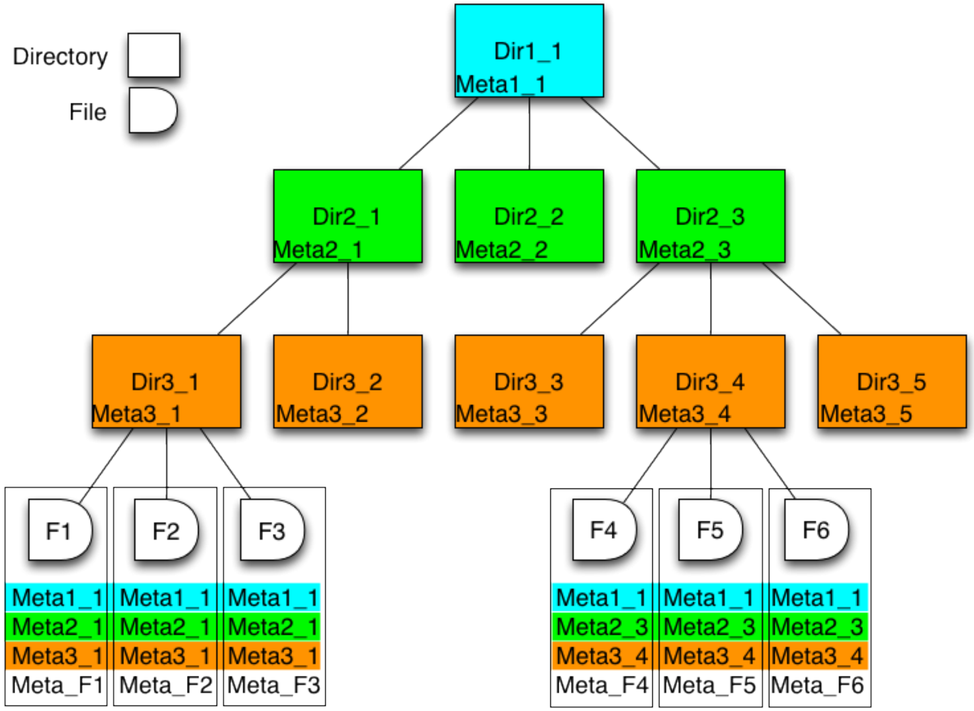
Transformation System

- Data driven workflows as chains of data transformations
  - Transformation: input data filter + recipe to create tasks
  - Tasks are created as soon as data with required properties is registered into the system
  - Tasks: jobs, data replication, etc
- Transformations can be used for automatic data driven bulk data operations
  - Scheduling RMS tasks
  - Often as part of a more general workflow
- Example: Automatic LHCb Raw Data distribution to T1 centers according to predefined quotas



File Catalog

- DFC is Replica and Metadata Catalog
  - User defined metadata
  - The same hierarchy for metadata as for the logical name space
    - Metadata associated with files and directories
    - Allow for efficient searches
  - Efficient Storage Usage reports
    - Suitable for user quotas



- Example query:

  -
    ```
    find /lhcb/mcdata LastAccess < 01-01-2012
    GaussVersion=v1,v2 SE=IN2P3,CERN Name=*.raw
    ```

Support for datasets

- Datasets defined as a resulting list of files from a given metaquery
  - Particular case: all the files under a given directory
- Dataset objects are stored in the same directory hierarchy as files
  - ACLs, ownership, show up in the output of ***ls*** command as for files
- Datasets can be frozen in order not to change from one query to another
  - Can be refreshed by an explicit command, a quick check if changed since the last update
- Datasets can be annotated
- Operations on datasets
  - Replica lookup for all the files in a dataset
  - Total size, number of files report
  - Replication, removal, etc

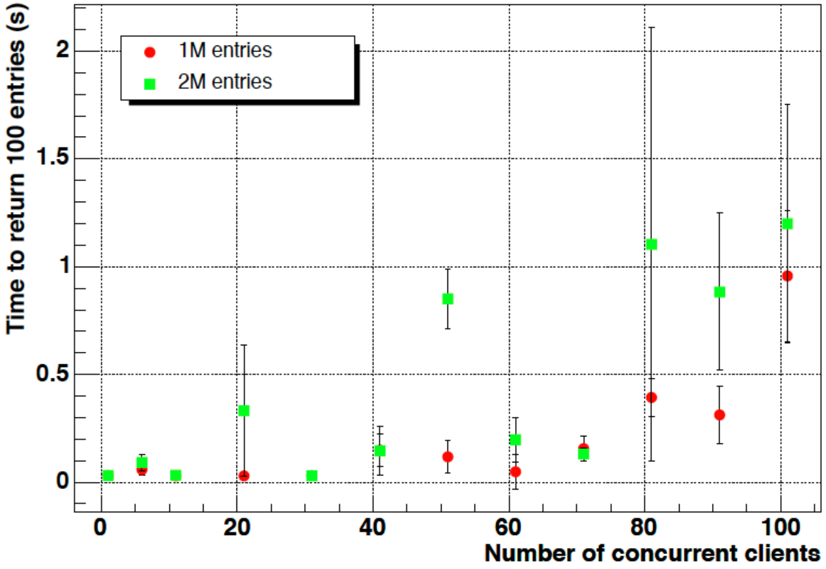
DMS usage

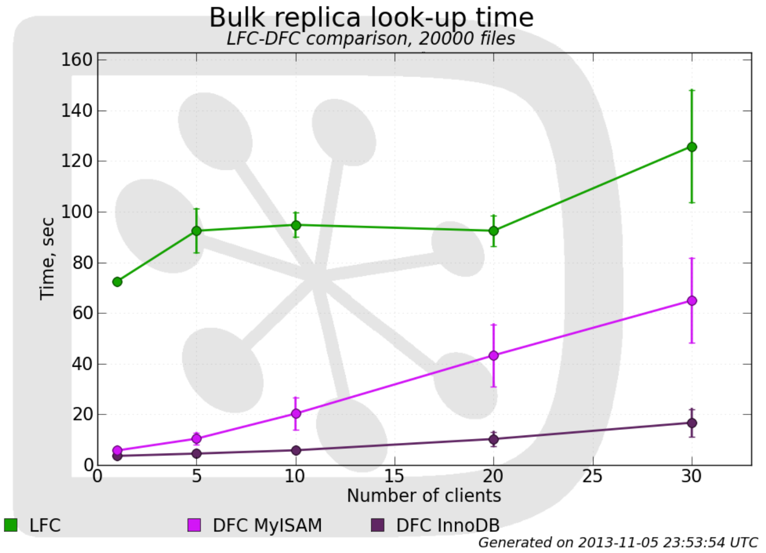
- LHCb accomplished migration from LFC to DFC
  - Needed to develop a specific ACL plugin where several DIRAC groups have same ACLs for a given data
  - Not using the Metadata features of the DFC except for the Storage Usage reports
  - Using Transformation System of DIRAC for bulk data driven operations (e.g. replication, processing tasks submission, etc)
- ILC, BES III, CTA use intensively DFC as both Replica and Metadata Catalog
  - BES III performed a detailed performance comparison with the AMGA metadata service
- Pierre Auger Observatory
  - Working on complex metadata queries and dataset algebra (dataset relations, intersections, unions, etc)

DIRAC File Catalog evaluation

- Tests with Auger data
  - ~30M files
  - Identical LFC and DFC server hardware

File search by metadata





- BES Collaboration made a thorough comparison of DFC vs AMGA
  - Similar performance
  - More suitable functionality

Interfaces

- Command line tools
  - Multiple dirac-dms-… commands
- COMDIRAC
  - Representing the logical DIRAC file namespace as a parallel shell
  - **dls, dcd, dpwd, dfind, ddu** etc commands
  - **dput, dget, drepl** for file upload/download/replication
- Web Interface
  - Using a standard file browser paradigm
    - Possibility to define metadata queries
  - Under development
    - Better connection to other systems (WMS)
    - Better support of the DIRAC "computer" paradigm

DIRAC for CTA: DIRAC File Catalog

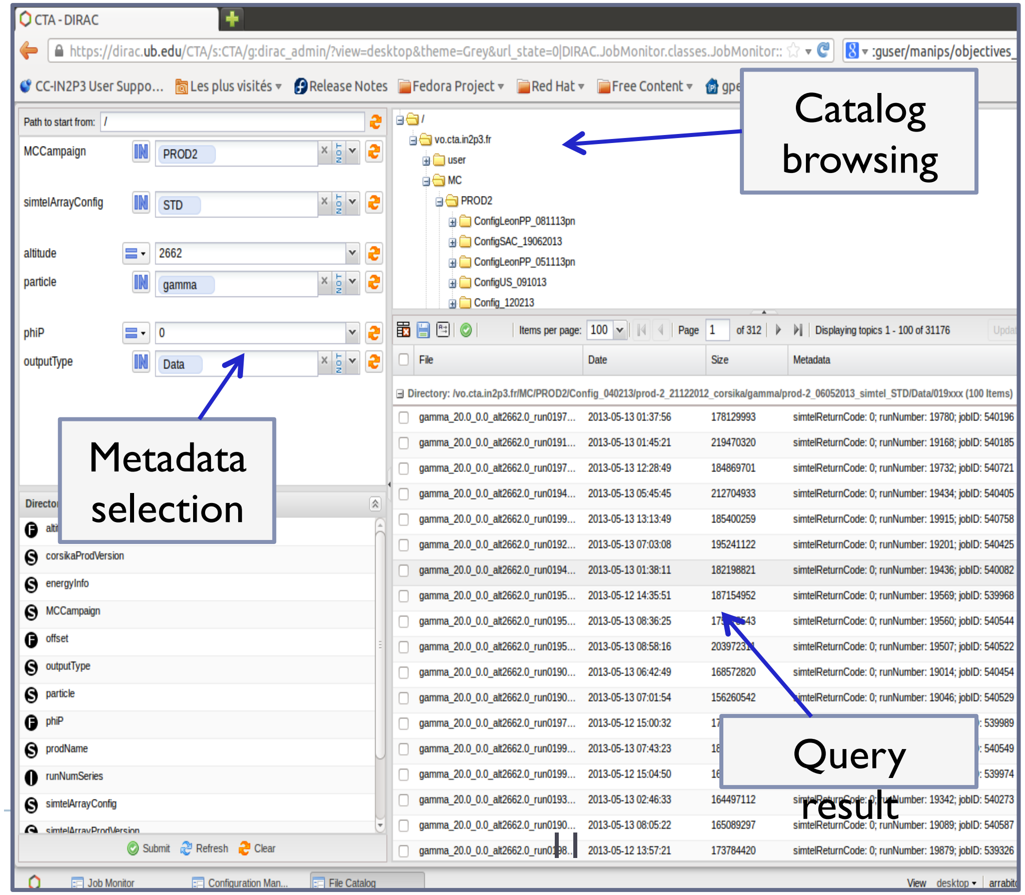
- In use since 2012 in parallel with LFC. Full migration to DFC in summer 2015
- More than 21 M of replicas registered
- About 10 meta-data defined to characterize MC datasets

Query example:

cta-prod3-query --site=Paranal --particle=gamma --tel_sim_prog=simtel --array_layout=hex --phiP=180 --thetaP=20 --outputType=Data

Typical queries return several hundreds of thousands of files

DFC web interface



Path to start from: /

MCCampaign	IN	PROD2
simtelArrayConfig	IN	STD
altitude	=	2662
particle	IN	gamma
phiP	=	0
outputType	IN	Data

Directory metadata: alti..., corsikaProdVersion, energyInfo, MCCampaign, offset, outputType, particle, phiP, prodName, runNumSeries, simtelArrayConfig, simtelArrayProdVersion

```
/
  vo.cta.in2p3.fr
    user
    MC
      PROD2
        ConfigLeonPP_081113pn
        ConfigSAC_19062013
        ConfigLeonPP_051113pn
        ConfigUS_091013
        Config_120213
```

Items per page: 100 | Page 1 of 312 | Displaying topics 1 - 100 of 31176

Directory: /vo.cta.in2p3.fr/MC/PROD2/Config_040213/prod-2_21122012_corsika/gamma/prod-2_06052013_simtel_STD/Data/019xxx (100 Items)

File	Date	Size	Metadata
gamma_20.0_0.0_alt2662.0_run0197...	2013-05-13 01:37:56	178129993	simtelReturnCode: 0; runNumber: 19780; jobID: 540196
gamma_20.0_0.0_alt2662.0_run0191...	2013-05-13 01:45:21	219470320	simtelReturnCode: 0; runNumber: 19168; jobID: 540185
gamma_20.0_0.0_alt2662.0_run0197...	2013-05-13 12:28:49	184869701	simtelReturnCode: 0; runNumber: 19732; jobID: 540721
gamma_20.0_0.0_alt2662.0_run0194...	2013-05-13 05:45:45	212704933	simtelReturnCode: 0; runNumber: 19434; jobID: 540405
gamma_20.0_0.0_alt2662.0_run0199...	2013-05-13 13:13:49	185400259	simtelReturnCode: 0; runNumber: 19915; jobID: 540758
gamma_20.0_0.0_alt2662.0_run0192...	2013-05-13 07:03:08	195241122	simtelReturnCode: 0; runNumber: 19201; jobID: 540425
gamma_20.0_0.0_alt2662.0_run0194...	2013-05-13 01:38:11	182198821	simtelReturnCode: 0; runNumber: 19436; jobID: 540082
gamma_20.0_0.0_alt2662.0_run0195...	2013-05-12 14:35:51	187154952	simtelReturnCode: 0; runNumber: 19569; jobID: 539968
gamma_20.0_0.0_alt2662.0_run0195...	2013-05-13 08:36:25	[illegible]	simtelReturnCode: 0; runNumber: 19560; jobID: 540544
gamma_20.0_0.0_alt2662.0_run0195...	2013-05-13 08:58:16	[illegible]	simtelReturnCode: 0; runNumber: 19507; jobID: 540522
gamma_20.0_0.0_alt2662.0_run0190...	2013-05-13 06:42:49	168572820	simtelReturnCode: 0; runNumber: 19014; jobID: 540454
gamma_20.0_0.0_alt2662.0_run0190...	2013-05-13 07:01:54	156260542	simtelReturnCode: 0; runNumber: 19046; jobID: 540529
gamma_20.0_0.0_alt2662.0_run0197...	2013-05-12 15:00:32	[illegible]	[illegible] 539989
gamma_20.0_0.0_alt2662.0_run0199...	2013-05-13 07:43:23	[illegible]	[illegible] 540549
gamma_20.0_0.0_alt2662.0_run0199...	2013-05-12 15:04:50	[illegible]	[illegible] 539974
gamma_20.0_0.0_alt2662.0_run0193...	2013-05-13 02:46:33	164497112	simtelReturnCode: 0; runNumber: 19342; jobID: 540273
gamma_20.0_0.0_alt2662.0_run0190...	2013-05-13 08:05:22	165089297	simtelReturnCode: 0; runNumber: 19089; jobID: 540587
gamma_20.0_0.0_alt2662.0_run0198...	2013-05-12 13:57:21	173784420	simtelReturnCode: 0; runNumber: 19879; jobID: 539326

Catalog browsing

Metadata selection

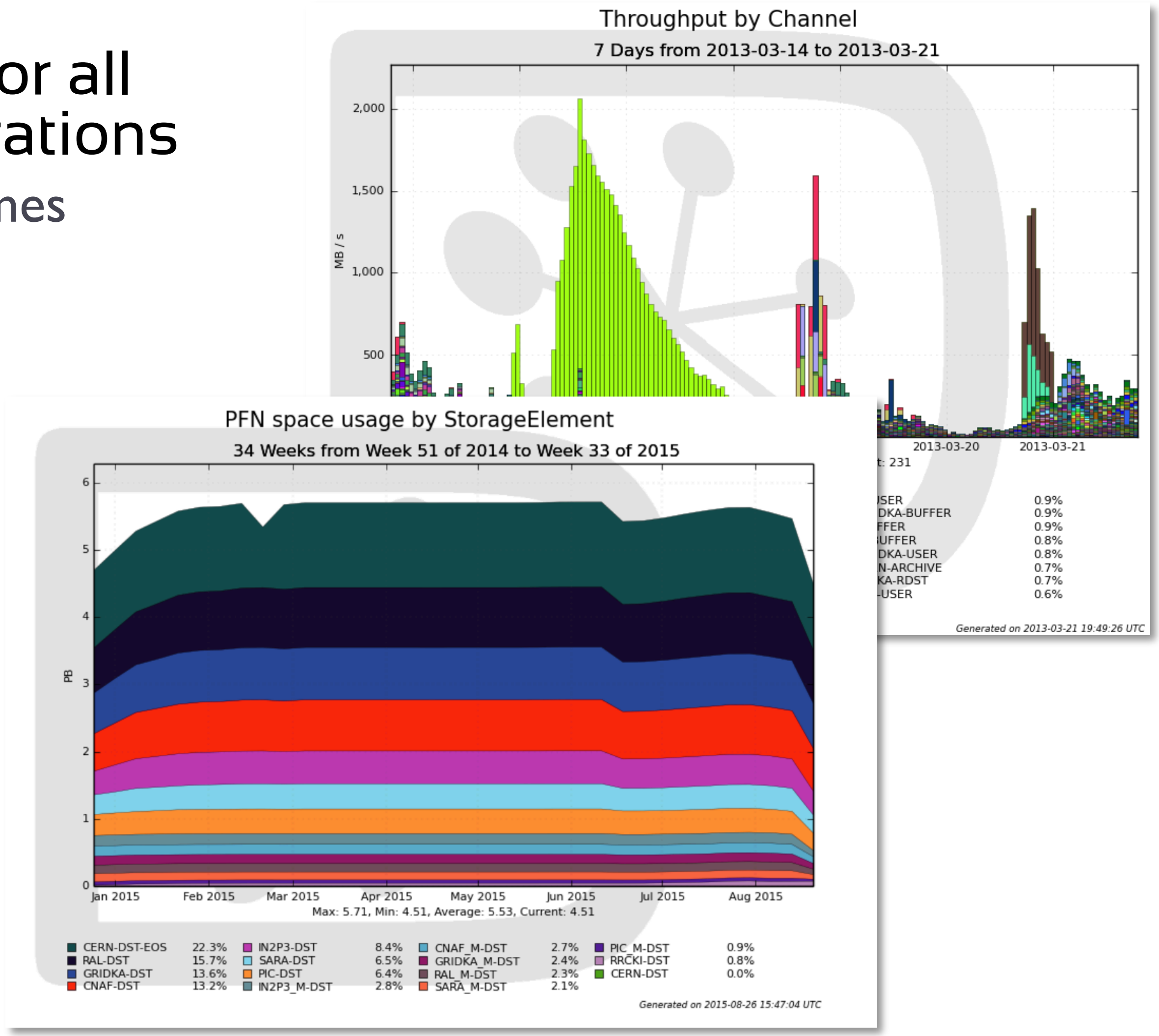
Query result

FSDIRAC

- DIRAC extension to mount the DIRAC File System as a local one
- Using FUSE, fuse-python
- Needs X509 credentials to browse data
- Files can be looked up in a standard file browser on Mac, Linux
- Simple if mounted read-only
- Not simple if mounted with write access enabled
  - Dealing with multiple replicas

Accounting

- DIRAC includes a general purpose Accounting System
  - For all the activities
- Accounting reports for all the data related operations
  - Transfer rates and volumes
  - Storage usage
  - Success/failure rates
  - Etc
- Plots selectable by
  - Storage Elements
  - Transfer channels
  - Owner of the data
  - Dates
  - Etc, etc



Throughput by Channel

7 Days from 2013-03-14 to 2013-03-21

Generated on 2013-03-21 19:49:26 UTC

PFN space usage by StorageElement

34 Weeks from Week 51 of 2014 to Week 33 of 2015

Generated on 2015-08-26 15:47:04 UTC

Other services

- Data Logging service
  - Each operation on a chosen subset of name space changing the status of the file is recorded
    - Storage, identity of the operation initiator, status, etc
  - Useful in debugging problems with the data flows
- Data Integrity service
  - Each file access problem can be reported and accumulated in the Data Integrity database
  - Problem resolution either automatically or manually
- Data Consistency service
  - Comparison of catalog and storage contents
    - Usually needs a dump of the storage name space
- Staging service
  - Bringing data online before job submission
    - Asynchronous staging requests with polling for progress status
    - Pin time management

DIRAC Framework

- DIRAC has a well defined architecture and development framework
  - Emphasis on modularity for easy extension
  - Standard rules to create DIRAC extension
    - LHCbDIRAC, BESDIRAC, ILCDIRAC, …
- Examples
  - Support for datasets first added to the BESDIRAC
  - LHCb has a custom Directory Tree module in the DIRAC File Catalog
- Allows to customize the DIRAC functionality for a particular application with minimal effort

Conclusions

- DIRAC combines various distributed computing and storage resources in a coherent system seen by the user as a single large computer

- The Data Management Model of DIRAC is organizing storage resources in a large distributed logical File System optimized for massive operations with data

- The DMS is built in the DIRAC general framework and is naturally linked to other services – Workload Management, Transformation, etc

- DIRAC DMS is extensible due to its modular architecture and can be easily adapted to the needs of particular applications