Distributed Computing Framework

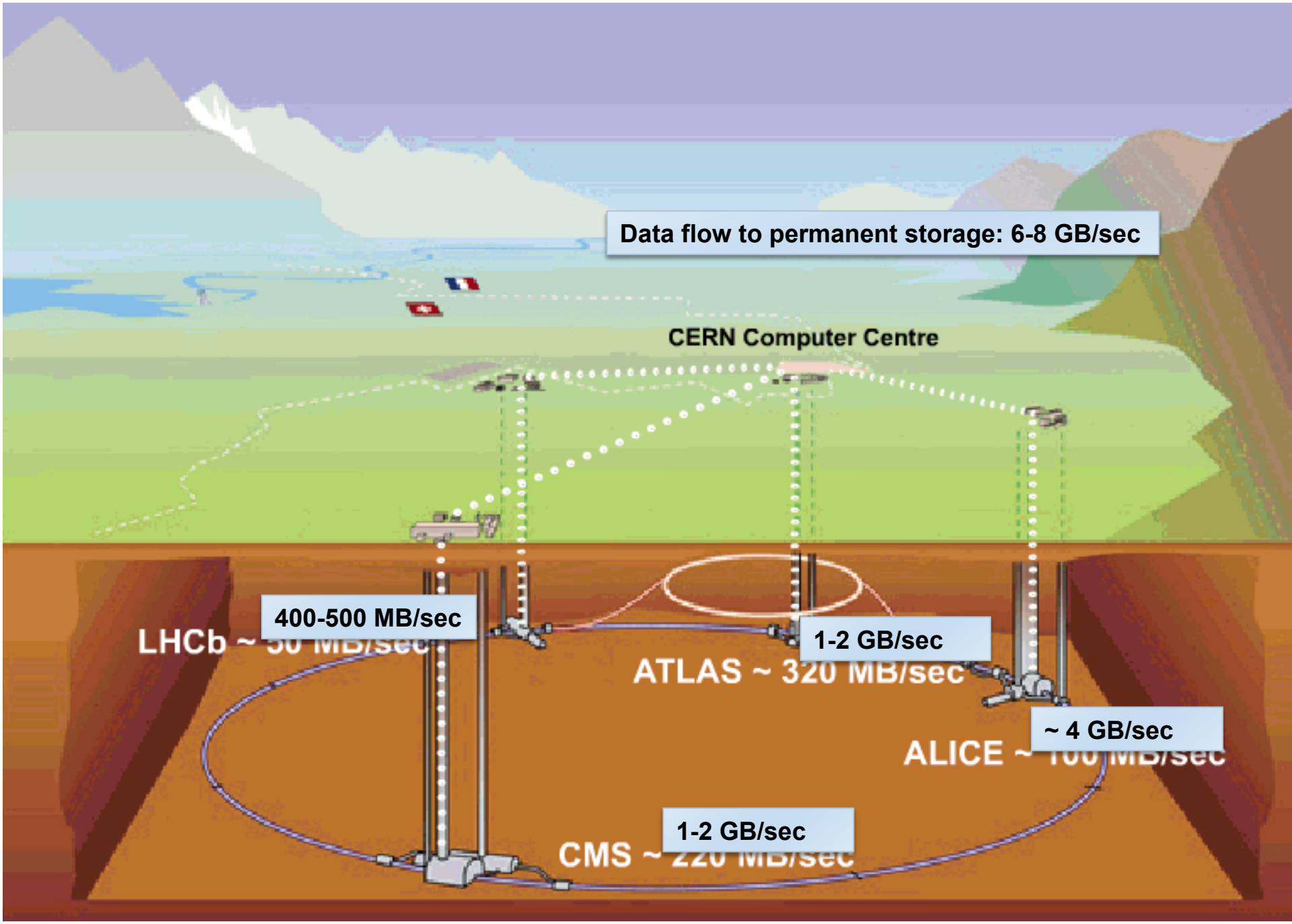
A. Tsaregorodtsev,
CPPM-IN2P3-CNRS, Marseille

EGI Webinar, 7 June 2016

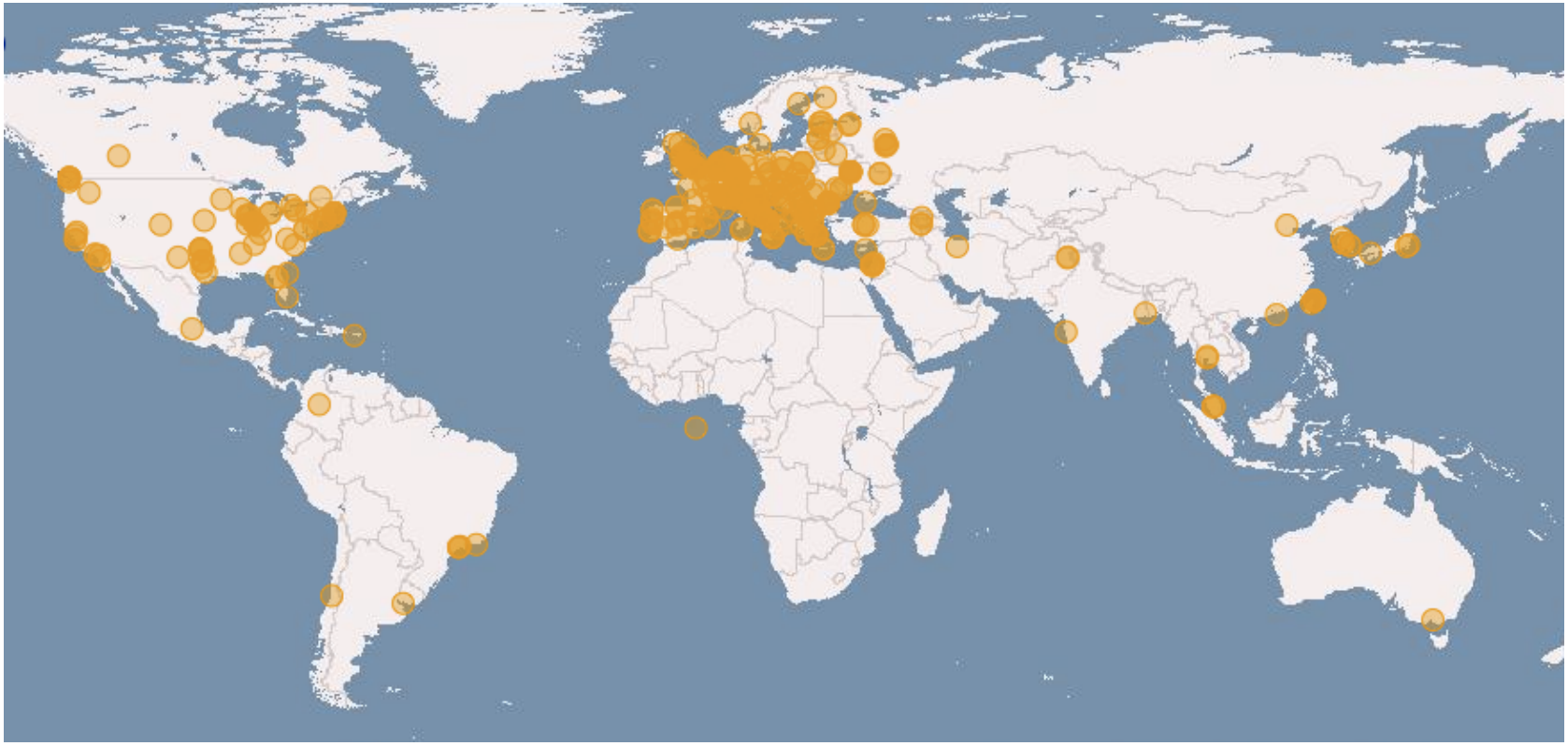
DIRAC
THE INTERWARE

Plan

- DIRAC Project
  - Origins
  - Agent based Workload Management System
  - Accessible computing resources
  - Data Management
  - Interfaces
- DIRAC users
- DIRAC as a service
- Conclusions


Data flow to permanent storage: 6-8 GB/sec
CERN Computer Centre
400-500 MB/sec
LHCb ~
1-2 GB/sec
ATLAS ~ 320 MB/sec
~ 4 GB/sec
ALICE ~
1-2 GB/sec
CMS ~

Worldwide LHC Computing Grid Collaboration (WLCG)



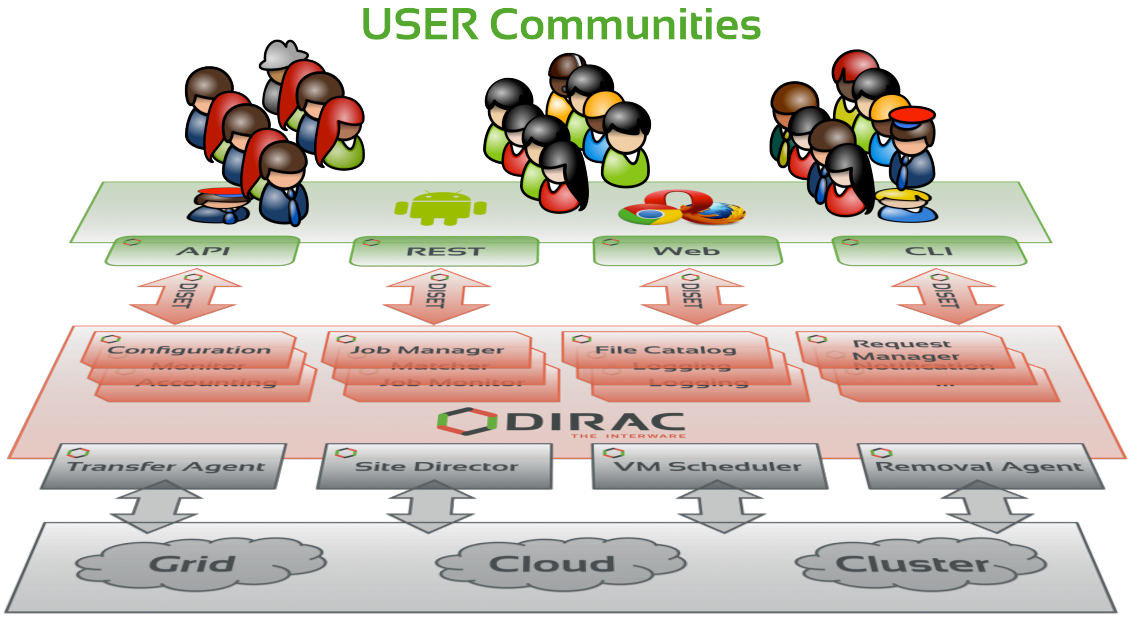
- >100 PB of data at CERN and major computing centers
- Distributed infrastructure of 150 computing centers in 40 countries
- 300+ k CPU cores (~ 2M HEP-SPEC-06)
- The biggest site with ~50k CPU cores, 12 T2 with 2-30k CPU cores
- Distributed data, services and operation infrastructure

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, IHEP, KEK
- The results of this work allow to offer DIRAC as a general purpose distributed computing framework

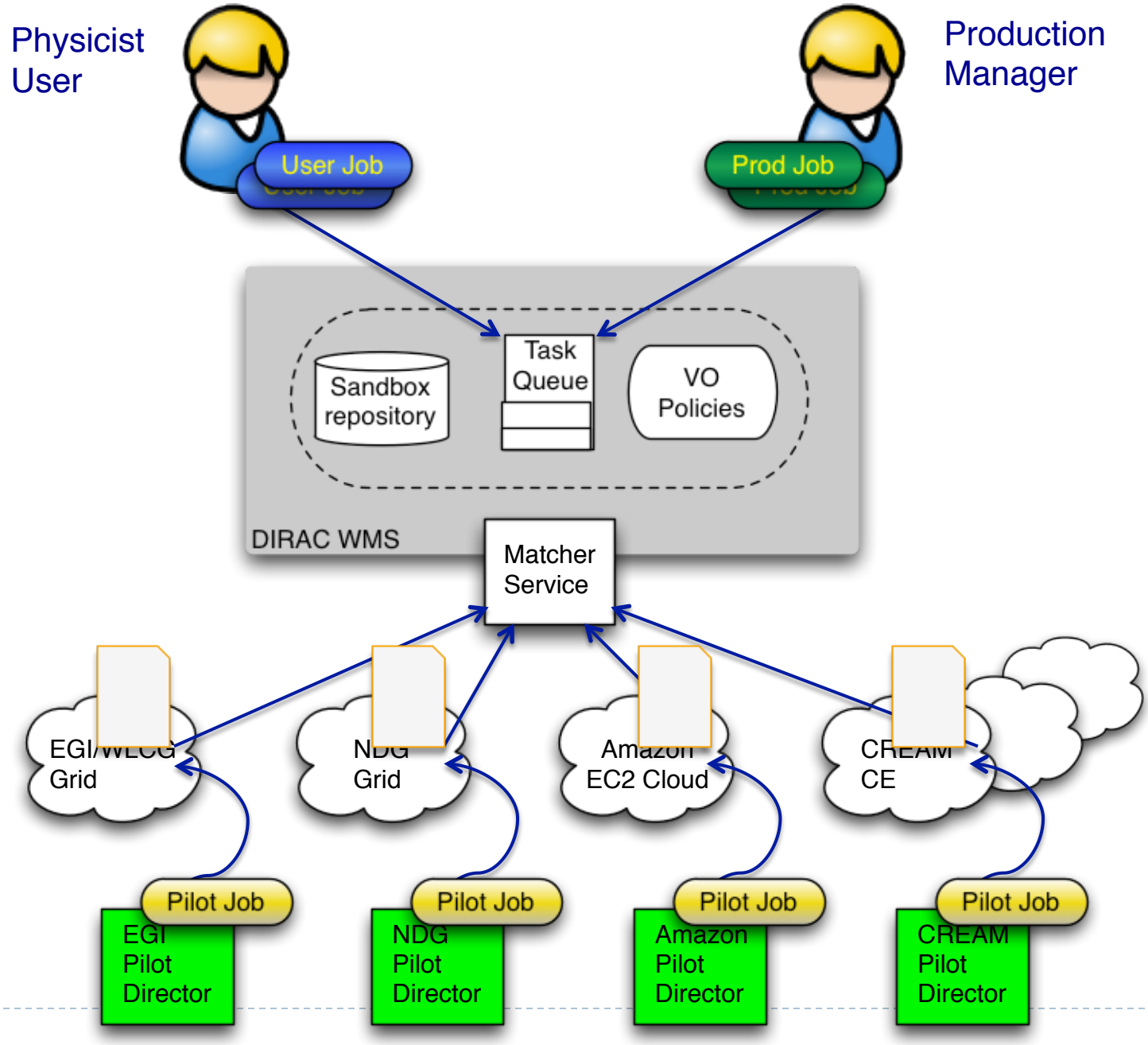
Interware

- 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

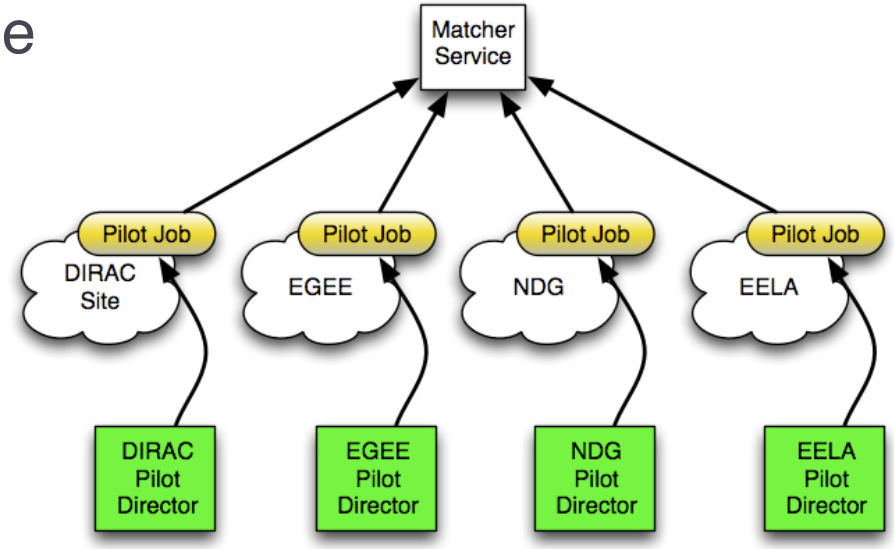
DIRAC Workload Management



DIRAC
THE INTERWARE

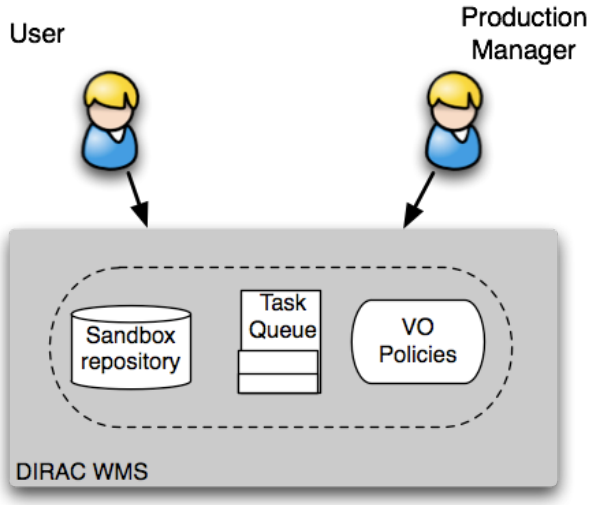
WMS: using heterogeneous resources

- Including resources in different grids and standalone clusters is simple with Pilot Jobs
  - Needs a specialized Pilot Director per resource type
  - Users just see new sites appearing in the job monitoring



WMS: applying VO policies

- In DIRAC both User and Production jobs are treated by the same WMS
  - Same Task Queue
- This allows to apply efficiently policies for the whole VO
  - Assigning Job Priorities for different groups and activities
  - Static group priorities are used currently
  - More powerful scheduler can be plugged in
    - demonstrated with MAUI scheduler



- Users perceive the DIRAC WMS as a single large batch system

DIRAC
THE INTERWARE

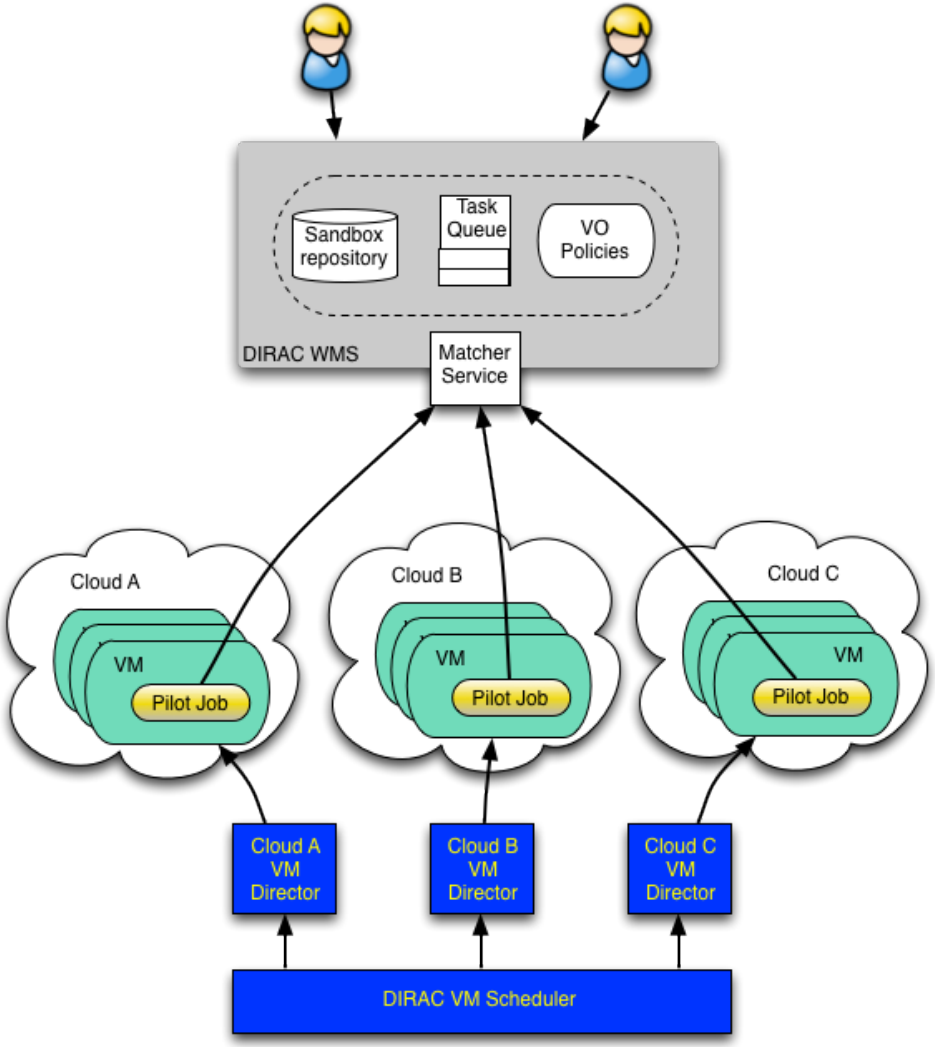
DIRAC computing resources

Computing Grids

- DIRAC was initially developed with the focus on accessing conventional Grid computing resources
  - WLCG grid resources for the LHCb Collaboration
- It fully supports gLite middleware based grids
  - European Grid Infrastructure (EGI), Latin America GISELA, etc
    - Using gLite/EMI middleware
  - Northern American Open Science Grid (OSG)
    - Using VDT middleware
  - Northern European Grid (NDGF)
    - Using ARC middleware
- Other types of grids can be supported
  - As long we have customers needing that

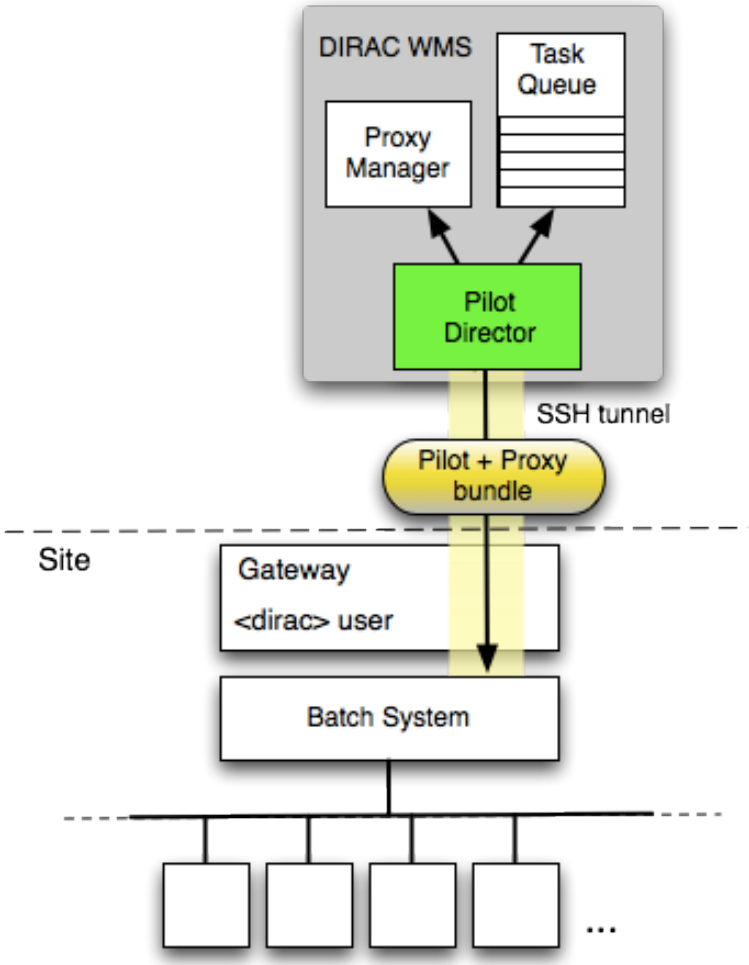
Clouds

- VM scheduler developed for Belle MC production system
  - Dynamic VM spawning taking Amazon EC2 spot prices and Task Queue state into account
  - Discarding VMs automatically when no more needed
- The DIRAC VM scheduler by means of dedicated VM Directors is interfaced to
  - OCCI compliant clouds:
    - OpenStack, OpenNebula
  - CloudStack
  - Amazon EC2



Standalone computing clusters

- Off-site Pilot Director
  - Site delegates control to the central service
  - Site must only define a dedicated local user account
  - The payload submission through the SSH tunnel
- The site can be a single computer or a cluster with a batch system
  - LSF, BQS, SGE, PBS/Torque, Condor, OAR, SLURM
    - HPC centers
  - More to come:
    - LoadLeveler. etc
- The user payload is executed with the owner credentials
  - No security compromises with respect to external services



DIRAC
THE INTERWARE

Data Management

DM Problem to solve

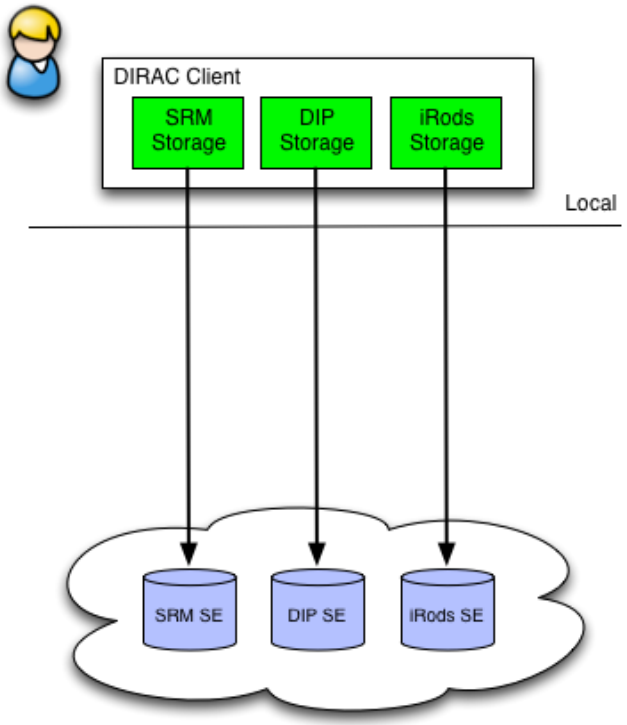
- Data is partitioned in files
- File replicas are distributed over a number of Storage Elements world wide

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

- Need for transparent file access for users
- Often working with multiple (tens of thousands) files at a time
  - Make sure that ALL the elementary operations are accomplished
  - Automate recurrent operations

Storage plugins

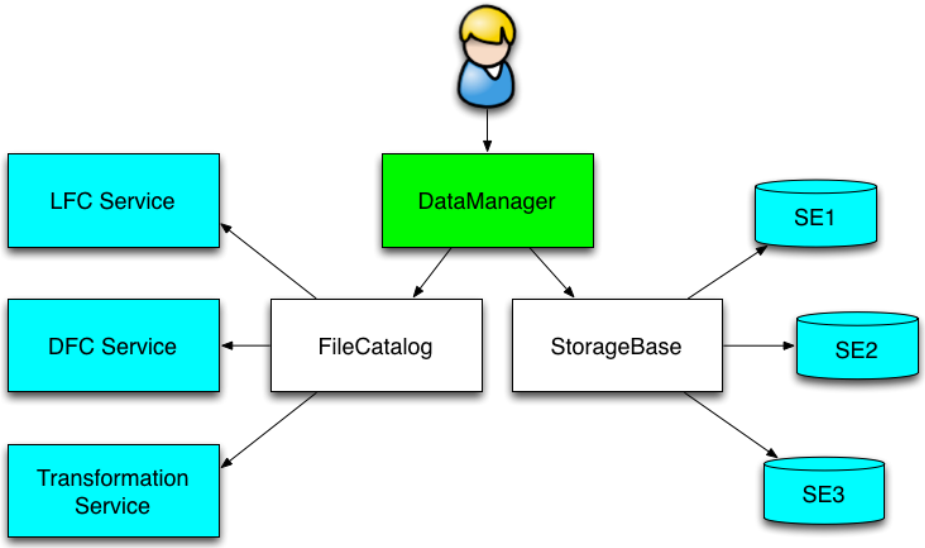
- Storage element abstraction with a client implementation for each access protocol
  - DIPS, SRM, XROOTD, RFIO, etc
  - gfal2 based plugin gives access to all protocols supported by the library
    - DCAP, WebDAV, S3, http, …
    - iRODS



- Each SE is seen by the clients as a logical entity
  - With some specific operational properties
  - SE's can be configured with multiple protocols

File Catalog

- Central File Catalog (DFC, LFC, …)
  - Keeps track of all the physical file replicas
- Several catalogs can be used together
  - The mechanism is used to send messages to "pseudocatalog" services, e.g.
    - Transformation service (see later)
    - Bookkeeping service of LHCb
  - A user sees it as a single catalog with additional features
- DataManager is a single client interface for logical data operations

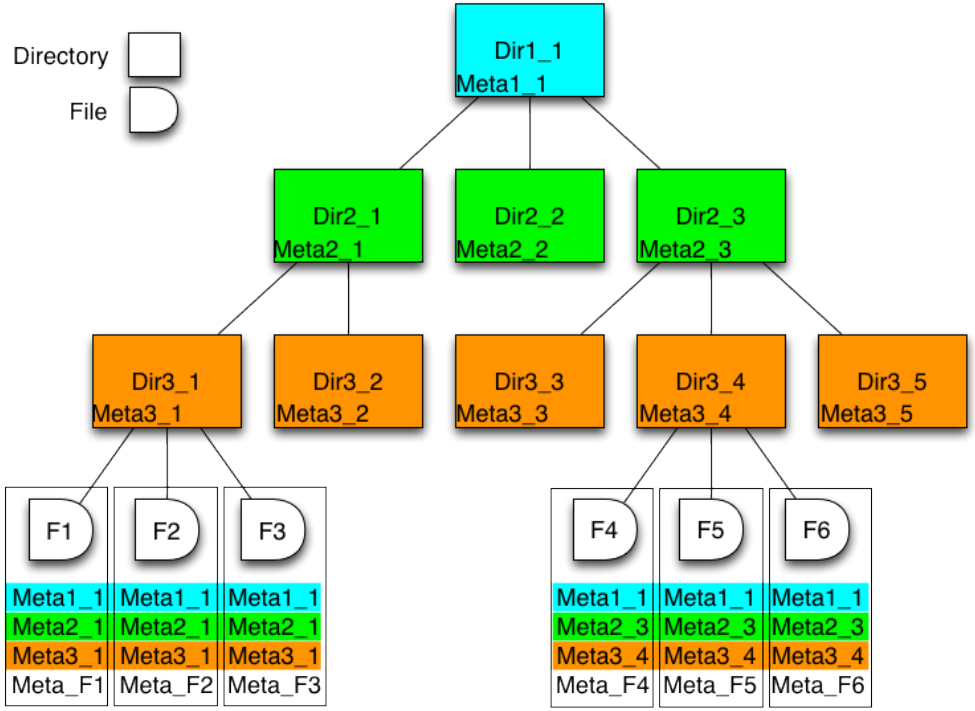


File Catalog

- DFC is the central component of the DIRAC Data Management system
- Defines a single logical name space for all the data managed by DIRAC
- Together with the data access components DFC allows to present data to users as single global file system
  - User ACLs
  - Rich metadata including user defined metadata

File Catalog: Metadata

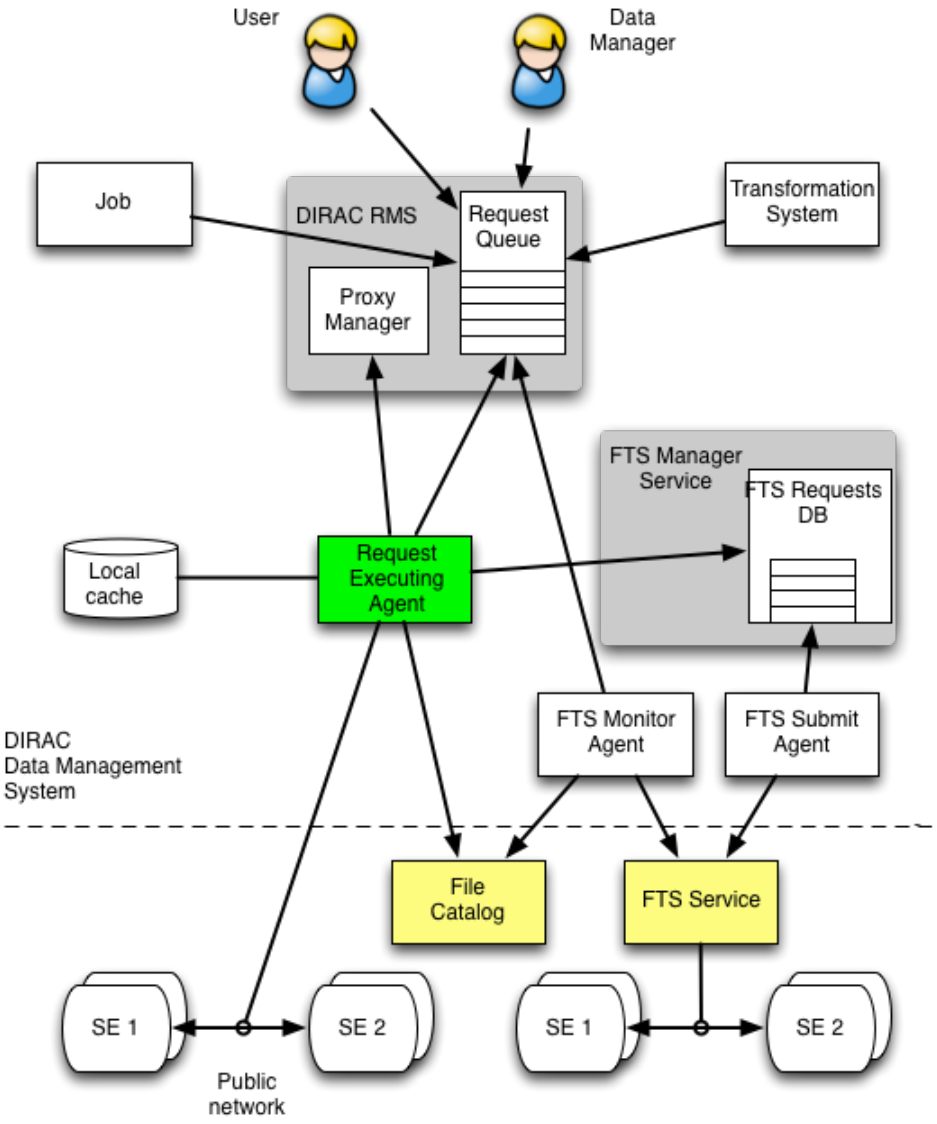
- DFC is Replica and Metadata Catalog
  - User defined metadata
  - The same hierarchy for metadata as for the logical name space
    - Metadata associated with files and directories
    - Allow for efficient searches
  - Efficient Storage Usage reports
    - Suitable for user quotas



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

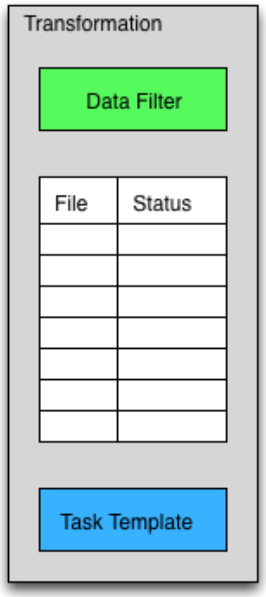
Bulk data transfers

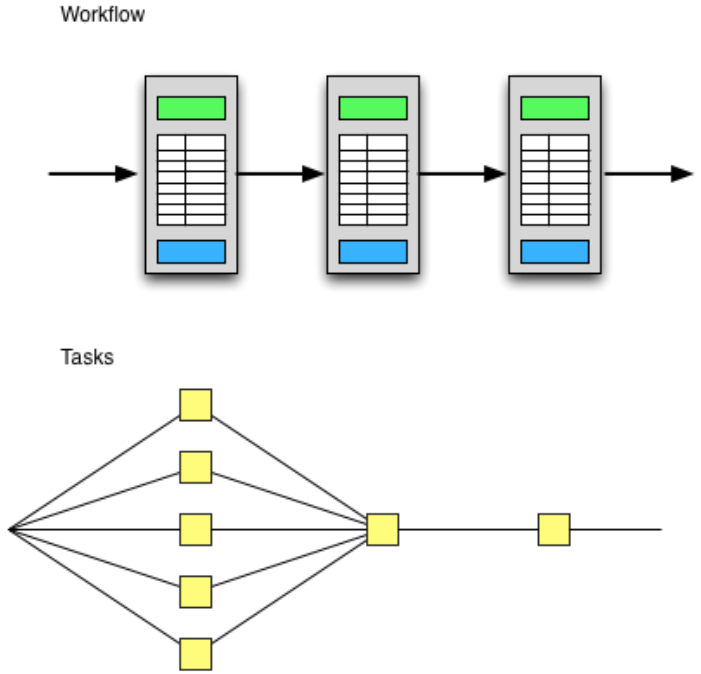
- Replication/Removal Requests with multiple files are stored in the RMS
  - By users, data managers, Transformation System
- The Replication Operation executor
  - Performs the replication itself or
  - Delegates replication to an external service
    - E.g. FTS
  - A dedicated FTSManager service keeps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



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





DIRAC
THE INTERWARE

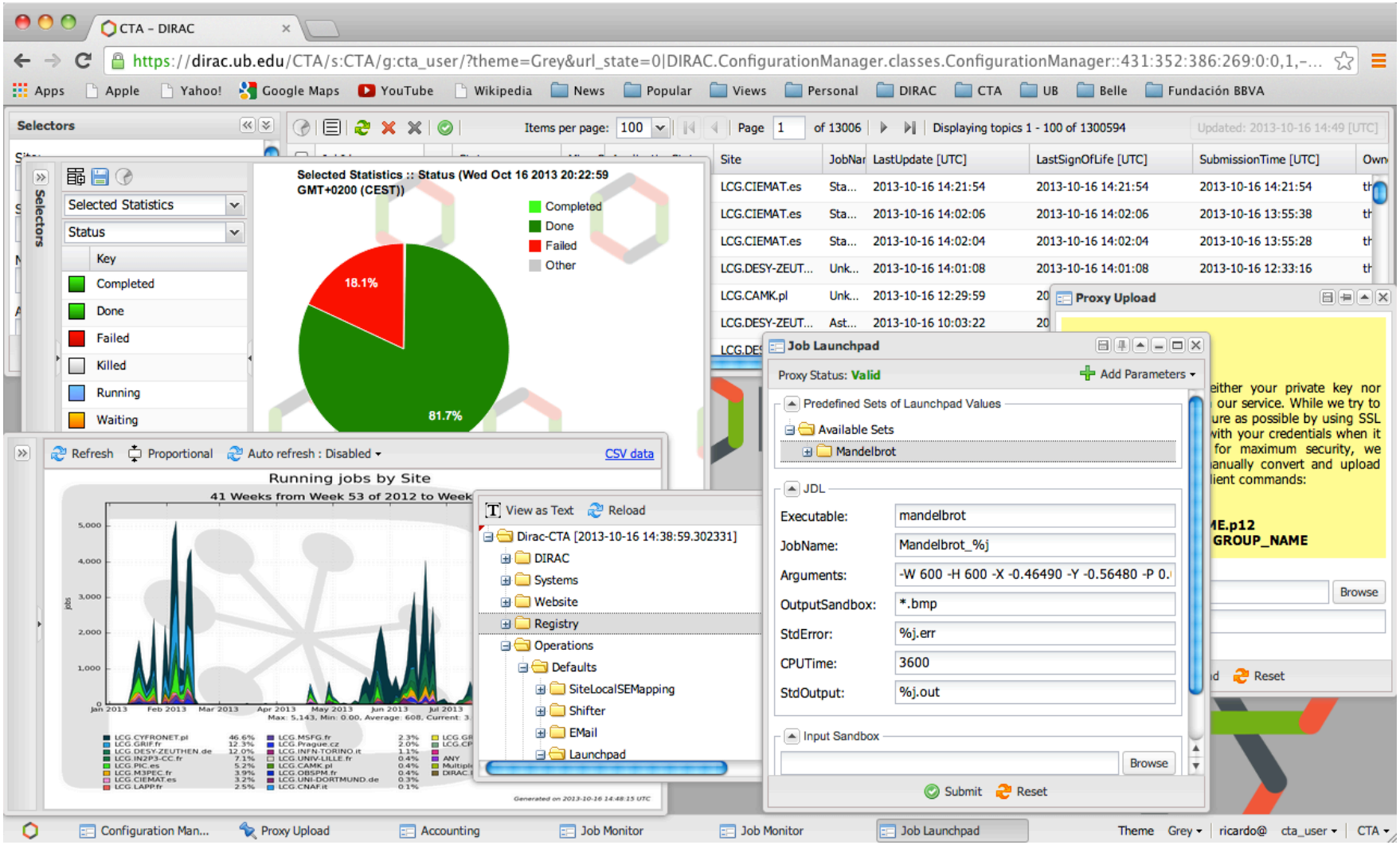
Interfaces

DM 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
- REST interface
  - Suitable for use with application portals
  - WS-PGRADE portal is interfaced with DIRAC this way

DIRAC
THE INTERWARE

Web Portal



DIRAC
THE INTERWARE

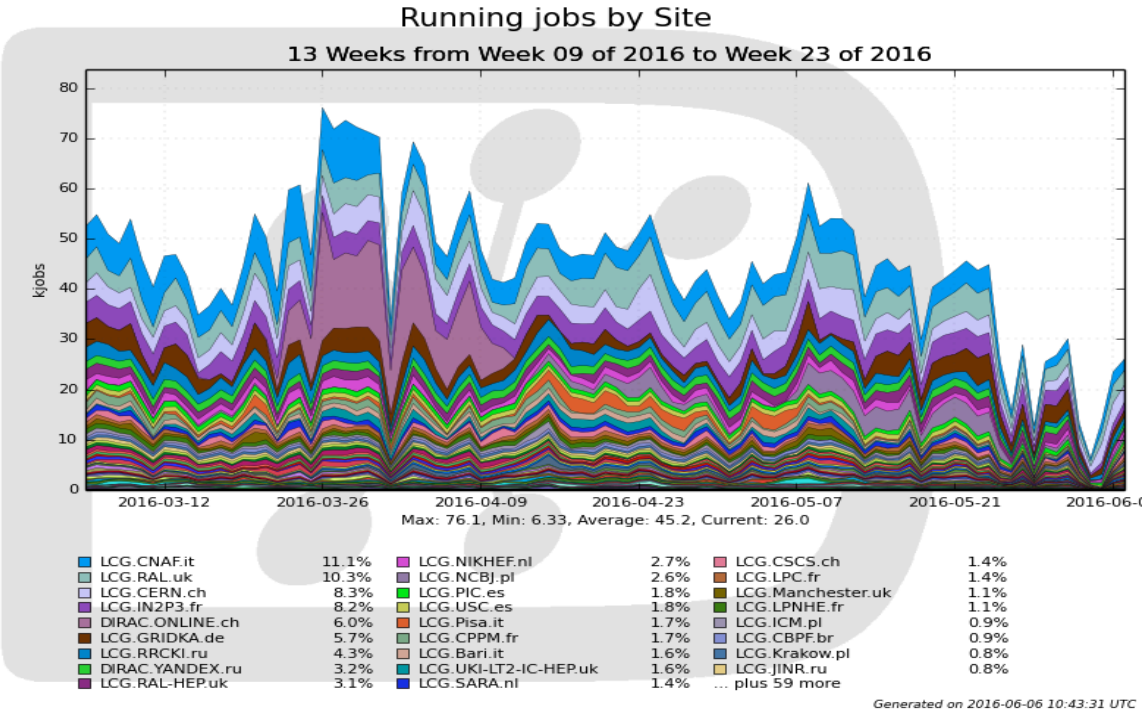
Distributed Computer

- DIRAC is aiming at providing an abstraction of a single computer for massive computational and data operations from the user perspective
  - Logical Computing and Storage elements (Hardware)
  - Global logical name space (File System)
  - Desktop-like GUI

DIRAC
THE INTERWARE

DIRAC Users

LHCb Collaboration



- Up to 100K concurrent jobs in ~120 distinct sites
  - Equivalent to running a virtual computing center with a power of 100K CPU cores
- Further optimizations to increase the capacity are possible
  - Hardware, database optimizations, service load balancing, etc

Community installations

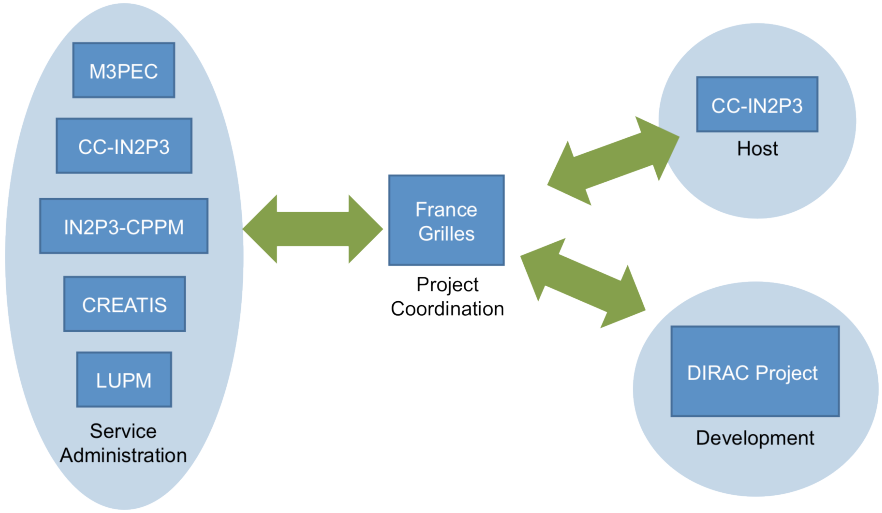
- Belle II Collaboration, KEK
  - First use of clouds (Amazon) for data production
- ILC/CLIC detector Collaboration, Calice VO
  - Dedicated installation at CERN, 10 servers, DB-OD MySQL server
  - MC simulations
  - DIRAC File Catalog was developed to meet the ILC/CLIC requirements
- BES III, IHEP, China
  - Using DIRAC DMS: File Replica and Metadata Catalog, Transfer services
  - Dataset management developed for the needs of BES III
- CTA
  - CTA started as France-Grilles DIRAC service customer
  - Now is using a dedicated installation at PIC, Barcelona
  - Using complex workflows
- Geant4
  - Dedicated installation at CERN
  - Validation of MC simulation software releases
- DIRAC evaluations by other experiments
  - LSST, Auger, TREND, Daya Bay, Juno, ELI, NICA, …
  - Evaluations can be done with general purpose DIRAC services

National services

- **DIRAC** services are provided by several National Grid Initiatives: France, Spain, Italy, UK, China, …
  - Support for small communities
  - Heavily used for training and evaluation purposes

- Example: France-Grilles DIRAC service
  - Hosted by the CC/IN2P3, Lyon
  - Distributed administrator team
    - 5 participating universities
  - 15 VOs, ~100 registered users
  - In production since May 2012
    - >12M jobs executed in the last year
      - At ~90 distinct sites

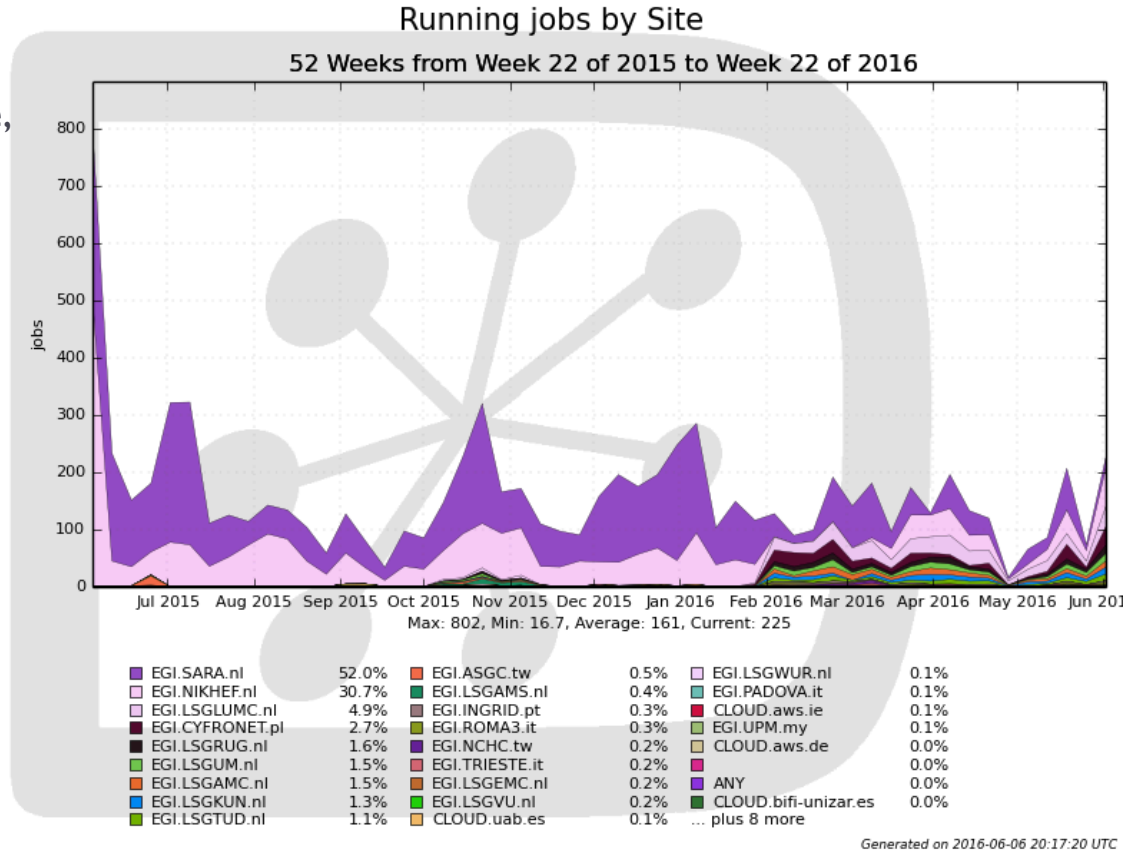
http://dirac.france-grilles.fr



DIRAC4EGI service

- In production since 2014
- Partners
  - Operated by EGI
  - Hosted by CYFRONET
  - DIRAC Project providing software, consultancy
- 10 Virtual Organizations
  - enmr.eu
  - vlemed
  - eiscat.se
  - fedcloud.egi.eu
  - training.egi.eu
  - …
- Usage
  - > 6 million jobs processed in the last year

DIRAC4EGI activity snapshot



DIRAC
THE INTERWARE

DIRAC Framework

DIRAC Framework

- DIRAC has a well defined architecture and development framework
  - Standard rules to create DIRAC extension
    - LHCbDIRAC, BESDIRAC, ILCDIRAC, …
- Large part of the functionality is implemented as plugins
  - Almost the whole DFC service is implemented as a collection of plugins
- 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

- Computational grids and clouds are no more something exotic, they are used in a daily work for various applications
- Agent based workload management architecture allows to seamlessly integrate different kinds of grids, clouds and other computing resources
- DIRAC is providing a framework for building distributed computing systems and a rich set of ready to use services. This is used now in a number of DIRAC service projects on a regional and national levels
- Services based on DIRAC technologies can help users to get started in the world of distributed computations and reveal its full potential

http://diracgrid.org

DIRAC
THE INTERWARE

Demo

DIRAC
THE INTERWARE

Demo

- Using EGI sites and storage elements
  - Grid sites
  - Fed Cloud sites
  - DIRAC Storage Elements
- Web Portal
- Command line tools
- Demo materials to try out off-line can be found here
  - https://github.com/DIRACGrid/DIRAC/wiki/Quick-DIRAC-Tutorial