DIRAC Distributed Computing Services

A. Tsaregorodtsev,
CPPM-IN2P3-CNRS

EGI Webinar, 22 March 2013

DIRAC
THE INTERWARE

Plan

- Motivation and brief history of the project
- DIRAC grid middleware
- Resources available to DIRAC users
- Communities using DIRAC
- DIRAC as a Service
- Conclusions

DIRAC
THE INTERWARE

Massive computations: HEP

- LHC experiments pioneered the massive use of computational grids
  - 10s of PBytes of data per year
  - 100s of thousands CPUs in 100s of centers
  - 100s of users from 100s of institutions
- CERN Director General Rolf Heuer about the Higgs discovery:

  "It was a global effort and it is a global success. The results today are only possible because of the extraordinary performance of the accelerators, including the infrastructure, the experiments, and the *Grid computing*."
- Other domains are catching up quickly with the HEP experiments
  - Life sciences, earth sciences, astrophysics, social sciences, etc

DIRAC
THE INTERWARE

Massive computations

- Large HEP experiments have dedicated teams of experts to build their computing systems
  - Largely relying on dedicated grid resources
- The computing expertise level in other scientific domains is relatively lower
  - Grouped around well known applications and scientific portals
  - New application development to run on the grid is still difficult
- Need for convenient tools for small research groups with no local grid gurus.
- The experience of the HEP experiment developers can be very useful for the non-HEP users

DIRAC
THE INTERWARE

User problems on the grid

- Complicated interfaces
  - Especially for non-computing experts
- Frustration with failing resources and middleware
  - Why my jobs worked yesterday and not today ?
- For small communities difficult to organize collective work
  - Lack of expertise in high level computing tasks
    - Massive jobs, massive data movement, etc
- Difficult to build custom services to orchestrate execution of particular applications
  - Example: workflow managers
- Small communities tend to become larger with time

Large community issues

- Large user communities (Virtual Organizations) have specific problems
  - Dealing with heterogeneous resources
    - Various computing clusters, grids, etc
  - Dealing with the intracommunity workload management
    - User group quotas and priorities
    - Priorities of different activities
  - Dealing with a variety of applications
    - Massive data productions
    - Individual user applications, etc
- HEP Experiments are typical examples

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 with the goals:
  - Integrate all the heterogeneous computing resources available to the community
  - Provide solution for both WMS and DMS tasks
  - Minimize human intervention at sites providers of resources
  - Make the grid convenient for the users:
    - Simpler intuitive interfaces
    - Fault tolerance, quicker turnaround of user jobs
    - Enabling Community policies

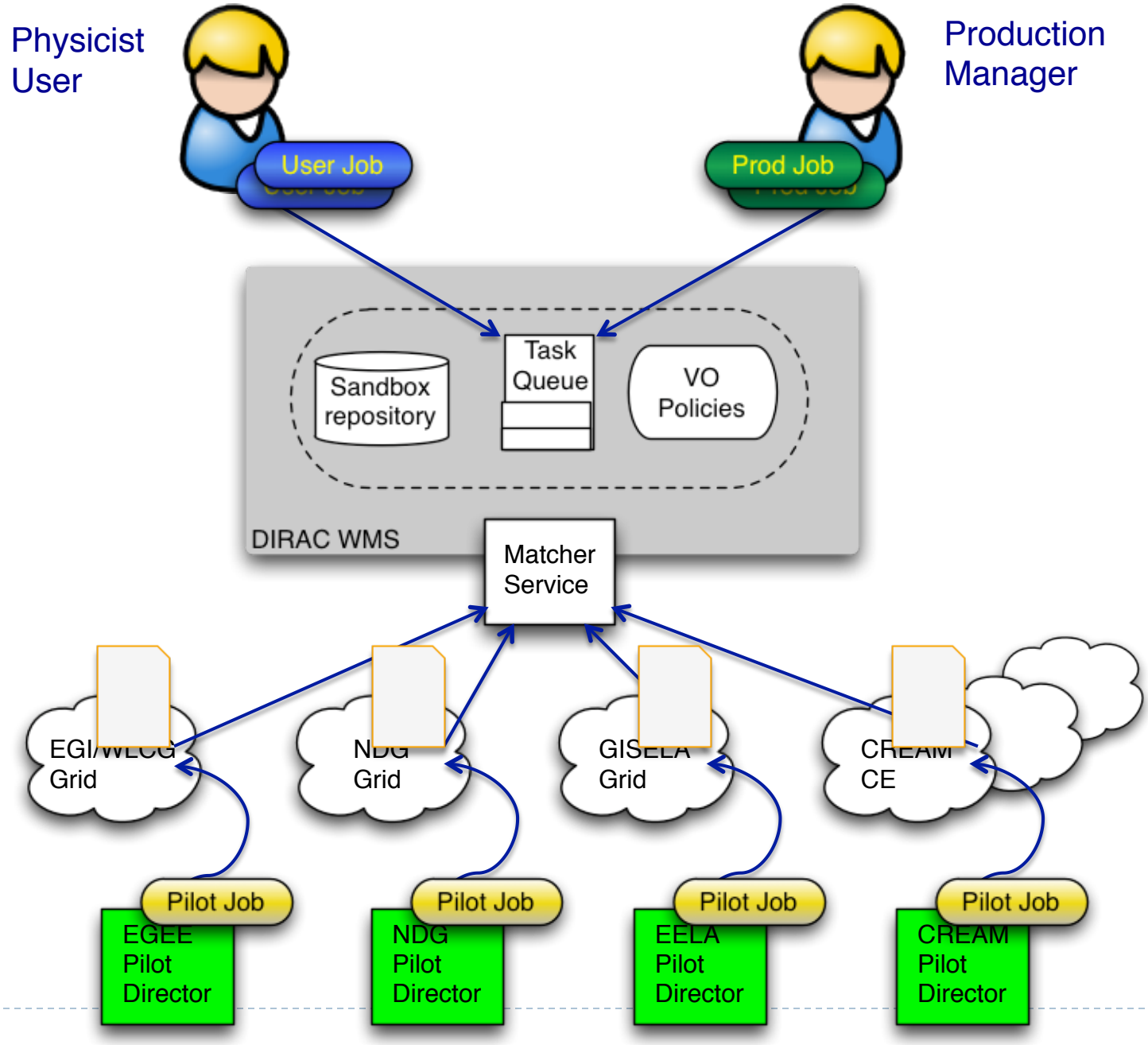
DIRAC
THE INTERWARE

Towards general purpose middleware

- 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.
  - Separate LHCb specific functionality into a set of extensions to the generic core libraries
  - Introduce new services to make it a complete solution
  - Support for multiple small groups by a single DIRAC installation
  - General refurbishing of the code, code management, deployment, documentation, etc
- The results of this work are presented in the following

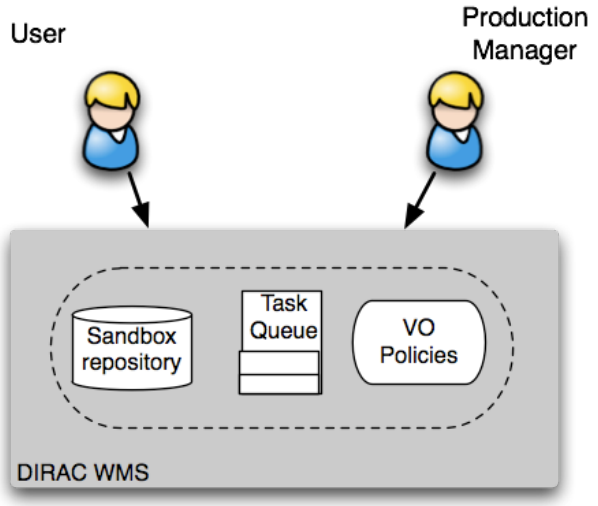
DIRAC
THE INTERWARE

DIRAC Workload Management


Physicist User
Production Manager
User Job
Prod Job
Sandbox repository
Task Queue
VO Policies
DIRAC WMS
Matcher Service
Grid
Grid
Grid
CE
Pilot Job
Pilot Job
Pilot Job
Pilot Job
EGEE Pilot Director
NDG Pilot Director
EELA Pilot Director
CREAM Pilot Director

WMS: applying VO policies

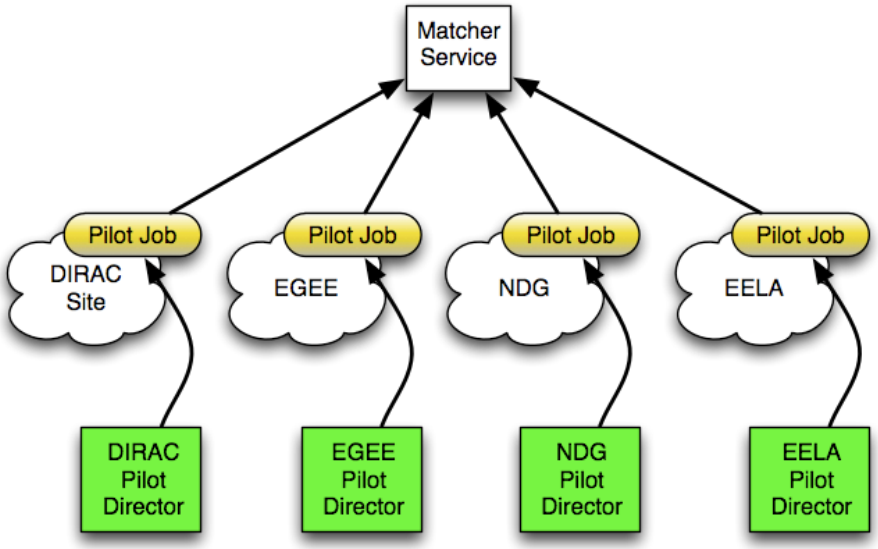
- In DIRAC both User and Production jobs are treated by the same WMS
  - Same Task Queue
- This allows to apply efficiently policies for the whole VO
  - Assigning Job Priorities for different groups and activities
  - Static group priorities are used currently
  - More powerful scheduler can be plugged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



- The VO policies application in the central Task Queue dictates the use of Multiuser Pilot Agents
  - Do not know apriori whose job has the highest priority at the moment of the user job matching
- DIRAC fully supports this mode of operation
  - Multiuser Pilots Jobs submitted with a special “pilot” VOMS role
  - Using glexec on the WNs to track the identity of the payload owner

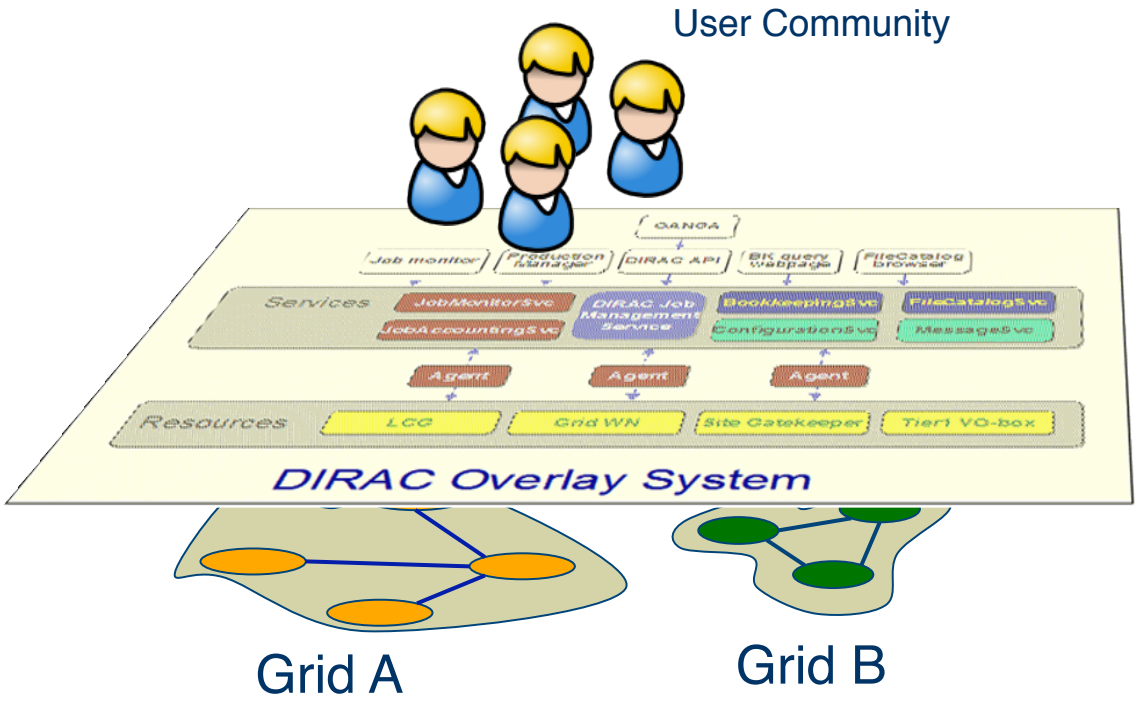
WMS: using heterogeneous resources

- 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



Interware

- DIRAC has all the necessary components to build ad-hoc grid infrastructures interconnecting computing resources of different types. This allows to speak about the DIRAC ***interware***.



DIRAC as a resource manager

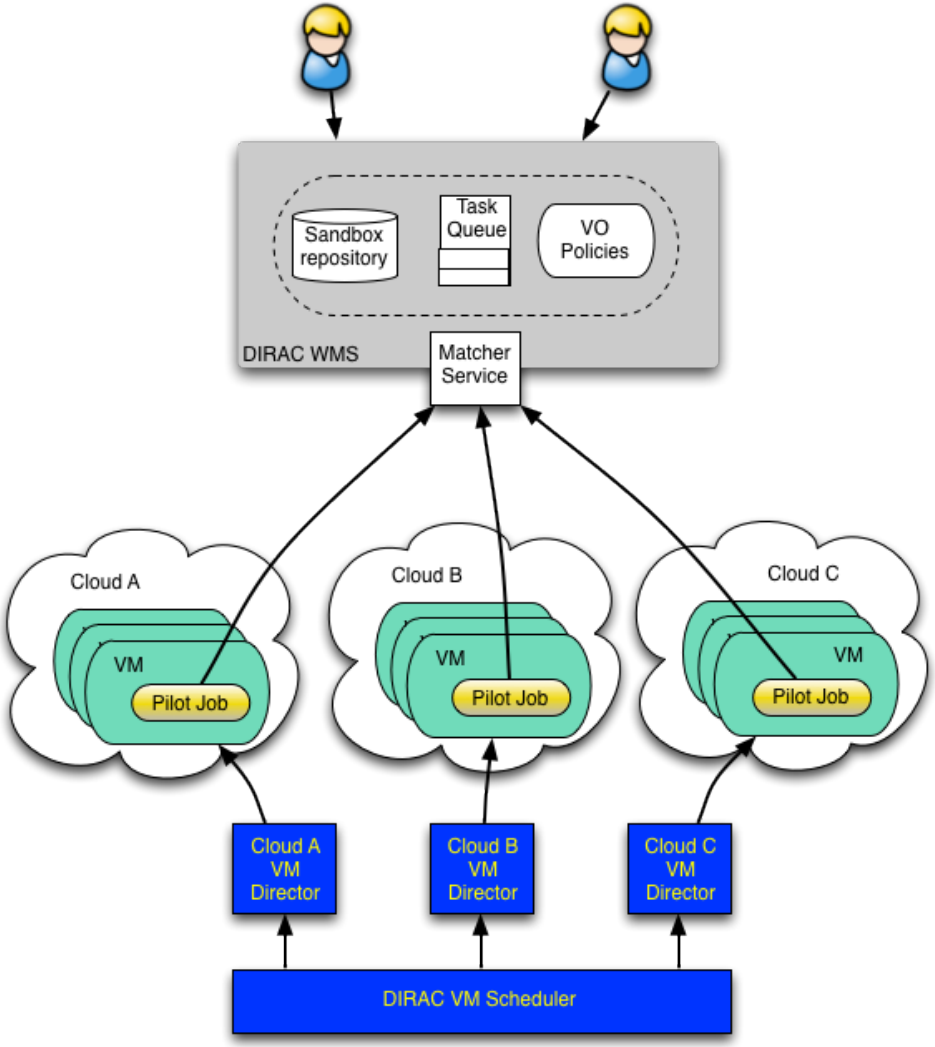
DIRAC
THE INTERWARE

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
  - EGI, GISELA, etc
    - Using gLite WMS or accessing CE's directly
  - OSG
- The work is in progress to support ARC middleware based grids
  - NorduGrid
  - A successful demonstration was already done
- Other types of grids can be supported
  - As long we have customers needing that

Clouds

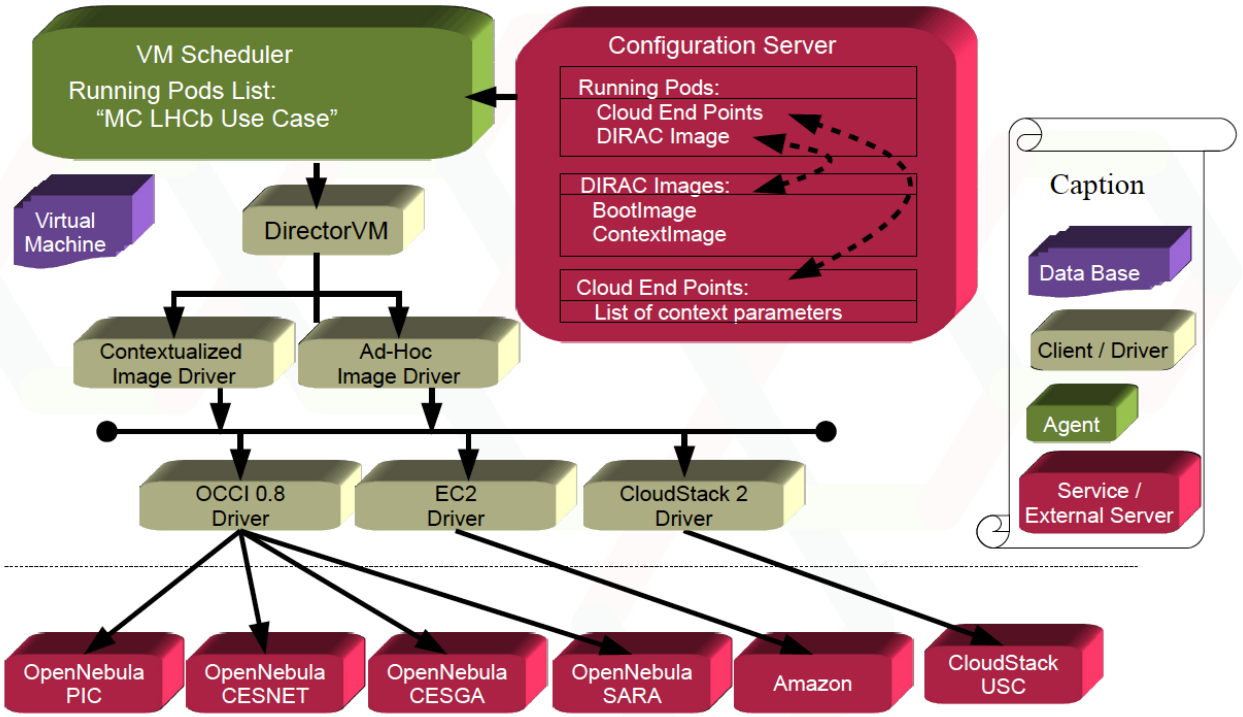
- VM scheduler developed for Belle MC production system
  - Dynamic VM spawning taking Amazon EC2 spot prices and Task Queue state into account
  - Discarding VMs automatically when no more needed
- The DIRAC VM scheduler by means of dedicated VM Directors is interfaced to
  - OCCI compliant clouds: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



Ongoing VMDIRAC work

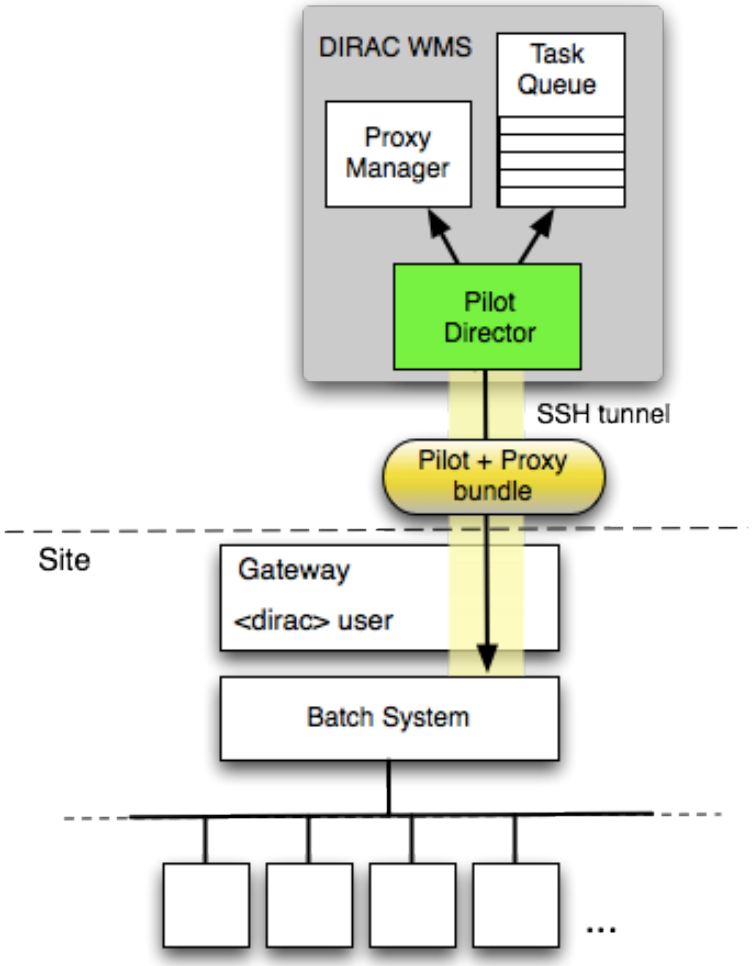
VMDIRAC – multiple cloud broker

- Gives a transparent access to multiple clouds with optimized dynamic allocation of Virtial Machines (VM)
- Intensive development now
  - different access methods, VM contextualization, VM scheduling policies
  - part of the EGI Cloud Task Force activities



Standalone computing clusters

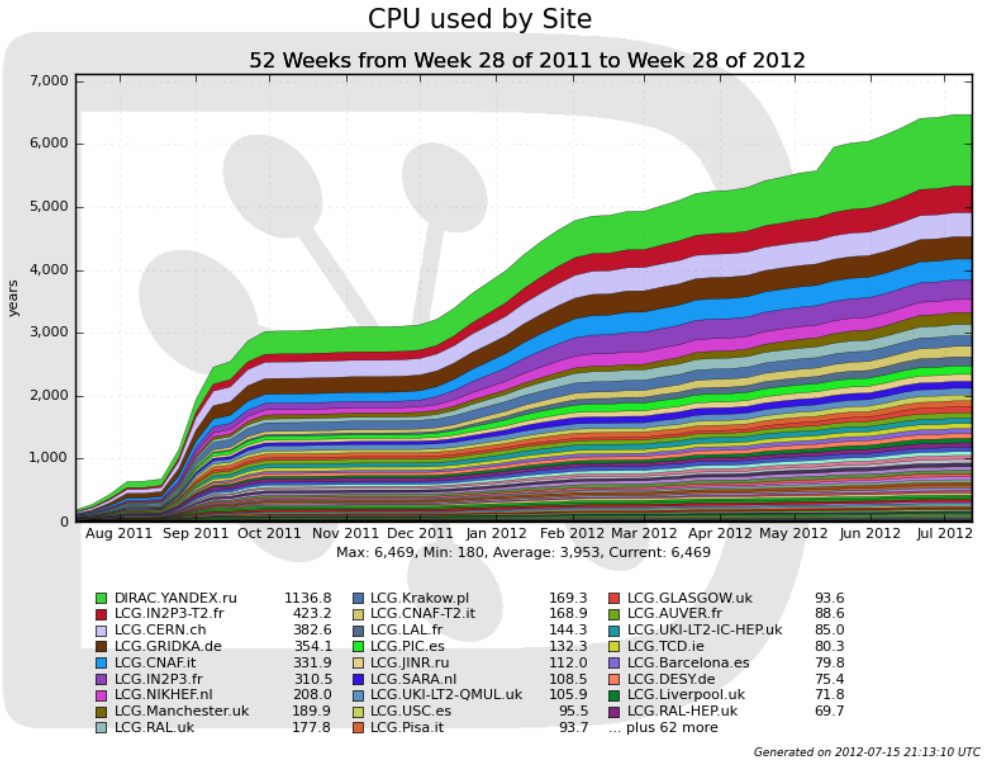
- Dedicated Pilot Director per group of sites
- Off-site Director
  - Site delegates control to the central service
  - Site must only define a dedicated local user account
  - The payload submission through the SSH tunnel
- The site can be a single computer or a cluster with a batch system
  - LSF,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



Standalone computing clusters

Examples:

- DIRAC.Yandex.ru
  - 1800 cores
  - Torque batch system, no grid middleware, access by SSH
  - Second largest LHCb MC production site
- LRZ Computing Center, Munich
  - SLURM batch system, GRAM5 CE service
  - Gateway access by GSISSH
  - Considerable resources for biomed community (work in progress)
- Mesocentre Aix-Marseille University
  - OAR batch system, no grid middleware, access by SSH
  - Open to multiple communities (work in progress)



Other types of resources

- Examples of other types of resources available via DIRAC
  - LHCb online filter farm
  - Volunteer grids based on BOINC+virtualization technology
    - EDGI resources
- DIRAC attempts to provide access to all kinds of existing resources and services useful for its users
- At the same time it provides its own solutions, e.g. catalogs, storage element, transfer service, etc.
- By design services from different providers can be used together in parallel not necessarily replacing one another
  - Example: use of DFC and LFC together, see below
- The final choice of the services to use is left to the user

DIRAC
THE INTERWARE

DIRAC Data Management

Data Management components

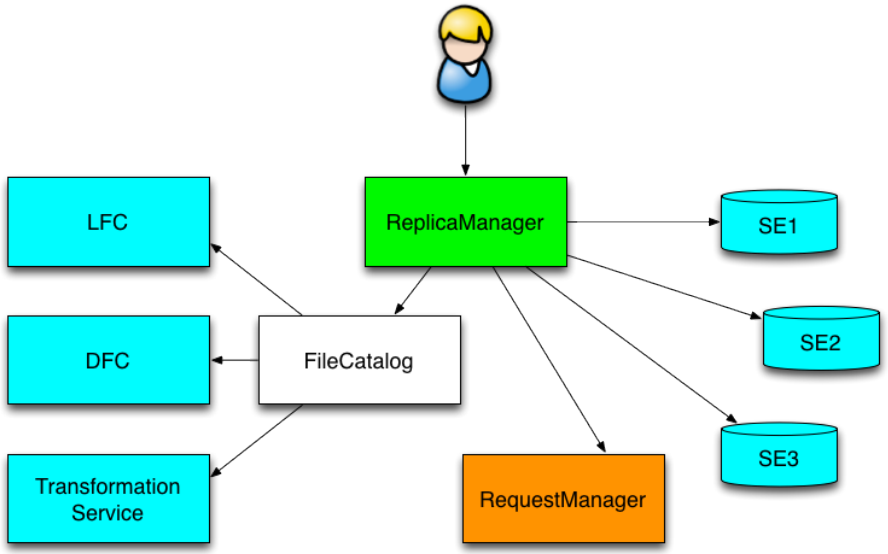
- Storage Elements
  - gLite/EGI Storage Elements
    - Standard SRM interface
    - Gridftp protocol
      - Need Globus libraries, limited number of platforms
    - Allow third party transfers between them
    - Managed by the site managers within EGI SLAs
  - DIRAC Storage Elements
    - DISET based components
    - DIPS (Dirac Secure Protocol)
    - Does not allow third party transfers
      - Replication through local cache
      - Third party transfers will be available in the future
  - More Storage Elements can be included
    - (F,SF,HT,BBF)TP servers

Data Management components

- File Catalogs
  - LCG File Catalog (LFC)
    - Part of the EGI middleware
    - Service provided by the NGI
      - ORACLE backend
    - Client tools: command line, Python API
      - Need Globus libraries
    - No User Metadata support
  - DIRAC File Catalog
    - DISET based components
    - Part of the DIRAC set of services
      - Community service
      - MySQL backend
    - Client tools: command line, CLI, Python API
    - Support of the User Metadata
      - Similar to AMGA metadata service
  - More Catalogs can be included
    - LHCb has developed several specific catalogs in the same framework
    - iRods ?

Data Management components

- For DIRAC users the use of any Storage Element or File Catalog is transparent
  - Up to a user community to choose components to use
  - Different SE types can be mixed together
  - Several File Catalogs can be used in parallel
    - Complementary functionality
    - Redundancy



- Users see depending on the DIRAC Configuration
  - Logical Storage Elements
    - e.g. DIRAC-USER, M3PEC-disk
  - Logical File Catalog

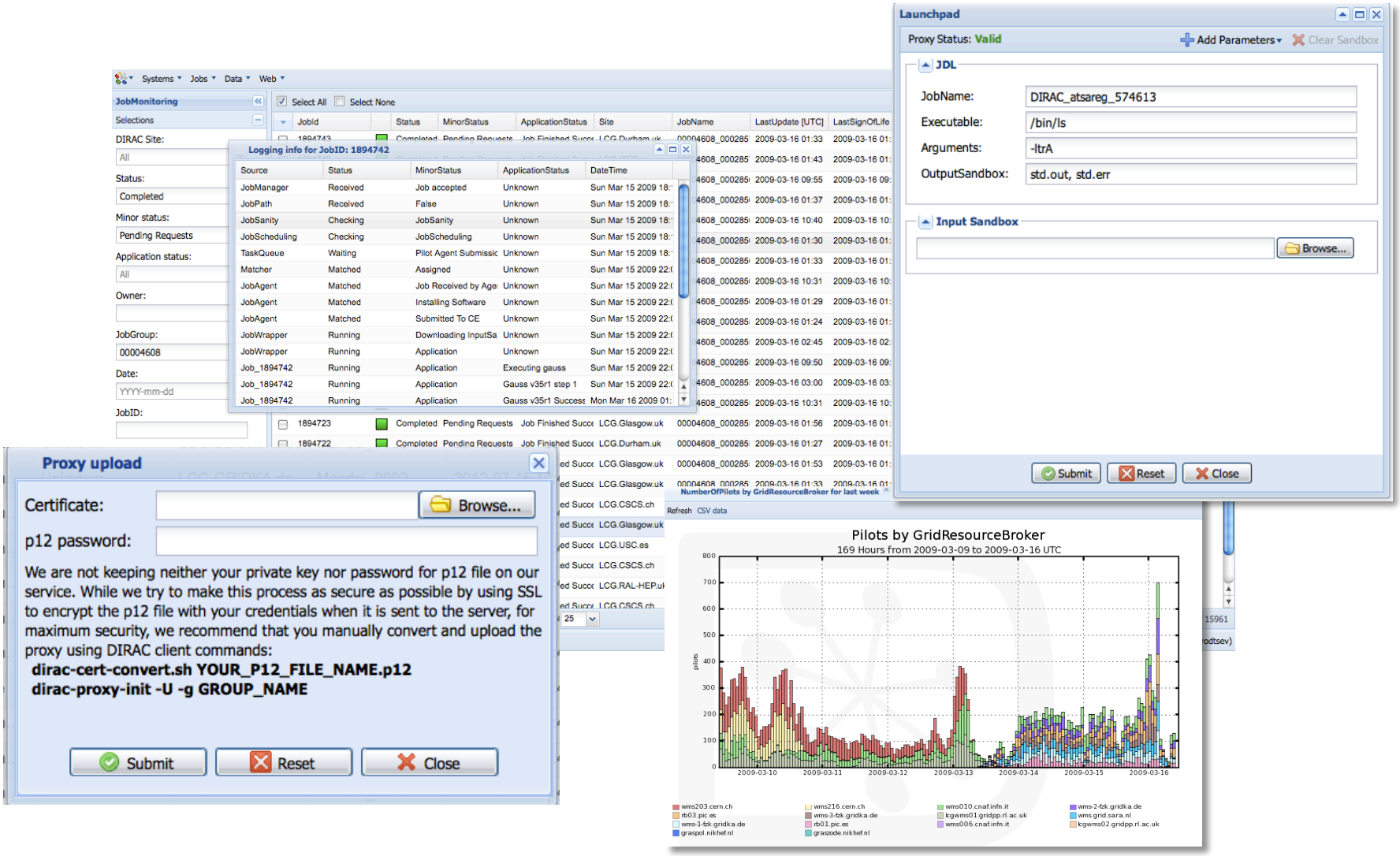
DIRAC
THE INTERWARE

Interfaces

DIRAC: Secure Web Portal

- Focus on the Web Portal as the main user tool for interactions with the grid
- Intuitive desktop application like interface
  - Ajax, Pylons, ExtJS Javascript library
- Monitoring and control of all activities
  - User registration, proxy upload
  - User job monitoring and manipulation, downloading results
  - Data manipulation and downloads
  - DIRAC Systems configuration and management
- Secure access
  - Standard grid certificates
  - Fine grained authorization rules

Web Portal: example interfaces



Application Web Portals

- Specific application portals can be built in the DIRAC Web Portal framework
  - Community Application Servers
- DIRAC RESTful interface
  - Language neutral
  - Suitable to use with portals written in Java, PHP, etc
- Other interfaces include
  - Extensive Python API
    - E.g. used by GANGA user front-end
  - A rich set of command line tools (>200 commands)

DIRAC
THE INTERWARE

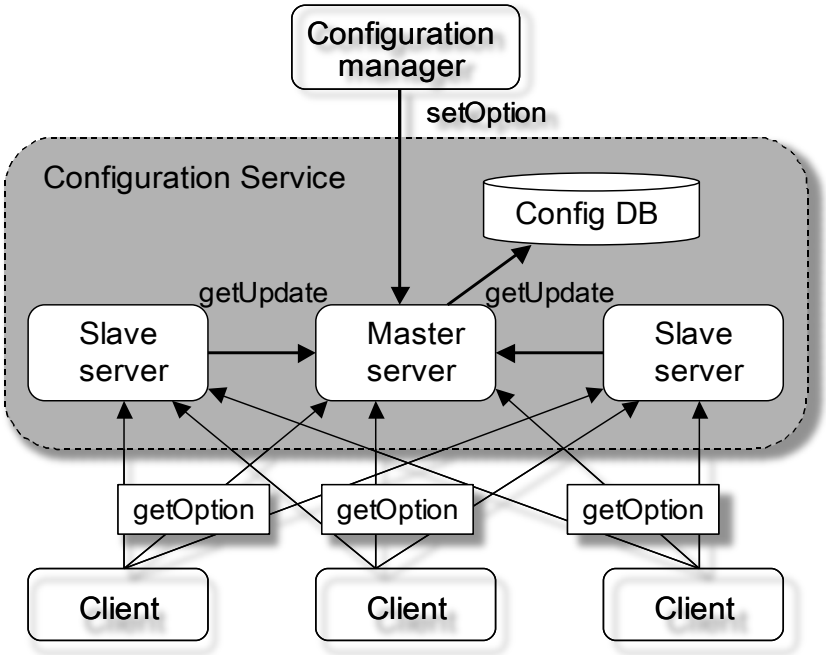
DIRAC Framework

DIRAC Framework

- DIRAC systems consist of well defined components with clear recipes for developing
  - Services, agents, clients, databases
- Framework allows to easily build these components concentrating on the business logic of the applications
  - Development environment: Python, MySQL
  - Using base services for configuration, monitoring, logging, etc
  - Specific functionality can be provided in many cases as plugin modules, e.g.
    - Data access policies
    - Job scheduling policies
- All the communications between the distributed components are secure
  - DISET custom client/service protocol
    - Focus on efficiency
    - Control and data transfer communications
  - X509, GSI security standards
  - Fine grained authorization rules

DIRAC base services

- Redundant Configuration Service
  - Provides service discovery and setup parameters for all the DIRAC components
- Full featured proxy management system
  - Proxy storage and renewal mechanism
  - Support for multiuser pilot jobs
- System Logging service
  - Collect essential error messages from all the components
- Monitoring service
  - Monitor the service and agents behavior
- Security Logging service
  - Keep traces of all the service access events

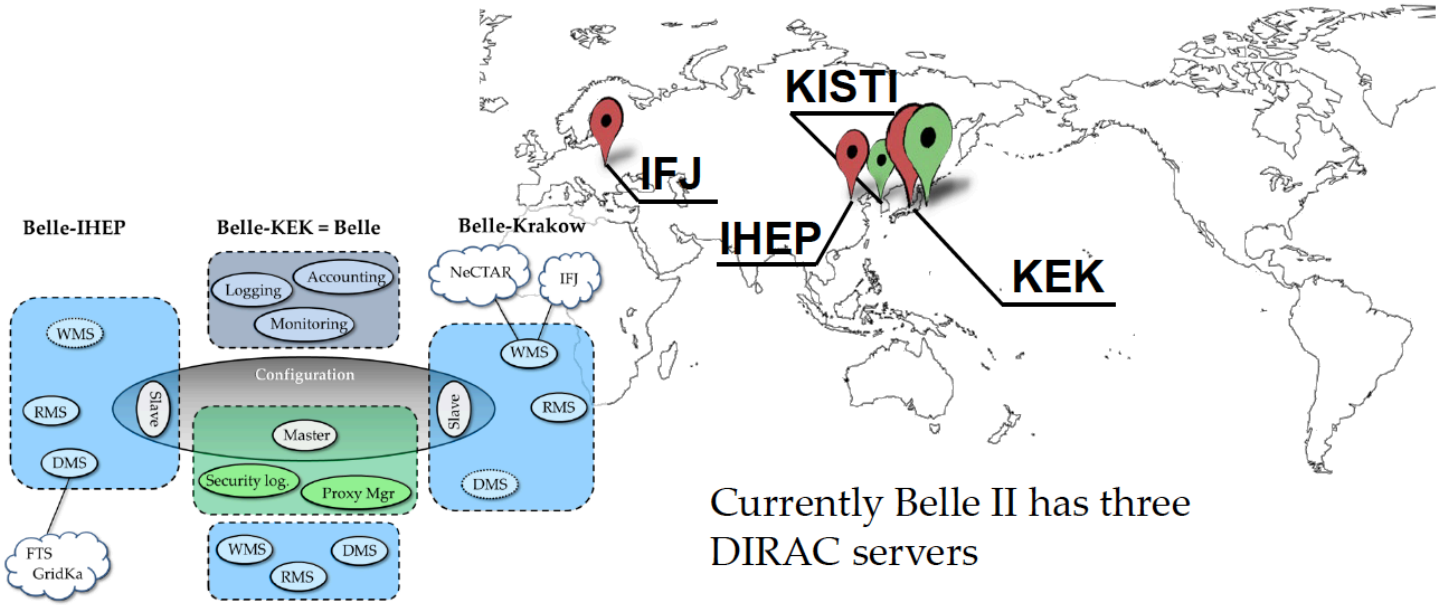


Belle II example

Distributed DIRAC servers:
KEK(Japan), IHEP(China),
IFJ(Poland)

DIRAC: master (KEK)
slave (IHEP, IFJ)

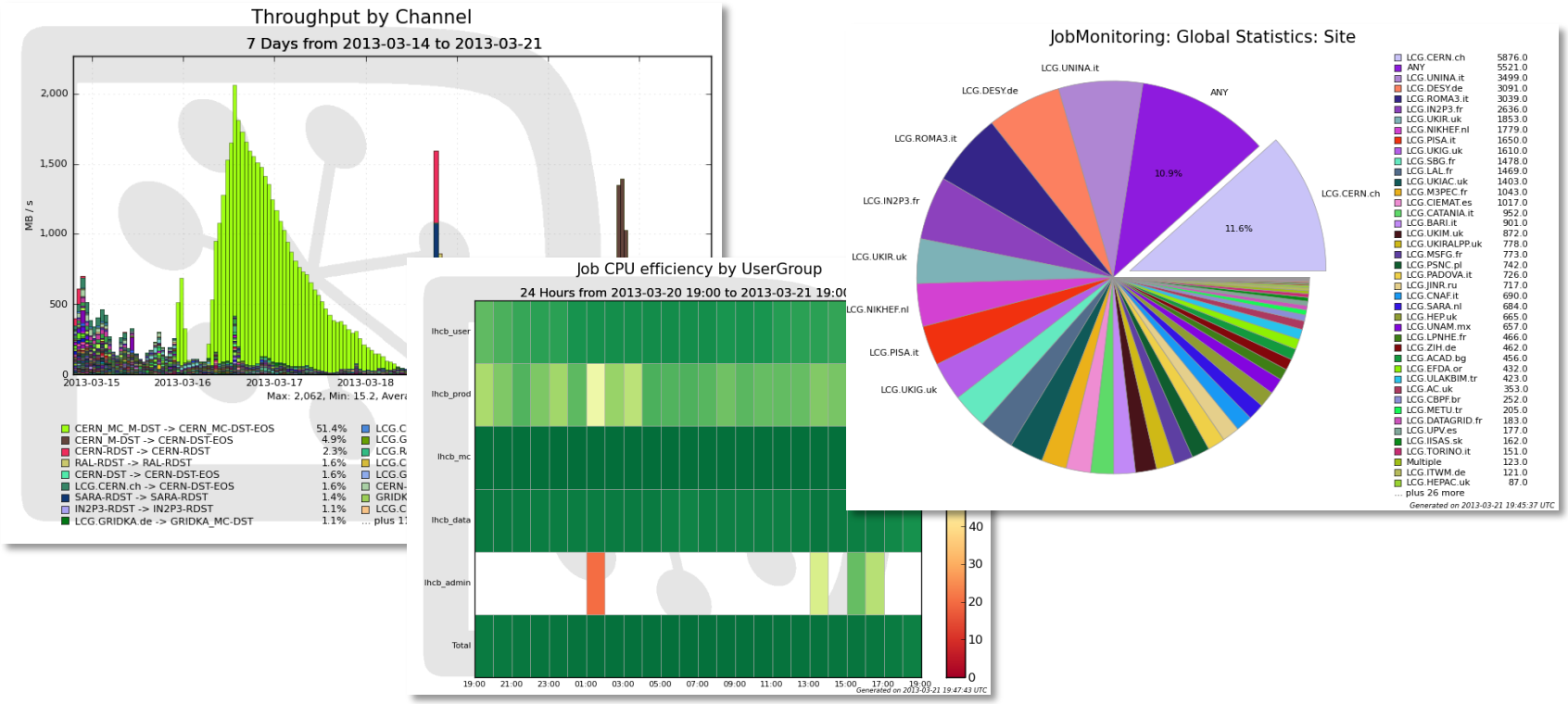
AMGA: master (KEK)
slave (KISTI)



Currently Belle II has three
DIRAC servers

Accounting

- Comprehensive accounting of all the operations

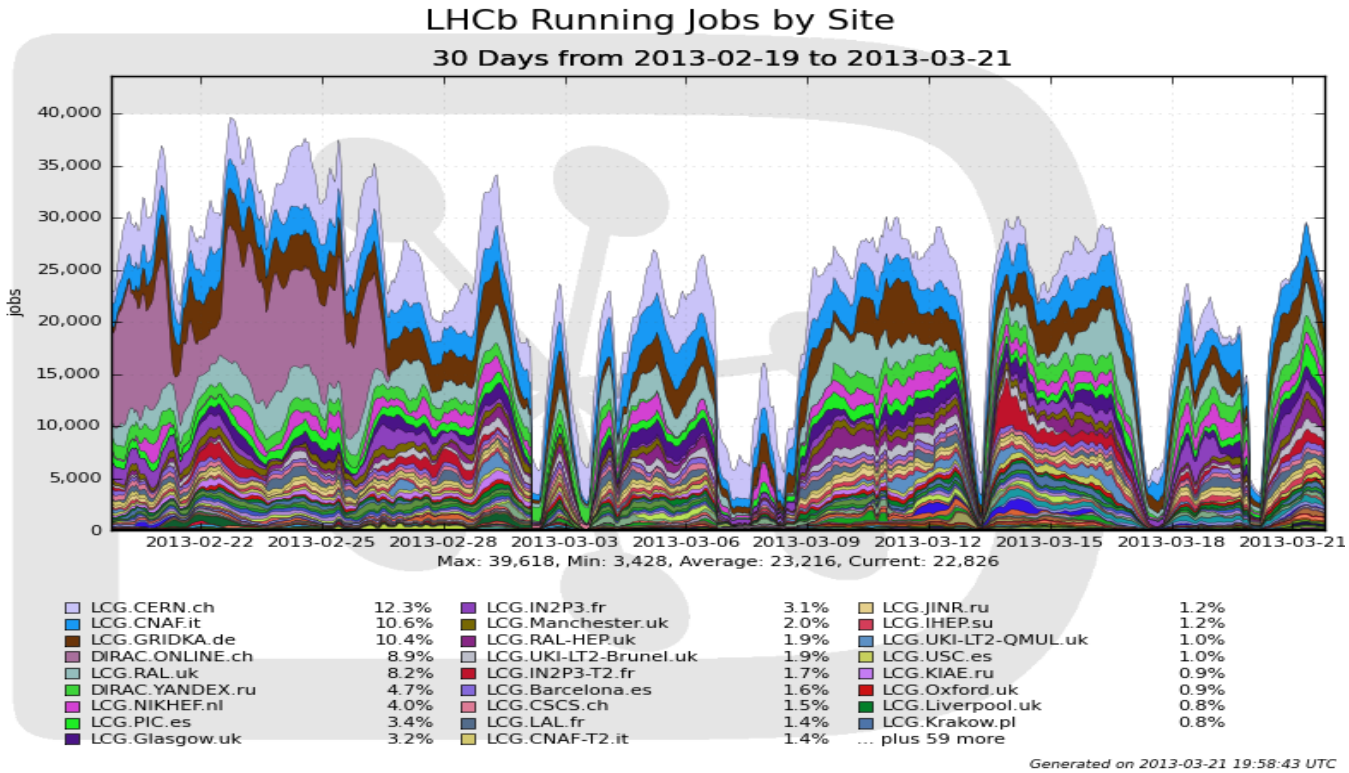


- Publication ready quality of the plots
  - Plotting service can be used by users for there own data

DIRAC
THE INTERWARE

DIRAC Users: large communities

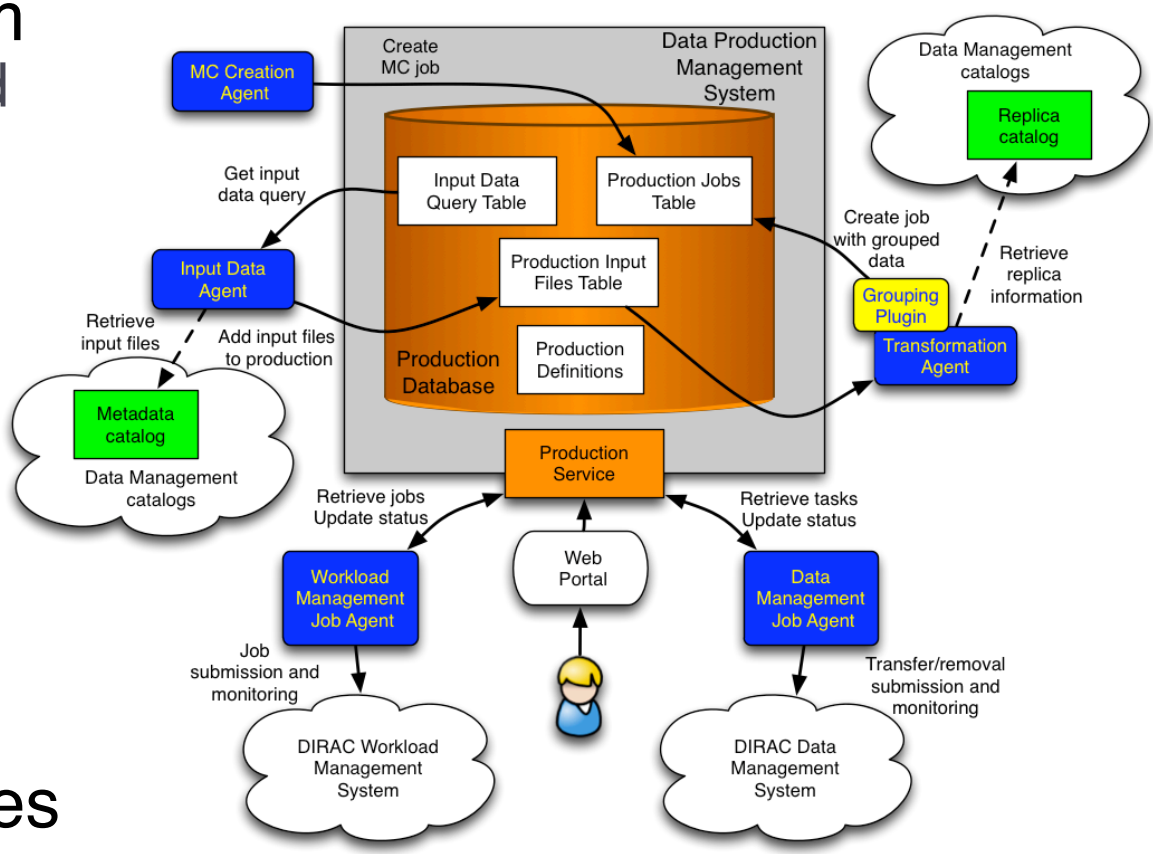
LHCb Collaboration



- Up to 40K concurrent jobs in ~120 distinct sites
  - Limited by the resources available to LHCb
- 10 mid-range servers hosting DIRAC central services
- Further optimizations to increase the capacity are possible
  - Hardware, database optimizations, service load balancing, etc

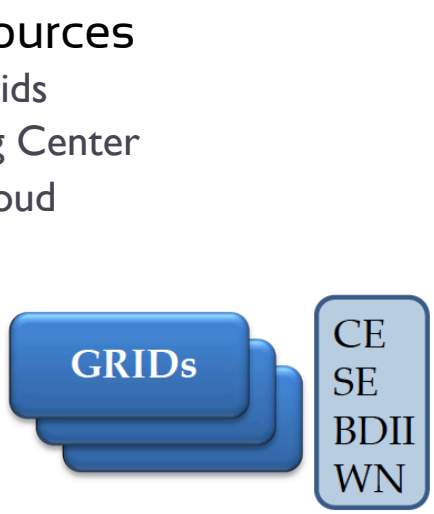
LHCb Production system

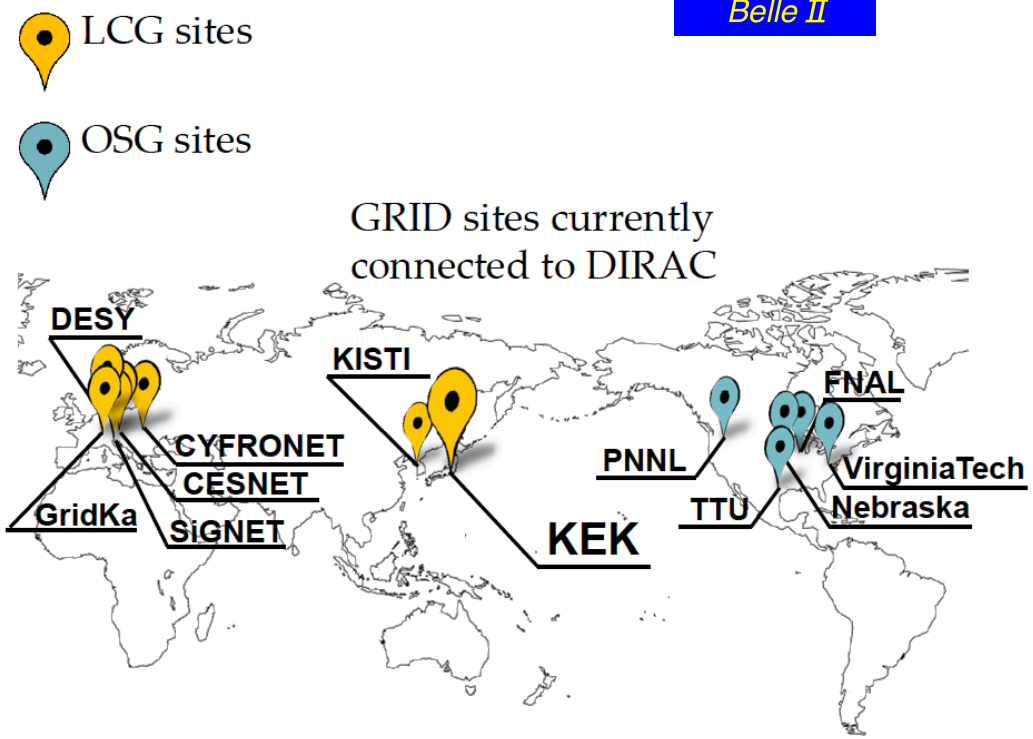
- Based on the DIRAC Transformation System
  - Multiple extensions and custom plugins
- Data driven payload generation based on templates
- Generating data processing and replication tasks
- LHCb specific templates and catalogs



Belle II

- Belle II, KEK, Japan
  - DIRAC is chosen as the basis of Computing Model for phase II of the experiment
  - 2GB/s DAQ rate
  - Combination of the non-grid, grid sites and (commercial) clouds is a requirement
- Belle II grid resources
  - WLCG, OSG grids
  - KEK Computing Center
  - Amazon EC2 cloud

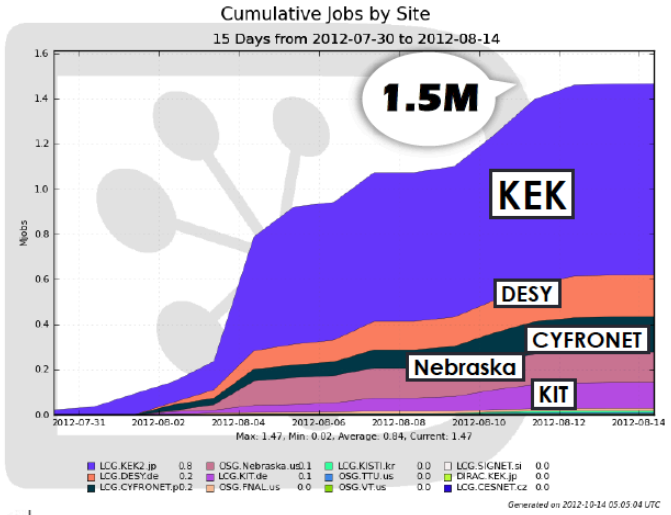


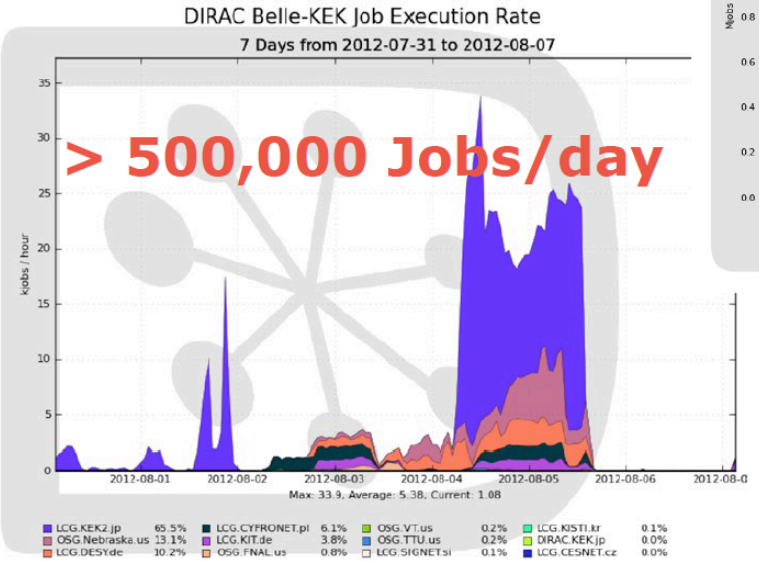


Hideki Miyake, KEK

DIRAC Scalability tests

- Random number generation (500/job) or just filling pilot job
  →no SE/AMGA used
- Good performance
  - Even saturated KEKCC GRID
- DIRAC itself was stable





Hideki Miyake, KEK

DIRAC dedicated installations

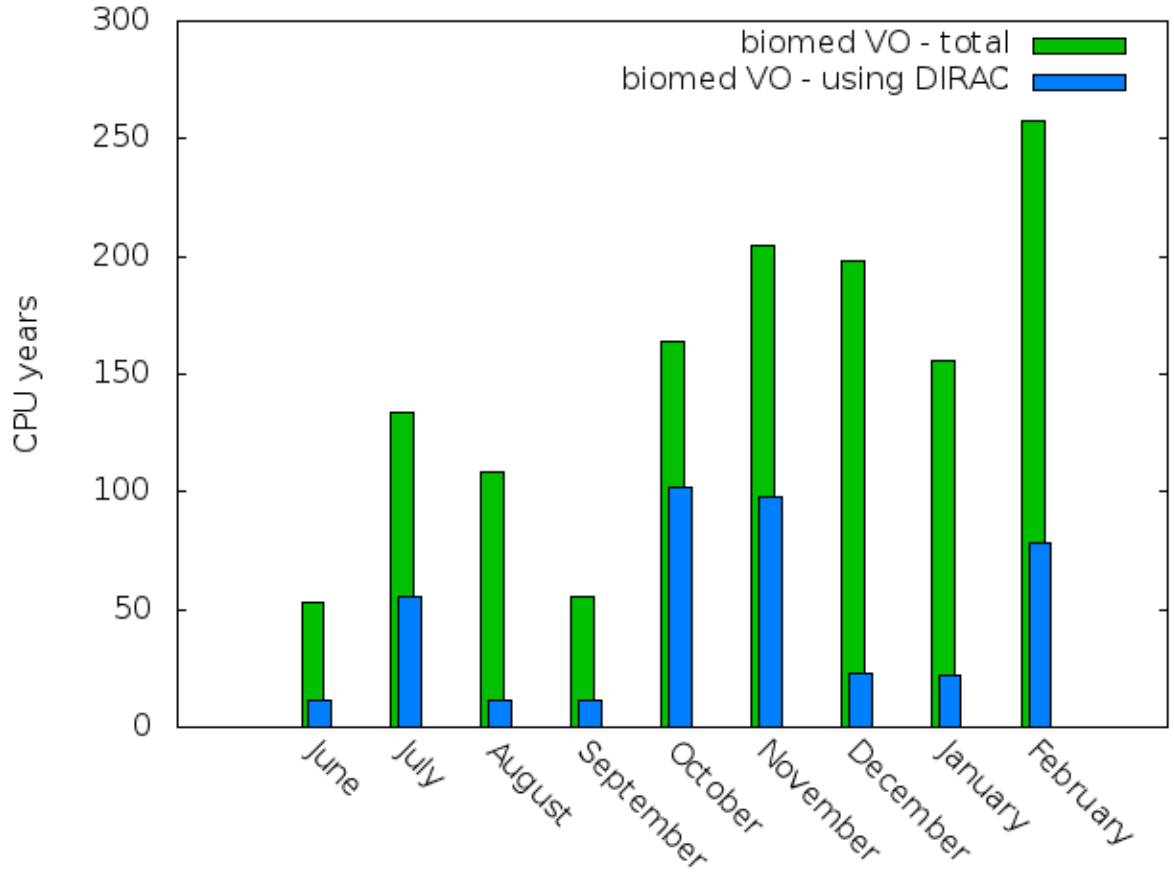
- ILC/CLIC detector Collaboration
  - Base production system on DIRAC
  - MC simulations
  - DIRAC File Catalog was developed to meet the ILC/CLIC requirements

- BES III, IHEP, China
  - DIRAC is chosen for the phase III
  - Using DIRAC DMS: File Catalog, Transfer services

- CTA
  - CTA started as FG-DIRAC customer for DIRAC evaluation
  - Now is using a dedicated installation at PIC, Barcelona
  - Using complex workflows

DIRAC in biomed

- Use of computing resources in the biomed grid community
  - DIRAC instance provided by France-Grilles since June 2012

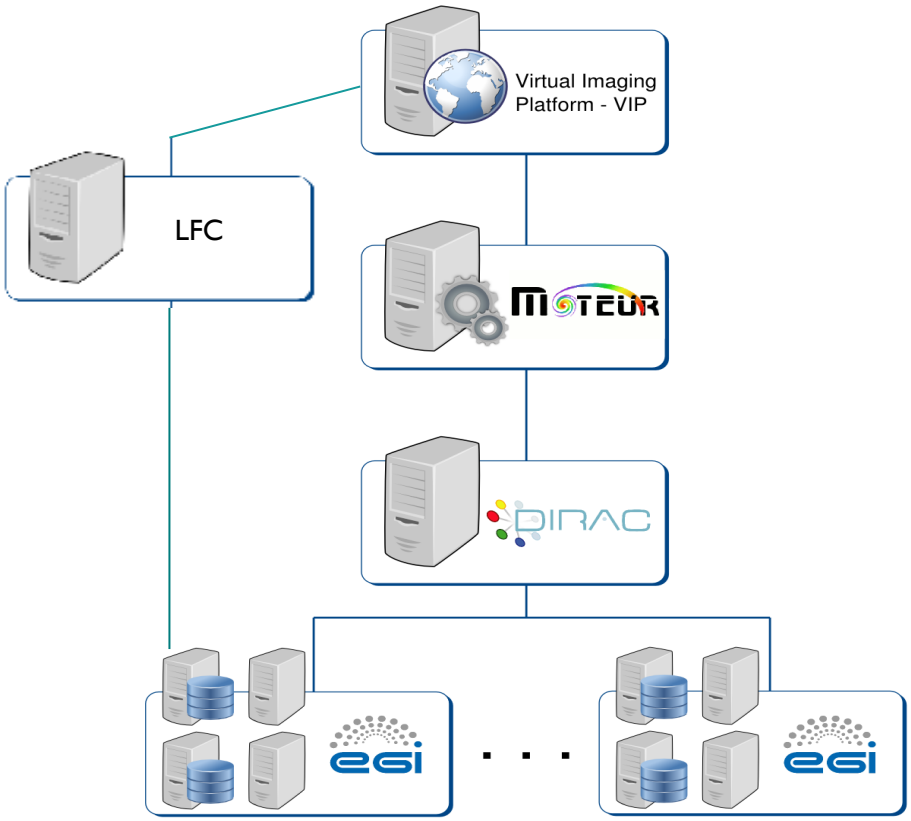


Tristan Glatard, CREATIS

source: https://accounting.egi.eu ; https://dirac.france-grilles.fr

Virtual Imaging Platform

- Platform for medical image simulations at CREATIS, Lyon
  - Example of a combined use of an Application Portal and DIRAC WMS



- Web portal with robot certificate
  - *File transfers, user/group/application management*
- Workflow engine
  - *Generate jobs, (re-)submit, monitor, replicate*
- DIRAC
  - *Resource provisioning, job scheduling*
- Grid resources
  - *biomed VO*

DIRAC
THE INTERWARE

DIRAC Services

DIRAC as a service

- DIRAC client is easy to install
  - Part of a usual tutorial
- DIRAC services are easy to install but
  - Needs dedicated hardware for hosting
  - Configuration, maintenance needs expert manpower
  - Monitoring computing resources
- Small user communities can not afford maintaining dedicated DIRAC services
  - Still need easy grid access
- Large grid infrastructures can provide DIRAC services for their users.

France-Grid DIRAC service

- Started as a support for user tutorials

- Several regional and university campus installations
  - Complex maintenance

- Joint effort to provide France-Grid DIRAC service
  - Hosted by the CC/IN2P3, Lyon, T1 center
    - 6 virtual servers, MySQL server
  - Distributed team of service administrators
    - 5 participating universities

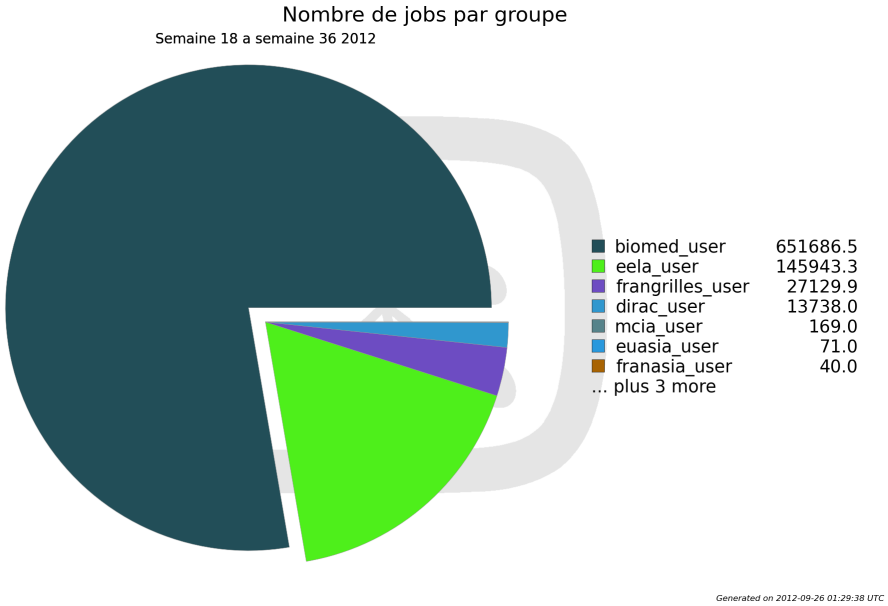
http://dirac.france-grilles.fr

Services in CC/Lyon

- Basic DIRAC services
  - WMS – managing users jobs
    - Job submission, monitoring, retrieval
    - Accounting of the resources consumed
  - DMS – managing user data basic tasks
    - Access to standard Grid Storage Elements
      - SRM, DIRAC
    - Replicating data between SEs
    - Providing Simple Storage Element in Lyon
    - DIRAC File Replica Catalog
    - DIRAC File Metadata Catalog
- Web Portal
- REST interface
  - OAuth2 authentication

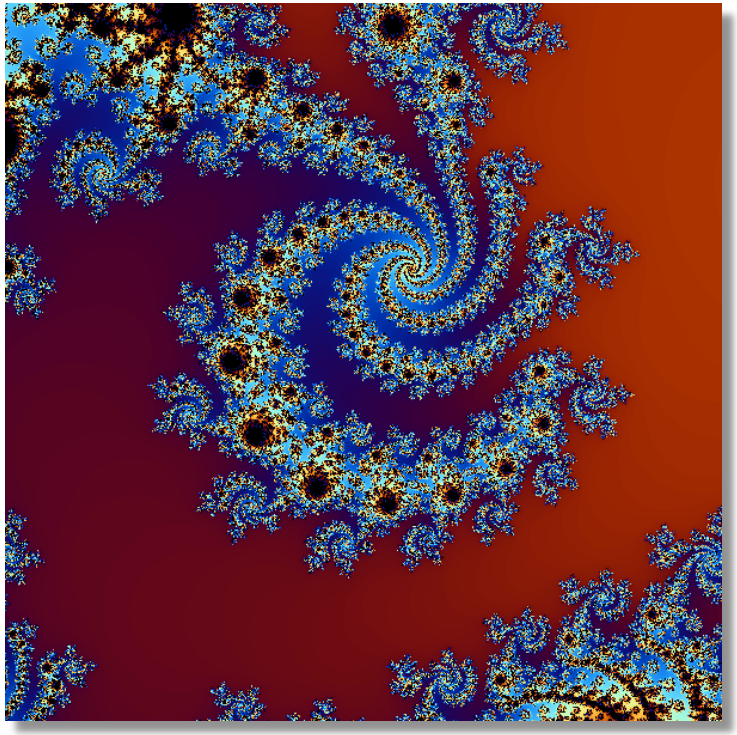
FG-DIRAC users

- France-Grilles users
  - 15 VOs, 88users registered
    - astro, auger, biomed, esr, euasia, gilda, glast.org, prod.vo.eu-eela.eu, vo.cta.in2p3.fr, vo.formation.idgrilles.fr, vo.france-asia.org, vo.france-grilles.fr, vo.msfg.fr, vo.mcia.org
    - 1 robot user VIP/GateLab Biomed
    - More VO's and users can be added as necessary
  - In production since May 2012
    - First ~3 millions jobs went through the system
      - Mostly biomed applications

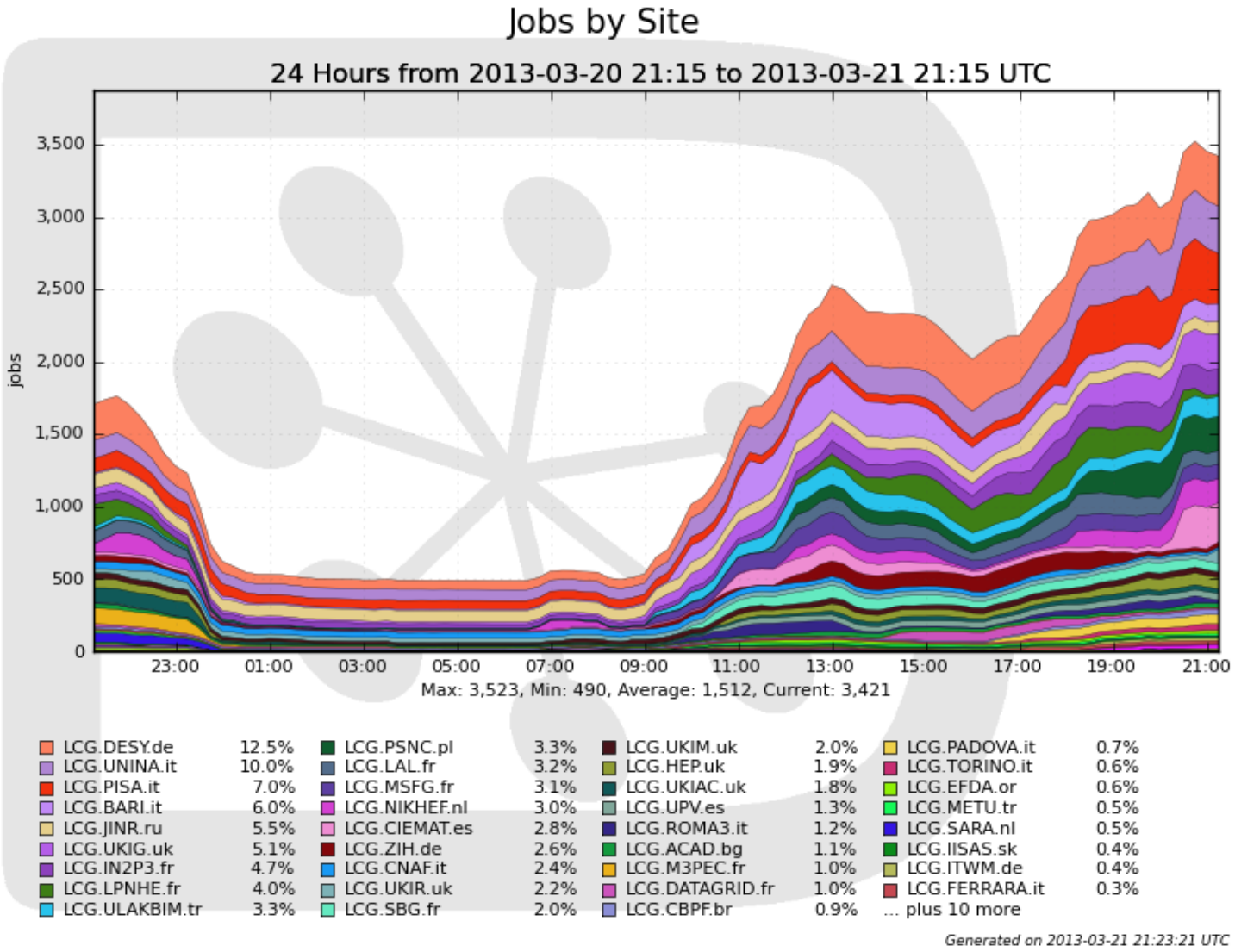


FG-DIRAC users

- Heavily used for the grid tutorials
  - Example of tutorial exercise -->
  - Using resources of the VO france-formation
- Support for applications
  - Help in porting applications to the grid
  - Help new communities to try out DIRAC for their production systems
    - Fermi-LAT
    - LSST
    - SuperB



Yesterday FG-DIRAC snapshot



Advanced services

- More advanced services can be made available in CC Lyon
  - Following the user demands
  - Transformation Service (automated job submission)
  - Replication Service (automated data replication)
  - Data integrity inspection
  - User storage consumption accounting
  - Support for MPI jobs
  - Others ?
- Hosting Community DIRAC services
  - Specific services developed for particular communities can be hosted in the same infrastructure

FG-DIRAC server configuration

- 6 virtual servers (3 physical hosts)
  - 8 cores, 16 GB RAM, 1TB disk
  - ***ccdirac01*** – secure services, configuration
  - ***ccdirac02*** – Workload Management
  - ***ccdirac03*** – Data Management
  - ***ccdirac04*** – StorageElement, Accounting, Monitoring
  - ***ccdirac05*** – Web Portal
    - http://dirac.france-grilles.fr
  - ***ccdirac06*** – REST Portal
- MySQL server
  - 30GB, 100 connections
- Redundant supporting services outside the CC in Lyon
  - CPPM, CREATIS, etc

DIRAC national services

Other national DIRAC installations

- GISELA Latin American grid
  - In production since 2010
  - Since 2012 GISELA DIRAC services are provided by France-Grid
- IberGrid Spanish NGI
- Including DIRAC services in order grid infrastructures are under discussion/constrcution:
  - China, Russia, Italy, …

DIRAC
THE INTERWARE

DIRAC Service

- The success of the France-Grilles and other DIRAC Services shows that they bring clear advantages to multiple users and whole user communities needing access to distributed computing resources
  - This is especially well seen during the user tutorials

- Therefore, we think that this can be a very useful facility for the users of any grid infrastructure project (NGI) and of the EGI project as a whole.

DIRAC
THE INTERWARE

Conclusions

- The computational grids are no more something exotic, they are used in a daily work for various applications
- Rich experience with using computational grids in the LHC experiments, as well as the developed tools, can now be shared with users in other experiments and in other scientific domains
- 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

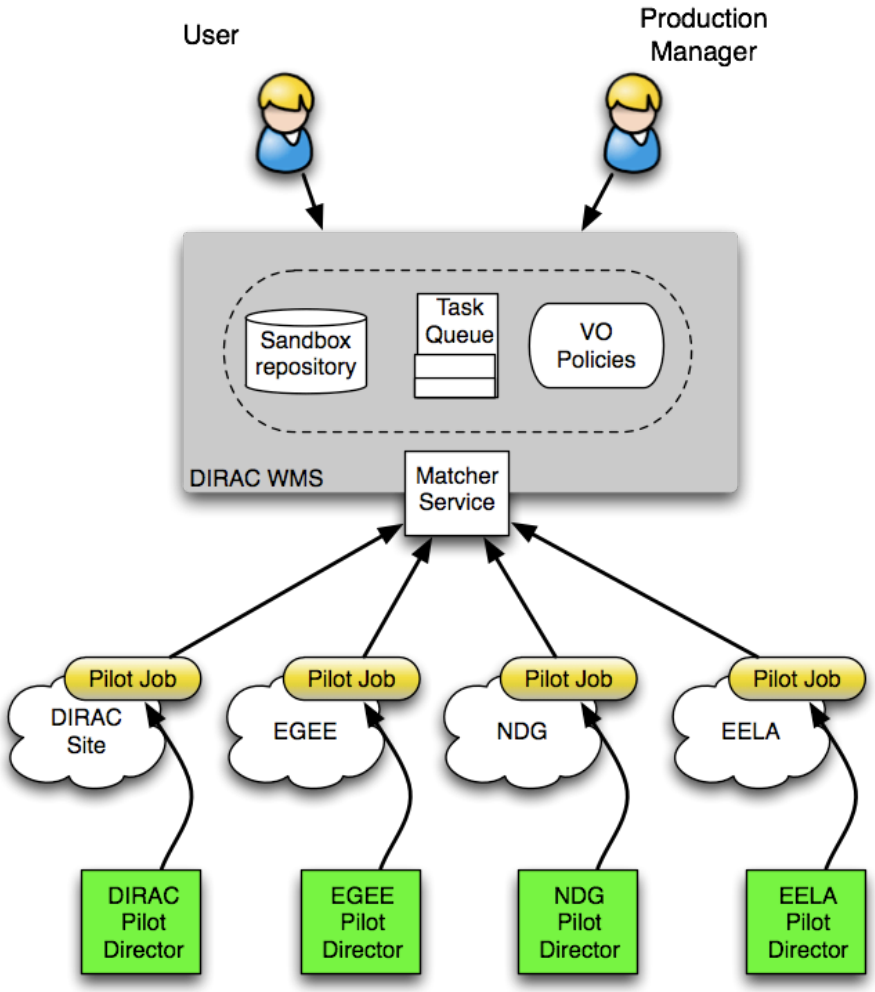
http://diracgrid.org

DIRAC
THE INTERWARE

Backup slides

DIRAC WMS

- Jobs are submitted to the DIRAC Central Task Queue with credentials of their owner (VOMS proxy)
- Pilot Jobs are submitted by specific Directors to a Grid WMS with credentials of a user with a special Pilot role
- The Pilot Job fetches the user job and the job owner's proxy
- The User Job is executed with its owner's proxy used to access SE, catalogs, etc

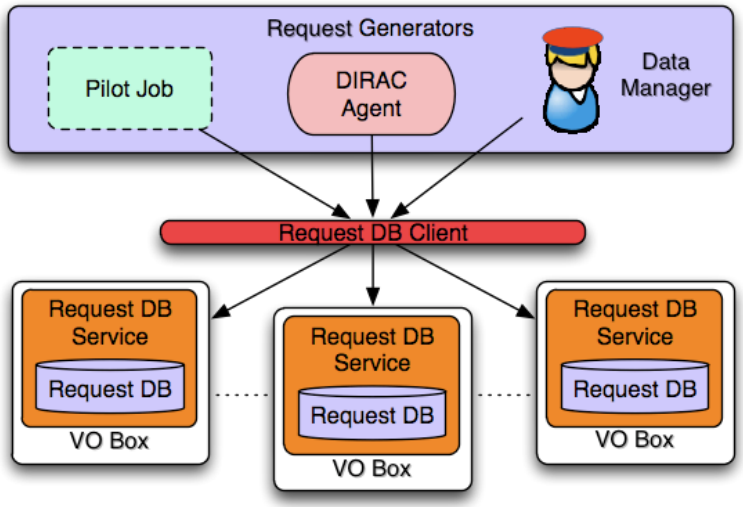


Advantages for site resources providers

- No need for a variety of local batch queues per VO
  - One long queue per VO would be sufficient
  - 24-48 hours queue is a reasonable compromise
    - Site maintenance requirements
  - Reduced number of grid jobs
- No need for specific VO configuration and accounting on sites
  - Priorities for various VO groups, activities
  - User level accounting is optional
- In the whole it can lower the site entry threshold
  - Especially useful for newcomer sites

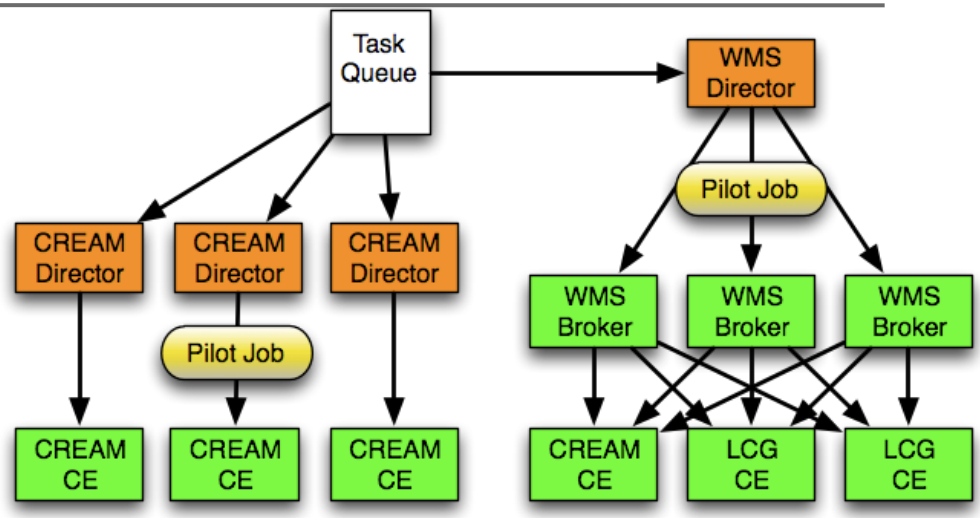
Request Management system

- A Request Management System (RMS) to accept and execute asynchronously any kind of operation that can fail
  - Data upload and registration
  - Job status and parameter reports



- Requests are collected by RMS instances at geographically distributed sites
  - Extra redundancy in RMS service availability
- Requests are forwarded to the central Request Database
  - For keeping track of the pending requests
  - For efficient bulk request execution

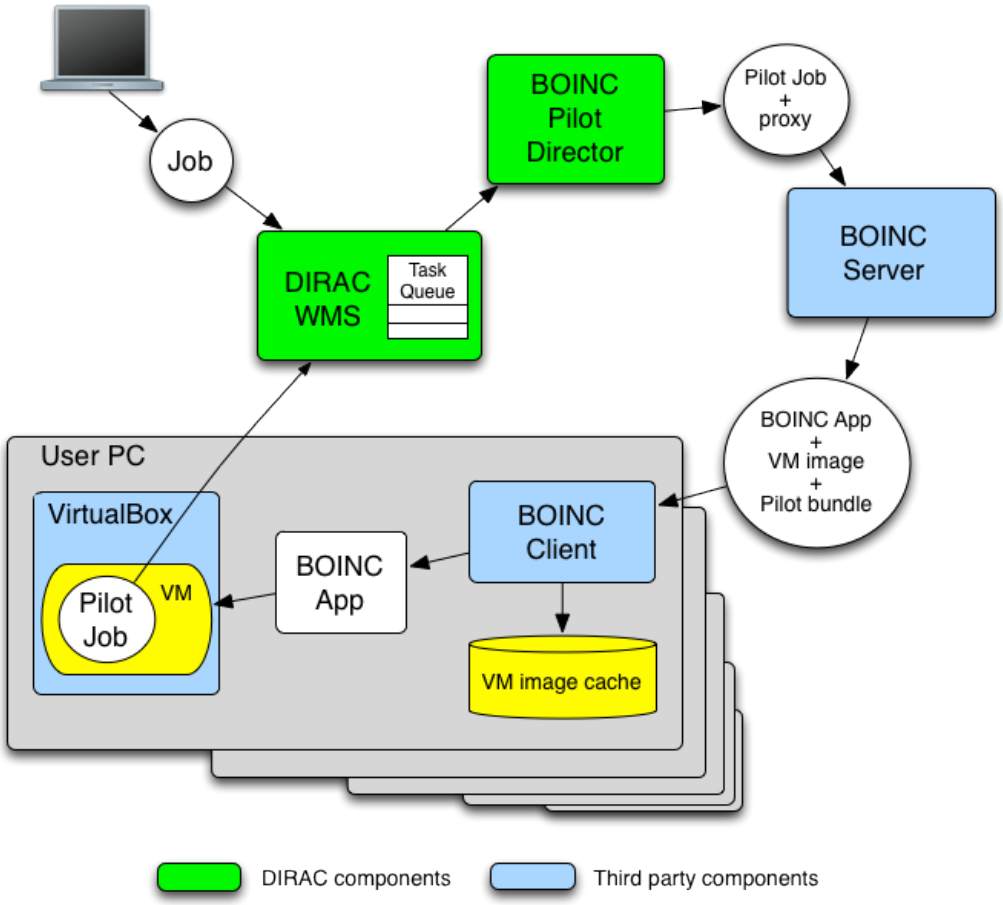
Direct submission to CEs



- Using gLite WMS now just as a pilot deployment mechanism
  - Limited use of brokering features
    - For jobs with input data the destination site is already chosen
  - Have to use multiple Resource Brokers because of ***scalability*** problems
- DIRAC is supporting direct submission to CEs
  - CREAM CEs or batch clusters through SSH tunnel
  - Can apply individual site policy
    - Site chooses how much load it can take (*Pull* vs *Push* paradigm)
  - Direct measurement of the site state watching the pilot status info
- This is a general trend
  - All the LHC experiments declared abandoning eventually gLite WMS

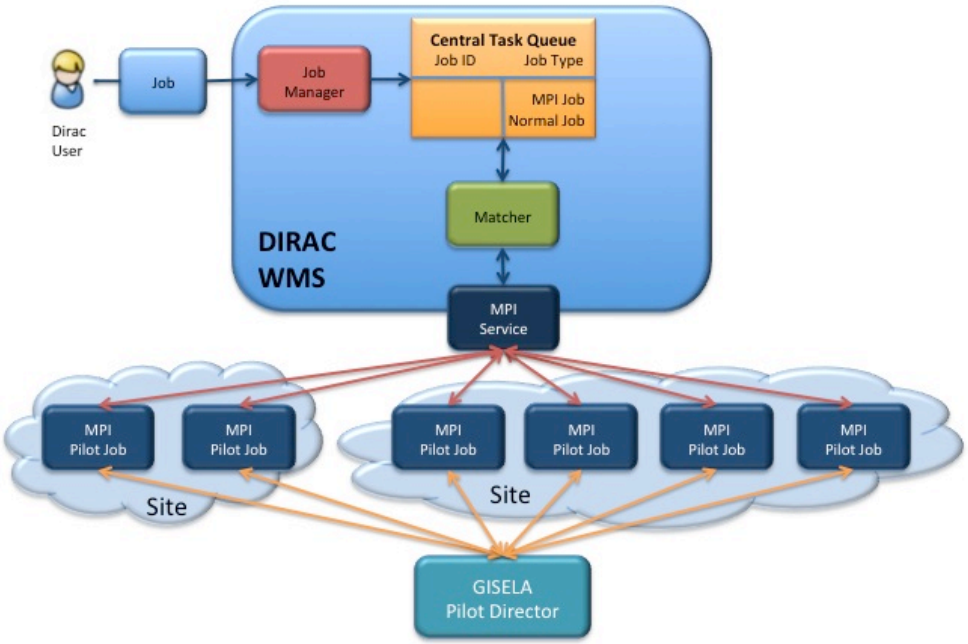
BOINC Desktop Grids

- On the client PC the third party components are installed:
  - VirtualBox hypervisor
  - Standard BOINC client
- A special BOINC application
  - Starts a requested VM within the VirtualBox
  - Passes the Pilot Job to the VM and starts it
- Once the Pilot Job starts in the VM, the user PC becomes a normal DIRAC Worker Node
- Work on interfacing DIRAC to EDGI resources is in progress



Support for MPI Jobs

- MPI Service developed for applications in the EELA/GISELA Grid
  - Astrophysics, BioMed, Seismology applications
  - No special MPI support on sites is required
    - MPI software is installed by Pilot Jobs
      - Possibility to use distributed systems, e.g. *Parrot*
  - MPI ring usage optimization
    - Ring reuse for multiple jobs
      - Lower load on the gLite WMS
    - Variable ring sizes for different jobs

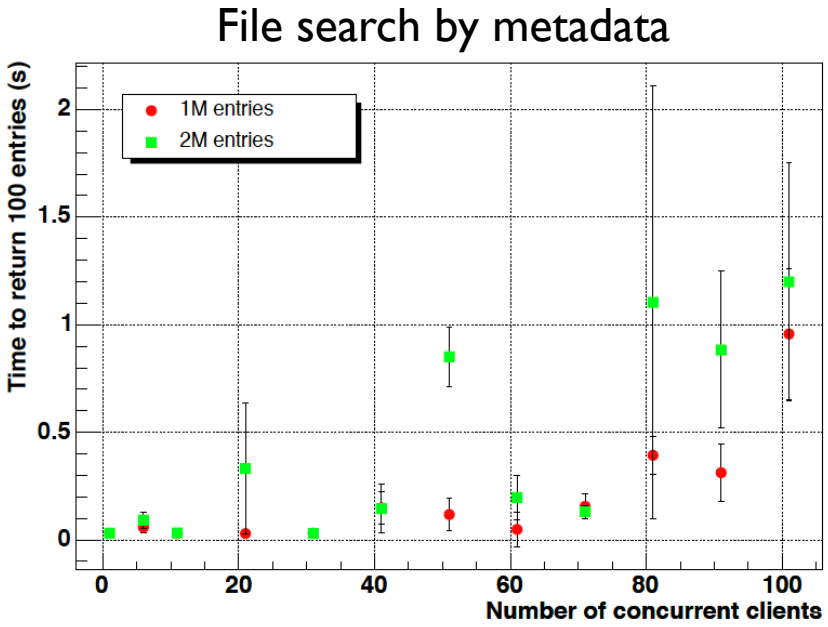


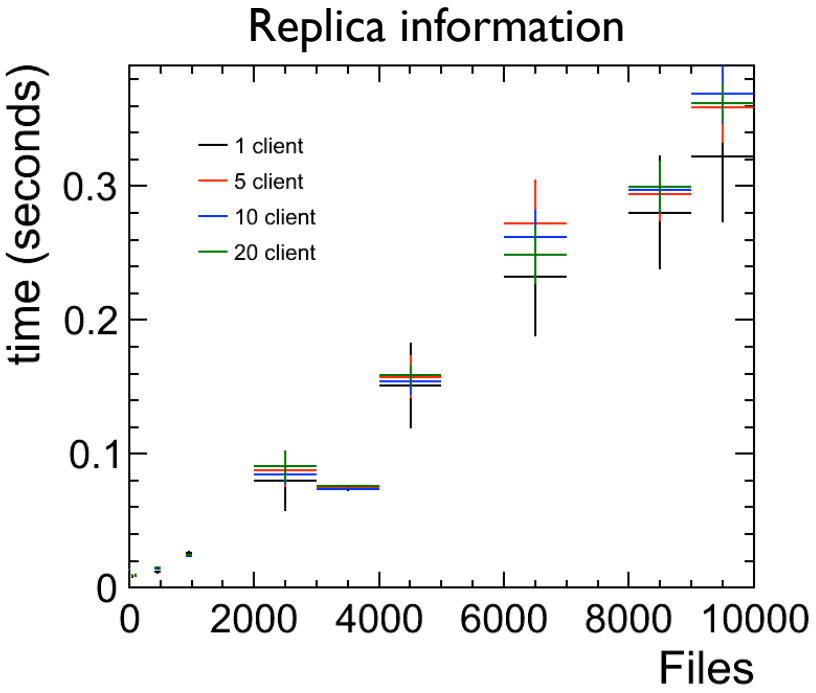
File Catalog Metadata

- Similar functionality with the AMGA metadata service
  - But coupled with the replica catalog to boost efficiency
- Metadata can be associated with each directory as key:value pairs to describe its contents
  - Int, Float, String, DateTime value types
- Some metadata variables can be declared indices
  - Those can be used for data selections
- Subdirectories are inheriting the metadata of their parents
- Data selection with metadata queries. Example:
  - `find . Meta1=Value1 Meta2>3 Meta2<5 Meta3=2,3,4`
- File metadata can also be defined

DIRAC File Catalog evaluation

- ILC/CLIC Collaboration experience
  - ~1M files
  - Intensive use of metadata, provenance data

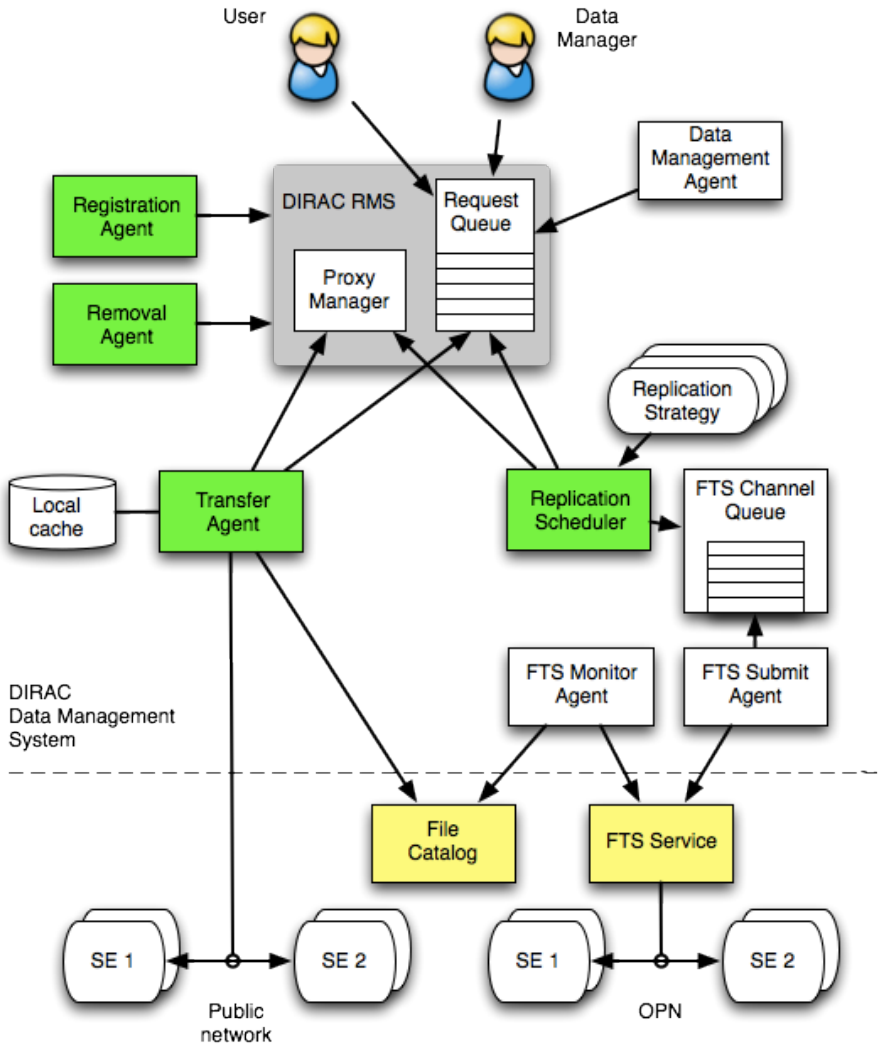




- BES Collaboration made a thorough comparison of DFC vs AMGA
  - Similar performance
  - More suitable functionality

Data Management

- Based on the Request Management System
  - Asynchronous data operations
  - transfers, registration, removal
- Two complementary replication mechanisms
  - Transfer Agent
    - user data
    - public network
  - FTS service
    - Production data
    - Private FTS OPN network
    - Smart pluggable replication strategies



DIRAC as a service

- Necessity to manage multiple VOs with a single DIRAC installation
  - Per VO pilot credentials
  - Per VO accounting
  - Per VO resources description
- Pilot directors are VO aware
  - Job matching takes pilot VO assignment into account
- This the work in progress