DIRAC: from particle physics to other scientific domains

A. Tsaregorodtsev,
CPPM-IN2P3-CNRS, Marseille
RO-LCG Conference, Magurele, 2016

DIRAC
THE INTERWARE

Plan

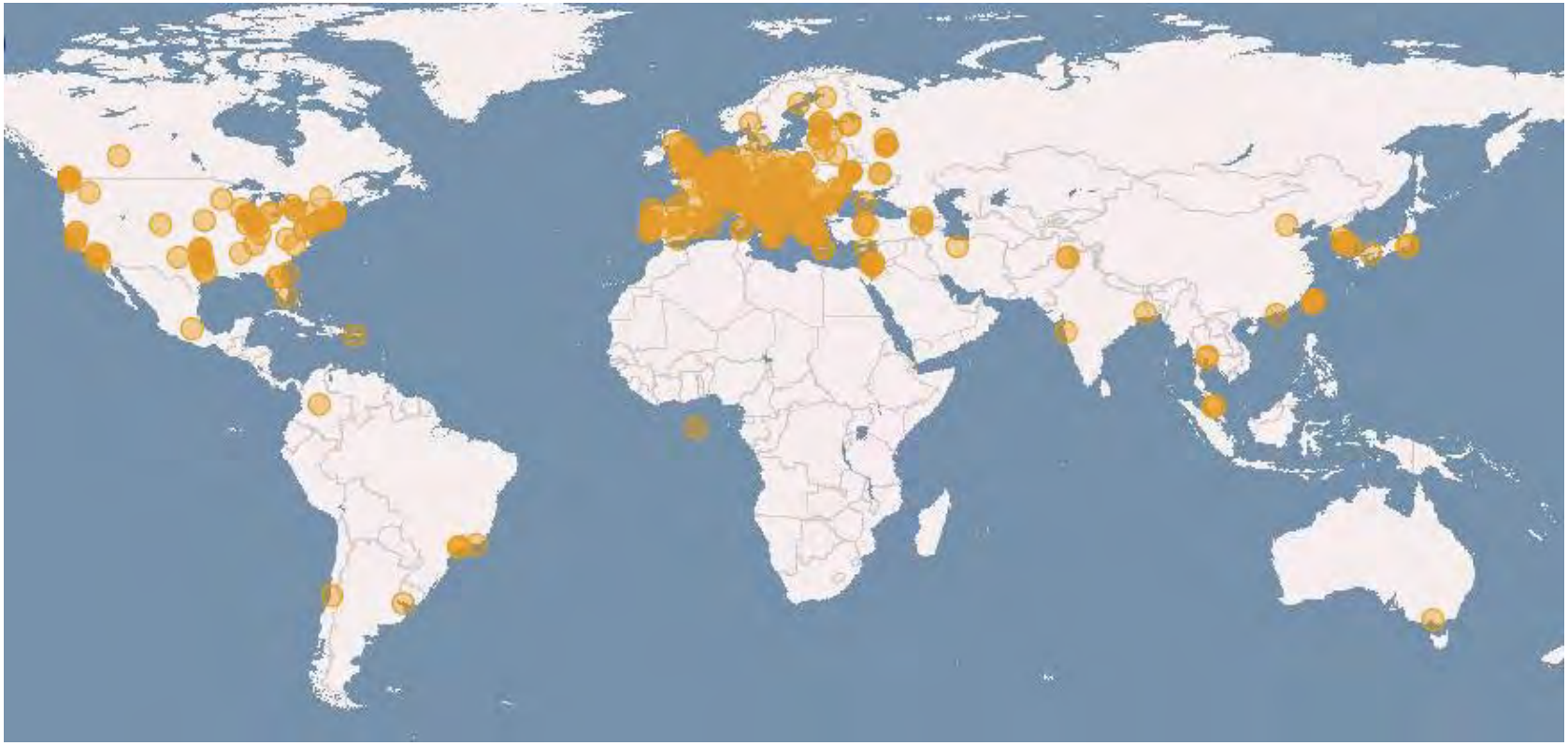
- The problem of high intensity scientific data processing
- DIRAC Project
  - Agent based Workload Management System
  - Accessible computing resources
  - Data Management
  - Interfaces
- DIRAC as a service
- Conclusions

Massive computations: HEP

- LHC experiments pioneered the massive use of computational grids as a solution to the High Energy Physics Big Data problem
  - Many 10s of PBytes of data per year
  - Many 100s of thousands CPUs in 100s of centers
  - Many 10s of Gbyte/s data transfers
  - Many 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

Worldwide LHC Computing Grid Collaboration (WLCG)



- >500 PB of data at CERN and major computing centers
- Distributed infrastructure of ~170 computing centers in ~40 countries
- 400+ k CPU cores (~ 5M HEP-SPEC-06 or ~5 PetaFlops)
- The biggest site with ~50k CPU cores, 12 T1 centers with 2-30k CPU cores
- Distributed data, services and operation infrastructure

DIRAC
THE INTERWARE

Computing grids

- Grids are providing common infrastructure and rules to integrate multiple computing centers
  - Common computing element access protocols
  - Common user task scheduling system
  - Common policies and security rules
    - Users are signing up once and have access to multiple computing centers
  - Common monitoring of the user jobs
  - Common planning and accounting of computing resources
  - Common user training and support
  - Common problem tracking system
- Grids are making all the constituent computing centers to be seen as one entity for the users
- Now most of the computing power for LHC is provided by grid systems

Virtualized Computing Resources

- Cloud technology allows to offer computing resources on demand according to the user specification and budget
  - Pay for what is consumed
  - Widely used as a platform for user defined services (PaaS)
  - Application portals (SaaS)
  - Amazon EC2 Cloud pioneered the field
- Since recently computing centers started to provide their resources using cloud technologies
  - Virtualized computing infrastructures, IaaS
  - Very flexible – provides resources adapted to the user needs
    - Operating systems, memory, CPU power, etc.
- However, large scientific collaborations can have access to multiple computational clouds
  - Dealing with independent clouds separately requires a huge management effort
  - Federating multiple clouds into a single coherent system is necessary to provide a transparent access for users.
    - Analogous to the grid infrastructures

DIRAC
THE INTERWARE

Other resources

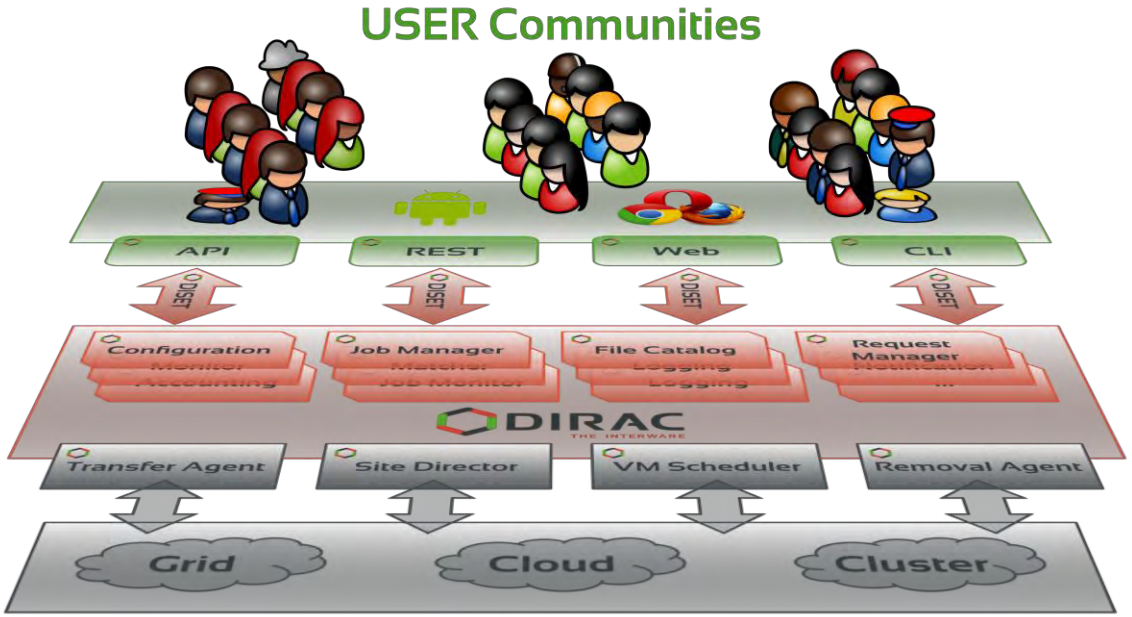
- Standalone computing clusters not included in any grid infrastructure
  - Resources available at Universities and scientific laboratories
- High Performance Computing (HPC) Centers, or Supercomputers
  - Computing centers oriented towards massively parallel applications using specialized hardware
- Volunteer Computing
  - Mostly based on BOINC technology
  - SETI@Home, LHC@Home, etc

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

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***.

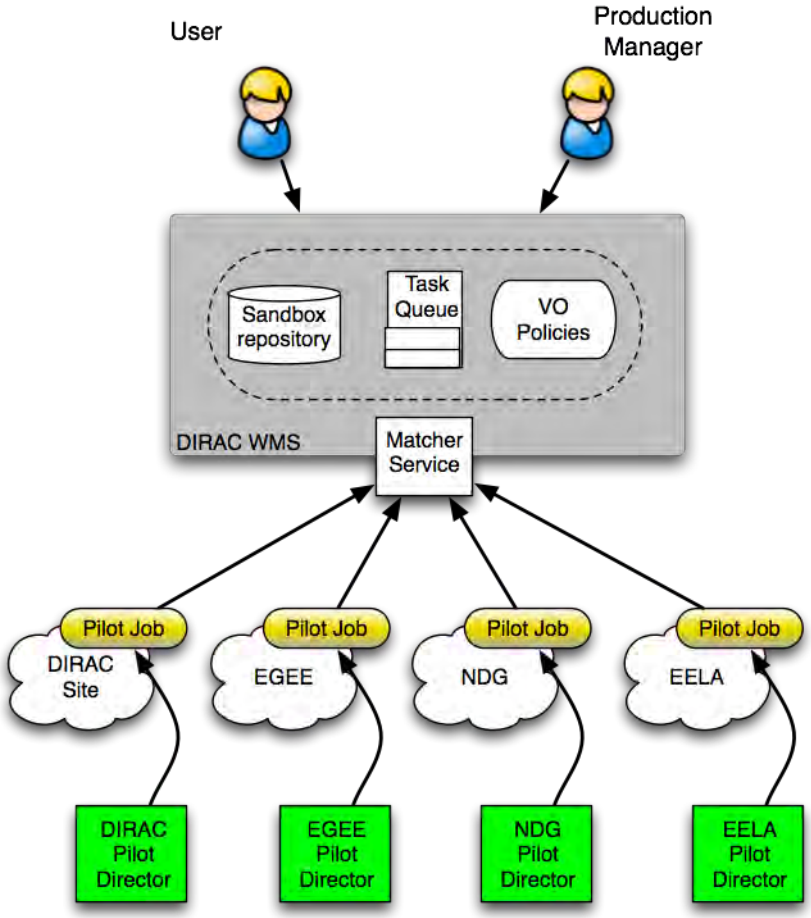


DIRAC
THE INTERWARE

DIRAC Workload Management

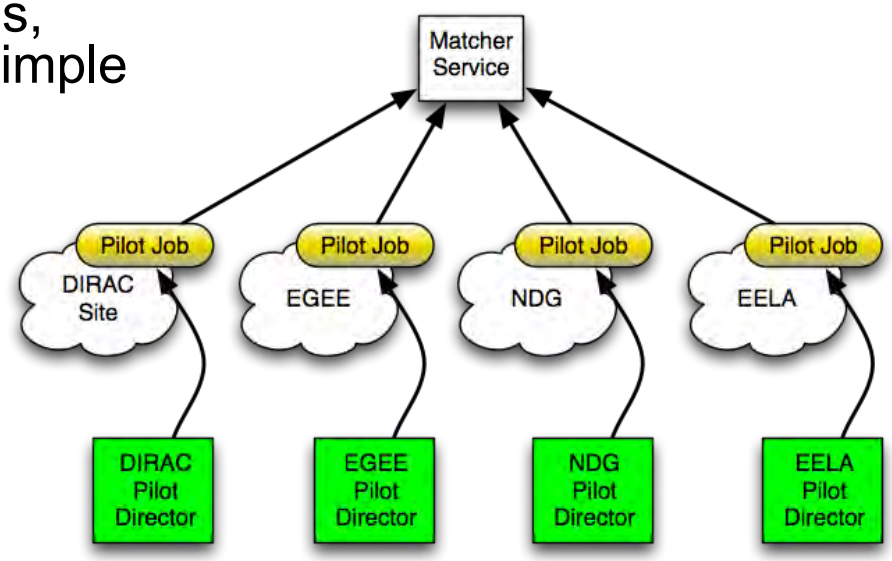
Job scheduling

- Pilot jobs are submitted to computing resources by specialized Pilot Directors
- After the start, Pilots check the execution environment and form the resource description
  - OS, capacity, disk space, software, etc
- The resources description is presented to the Matcher service, which chooses the most appropriate user job from the Task Queue
- The user job description is delivered to the pilot, which prepares its execution environment and executes the user application
- In the end, the pilot is uploading the results and output data to a predefined destination



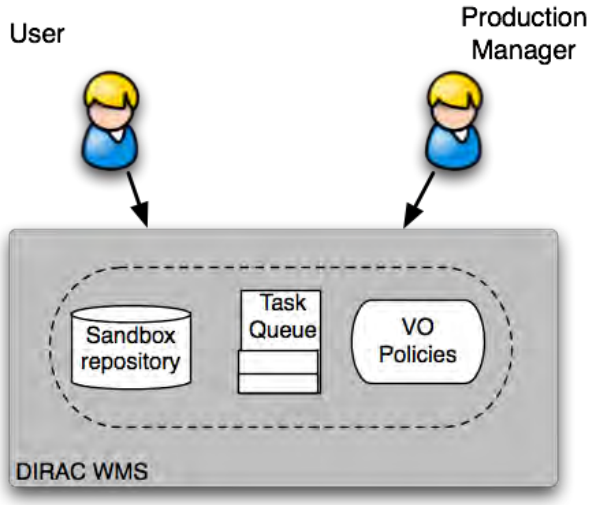
WMS: using heterogeneous resources

- Pilot based Workload Management provides abstraction of Computing Resources
  - Allows to combine heterogeneous resources in a transparent way
- Including resources in different grids, clouds and standalone clusters is simple with Pilot Jobs
  - Needs a specialized Pilot Director per resource type
  - Users just see new logical sites appearing
- Similar patterns are applied also for the Data Management System of DIRAC



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

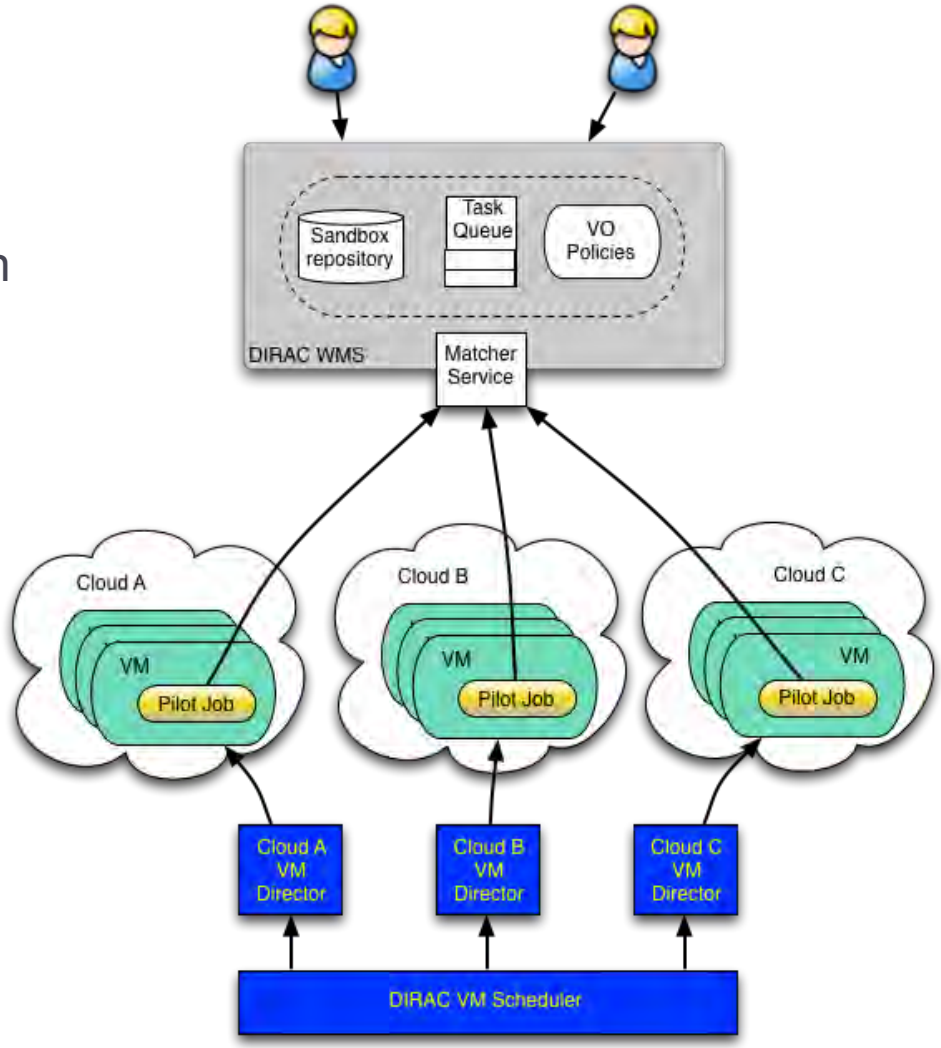
DIRAC computing resources

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

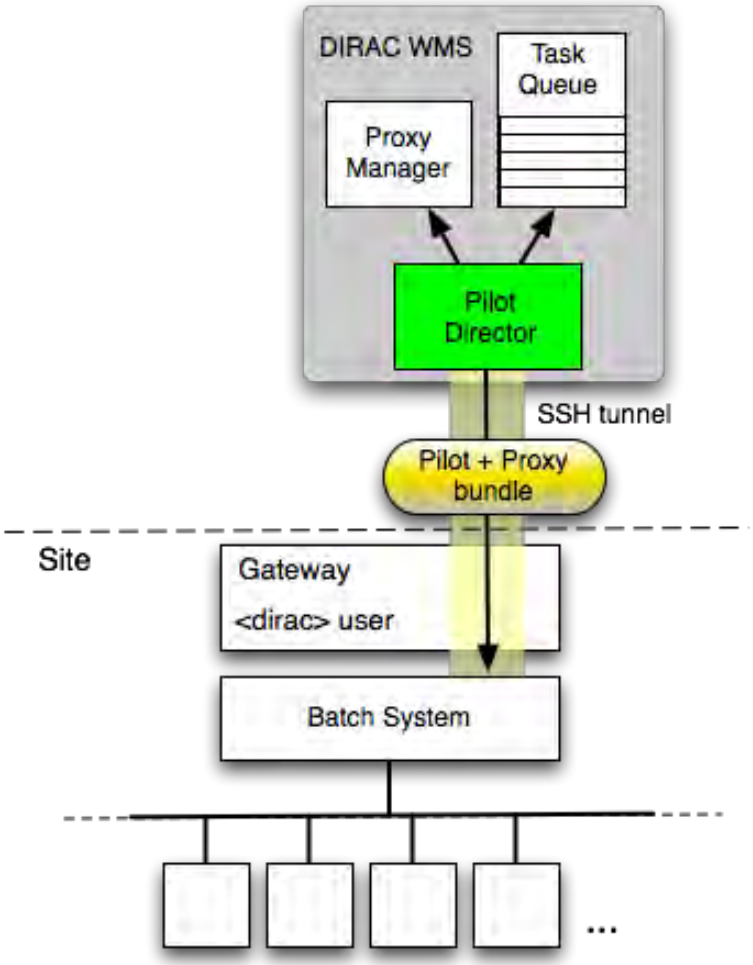
Clouds

- VM scheduler
  - Dynamic VM spawning taking 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
  - Apache ***cloudlib*** compliant clouds
  - ***OCCI*** compliant clouds:
    - OpenStack, OpenNebula
  - CloudStack
  - Amazon EC2
  - Vcycle, VAC
  - …
- ***cloudinit*** contextualization



Standalone computing clusters

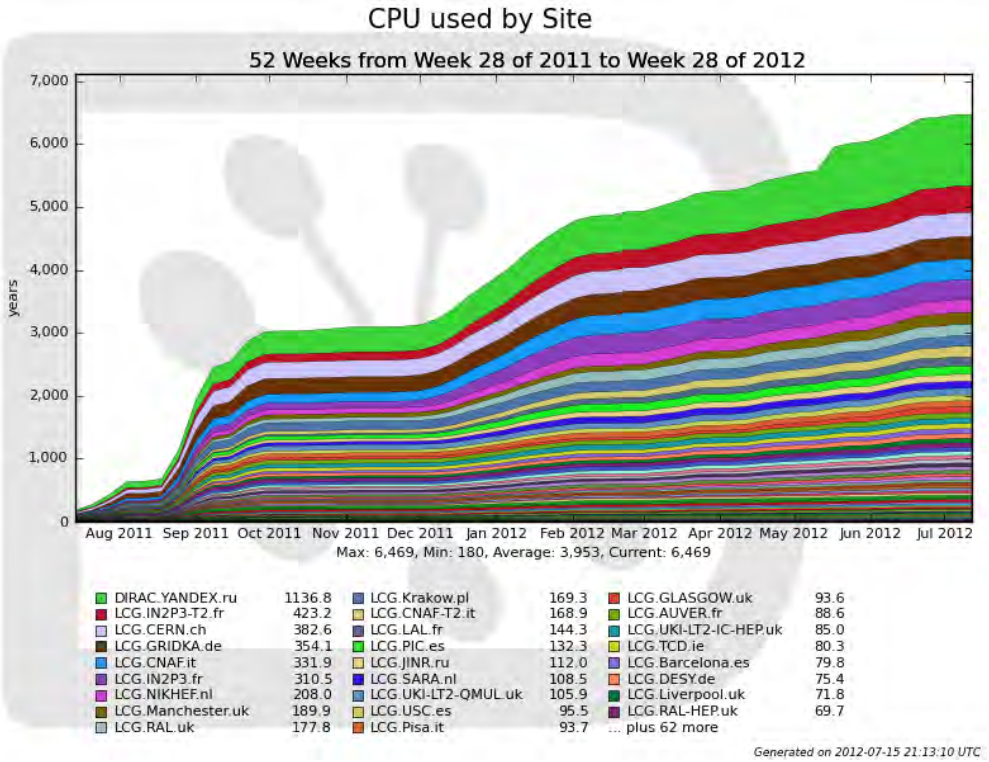
- Off-site Pilot Director
  - Site delegates control to the central service
  - Site must only define a dedicated local user account
  - The payload submission through an SSH tunnel
- The site can be:
  - a single computer or several computers without any batch system
  - a computing cluster with a batch system
    - LSF, BQS, SGE, PBS/Torque, Condor
      - Commodity computer farms
    - OAR, SLURM
      - HPC centers
- The user payload is executed with the owner credentials
  - No security compromises with respect to external services



Standalone computing clusters

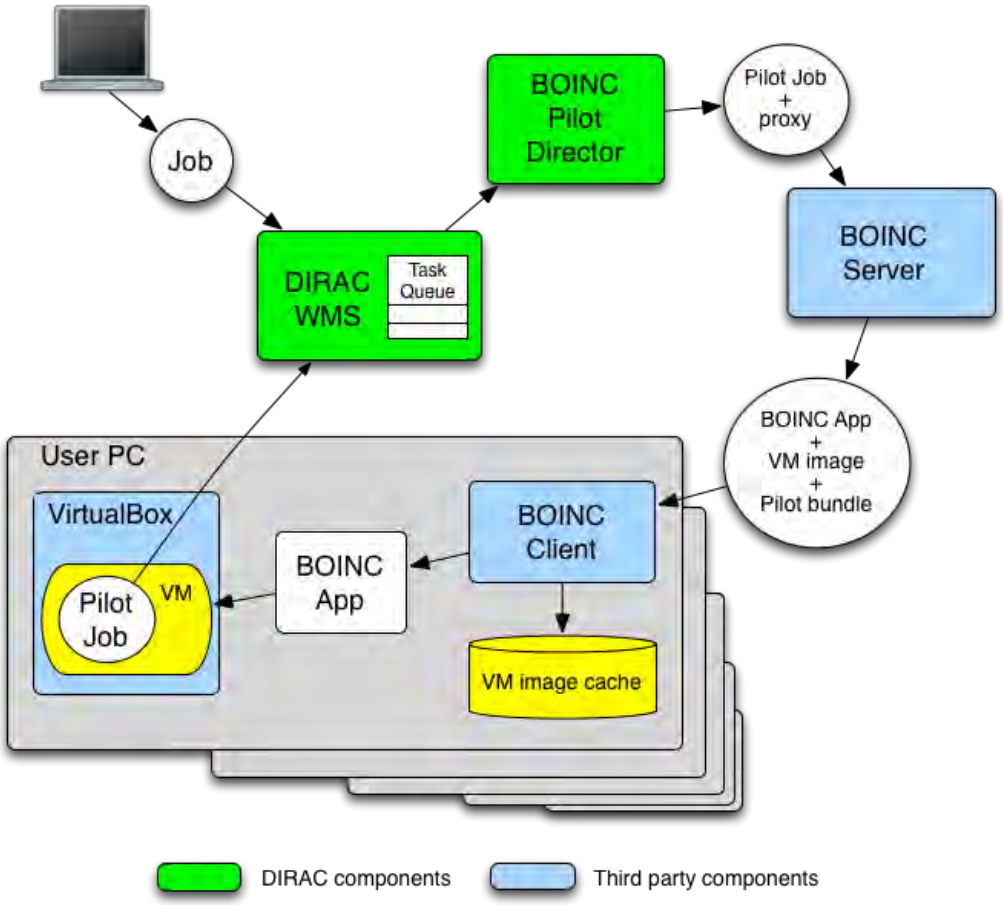
Examples:

- DIRAC.Yandex.ru
  - >2000 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)



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

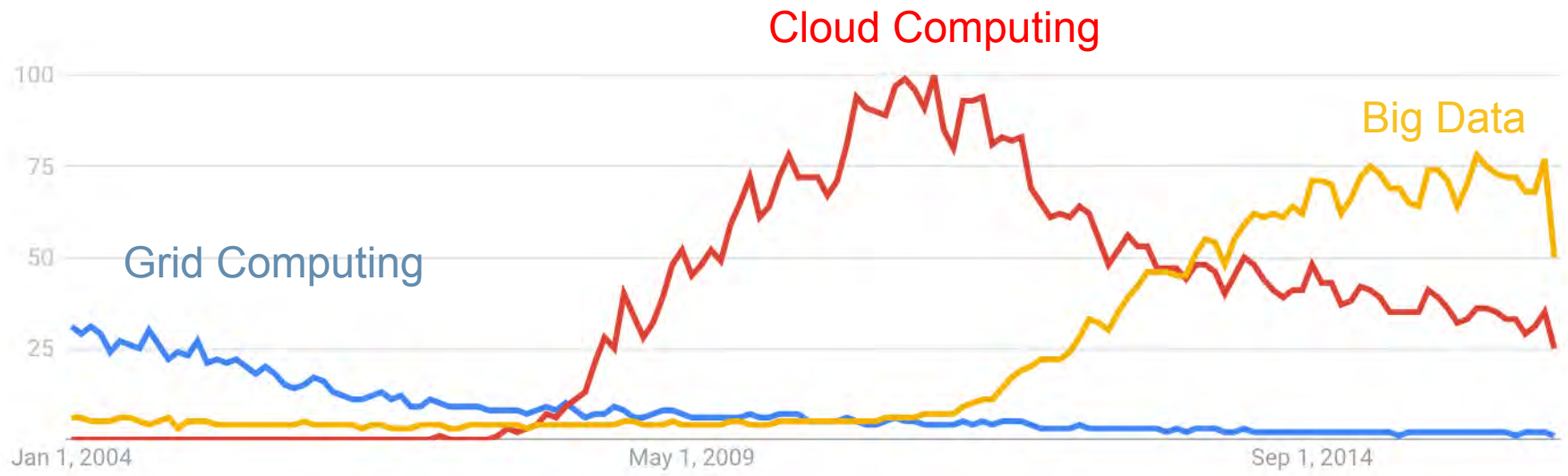


DIRAC
THE INTERWARE

Data Management

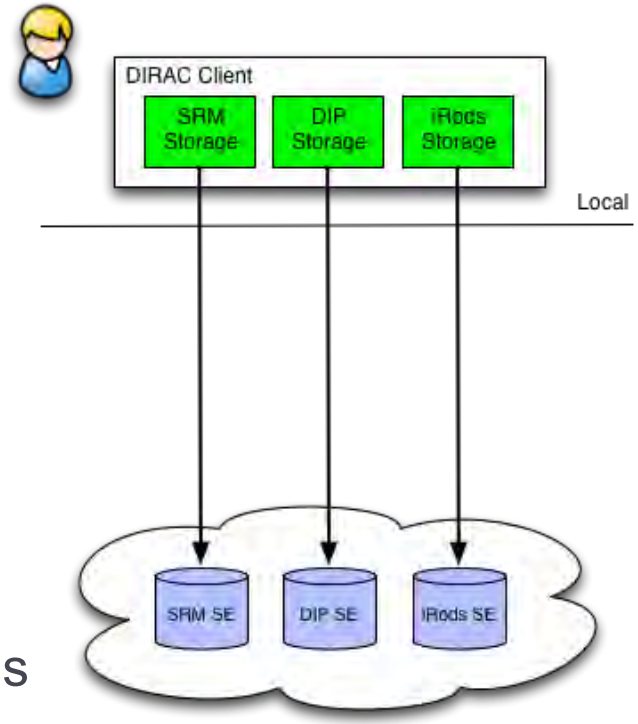
Big Data

- Data that exceeds the boundaries and sizes of normal processing capabilities, forcing you to take a non-traditional approach for the treatment
- Google trends:



Storage plugins

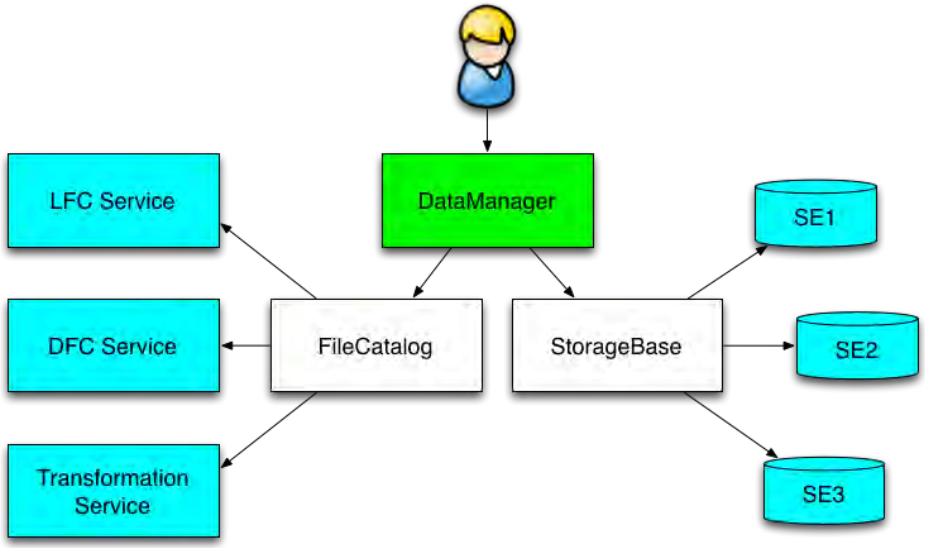
- Storage element abstraction with a client implementation for each access protocol
  - DIPS, SRM, XROOTD, RFIO, etc
  - gfal2 based plugin gives access to all protocols supported by the library
    - DCAP, WebDAV, S3, …



- 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, …) is maintaining a single global logical name space
- 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

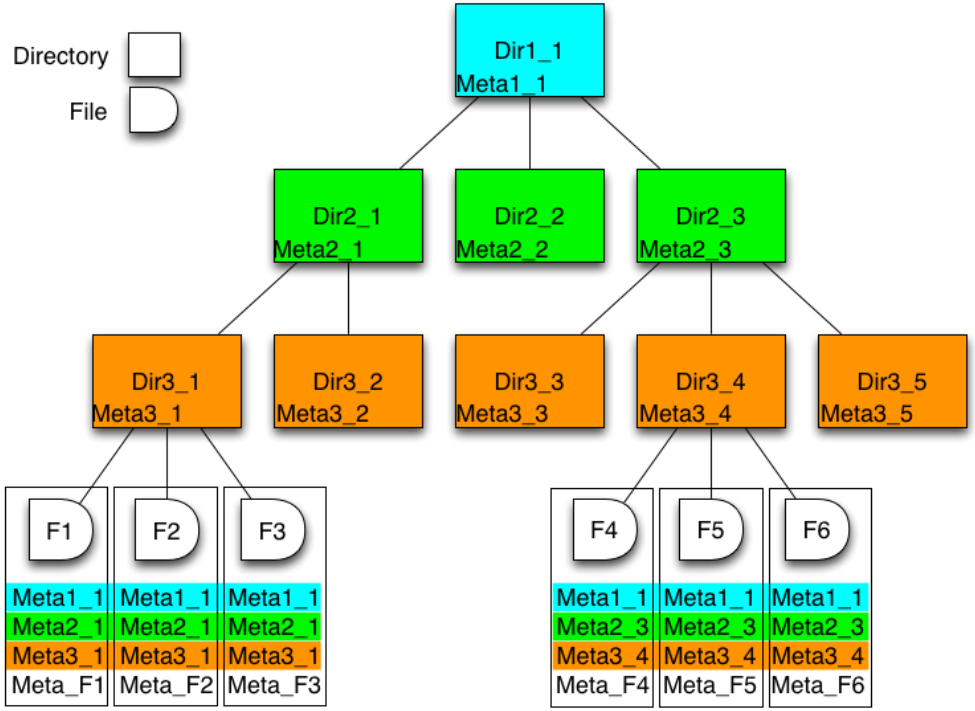


File Catalog

- DFC is the central component of the DIRAC Data Management system
- Defines the 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

File Catalog: Metadata

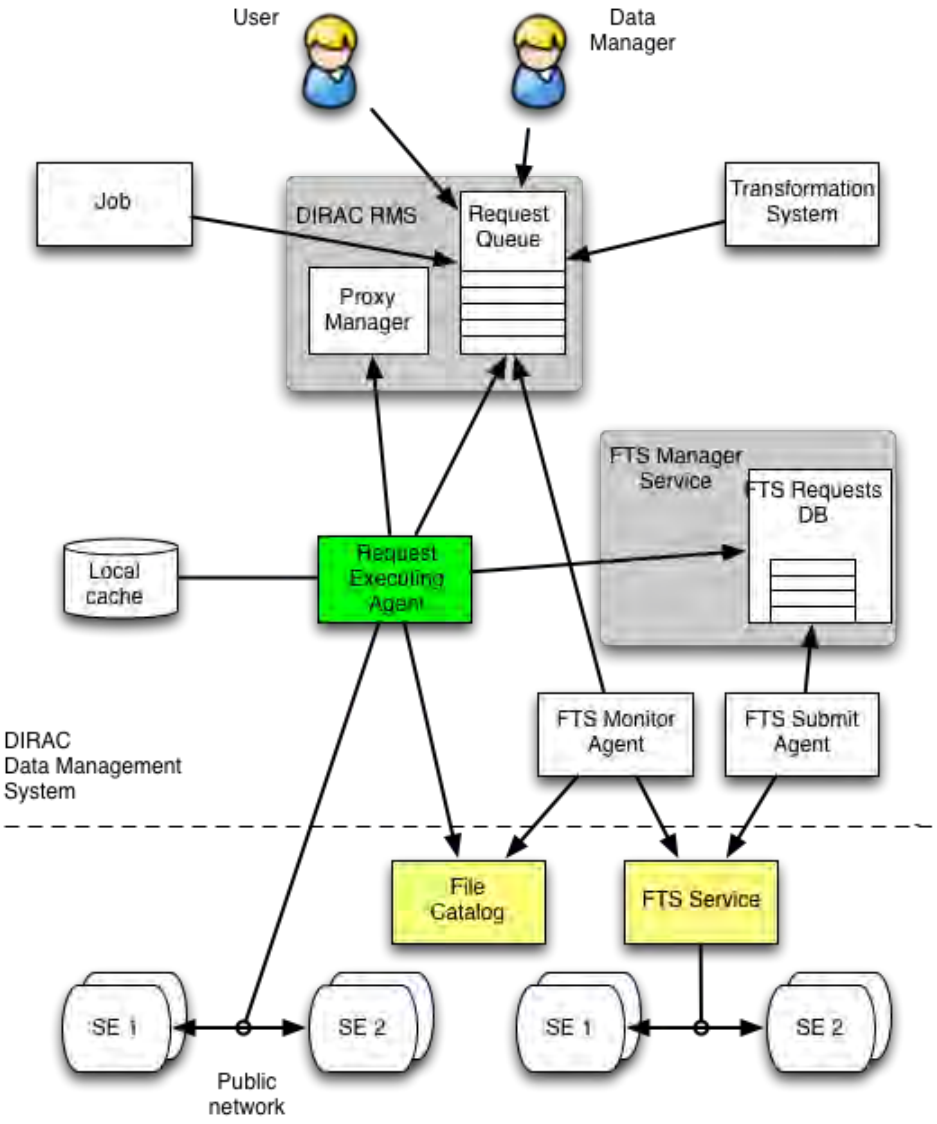
- DFC is Replica and Metadata Catalog
  - User defined metadata
  - The same hierarchy for metadata as for the logical name space
    - Metadata associated with files and directories
    - Allow for efficient searches
  - Efficient Storage Usage reports
    - Suitable for user quotas



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

Bulk operations

- Asynchronous data operations using Request Management System (RMS)
  - Placement, replication, removal
- Data driven operations using Transformation System (TS)
  - Automation of recurrent tasks
- The Replication Operation executor
  - Performs the replication itself or
  - Delegates replication to an external third party service, e.g.
    - FTS (developed at CERN)
    - EUDAT
    - OneData



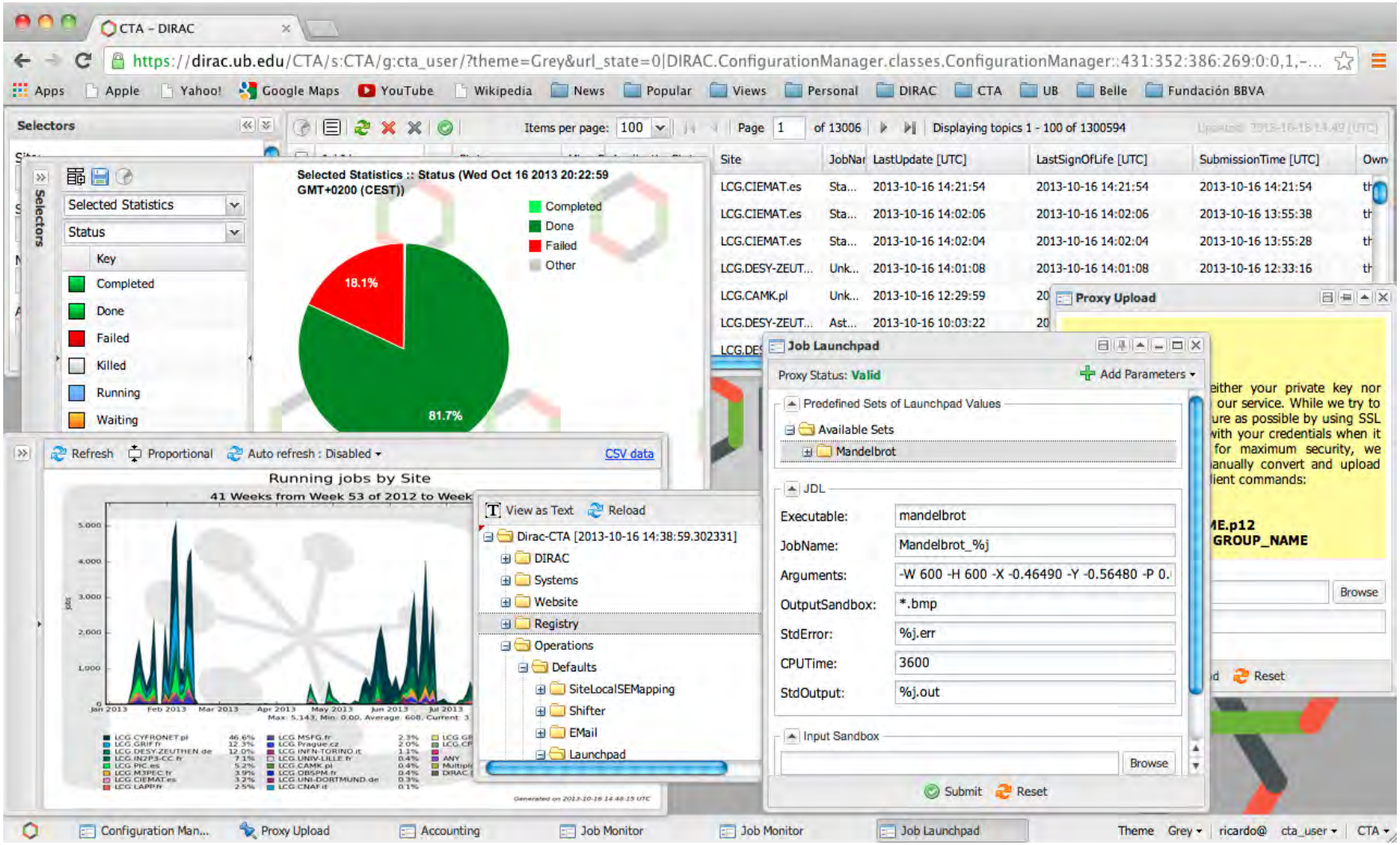
DIRAC
THE INTERWARE

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
- 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

Web Portal examples



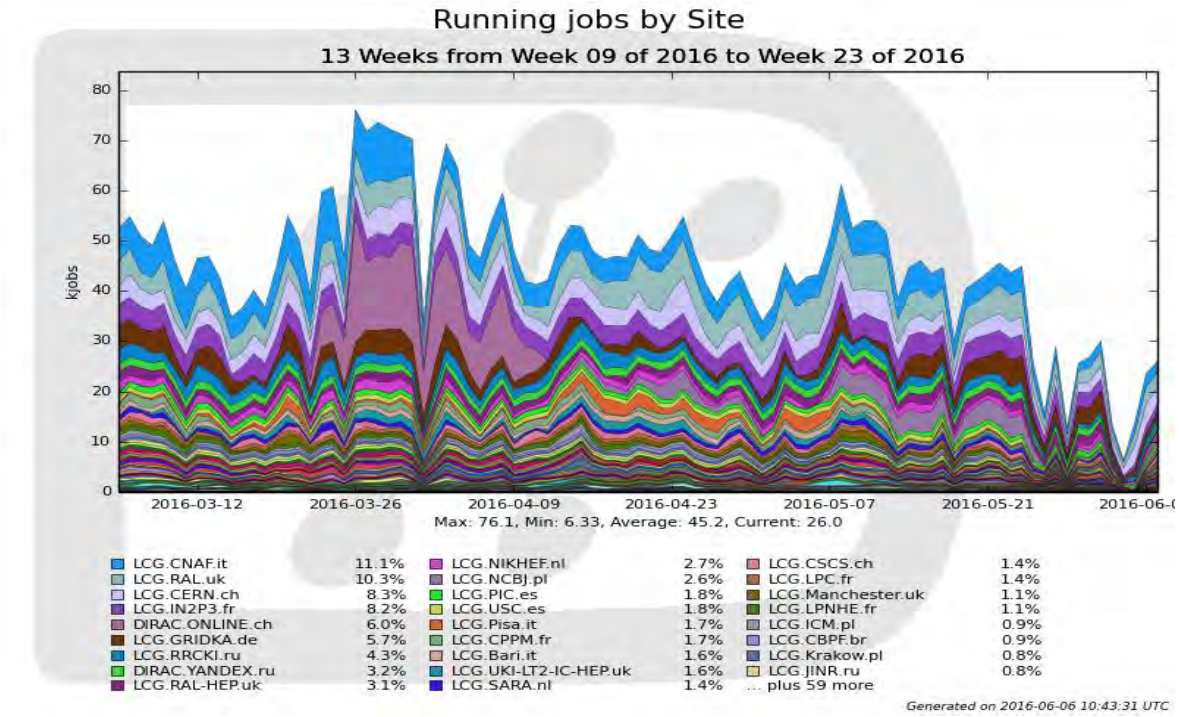
Distributed Computer

- 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
    - Standard tools
    - Application specific applications

DIRAC
THE INTERWARE

DIRAC Users: large communities

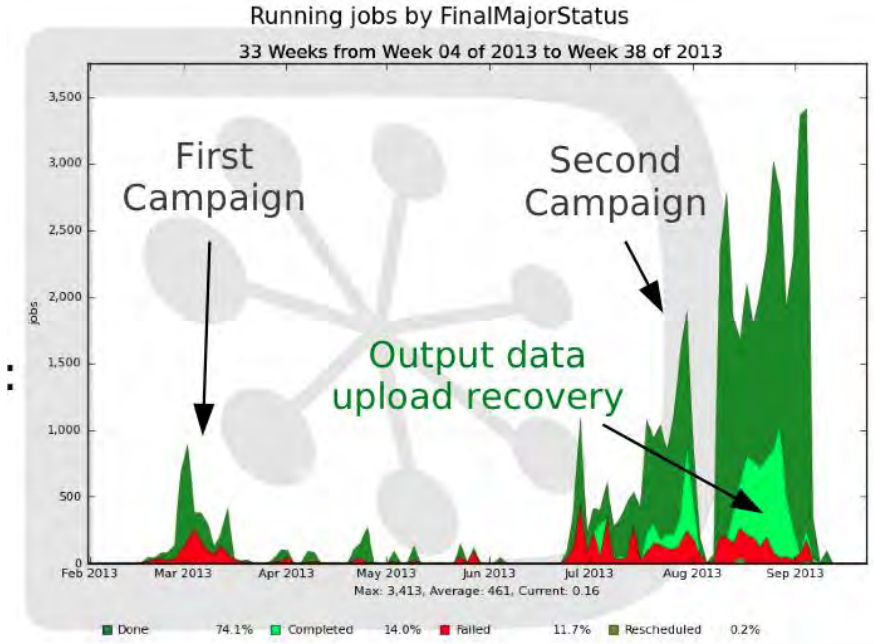
LHCb Collaboration



- About 600 researchers from 40 institutes
- Up to 100K concurrent jobs in ~120 distinct sites
  - Equivalent to running a virtual computing center with a power of 100K CPU cores, which corresponds roughly to ~ 1PFlops
  - Limited mostly by available capacity
- Further optimizations to increase the capacity are possible
  - Hardware, database optimizations, service load balancing, etc

Experiments: Belle II

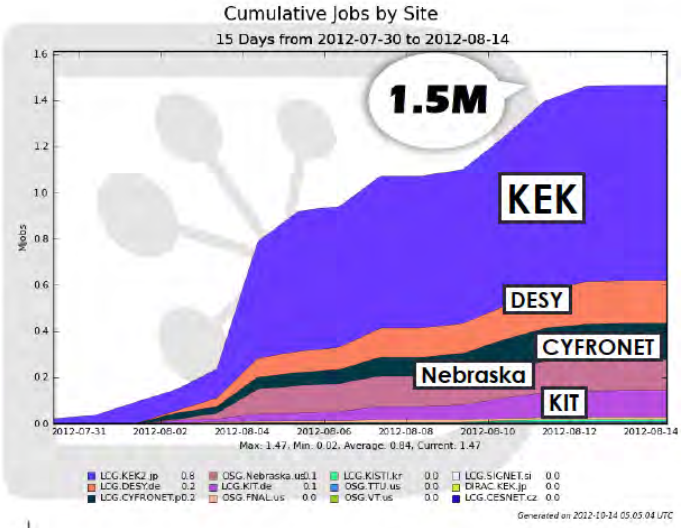
- Combination of the non-grid, grid sites and (commercial) clouds is a requirement
- 2 GB/s, 40 PB of data in 2019
- Belle II grid resources
  - WLCG, OSG grids
  - KEK Computing Center
  - Amazon EC2 cloud

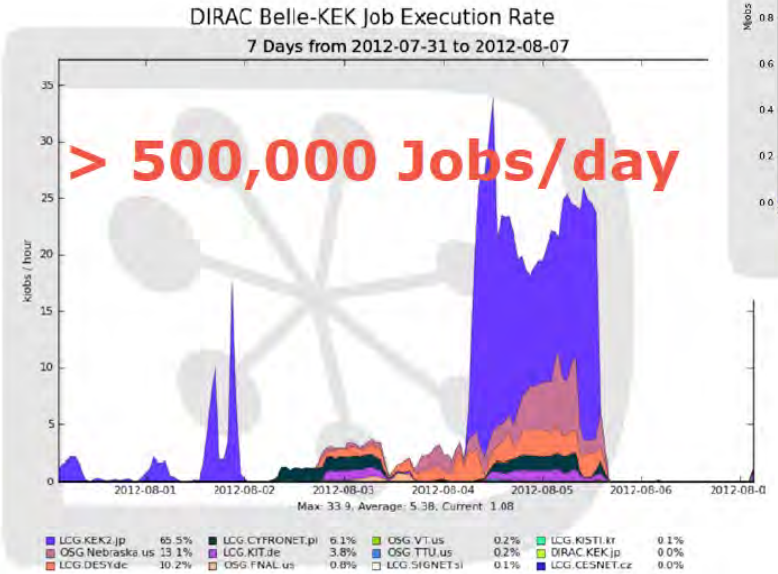


Thomas Kuhr, Belle II

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

Community installations

- 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
  - Basis for a multi-community service: Juno, CEPC
- 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

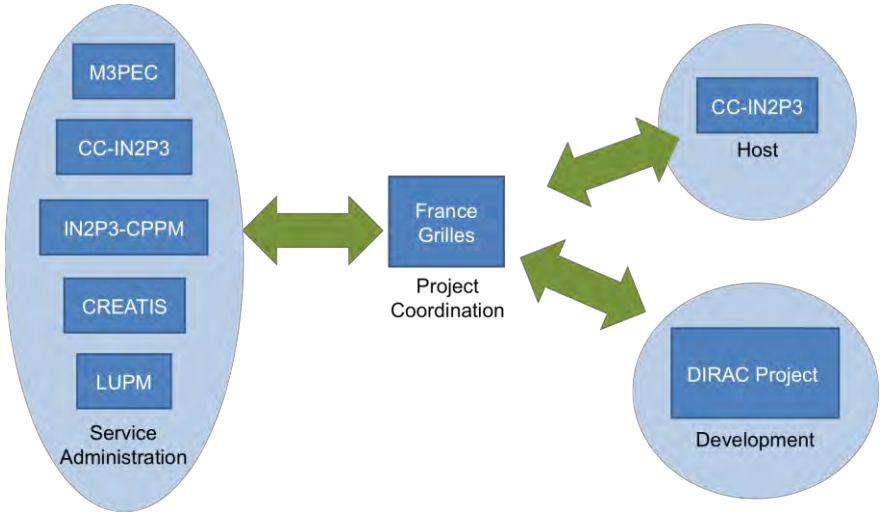
- DIRAC evaluations by other experiments
  - LSST, Pierre Auger Observatory, TREND, Juno, CEPC, NICA, ELI, …
  - Evaluations can be done with general purpose DIRAC services

DIRAC
THE INTERWARE

DIRAC as a Service

National services

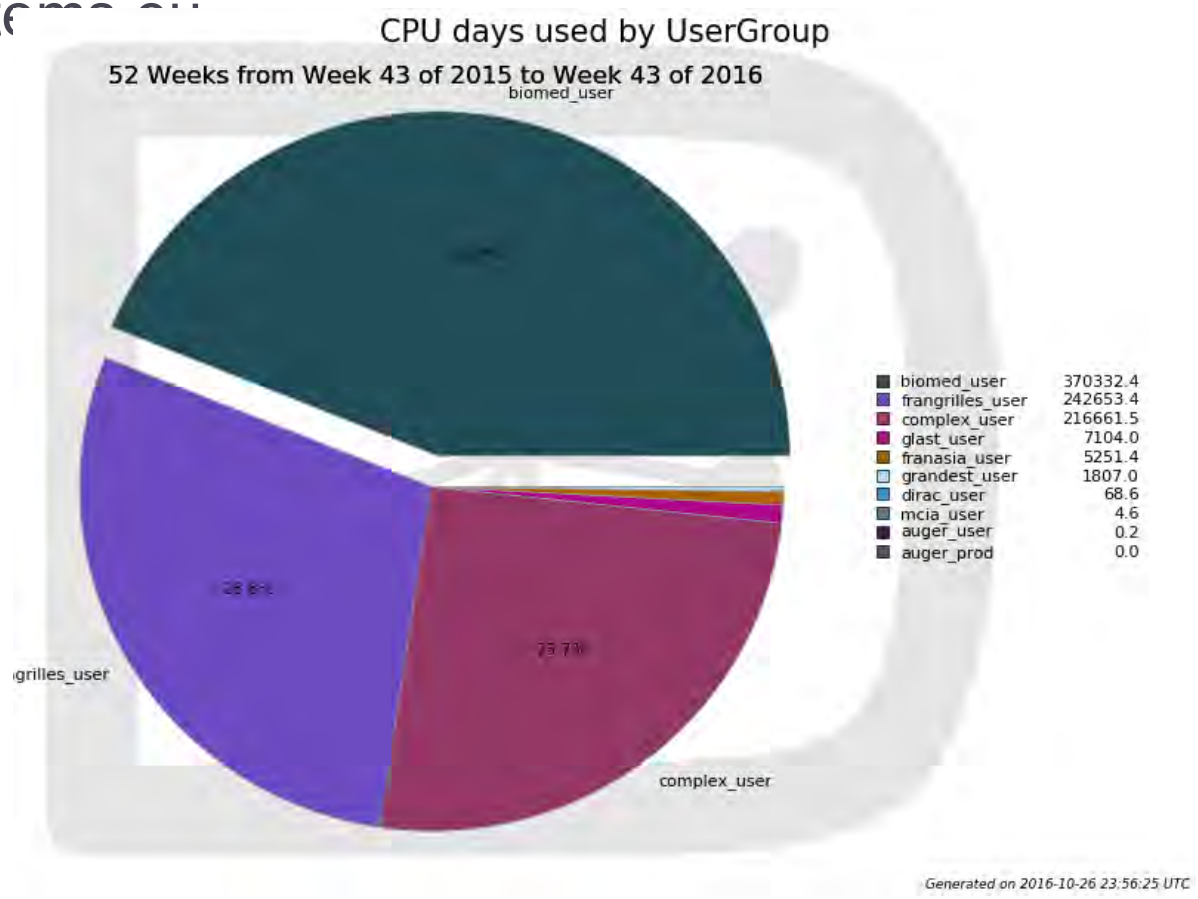
- **DIRAC** services are provided by several National Grid Initiatives: France, Spain, Italy, UK, China, Russia, …
  - Support for small communities
  - Heavily used for training and evaluation purposes
- Example: France-Grilles DIRAC service
  - Hosted by the CC/IN2P3, Lyon
  - Distributed administrator 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



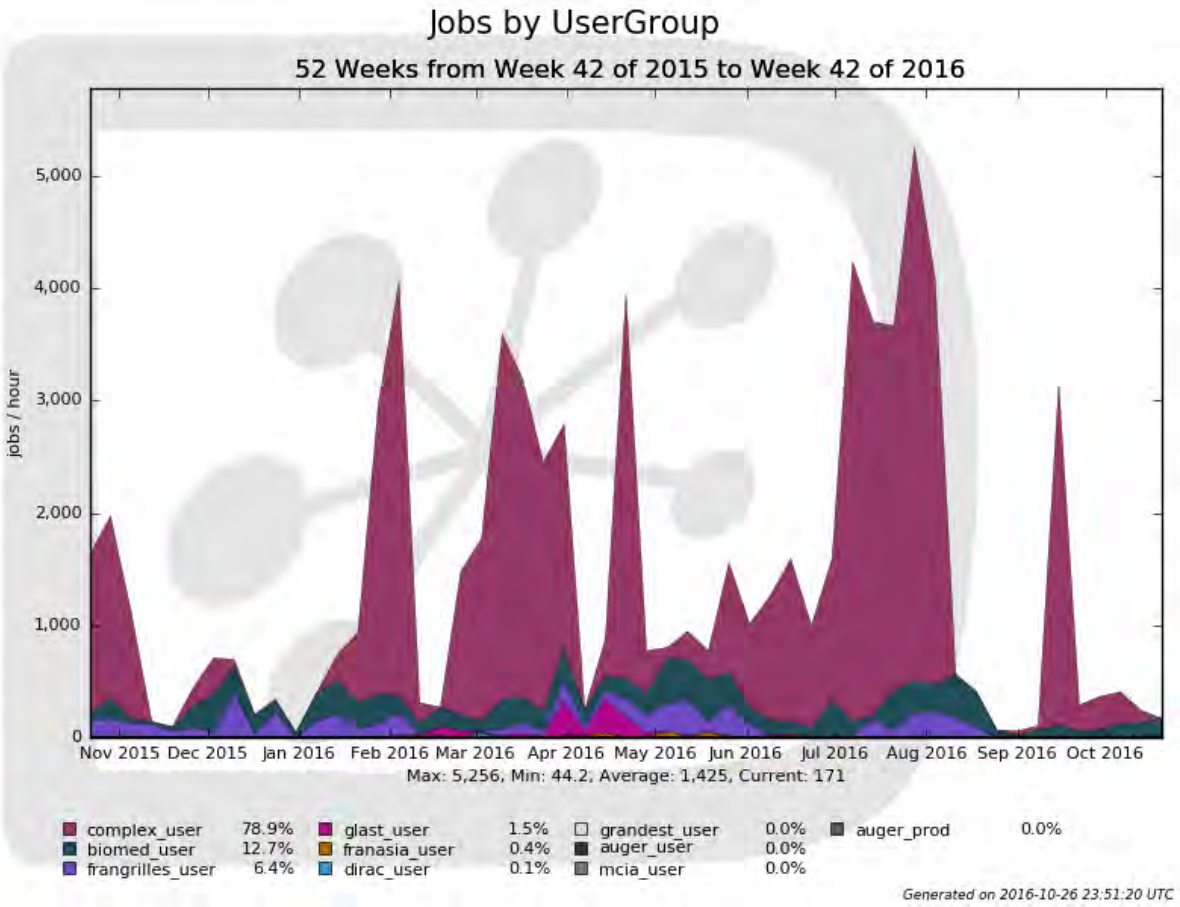
http://dirac.france-grilles.fr

FG-DIRAC CPU consumption

- > 2000 CPU years in the last year
  - Largest VO's: biomed, vo.france-grilles.fr, complex-syste



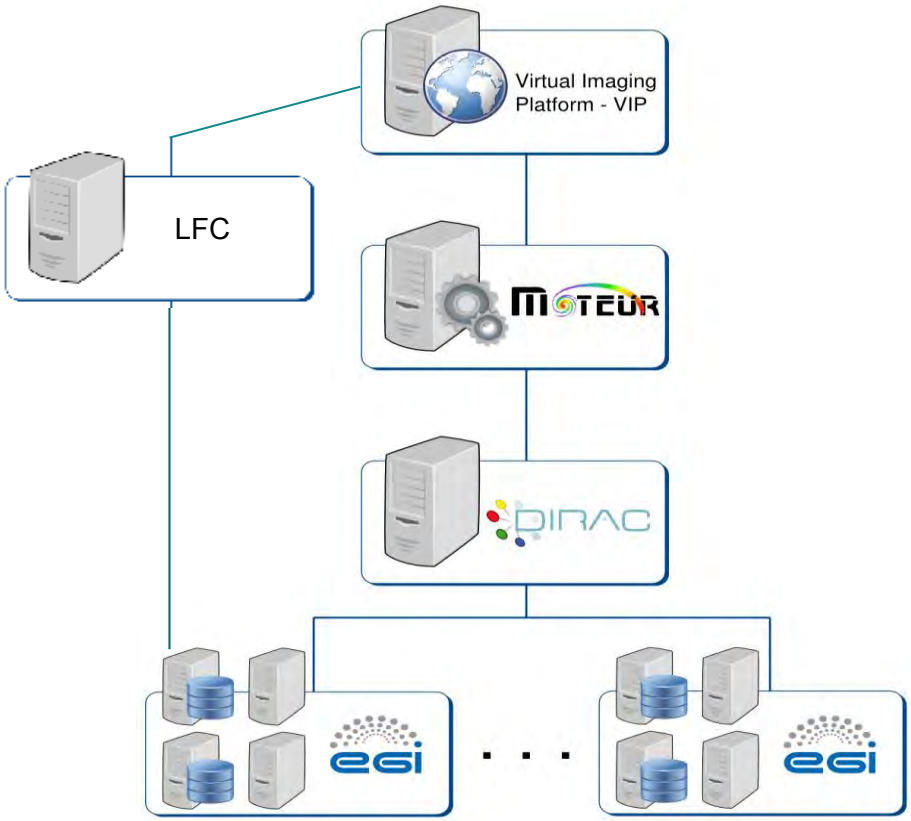
Job execution rate



- Up to 2 Hz job execution rate
  - 200K jobs per day
  - VO **complex-systems.eu**: workflows with large numbers of small jobs

Virtual Imaging Platform

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

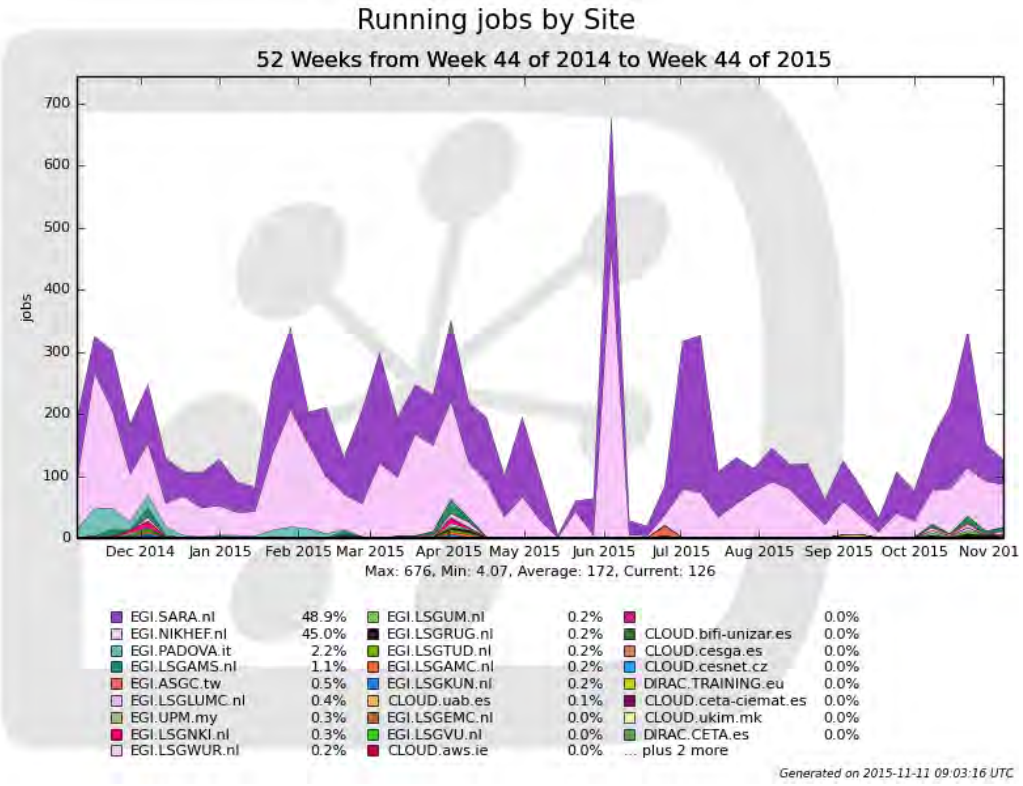


- Web portal with robot certificate
  - *File transfers, user/group/application management*
- Workflow engine
  - *Generate jobs, (re-)submit, monitor, replicate*
- DIRAC
  - *Resource provisioning, job scheduling*
- Grid resources
  - *biomed VO*

DIRAC4EGI service

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

DIRAC4EGI activity snapshot



DIRAC
THE INTERWARE

EGI ACCOUNTING PORTAL

Normalised CPU time [units 1K.SI2K.Hours] by DATE and VO												
DATE	alice	atlas	belle	biomed	cms	compchem	ilc	lhcb	virgo	vo.cta.in2p3.fr	Total	%
Nov 2015	83,043,071	213,187,021	29,633,040	2,992,249	107,998,028	812,409	3,051,240	44,495,710	365,193	5,203,790	**490,781,751**	**8.60%**
Dec 2015	81,681,064	167,642,164	30,755,315	2,771,463	81,200,999	1,197,402	10,250,775	42,772,247	4,370	9,643,804	**427,919,603**	**7.50%**
Jan 2016	100,472,899	212,596,116	8,254,706	2,221,994	99,768,667	2,869,544	3,904,455	32,614,451	329,113	8,746,790	**471,778,735**	**8.27%**
Feb 2016	80,340,391	202,531,157	48,965	1,312,309	100,330,129	1,220,127	2,704,948	44,547,976	1,962,465	5,563,528	**440,561,995**	**7.72%**
Mar 2016	108,810,699	172,663,251	3,412,262	2,286,939	75,113,354	1,623,540	2,049,130	83,154,401	1,917,611	1,539,919	**452,571,106**	**7.93%**
Apr 2016	111,707,745	211,516,946	496,969	1,622,314	67,855,621	1,970,394	3,051,624	78,821,567	3,517,152	3,079,316	**483,639,648**	**8.47%**
May 2016	88,434,699	229,055,135	457,771	3,055,283	64,161,648	3,990,478	4,366,309	70,550,242	11,311,493	669,299	**476,052,357**	**8.34%**
Jun 2016	91,963,895	220,222,321	10,039,317	1,375,916	104,040,606	1,755,334	2,097,169	66,545,602	2,558,741	1,103,183	**501,702,084**	**8.79%**
Jul 2016	113,408,142	187,198,001	3,614,046	2,152,445	104,373,741	1,614,892	1,596,155	65,898,735	8,005,698	7,794,153	**495,656,008**	**8.69%**
Aug 2016	88,278,412	212,942,846	34,225	6,500,219	51,366,225	3,474,177	5,538,912	72,803,805	2,919,127	5,410,036	**449,267,984**	**7.87%**
Sep 2016	88,164,653	309,040,532	7,314,602	514,897	90,018,815	2,602,763	3,297,430	106,365,999	1,770,213	6,487,567	**615,577,471**	**10.79%**
Oct 2016	68,902,764	167,532,717	1,528,430	467,733	82,329,281	1,301,416	5,324,702	71,019,670	2,752,272	104,325	**401,263,310**	**7.03%**
Total	**1,105,208,434**	**2,506,128,207**	**95,589,648**	**27,273,761**	**1,028,557,114**	**24,432,476**	**47,232,849**	**779,590,405**	**37,413,448**	**55,345,710**	**5,706,772,052**	
Percentage	**19.37%**	**43.91%**	**1.68%**	**0.48%**	**18.02%**	**0.43%**	**0.83%**	**13.66%**	**0.66%**	**0.97%**		

- 5 out of Top-10 EGI communities used heavily DIRAC for their payload management in the last year
  - 4 out of 6 top communities excluding LHC experiments
    - belle, biomed, ilc, vo.cta.in2p3.fr
  - VO **compchem** is a pilot community to try out gLite WMS replacement by the DIRAC WMS

Conclusions

- The computational grids, clouds, HPC and volunteer 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 and storage 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