DIRAC
FG-DIRAC

A. Tsaregorodtsev,
CPPM-IN2P3-CNRS, Marseille
Workshop Opérations France-Grilles
8-10 novembre 2016

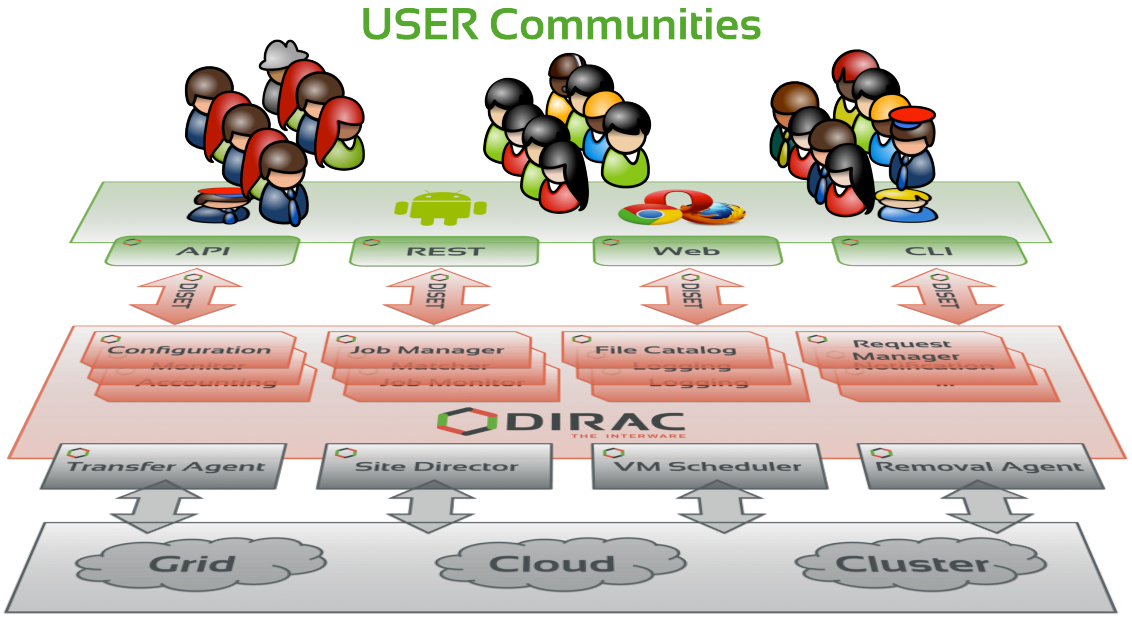
DIRAC
THE INTERWARE

Plan

- DIRAC Overview
- France-Grilles DIRAC service
- EGI DIRAC service
- Conclusions

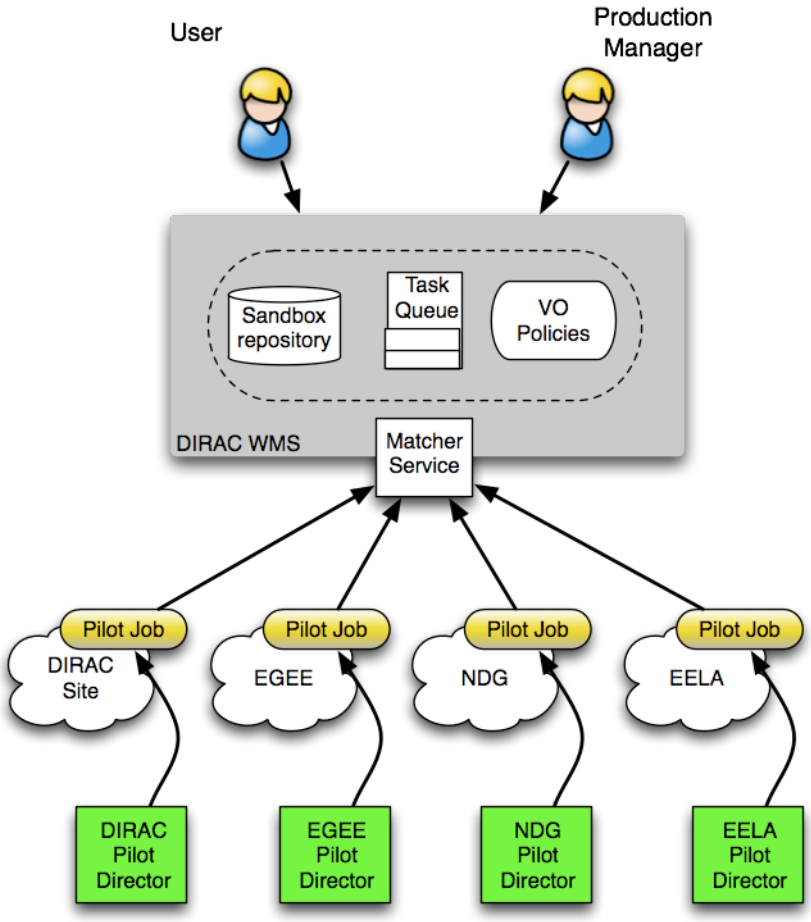
Interware

- 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***.



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

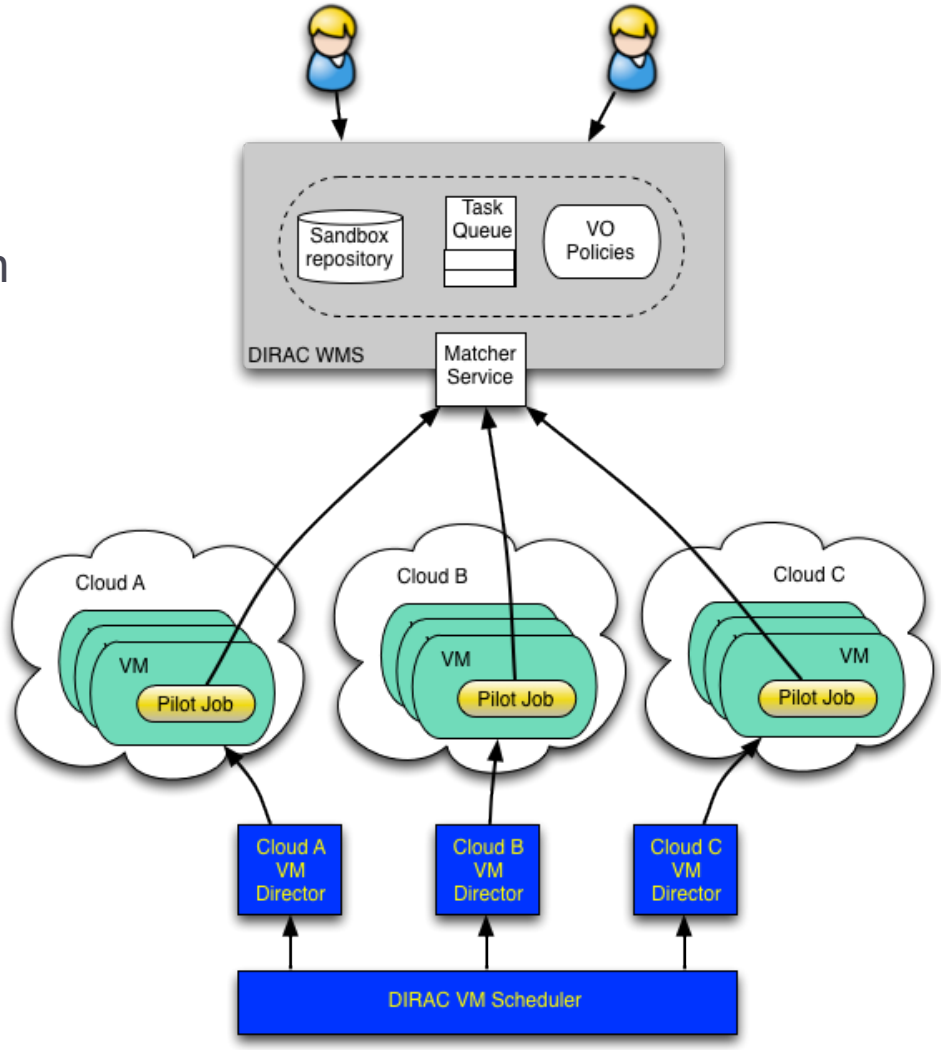


Computing Grids

- 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

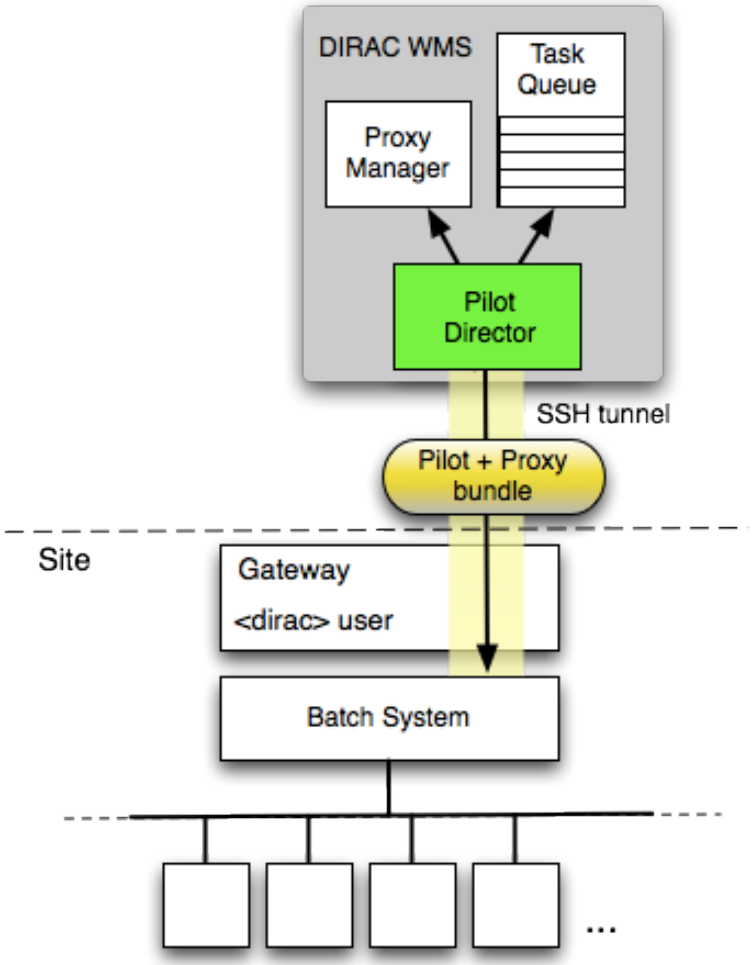
Clouds

- 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

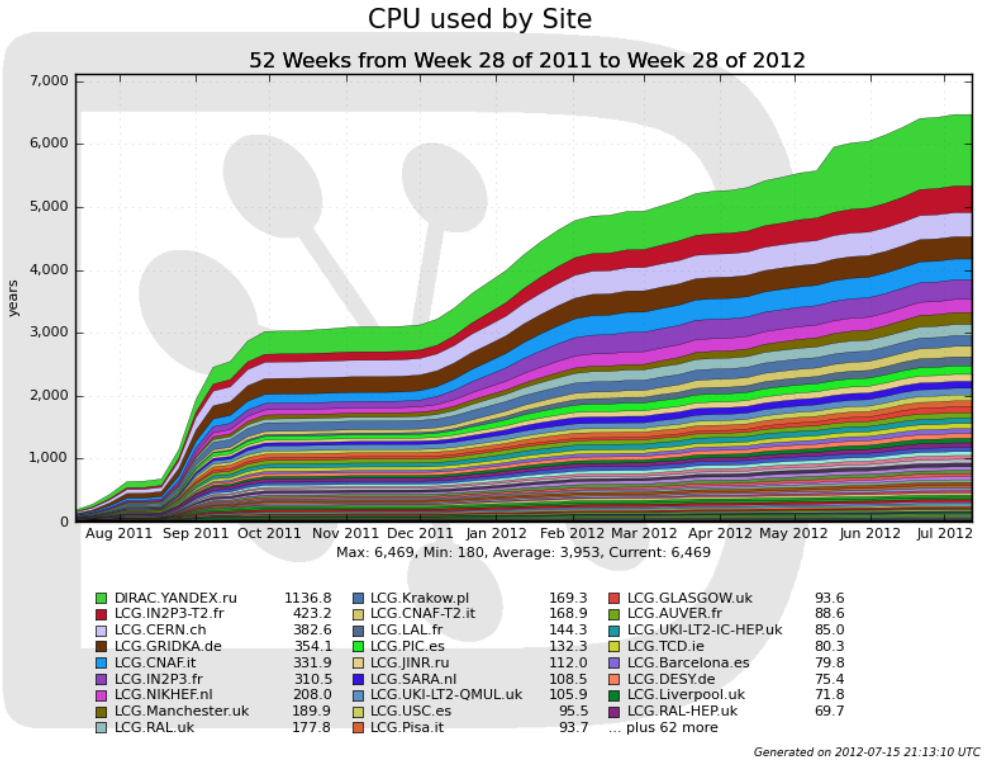
- Off-site Pilot Director
  - Site delegates control to the central service
  - Site must only define a dedicated local user account
  - The payload submission through an SSH tunnel
- The site can be:
  - a single computer or several computers without any batch system
  - a computing cluster with a batch 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

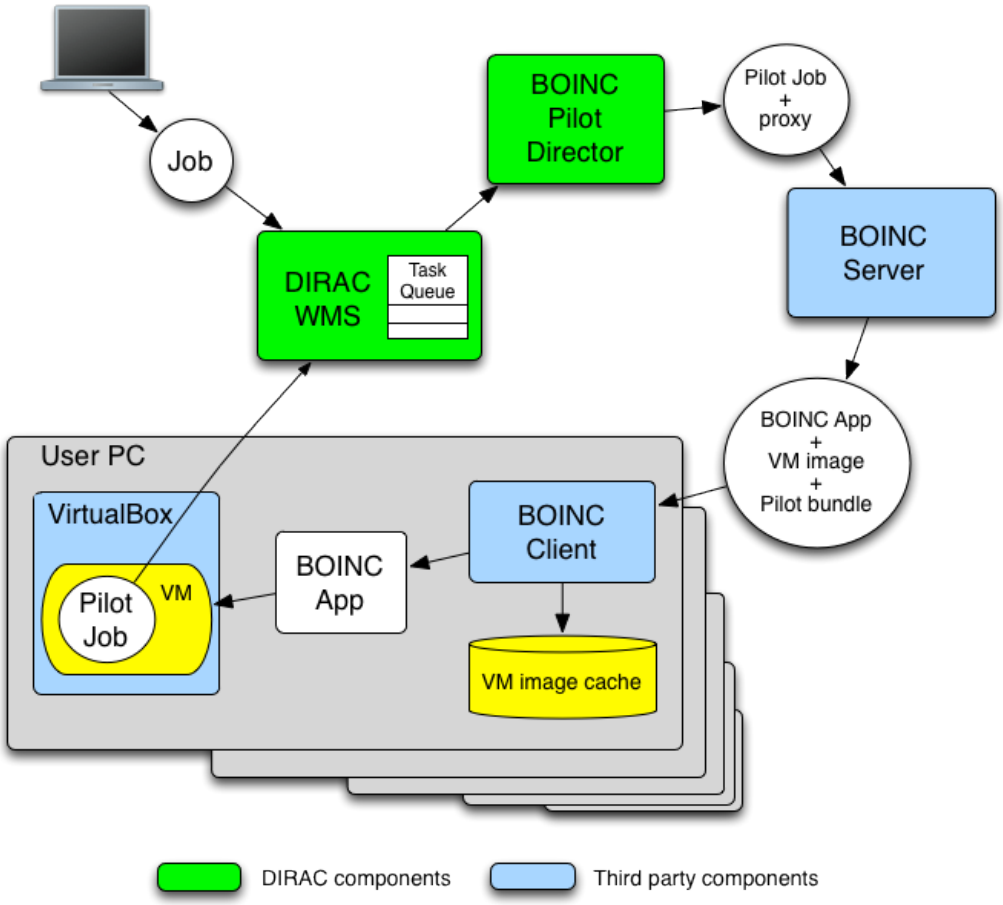
Examples:

- 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)



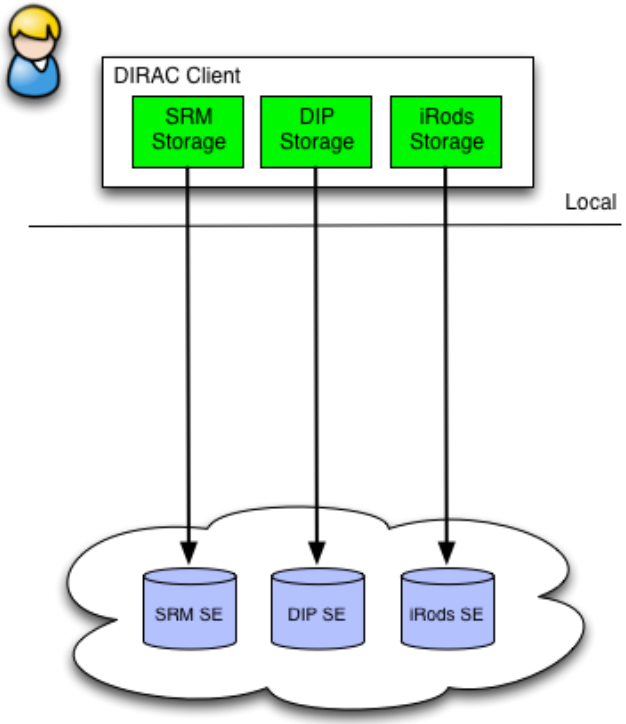
BOINC Desktop Grids

- 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



Storage plugins

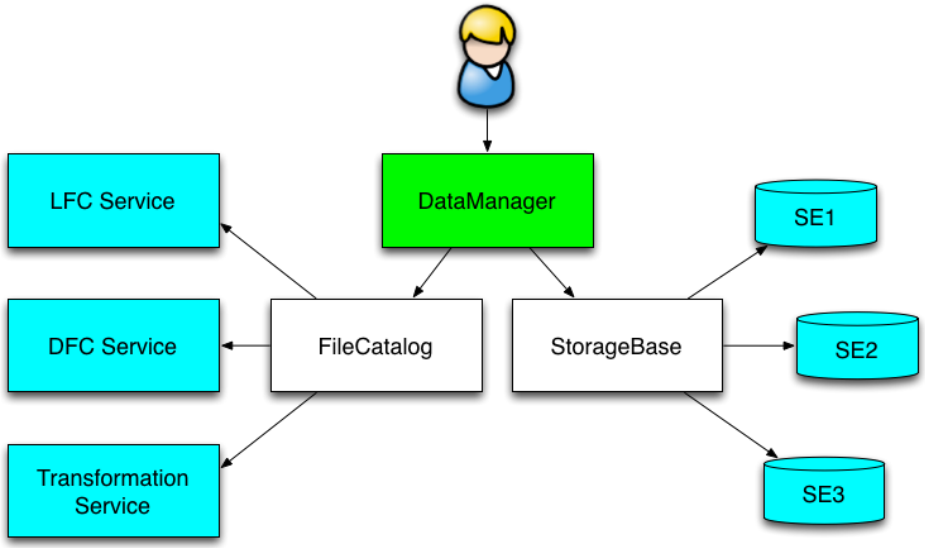
- Storage element abstraction with a client implementation for each access protocol
  - DIPS, SRM, XROOTD, RFIO, etc
  - gfal2 based plugin gives access to all protocols supported by the library
    - DCAP, WebDAV, S3, …



- 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



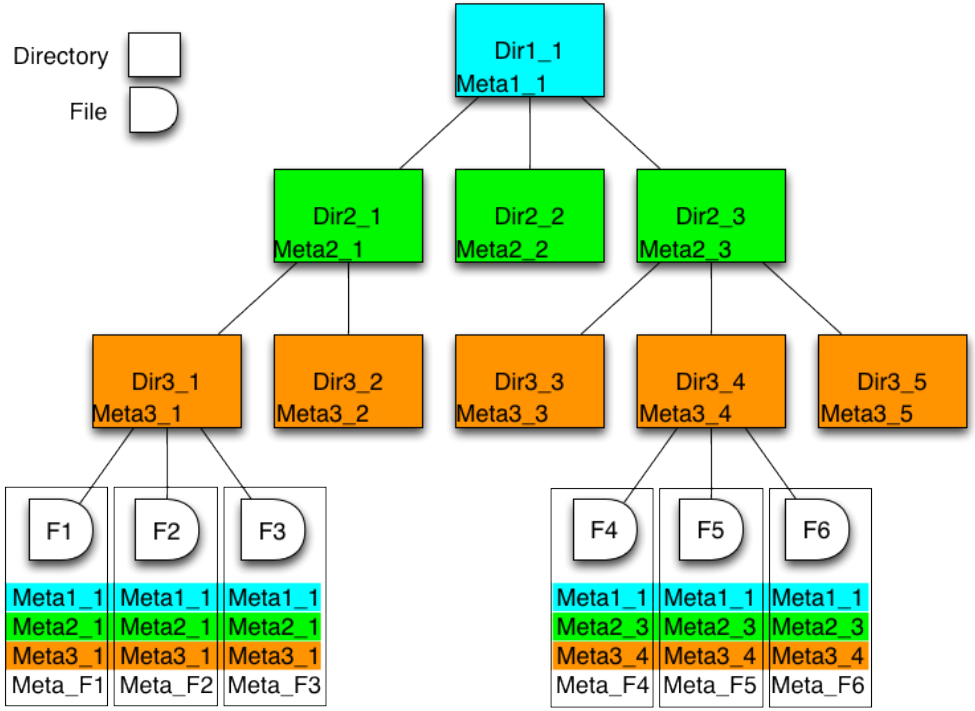
DIRAC
THE INTERWARE

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

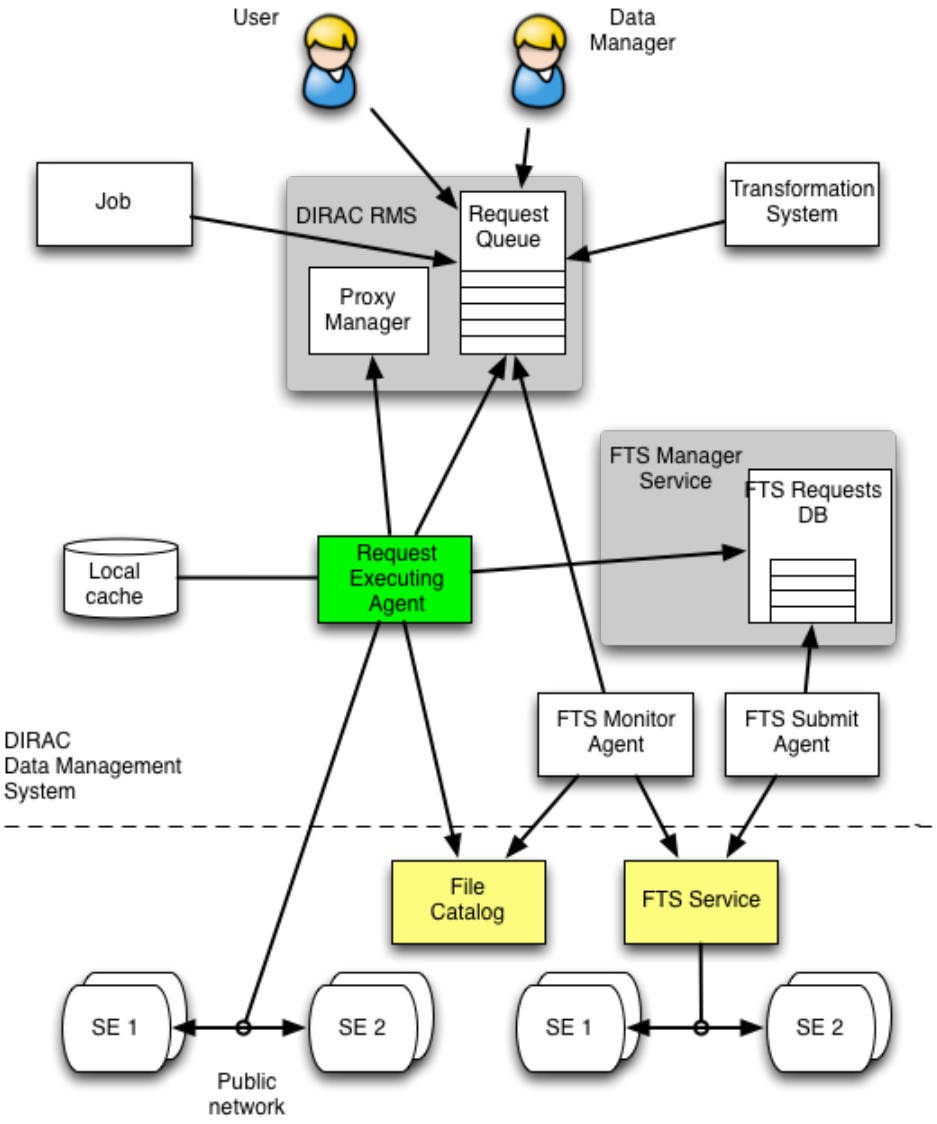
- DFC is Replica and Metadata Catalog
  - User defined metadata
  - The same hierarchy for metadata as for the logical name space
    - Metadata associated with files and directories
    - Allow for efficient searches
  - Efficient Storage Usage reports
    - Suitable for user quotas



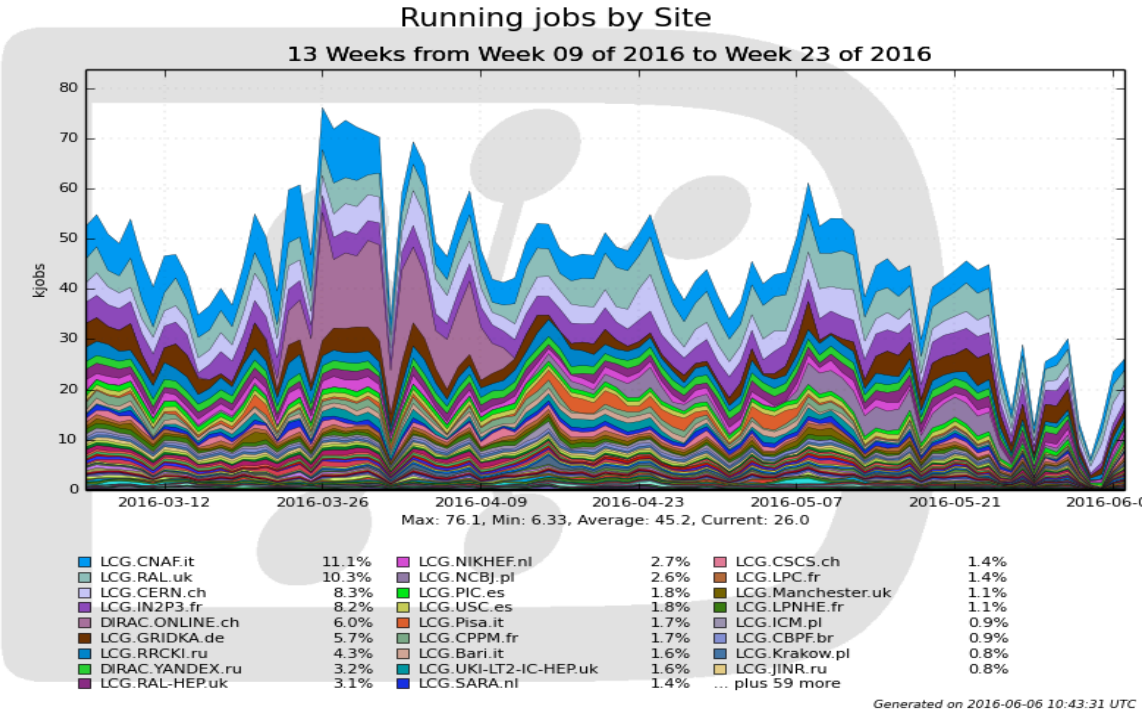
- Example query:
  - `find /lhcb/mcdata LastAccess < 01-01-2012 GaussVersion=v1,v2 SE=IN2P3,CERN Name=*.raw`

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



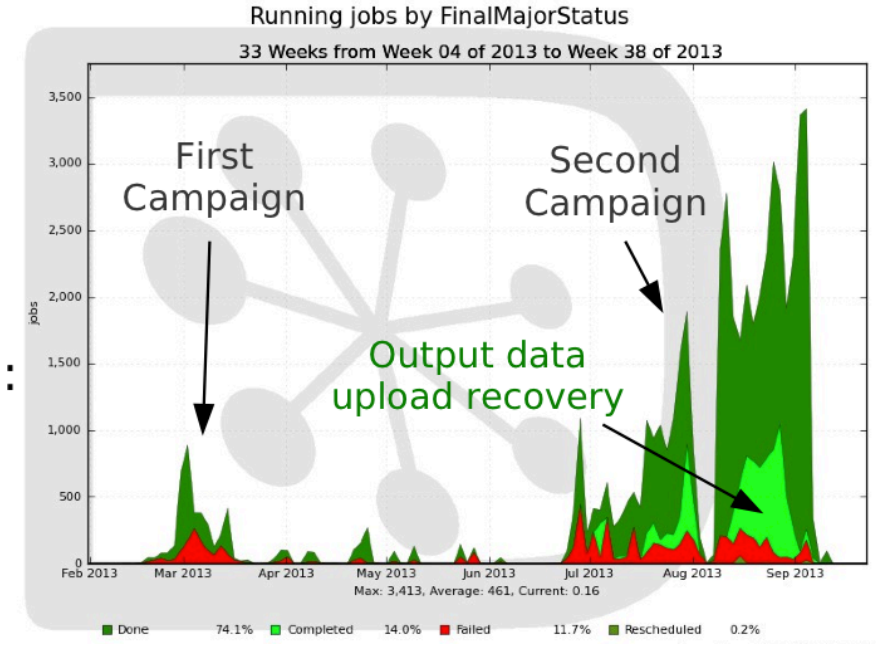
LHCb Collaboration



- 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

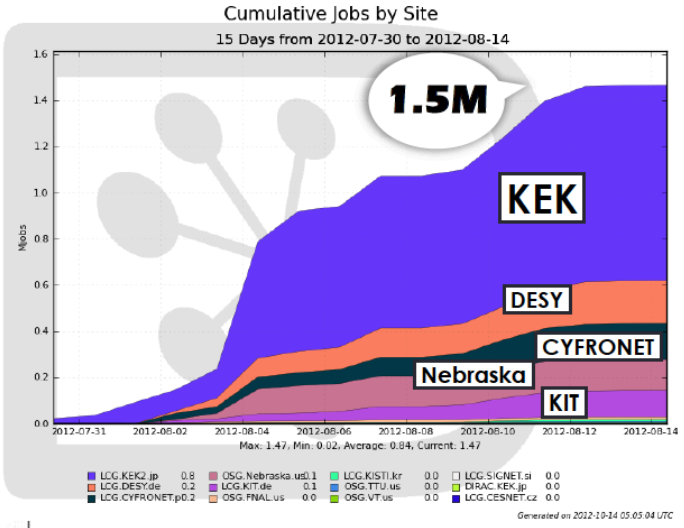
- Combination of the non-grid, grid sites and (commercial) clouds is a requirement
- 2 GB/s, 40 PB of data in 2019
- Belle II grid resources
  - WLCG, OSG grids
  - KEK Computing Center
  - Amazon EC2 cloud

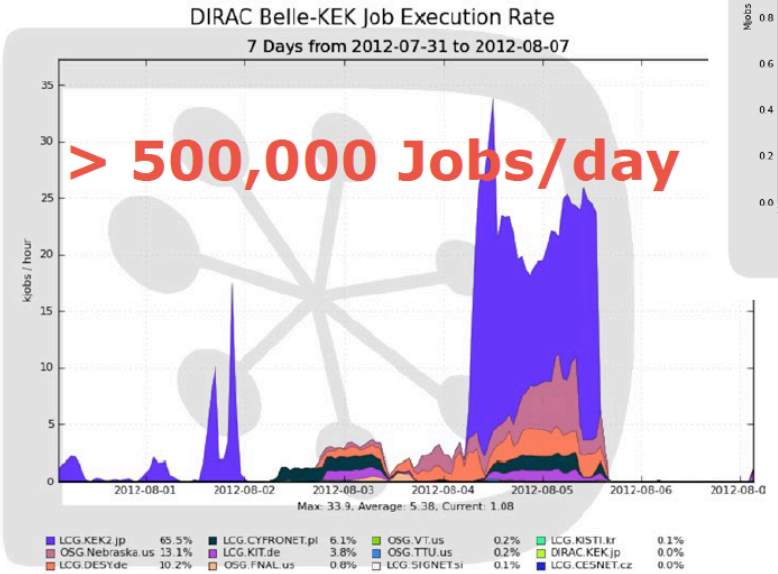


Thomas Kuhr, Belle II

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

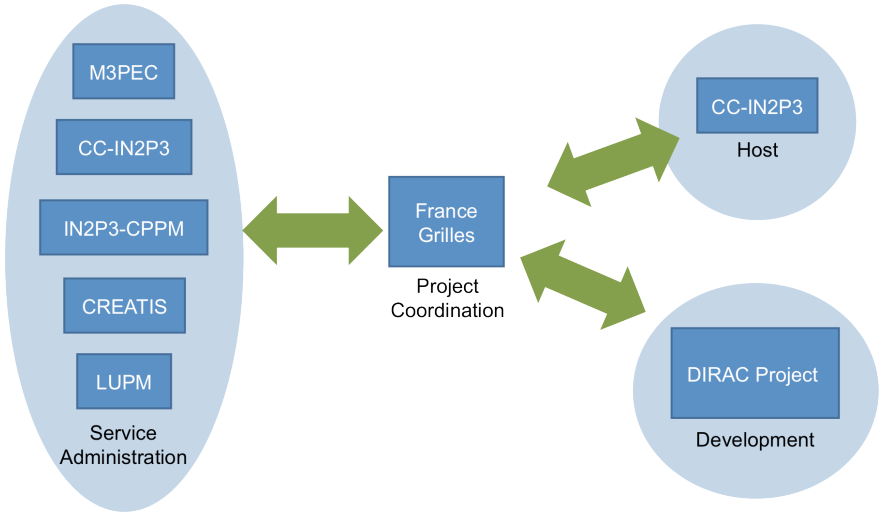
DIRAC
THE INTERWARE

DIRAC as a Service

National services

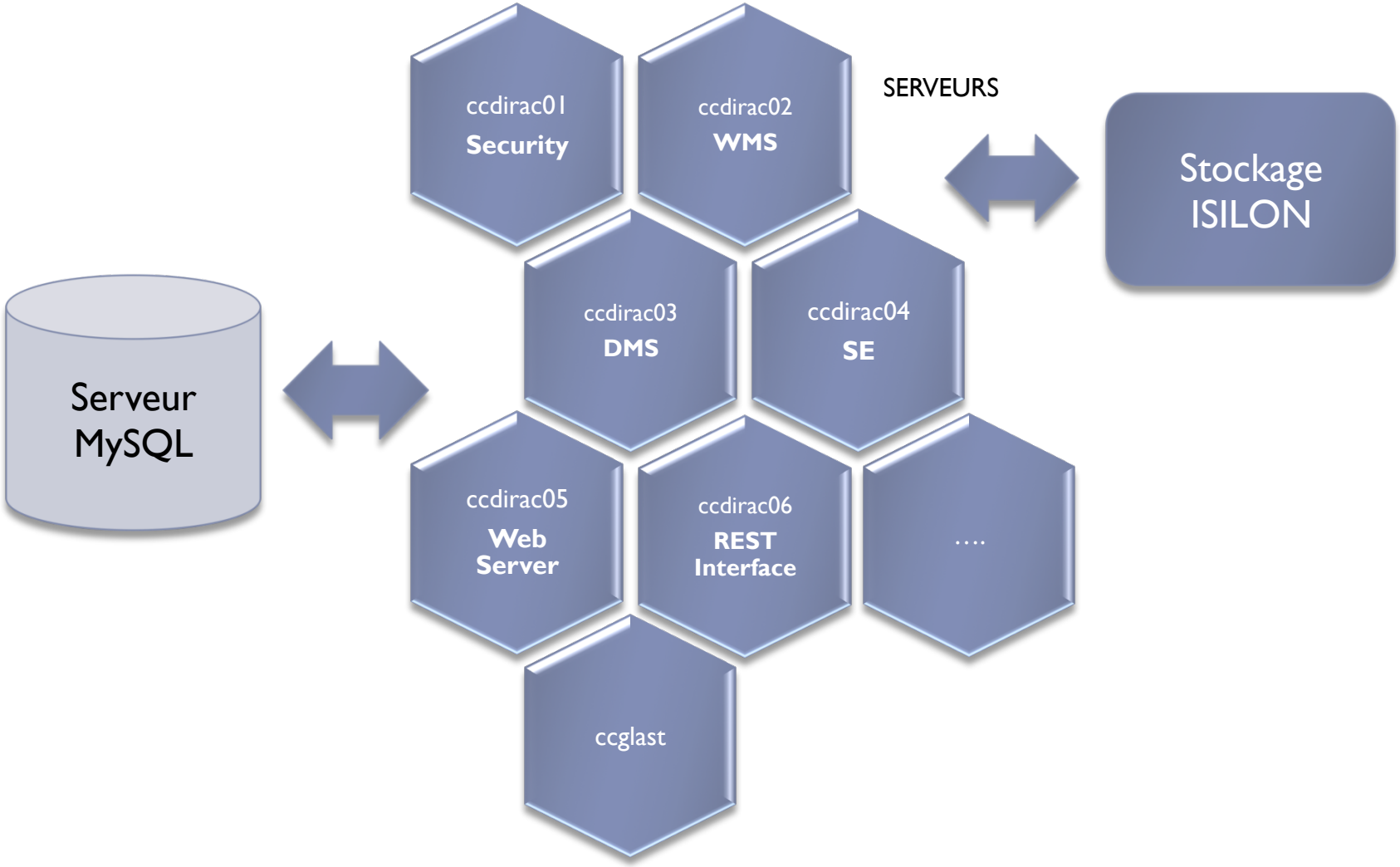
- **DIRAC** services are provided by several National Grid Initiatives: France, Spain, Italy, UK, China, Russia, …
  - Support for small communities
  - Heavily used for training and evaluation purposes

- Example: France-Grilles DIRAC service
  - Hosted by the CC/IN2P3, Lyon
  - Distributed administrator team
    - 5 participating universities
  - 23 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

Hosting hardware in CC/IN2P3

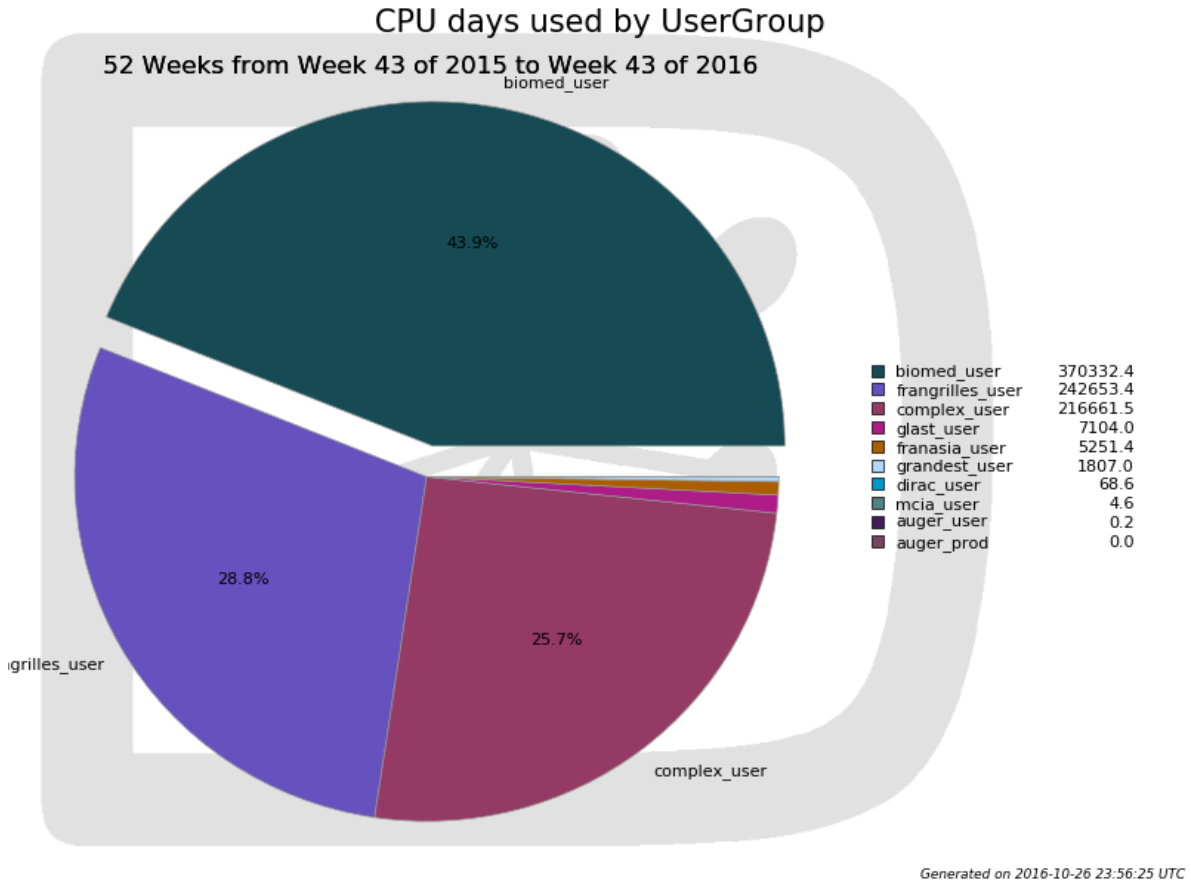


Services

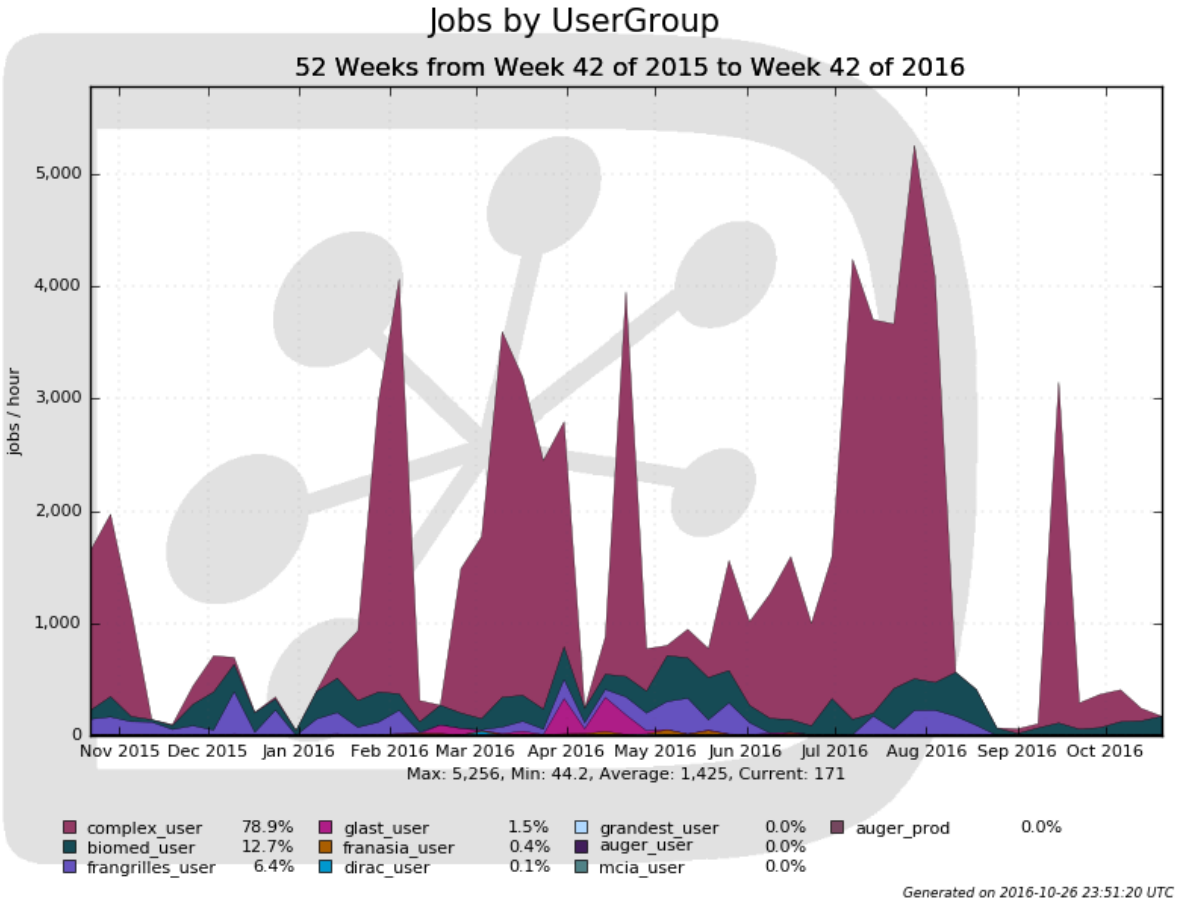
- 6 virtual hosts
  - Moving from Vmware to OpenStack infrastructure
  - 8 core, 16 GB RAM, 200 GB HDD
- 6TB storage
  - NFS mounted
  - DIRAC Storage Element
  - Job sandboxes
- MySQL service of CC/IN2P3
  - ~200 GB
  - ~1000 simultaneous connections
- More hosts will be needed
  - ElasticSearch DB (monitoring system activities)
  - Rabbit MQ server
- Nagios based services monitoring
  - Discussion/work in progress

FG-DIRAC CPU consumption

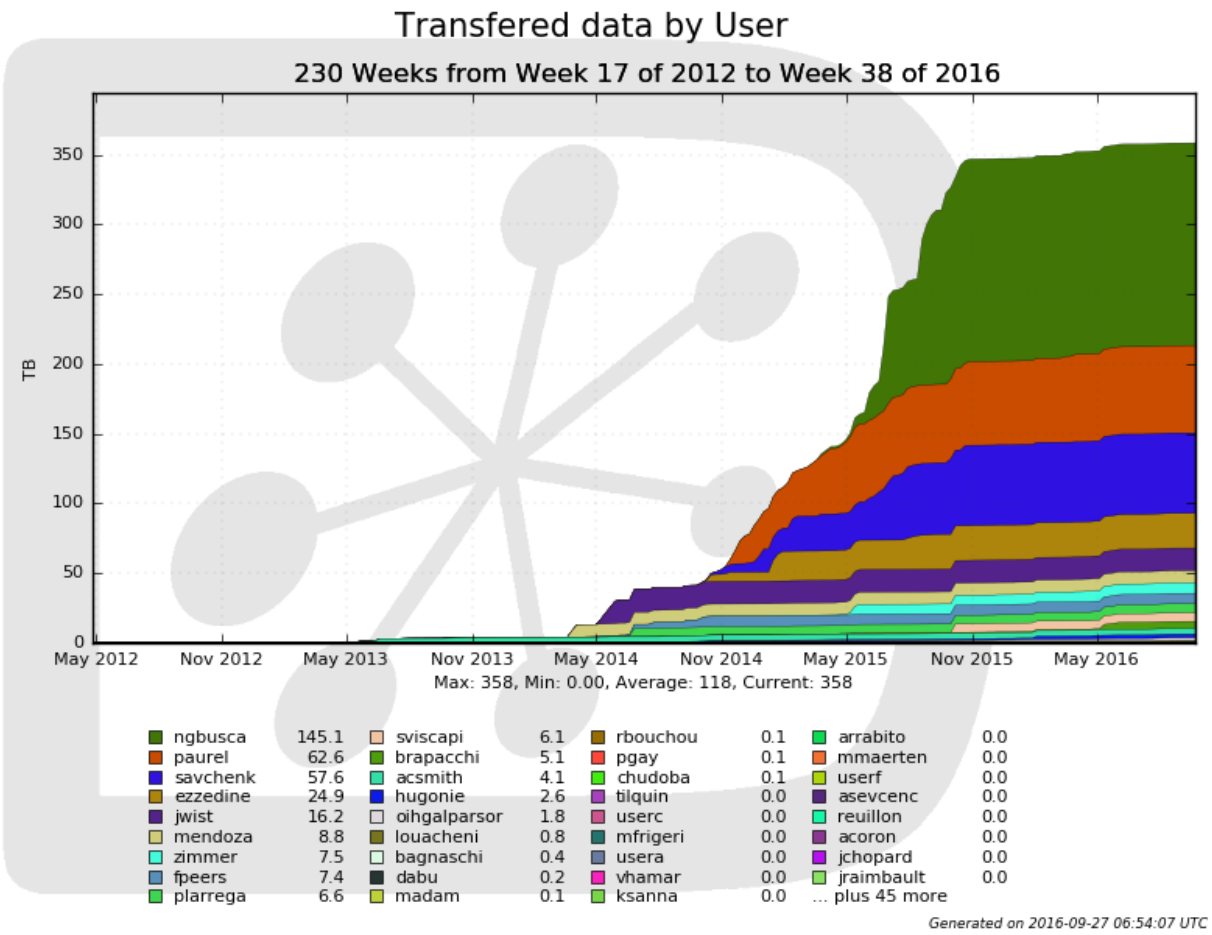
- > 2000 CPU years in the last year
  - Largest VO's: biomed, vo.france-grilles.fr, complex-systems.eu



Job execution rate

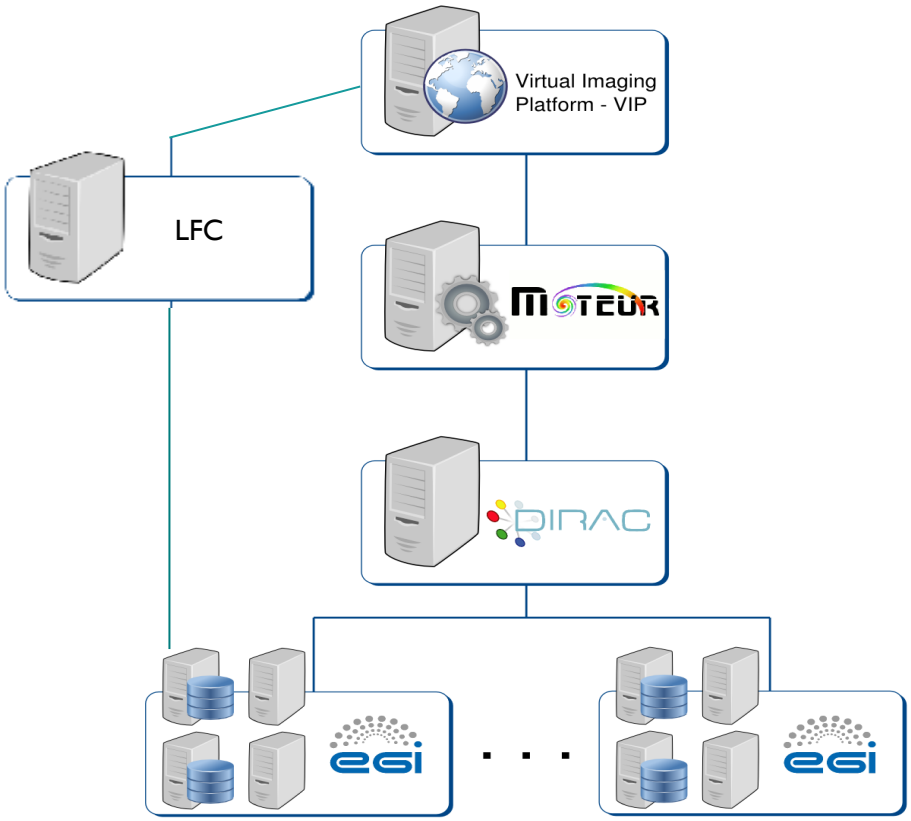


- Up to 2 Hz job execution rate
  - 200K jobs per day
  - VO **complex-systems.eu**: workflows with large numbers of small jobs


Transfered data by User
230 Weeks from Week 17 of 2012 to Week 38 of 2016
TB
350
300
250
200
150
100
50
0
May 2012
Nov 2012
May 2013
Nov 2013
May 2014
Nov 2014
May 2015
Nov 2015
May 2016
Max: 358, Min: 0.00, Average: 118, Current: 358
ngbusca 145.1
paurel 62.6
savchenk 57.6
ezzedine 24.9
jwist 16.2
mendoza 8.8
zimmer 7.5
fpeers 7.4
plarrega 6.6
sviscapi 6.1
brapacchi 5.1
acsmith 4.1
hugonie 2.6
oihgalparsor 1.8
louacheni 0.8
bagnaschi 0.4
dabu 0.2
madam 0.1
rbouchou 0.1
pgay 0.1
chudoba 0.1
tilquin 0.0
userc 0.0
mfrigeri 0.0
usera 0.0
vhamar 0.0
ksanna 0.0
arrabito 0.0
mmaerten 0.0
userf 0.0
asevcenc 0.0
reuillon 0.0
acoron 0.0
jchopard 0.0
jraimbault 0.0
... plus 45 more
Generated on 2016-09-27 06:54:07 UTC

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*

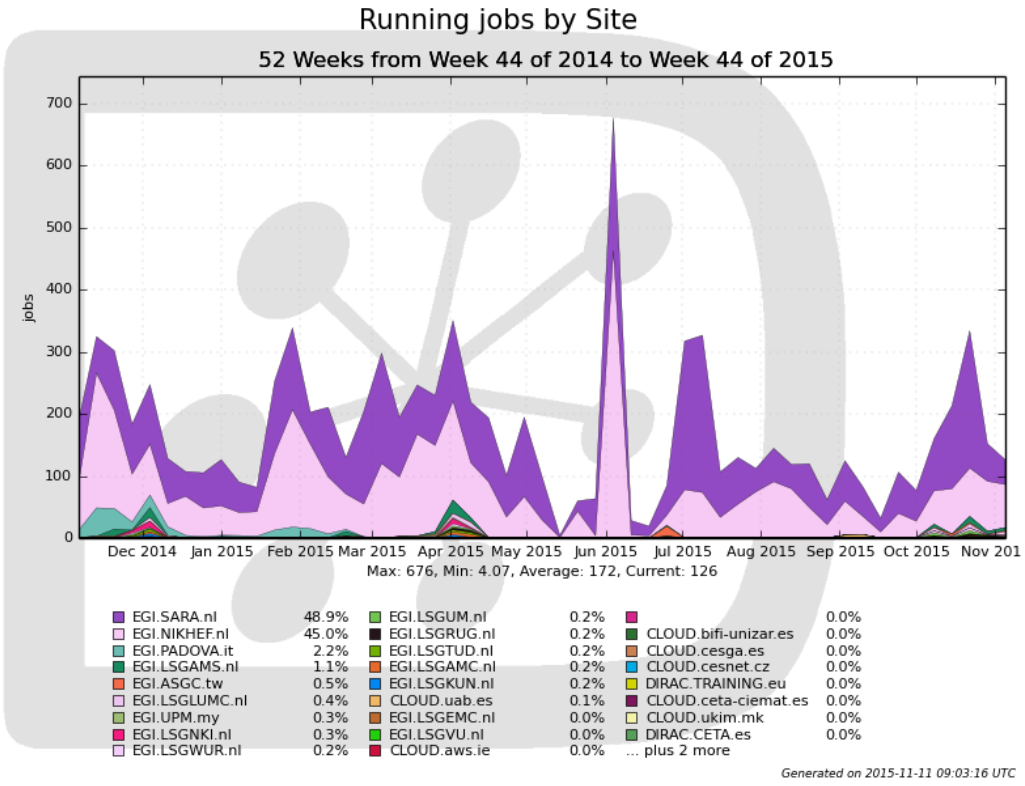
Using cloud sites

- 4 sites configured
  - CC, LUPM, IPHC, CPPM
  - Max 4-6 instances of 8 cores are allowed
  - APIs: apache-libcloud, rocci
- The sites were all operational in summer
  - Multiple operational problems
    - Changes (sometime subtle) in site configurations
    - Need for continuous attention, follow-up
    - Need for a unique DIRAC image across sites
  - Other DIRAC installations are using cloud resources routinely
    - E.g. BES III, Belle II, LHCb
- Certain interest from applications
  - Multi-core parallel applications
  - Special software, e.g. Docker
  - Try to exploit cloud provided flexibility

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
    - compchem will likely join the club

Replacement for gLite WMS

- DIRAC4EGI is included into the list of the EGI Core services
  - Counting on EGI support for the service management
    - Responding to a call for providing EGI services for Phase III 2018-2020
- EGI is interested to provide a viable replacement for the gLite WMS for their communities
  - Using DIRAC WMS in a "gLite" downgraded mode:
    - No use of generic pilots
    - Needs some development to restore this (obsoleted) functionality
  - User certificates or PUSP proxies or …
  - VO ***compchem*** is identified to be the first to try it out
- What about FG-DIRAC ?
  - Role in DIRAC4EGI service
  - Merging FG0DIRAC and DIRAC4EGI services

Conclusions

- 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
- FG-DIRAC service is fully operational. Need more effort to integrate cloud resources and define policies of their usage
- DIRAC4EGI service is becoming a Core EGI service intended to replace the gLite WMS

http://diracgrid.org