Distributed Infrastructure for Scientific Applications



A.Tsaregorodtsev, *Aix Marseille Univ, CNRS/IN2P3, CPPM* Belle II Computing Workshop, CC-IN2P3, Lyon 15 January 2020



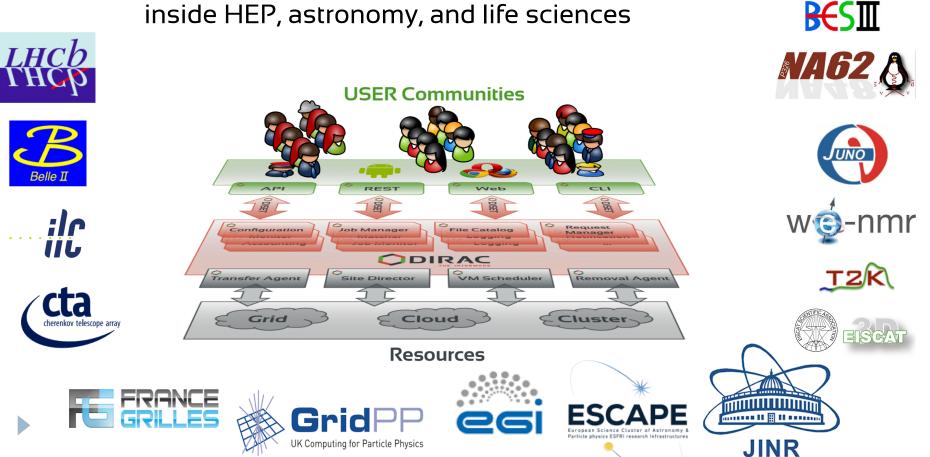


- DIRAC Project
- DIRAC Interware
 - WMS and computing resources
 - DMS and storage resources
 - Interfaces
 - Framework for distributed computing systems
- Conclusions



Interware

- A software framework for distributed computing
- A **complete** solution to one (or more) <u>user community</u>
- Builds a layer between users and <u>resources</u>
- A *framework* shared by multiple experiments, both inside HEP, astronomy, and life sciences





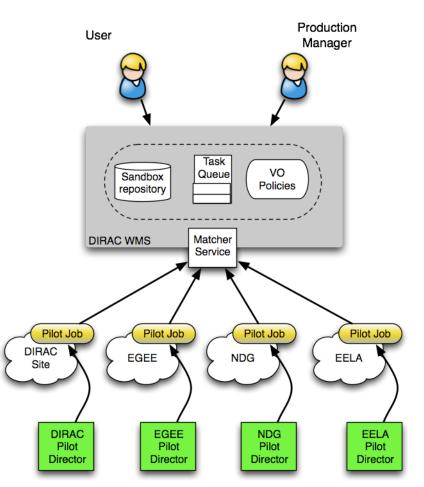
- Started as an LHCb project, became experimentagnostic in 2009
 - First users (after LHCb) end of 2009
- Developed by communities, for communities
 - Open source (GPL3+), <u>GitHub</u> hosted, python 2.7
 - No dedicated funding for the development of the "Vanilla" project
 - Publicly <u>documented</u>, active <u>assistance forum</u>, yearly <u>users workshops</u>, open <u>developers meetings</u>
 - 4 FTE as core developers, a dozen contributing developers
- The DIRAC <u>consortium</u> as representing body
 - CNRS, CERN, IHEP, KEK
 - PNNL, University of Montpellier, Imperial College



Workload Management

Job scheduling

- Pilot jobs are submitted to computing resources by specialized Pilot Directors
- Pilots retrieve user jobs from the central Task Queue and steer their execution on the worker nodes including final data uploading
- Pilot based WMS advantages:
 - increases efficiency of the user job execution
 - allows to apply efficiently community policies at the Task Queue level
 - allows to integrate heterogeneous computing resources







Computing resources

- Computing Grids
 - CREAM/HTCondorCE/ARC Computing Elements
 - Globus ComputingElement (OSG)

Clouds

- Dynamic VM life cycle management
- Openstack, OpenNebula
- Amazon EC2

Standalone computing clusters

- Access through SSH tunnel
- LSF, BQS, SGE, PBS/Torque, Condor
 - Commodity computer farms
- OAR, SLURM
 - HPC centers

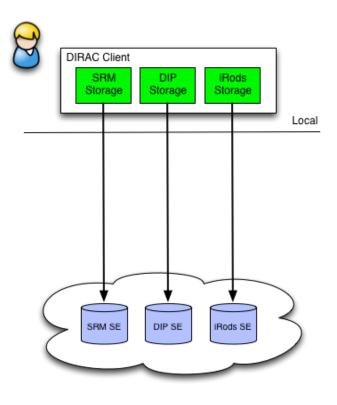


Data Management



Storage access: plugins

- Storage element abstraction with a client implementation for each access protocol
 - DIPS DIRAC data transfer protocol
 - FTP, HTTP, WebDAV
 - SRM, XROOTD, RFIO, DCAP, etc
 - HEP centers specific protocols
 - Using gfal2 library developed at CERN
 - S3, Swift, CDMI: cloud specific data access protocols
- Like with CE's, each SE is seen by the clients as a logical entity
 - With some specific operational properties
 - Archive, limited access, etc
 - SE's can be configured with multiple protocols
- Including new data access technologies requires creating new specific plug-in



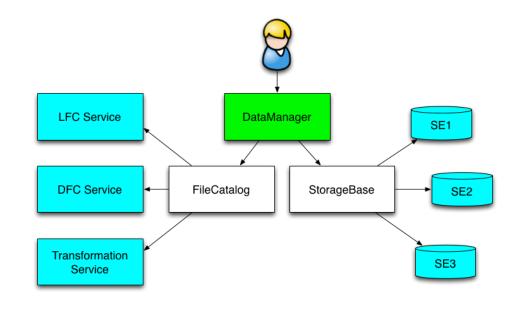


File Catalog Service

- File Catalog is a service to keep track of all the physical file replicas in all the SE's
 - Stores also file properties:
 - Size, creation/modification time stamps, ownership, checksums
 - User ACLs
- DIRAC relies on a *central* File Catalog
 - > Defines a single logical name space for all the managed data
 - Organizes files hierarchically like in common file systems
 - Other projects, e.g. distributed file systems, keep file data in multiple distributed databases
 - More scalable
 - Maintaining data integrity is very difficult



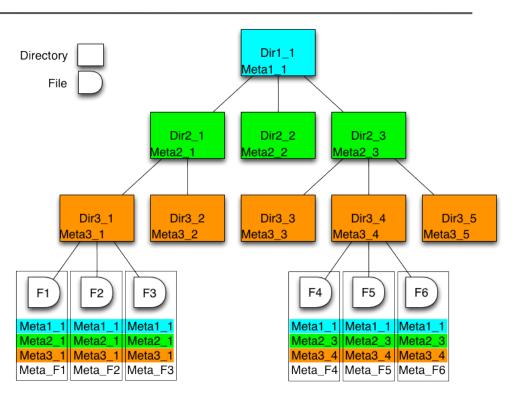
- Together with the data access components DFC allows to present data to users as a single global file system
- DataManager API is a single client interface for logical data operations





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</pre>
- Result of file search is a precise list of corresponding files
 - Unlike Google index





StorageManagement System

- Support for data staging
- RequestManagement System massive data operations
 - Data replication, removal, etc
 - Asynchronous execution
 - Automatic failure recovery

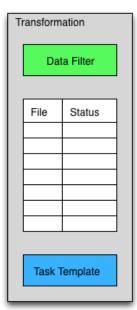
Transformation System

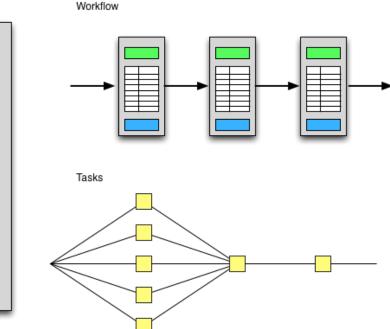
Automated data driven workflows



Transformation System for data driven workflows

- 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

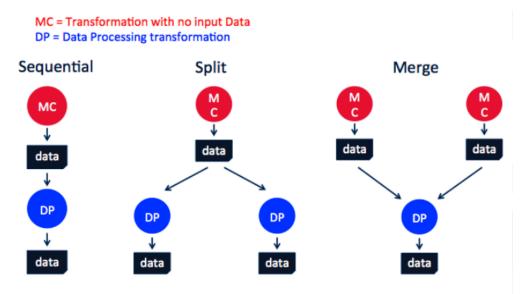






Production System

- Available in DIRAC version v7r0
- Automatic transformation instantiation based on the production definition
- Fully data-driven
- Tested for simple workflow schemes and their combinations





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
- Multiple application portals are interfaced to DIRAC this way



Web Portal applications

Apps 🗋 Apple	s://dirac.ub.edu/CTA/s:CTA/g:cta_user] Yahoo! 🔮 Google Maps 🗈 YouTube [☆
electors		Items per page: 100 💌	Page	1 of 13006	▶ ▶ Displaying topi	ics 1 - 100 of 1300594	Updated: 2013-10-16 14	:49 [UTC]
			Site	JobNar La	stUpdate [UTC]	LastSignOfLife [UTC]	SubmissionTime [UTC]	Ow
	Selected Statistics :: Status GMT+0200 (CEST))	(Wed Oct 16 2013 20:22:59	LCG.CIEM	AT.es Sta 20	13-10-16 14:21:54	2013-10-16 14:21:54	2013-10-16 14:21:54	tł
Selected Statistic Status		Completed Done	LCG.CIEM	AT.es Sta 20	13-10-16 14:02:06	2013-10-16 14:02:06	2013-10-16 13:55:38	ť
		Failed	LCG.CIEM	AT.es Sta 20	13-10-16 14:02:04	2013-10-16 14:02:04	2013-10-16 13:55:28	ť
Key		Other	LCG.DESY	-ZEUT Unk 20	13-10-16 14:01:08	2013-10-16 14:01:08	2013-10-16 12:33:16	tł
Completed	18.1%		LCG.CAM	K.pl Unk 20	13-10-16 12:29:59	20 📰 Proxy Upload	E	
Done			LCG.DESY	-ZEUT Ast 20	13-10-16 10:03:22	20		
Failed			LCG.DES	Job Launchpad				
Killed				Proxy Status: Valid		🕂 Add Paramete		lene no
Running				Predefined Sets	of Launchpad Values -		either your private our service. While	we try to
Waiting	81.79			🖃 😑 Available Set	ts		vith your credentials	
A Defeat to a								
Refresh 🖵 Pro	oportional 🛛 👌 Auto refresh : Disabled 🗸	CSV data		🗄 🧰 Mandelbr	ot		for maximum secu	
Refresh 🖵 Pro	Running jobs by Site				ot			
					ot		for maximum secu- anually convert and lient commands:	
Refresh ↓ Pro 5.000	Running jobs by Site		02331]	JDL Executable:	mandelbrot		for maximum secularually convert and	
	Running jobs by Site	ek [T] View as Text	02331]	JDL Executable: JobName:	mandelbrot Mandelbrot_%j	-0 46400 -V -0 56480 -D 0 1	for maximum securation in the securation of the	
5.000 -	Running jobs by Site	ek [T] View as Text [™] Reload [™] Dirac-CTA [2013-10-16 14:38:59 [™] DIRAC [™] Systems	02331]	JDL Executable: JobName: Arguments:	mandelbrot Mandelbrot_%j -W 600 -H 600 -X	-0.46490 -Y -0.56480 -P 0.1	for maximum securation in the securation of the	d upload
5.000	Running jobs by Site	ek [T] View as Text	02331]	jDL Executable: JobName: Arguments: OutputSandbox:	mandelbrot Mandelbrot_%j -W 600 -H 600 -X - *.bmp	-0.46490 -Y -0.56480 -P 0.I	for maximum securation in the securation of the	d upload
5.000 -	Running jobs by Site	ek T View as Text Reload Dirac-CTA [2013-10-16 14:38:59.] Dirac-CTA [2013-10-16 14:38:59.] Dirac-CTA [2013-10-16 14:38:59.] Dirac-CTA [2013-10-16 14:38:59.] Dirac-CTA [2013-10-16 14:38:59.]	02331]	JDL Executable: JobName: Arguments:	mandelbrot Mandelbrot_%j -W 600 -H 600 -X *.bmp %j.err	-0.46490 -Y -0.56480 -P 0.1	for maximum securation in the securation of the	
5.000	Running jobs by Site	EK T View as Text Reload T View as Text Reload T Dirac-CTA [2013-10-16 14:38:59.] T DIRAC T Systems T Website T Registry T Operations T Defaults	02331]	jDL Executable: JobName: Arguments: OutputSandbox:	mandelbrot Mandelbrot_%j -W 600 -H 600 -X - *.bmp	-0.46490 -Y -0.56480 -P 0.I	for maximum securation in the securation of the	d upload
5.000	Running jobs by Site	ek [T] View as Text Reload Dirac-CTA [2013-10-16 14:38:59.] Dirac-CTA [2013-16 14:38:59.] Dirac-CTA [2014	02331]		mandelbrot Mandelbrot_%j -W 600 -H 600 -X *.bmp %j.err	-0.46490 -Y -0.56480 -P 0.1	for maximum secu anually convert and lient commands: 4E.p12 GROUP_NAME	d upload
5.000 4.000 2.000 1.000 jan 2013 Feb	Running jobs by Site 41 Weeks from Week 53 of 2012 to We 2013 Mar 2013 Apr 2013 May 2013 Jun 2013 Jun 2013 Max 5.143, Min: 0.00, Average: 606, Current		02331]	jDL Executable: JobName: Arguments: OutputSandbox: StdError: CPUTime: StdOutput:	mandelbrot Mandelbrot_%j -W 600 -H 600 -X *.bmp %j.err 3600	-0.46490 -Y -0.56480 -P 0.	for maximum secu anually convert and lient commands: 4E.p12 GROUP_NAME	d upload
5,000 4,000 2,000 1,000	Running jobs by Site 41 Weeks from Week 53 of 2012 to We 41 Weeks from Week 53 of 2012 to We 2013 Mar 2013 Apr 2013 Apr 2013 Mar 2013 Apr 2013 Mar 2014 Mar 2015 Mar 2015 Mar 2016 Mar 2017 Mar 2018 Mar 2018 Mar 2018 Mar 2018 Mar 2018 Mar 2	Ek [T] View as Text Reload Dirac-CTA [2013-10-16 14:38:59.] DIRAC Systems Website Registry Operations Defaults StitLocalSEMapping Shifter EMail Default	02331]	jDL Executable: JobName: Arguments: OutputSandbox: StdError: CPUTime:	mandelbrot Mandelbrot_%j -W 600 -H 600 -X *.bmp %j.err 3600	-0.46490 -Y -0.56480 -P 0.	for maximum secu anually convert and lient commands: 4E.p12 GROUP_NAME	d uploa



Development Framework



DIRAC software

- Python 2.7
 - Python 3 migration is planned
- GPLv3 license
- GitHub repository
- External dependencies are compiled and provided as part of the DIRAC distribution
 - DIRACOS package
- DIRAC extensions are managed by communities
 - E.g. CTADIRAC
 - GitHub repository
- External service used
 - MySQL, ElasticSearch, Stomp enabled MQ's
- Software is available as
 - installable tar archives
 - dirac-install tool
 - preinstalled in CVMFS
 - Docker containers



- Modular architecture
 - Much functionality is implemented as plugins, e.g.
 - Job policies
 - Data access rules
 - Web applications
- Multiple community specific extensions exist already
 - BelleDIRAC, LHCbDIRAC, CTADIRAC, ILCDIRAC, etc
- Community specific services can run as part of a general purpose DIRAC service
 - E.g. Eiscat 3D specific File Catalog as part of the EGI DIRAC service



- Support for OAuth2/OIDC based AuthN/AuthZ frameworks
 - Federated Identity Providers, Proxy Providers
- Replacing the custom DIRAC client/server protocol by an HTTP based protocol
 - Standards based, better flexibility, and scalability
- Python 3 migration
- Rucio interface (RucioFileCatalog)



Conclusions

- DIRAC is a complete solution for a large scientific community to build its distributed computing system
- DIRAC is continuously developed to keep it up to date with the technology advancements
- DIRAC together with some specific extension can suit the needs of the Belle II experiment



Backup slides



- DIRAC was initially developed with the focus on accessing conventional Grid computing resources
 - WLCG grid resources for the LHCb Collaboration
- It fully supports different middleware based grids
 - European Grid Infrastructure (EGI), WLCG, etc
 - DIRAC is an officially supported WMS service for the EGI infrastructure
 - CREAM/HTCondorCE/ARC ComputingElement's
 - Northern American Open Science Grid (OSG)
 - GlobusComputingElement's
- Other types of grids can be supported
 - As long we have customers needing that



account

needed

Public:

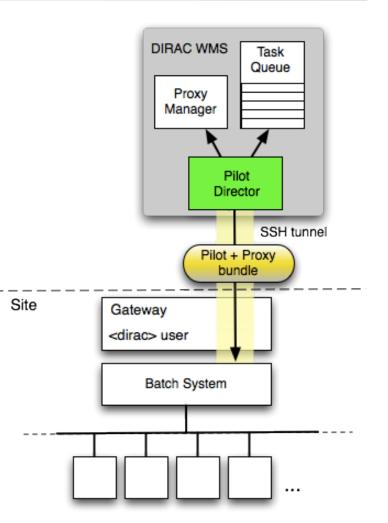
Resources: Clouds

VM scheduler Dynamic VM spawning taking Task Queue state into Task VO Queue Sandbox Policies repository **Discarding VMs** Matcher DIRAC WMS Service automatically when no more Cloud C Cloud B The DIRAC VM scheduler Cloud A by means of dedicated VM VM VM Pilot Job Pilot Job Directors is interfaced to Pilot Job OpenStack, OpenNebula Cloud A Cloud B Cloud C VM VM VM Director Director Director Amazon EC2 DIRAC VM Scheduler



Resources: standalone computing clusters

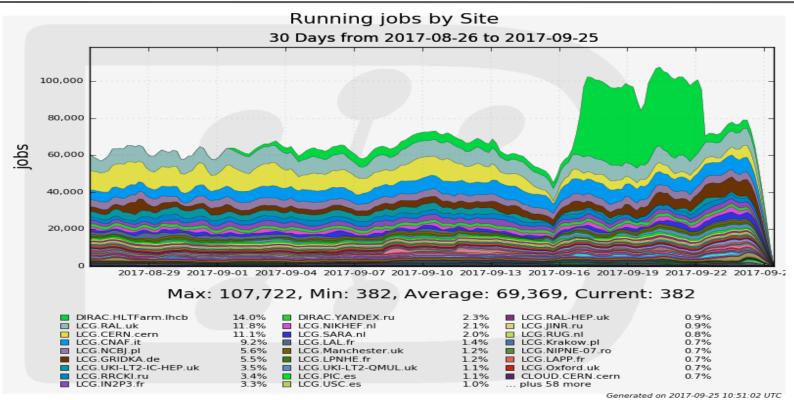
- Off-site Pilot Director
 - Site delegates control to the central service
 - Site must only define a dedicated local user account
 - The payload submission through an SSH/GSISSH tunnel
- The site can be:
 - a single computer or several computers without any batch system
 - a computing cluster with a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







LHCb Collaboration

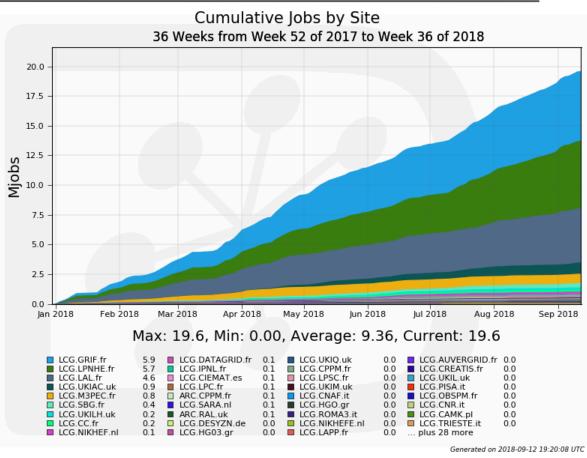


- More than 100K concurrent jobs in ~120 distinct sites
 - Limited by available resources, not by the system capacity
- Further optimizations to increase the capacity are possible
 - Hardware, database optimizations, service load balancing, etc





- Hosted by the CC/IN2P3, Lyon
 - dirac.france-grilles.fr
- Distributed administrator team
 - 5 participating universities
- In production since May 2012



- About 5 active communities complexsystems, biomed, vo.france-grilles.fr, ...
- > > 20M jobs executed this year at 90 sites



DIRAC4EGI service



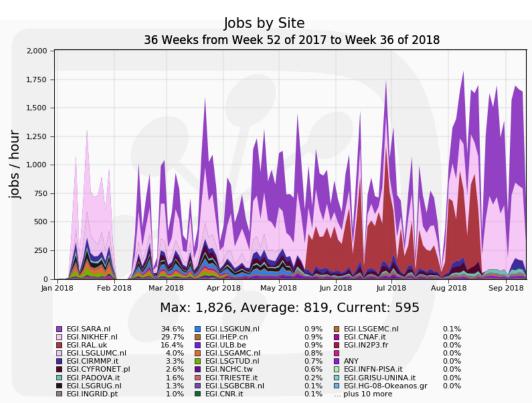
- Partners
 - Operated by EGI
 - Hosted by CYFRONET, Krakow
 - DIRAC Project providing software, consultancy
 - Supported via the EOSC-Hub H'2020 grant
 - dirac.egi.eu
- > 20 Virtual Organizations
 - enmr.eu
 - virgo
 - eli-beams
 - eiscat.se
 - fedcloud.egi.eu
 - ...

Usage

30

> 5 million jobs processed this year

DIRAC4EGI activity snapshot



Generated on 2018-09-12 19:23:21 UTC