Managing User Communities Workflows With DIRAC

SW and Computing round table, <u>JLab</u>, May 3rd 2022



Federico Stagni DIRAC technical coordinator federico.stagni@cern.ch





- What's DIRAC
- What are the features and attributes that make DIRAC popular
- Why it is not more popular
- Current, upcoming and future developments



What's DIRAC?

- A software framework for distributed computing
- A complete solution to one (or more) user community
- Builds a layer between users and <u>resources</u>



- Started as an LHCb project, experiment-agnostic in 2009
- Developed by communities, for communities
 - Open source (GPL3+), <u>GitHub</u> hosted
 - Python 3 (python 2.7 kept for current production release)
 - Publicly <u>documented</u>, active <u>assistance forum</u>, yearly <u>users</u> <u>workshops</u>, open <u>developers meetings</u> and <u>hackathons</u>
- The DIRAC <u>consortium</u> as representing body



... a few examples of what DIRAC can be used for

- sending jobs to "the Grid"
 - the obvious one...
- interfacing with different sites
 - with different computing elements
 - and batch systems
 - with different storage elements
- interfacing with different information systems
- interfacing with different catalogs
- interfacing with different *MQs*, *DBs*
- authenticate through different *identity providers* (in preparation)
- managing "productions" (e.g. reconstruction, simulation...)
- managing dataset transfers and removals
 - or delegate the task
- interacting with FTS
- providing a failover system
 - your jobs won't fail because a certain SE is down, nor because of central service are down
- transfer data from the experiment (the "online") to a Grid SE (the "offline")
- monitor your resources with a policy-based system
- ... and more





Installations and communities

(that I know)







A *framework* shared by multiple experiments/projects, both inside HEP, astronomy, and life science

Experiment agnostic Extensible Flexible





Why is DIRAC popular?

- **DIRAC as-a-service** (1 installation, several VOs) available since long time
 - good for medium-small communities
 - national initiatives to provide DIRAC as-a-service (e.g. GridPP, DutchGrid) very few DIRAC administrators for possibly dozens of communities
- Feature-rich, all-in-one (WMS, DMS, but also Productions and Dataset management, and monitoring)
 - again, good for limited-manpower communities
- Tightly-integrated DIRAC WebApp
- Actively developed and maintained





[WMS] basics

Pilots-based WMS, with late binding

- Users define and submit jobs. Jobs have requirements. Job descriptions are stored in DIRAC's Job DB.
- Independently, Pilots are started (1) on the sites' worker nodes (WN)
- Pilots will try to **match** (2) the worker nodes' capabilities to the jobs requirements.
- Jobs are started on WNs.
 DIRAC monitors their progress.





[WMS] Computing resources

where to run the pilots

P04: Failed Pilots / Site

- Grids (nowadays: HTCondor, ARC)
- **Clusters** behind a BS
 - a really thin layer that we call "SSH CE" \bigcirc
 - and then talks with batch system Ο

VMs scheduler:

- Based on apache libcloud 0
- Contextualization from standard images
 - with, at least, the DIRAC pilot

Vacuum:

- VAC/vcycle resources 0
- **BOINC Volunteer resources** \cap
- HLT farm (LHCb) 0

HPC sites

it often means at least SSH+Slurm 0 - more later on





[DMS] Data Management

Basics of **DIRAC** DMS:

LFNs: unique identifier within DIRAC of a file

Logical File Name (described as paths)

• LFNs are registered in **catalog**(s).

and there are implementations like the DFC

LFNs may have PFNs, stored in SEs.

Physical File Name on Storage Elements (and SEs are monitored, within the DIRAC Resource Status System) You can access those PFNs with several protocols.

- DIRAC catalogs for implementing namespace functionalities
 - Several <u>catalogs</u> can live in parallel
 - \rightarrow DFC: full replica and metadata catalog used by all DIRAC installation but Belle2
 - \rightarrow plugins for LFC, Rucio, DIRAC TS, LHCb Bookkeeping, Belle2 AMGA (?)











- Some VOs using DIRAC would like to use Rucio as DMS
 and maybe some VOs using Rucio would like to use DIRAC
- Discussions started at the 8th DIRAC workshop (May 2018)
- Since January 2021 Belle2 uses DIRAC and Rucio
 - from LCG file catalog to Rucio FC
- Developments were (and are being) done on both sides
 - not many lines of code
 - available: integration of (multi-VO) DIRAC with (multi-VO) Rucio, working without a rucio.cfg



Job Productions and datasets management with the "Transformation System"

The Transformation System (TS) is used to automate common tasks related to production activities

- A *"production"* is a transformation managed by the TS that is a "Data Processing" transformation (e.g. Simulation, Merge, DataReconstruction...). A Production ends up creating jobs in the WMS.
- A "Data Manipulation" transformation replicates, or remove, data from storage elements. A "Data Manipulation" transformation ends up creating requests in the RMS (Request Management System), which feeds the DMS.



The TransformationSystem is finely tuned and can manage millions of jobs and files daily





- Web apps available for monitoring jobs, files, productions, etc.
- Configurable
- Extendable

Menu 📀									
A 📃 O	Configuration Manager [Untitled 1] $ \times $	Plict	Monitor (Untitled)	2] ×	Transform	mation Monitor [Untitled 3] × Accounting [Untitled 4] >	Job Monitor (Uni	itled 5] × Syst	em Administration [Untitled 6]
Desktops&Applications	Selectors 🛞 😔		e x x	•			Items per page:	200 🗸 🔍 🤇	Page 1 of 1 >
> 🛅 VM tools	Site:	0	Jobid 👃		Status	MinorStatus	Application	Site	JobName
> 🛅 Tools		0	4732	1	Done	Execution Complete	exe-script	CLOUD.U	helioCloudWorld
—	Status:	0	4728		Done	Execution Complete	Unknown	LCG.UKI-L	MultiVOTest
Accounting			4727	- 2	Done	Execution Complete	Unknown	LCG.UKI-S	Mult/VOTest
Component History Configuration Manager	Minor Status:		4726	Ē		Job stalled: pilot not running	Unknown	DIRAC.Jen	job/WithOutputs
Downtmes		0	4725	_	Done	Execution Complete	exe-script	LCG.UKH	job/WthOutput
Elle Catalog	Application Status:	0	4724	_	Done	Execution Complete	exe-script	LCG.UKHL	parametricJob
Job Monitor			4723	_				LCG.UKHL	
Job Summary	Owner:			-	Done	Execution Complete	exe-script		parametricJob
Pilot Monitor			4722	_	Done	Execution Complete	exe-script	LCG.UKI-L	parametricJob
Pilot Summary	OwnerGroup:		4721	_	Waiting	Pilot Agent Submission	Unknown	ANY	wholeNodeJob
Proxy Manager			4720	_	Killed	Job stalled: pilot not running	Unknown	DIRAC.Jen	min2max4Job
Public State Manager	Job Group:		4719		Done	Execution Complete	mpTest.py	LCG.UKI-S	min2max4Job
Registry Manager	sub cloup.		4718		Done	Execution Complete	mpTest.py	LCG.UKI-S	doLqm
Resource Summary	Job Type:		4717		Waiting	Pilot Agent Submission	Unknown	DIRAC.Jen	helloWorldSSHBatch
Site Summary	Job Type:		4716		Waiting	Pilot Agent Submission	Unknown	LCG.GRIF.fr	helioWorldGRIF
Space Occupancy			4715		Done	Execution Complete	exe-script	LCG.GRID	helloWorldGRIDKA
System Administration	Time Span:		4714		Done	Execution Complete	exe-script	LCG.NCBJ.pl	helloWorldNCBJ
Transformation Monitor	For all time?	0	4713		Done	Execution Complete	exe-script	LCG.CER	helloWorldCERN
VMDIRAC Monitor			4712	Ē	Killed	Job stalled: pilot not running	Unknown	DIRAC.Jen	helloWorld
> 🗮 My Desktops	JobID(s):		4711		Failed	Job forced to Failed	Workflow s	LCG.UKI-L	00000252 00000003
> 🖽 Shared			4710	_	Failed	Job forced to Failed	Workflow s		00000252 00000002
	Pilot Job Reference(s):		4709	ī	Failed	Job forced to Failed	Workflow s	LCG.CER	00000252_00000001
		0	4708		Killed	Job stalled: pilot not running	Unknown	DIRAC.Jen	jobWithOutputs
			4707		Killed	Job stalled: pilot not running	Unknown	DIRAC.Jen	jobWithOutput
			4706		Killed	Job stalled: pilot not running	Unknown	DIRAC.Jen	parametric.Job
			4705	_	Killed	Job stalled: pilot not running	Unknown	DIRAC.Jen	parametricJob
			4704		Killed	Job stalled: pilot not running	Unknown	DIRAC.Jen	parametric.Job
			4703		Waiting	Pilot Agent Submission	Unknown	ANY	wholeNodeJob
			4702		Killed	Job stalled: pilot not running	Unknown	DIRAC.Jen	min2max4Job
		0	4701		Done	Execution Complete	mpTest.py	LCG.UKI-S	min2max4Job
			4700		Done	Execution Complete	mpTest.py	LCG.UKI-S	mpJob
			4699		Waiting	Pilot Agent Submission	Unknown	DIRAC.Jen	helloWorldSSHBatch
			4698		Waiting	Pilot Agent Submission	Unknown	LCG.GRIF.fr	helloWorldGRIF
	 Submit 2 Reset 2 Refresh 		4697		Done	Execution Complete	exe-script	LCG.GRID	hello/WorldGRIDKA
Settings	Default ×								





DIRAC extensions







Why not more popular? (a bit more than a guesswork)

- complex, with high entrance bar
- somewhat cumbersome deployment
- known for (only) its WMS functionalities
- often a bit late on "standards"
 - http services
 - o tokens
 - monitoring
- still considered "an LHCb thing"
- "old"-ish design



Some DIRAC developments

- always valid: Integrating DIRAC workflows in HPCs
- always valid: DMS advancements
- Done: Python 3
 - py3 clients supported since version 7.2 (pip installable)
 - py3 server supported since version 7.3 (production)
 - py2 support ends with 8.0 (release is few weeks away)
 - with some obvious exceptions of part of pilots code
- Ongoing: dips:// → https://
 - dips: DIRAC proprietary protocol for RPC calls
 - http: based on <u>tornado</u>
 - several DIRAC services already available using HTTP, and adding more
 - http will be the default for all the DIRAC services from version 8.1
- Ongoing: token support, and IdP (IaM, Check-in)
- Ongoing: ES/kibana/grafana dashboards
- Started: running on kubernetes (goal: define a helm chart)
- Started: using celery and RabbitMQ (retiring part of DIRAC framework)





HPC

Running on HPCs

Different solutions must be adopted for different HPCs. DIRAC can only take care of the distributed computing issues, such as:

- For MP WNs: logical partitioning using DIRAC "inner" PoolCE
- Interfacing with the batch system (mostly slurm)
- Interfacing with WNs with limited network connectivity using the *PushJobAgent*

VO actions often needed.

On software challenges: the good news is that py3 versions of DIRAC clients (so, DIRAC pilots) can be installed on *ppc64le* and *aarch64*

 \rightarrow but, you need to be able to start the pilots



- Current production release depends on VOMS
- DIRAC v8 (due in few weeks) rationalizes many aspects related to AuthN, AuthZ, Tokens and OAuth2 support, and will add *experimental* support to new Identity Providers (IaM and CheckIn)
- Longer term goals include:
 - externalize (to IdPs) users' management
 - use tokens (and/or proxies) for interfacing with computing and storage resources (v8.1)



Accounting and Monitoring

Accounting:

- For historic data
 - \circ Jobs
 - Pilots
 - Data Operations
 - Storage
- MySQL backend
- DIRAC Web App dashboard



Monitoring:

- Real Time monitoring and not only
- ElasticSearch (OpenSearch) backend
- Visualize in kibana, grafana, and (partially) DIRAC WebApp
- largely improved within DIRAC v8





Part of the communities

- Participation in WLCG activities (mostly DOMA)
- EGI
 - Check-In
 - Jupyter notebooks





Development and testing

92 17 Pull requests 13 O Actions III Projects 0 EI Wiki III Security Idi Insights O Settings ript with exec · rr80135 Integration tests / Integration (5,7, slc6) *** Thu Oct 17 08:49:09 UTC 2019 **** FRAMEWORK TESTS (partially skipped) **** MARNING: assertions not in test modules or pluging will be ignored because assert statements are not executed ======= test session starts === sis profile 'default' -> database+DirectoryBase alledComponentsDB.pv::ComponentMonitoringClie === 4 passed in 2.22 seconds * Thu Oct 17 88:52:69 UTC 2019 **** 855 TESTS ** initial provide definition of the second provide the second provi IDIR/DIRAC/tests/Integration/ResourceStatusSystem ceManagement.ov:ResourceManagementClientChain:t

~6 FTE as core developers, a dozen contributing developers

Tests, certification, integration process is a daily work.

- We use (lots of) GitHub Actions, and Jenkins for some bits
- We run certification hackathons every 2nd week

				vitero-prot Trolio - Mazilla Pirefox				000
	😳 🧏 Python 🖃 Oracii 🖬 🕅		de 💦 and ny 🛣 Discovi 🌋 K	bene curth: 🧕 CCRN (🤮 xG	96-1 🗖 DIGAL 🛛 (1) Tending 🗍 🚯 Docker 🔤 1	A LHChi — 💗 el-Gri — 💗 templa 🕺 k	bana @ 153311 @ egiche 당 ☆	± N. CD © =
			Kibene 🗘 Computing Dements 🚦	1 HCh Marting, Indico				INUVI
	a			G Felle				+ 0 🗛 🙆
r Dictorial C	etification team 1886 A Team Vi							** Show Menu
	Storted -	Pessing ···	Not Passing ···	Not Pessing - fixed with PR/MR				
	Integration bests from CCT	DrikkC micease mody	Sabmit MC Production - use all templates	Submit replication transformation and removal				
d/opRece	Integration SLC6	Cristo LHCEDIRAC tags	+ Add another card	IF (0.) + Addensible card				
	Submit User Jobs from old LHCbDRMC version	Pylist + tost	* Add and the card	T AGE BROVE CET				
rs -	-	UNCO LCC bendle for DIRACOSI ready						
liert and server	Basic Client Lents (using previous UHCLOBEAC version) IP (0200	© © 1						
_	Unchrister, submittendMatch	Deploy release on server						
NFS	+ Add and the card	Install a client on CCP						
peline (CVMPS)		-						
on CVMPS		Made client tests (client and server serve version) IP 10:1/2						
deployed		instal a cleat on S.C.						
a comes		Verify Monitoring application						
ARE ICVMPS		Try -ext R007 extension						
etion		Deale imports for client and servers						
60e		Change pilot version						
dection.		P						
		Pliata (comit)						
		Plati (ormi)						
genta		P D W						
		_						





DIRAC Users' Workshop

The 2022 DUW is next week! <u>https://indico.cern.ch/e/DUW11</u>

(on zoom, free, CEST mornings)

we'll record the sessions



Questions/comments?

- <u>dirac.readthedocs.io</u>
 - including <u>code documentation</u>
- Ops and general questions: Google <u>forum</u> but we prefer <u>github discussions</u>
- Dev and DevOps issues: on github
- Bi-weekly developers meetings (and/or hackathons): <u>BILD</u>

backup



[DMS] Storage Elements

• **<u>Storage Elements</u>**: abstraction of the storage endpoints

- fully described in Configuration Service (CS)
- several abstractions of the same physical endpoint are possible

• Multi-protocol

- DIP: DIRAC custom protocol
- File: offers an abstraction of the local access as an SE.
- RFIO (deprecated): for the rfio protocol.
- Proxy: to be used with the StorageElementProxy.
- S3: for S3 (e.g. AWS, CEPH)
- GFAL2_SRM2: for srm
- GFAL2_XROOT: for xroot
- GFAL2_HTTPS: for https
- GFAL2_GSIFTP: for gsiftp
- SpaceOccupancy plugins:
 - BDIIOccupancy, WLCGAccountingJson, WLCGAccountingHTTPJson
- SEs definitions are sync-ed from DIRAC CS to Rucio RSE via a DIRAC agent

CERN-BUFFER

```
BackendTvpe = Eos
AccessProtocols = root, gsiftp, https
WriteProtocols = root, gsiftp, https
SEType = T0D1
SpaceReservation = LHCb-EOS
OccupancyLFN = /eos/lhcb/proc/accounting
OccupancyPlugin = WLCGAccountingJson
GFAL2 XROOT
      Host = eoslhcb.cern.ch
      Protocol = root
      Path = /eos/lhcb/grid/prod
      Access = remote
      Path = /eos/lhcb/grid/prod/lhcb/buffer
GFAL2_GSIFTP
      Host = eoslhcbftp.cern.ch
      Protocol = gsiftp
      Path = /eos/lhcb/grid/prod
      Access = remote
      Path = /eos/lhcb/grid/prod/lhcb/buffer
GFAL2 HTTPS
      Host = eoslhcb.cern.ch
      Protocol = https
      Path = /eos/lhcb/grid/prod
      Access = remote
      Path = /eos/lhcb/grid/prod/lhcb/buffer
```



[RSS] Resource Status System

- Stores info on the status of Resources (e.g. SEs)
- An autonomic computing tool evaluates a few policies to determine the status of the resources. E.g.:
 - \circ space left < threshold \rightarrow ban for writing
 - endpoint in downtime in GocDB \rightarrow ban r/w
 - o ...
- DIRAC SEs states are sync-ed from DIRAC RSS to Rucio via a DIRAC agent





A generic system, which can be used for queueing (also) DMS operations



Operation types:

- ReplicateAndRegister (e.g. using FTS)
- RemoveFile/RemoveReplica
- ... others (not useful for this pres)
- ...add your own (e.g. ReplicateUsingAnotherExternalSystem)





DIRAC <u>TS</u>

A <u>generic</u> system for queueing similar *operation types* on certain *datasets* and forward them to the appropriate *systems*

 An operation type can be, e.g.: a simulation workflow a reconstruction workflow a replication a removal 	 A <i>dataset</i> is split into groups, based on criterias defined by <i>plugins</i>, e.g.: split by size by destination by metadata [code it] 	A <i>system</i> is either (today) the DIRAC WMS (for productions) or the DIRAC RMS (for dataset management operation types)
---	--	--

E.g. Take all my holidays pictures from 2018 with tag='sunset', make sure that there is one copy on tape and one on disk, distributed on all the sites according to free space, and group the operations by group of at most 100 files.