

GANGA TUTORIAL FOR LHCb

Prepared by GANGA_Team::Mike_Williams (ICL, London)
Presented by GANGA_User::Jibo_He (LAL, Orsay)

March 25th, 2011

59th LHCb Week

OUTLINE

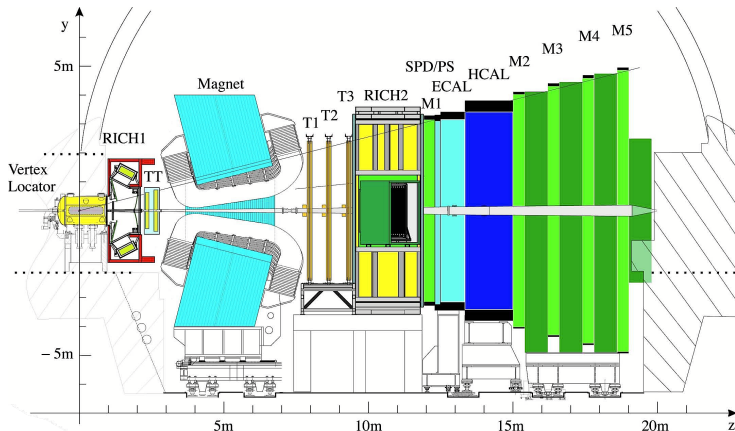
- 1 INTRODUCTION
- 2 DISTRIBUTED ANALYSIS @ LHC*b*
- 3 GANGA
- 4 ETC.
- 5 SUMMARY

OUTLINE

- 1 INTRODUCTION
- 2 DISTRIBUTED ANALYSIS @ LHC*b*
- 3 GANGA
- 4 ETC.
- 5 SUMMARY

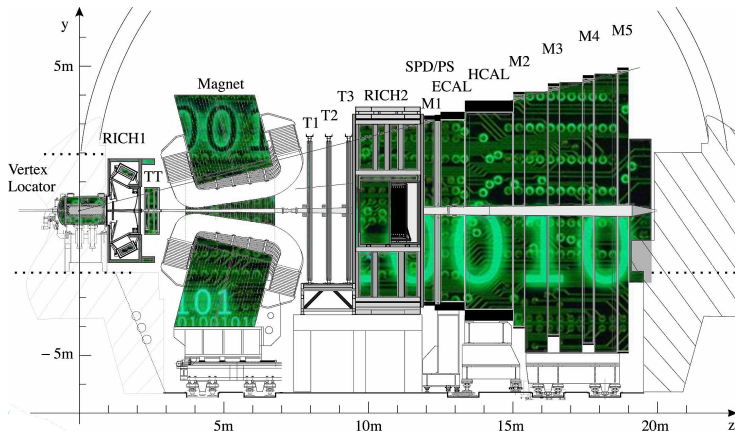
LHCb DATA TAKING

LHCb takes data @ $\mathcal{O}(100)$ MB/s & expects to collect $\mathcal{O}(1/2)$ PB in 2011.



LHCb DATA TAKING

LHCb takes data @ $\mathcal{O}(100)$ MB/s & expects to collect $\mathcal{O}(1/2)$ PB in 2011.

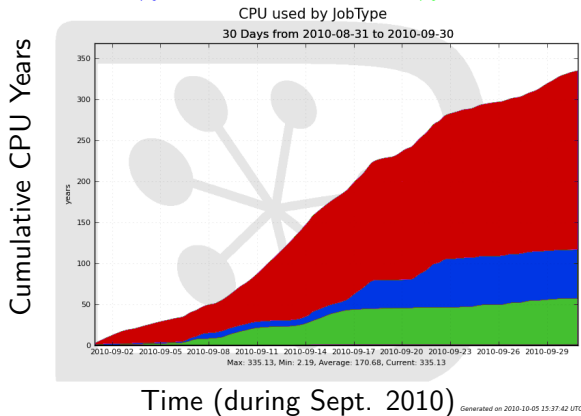


LHCb is a super bit factory (not to be confused w/ a Super *B* factory).

LHCb COMPUTING RESOURCE USAGE

LHCb used over 300 CPU years (shared resources only) in Sept. 2010.

65% User Jobs 18% MC Production 17% Data Reconstruction



We've also used close to 700 TB of disk space in Sept. 2010. So, that's 10 CPU years and 35 TB every day. We need *The Grid!*

WHAT IS THE GRID?

The Grid is a collection of computing resources located at sites around the world and consists of computing and storage elements (CEs and SEs).

Only a single *login* is required to access the system. After ID, security is handled by the system.

There are several flavors of the Grid; however, in *LHCb* we only use the LHC Computing Grid (LCG).



GRID RESOURCES

Your *grid certificate* is what gives you a unique identification on the Grid (2 files in your `.g1obus` directory). By joining a *virtual organization* (VO), you gain access to the resources available to the VO*.

By sending a *grid proxy* along with your Grid jobs, you allow computers to act on your behalf for a limited time. This lets your jobs run at LCG sites and read(write) files from(to) LCG SEs.

If your proxy expires while some of your jobs are running on the LCG, the jobs will continue to run; however, you will not be able to access the results w/o renewing your grid proxy.

*You all should have joined the LHCb VO!

OUTLINE

- 1 INTRODUCTION
- 2 DISTRIBUTED ANALYSIS @ LHC*b***
- 3 GANGA
- 4 ETC.
- 5 SUMMARY



GANGA is a user-friendly frontend that handles job definition and management for LHCb users.

GANGA's main goal is to ensure users are able to efficiently access all available resources (local, batch, grid, *etc.*).

DIRAC is the workload/data mgmt. system (WMS/DMS) for LHCb. It does the *heavy lifting* for all DA in LHCb.

DIRAC's main goal is to insure that the VO uses its resources efficiently and to enforce job prioritization.

THE DIRAC WMS/DMS

Distributed Infrastructure w/ Remote Agent Control

DIRAC provides us with the following benefits (not an exhaustive list):

- job monitoring via web portal;
- DIRAC's many failover mechanisms greatly increase user success rates;
- user & production jobs happily coexist;
- having only one central task queue means that the VO's highest priority jobs always run first;



–NOT–

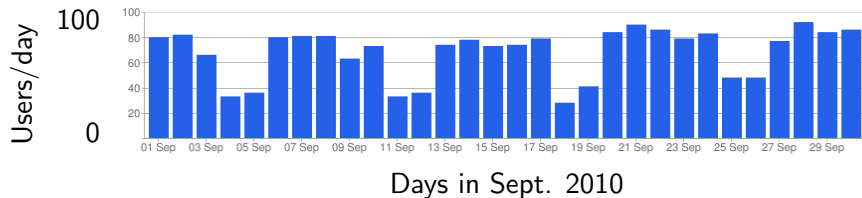


"This is the one thing I didn't do."

and, of course, all of the behind-the-scenes work the DIRAC team does investigating problems w/ sites, production jobs, etc.

GANGA USAGE IN LHCb

Almost 100% of LHCb grid users used GANGA in Sept. 2010



GANGA provides a *complete* analysis environment for LHCb and greatly simplifies the user experience (the topic of the rest of this talk). Thus, the vast majority of LHCb users choose to use GANGA for most tasks.

N.b., you can use DIRAC directly; however, the DIRAC team actually prefers that you use GANGA unless you really know what you're doing.

OUTLINE

- 1 INTRODUCTION
- 2 DISTRIBUTED ANALYSIS @ LHC*b*
- 3 GANGA**
- 4 ETC.
- 5 SUMMARY

EFFICIENT USAGE OF COMPUTING RESOURCES (USERS)

Users (should) want:

- development on their laptop/desktop;
- full analysis utilizing all available resources (wherever they are);
- to get results quickly and easily;
- a familiar and consistent UI for all resources.



Users don't want:

- to know all of the details about the Grid or any other resources;
- to learn yet another tool to access a resource;
- to have to reconfigure their application to run on different resources.



The GANGA mantra: Configure once, run anywhere!



GANGA was developed to meet the needs of ATLAS & LHC*b* for a grid user interface and is now used by many other groups as well. Usage: 45% ATLAS, 45% LHC*b*, 10% other.

550+ unique users, 40k+ sessions, run at 70+ sites (all in Sept. 2010!)

GANGA FEATURES

GANGA handles the complete life cycle of a job:

Build → Configure → Split → Submit → Monitor → Merge



GANGA does the following (and much more) for the user:

- builds/compiles applications;
- configures jobs, including building input sandboxes, to run on user-specified backends;
- submits jobs locally, to batch systems and to the grid;
- monitors jobs and updates the user on any status changes;
- automatically retrieves output when jobs complete;
- merges output (if requested).

GANGA LHC***b*** FEATURES

Loading the LHC***b*** plug-in adds the following features to GANGA:

- DIRAC backend and ability to contact the DIRAC server;
- many built-in DIRAC-based methods, e.g. `Dirac().checkSites()`;
- automatic collection of user-modified LHC***b*** software for sandbox;
- input data site-based job splitting (`DiracSplitter`);
- LHC***b*** data file (DST) merger (`DSTMerge`);
- automatic output file discovery (from application options);
- ability to checkout and build LHC***b*** software packages;
- *etc.* (too many to list them all here).

The *automatic* features are truly that; *i.e.*, the user is often not even aware of them. *E.g.* many users forget to add their output to the GANGA job definition for LHC***b*** applications. GANGA notices this and automatically adds the output for them (ignorance is bliss).

RUNNING GANGA

Since version 5.4.0, GANGA is now part of the LHC*b* software framework; thus, to set up the environment you should do:

```
[you@computer] SetupProject Ganga
```

To run GANGA interactively ($\sim 50\%$ of usage), do:

```
[you@computer] ganga
```

To run GANGA on a script ($\sim 50\%$ of usage), do:

```
[you@computer] ganga your-script.gpi
```

To run the GANGA GUI ($\sim 0\%$ of usage), do:

```
[you@computer] ganga --gui
```

THE GANGA PROMPT & CONFIGURATION

IP[y]: GANGA is written in Python and has an enhanced Python prompt (IPython) that supports:

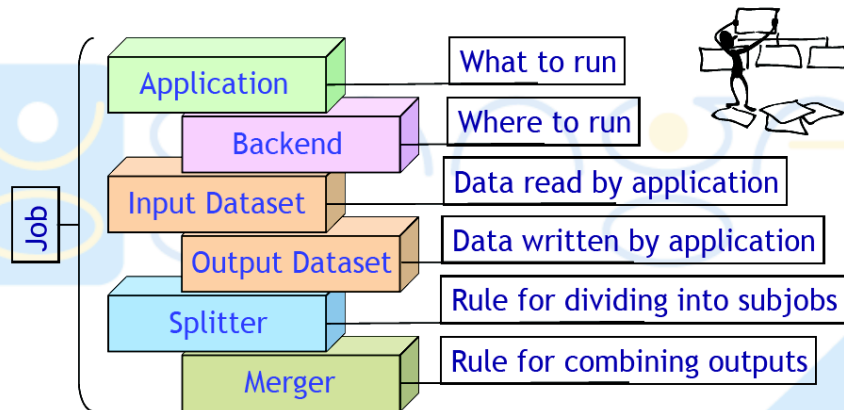
- Python syntax;
- Shell commands;
- TAB completion, scrolling thru your history, *etc.*

It's similar to working on the command line except Python syntax is valid and TAB completion works for Python objects, methods, variables, *etc.*

GANGA allows the user to configure many of its settings. To *permanently* change a setting (*i.e.*, to change it for the current and future sessions), simply edit it in your `.gangarc` file. Settings can also be viewed/changed in the current session by accessing the `config` object (these changes are not persisted).

GANGA JOBS

GANGA jobs are handled by the Job object.



JOB BASICS

To create a GANGA job, simply do:

```
In[1]:j = Job()
```

You can then edit its properties (*application*, *backend*, *etc.*); thus, you can configure the job to do what you want.

To submit the job to whatever backend you've chosen to run on, do:

```
In[2]:j.submit()
```

GANGA will monitor the job and let you know when it's done. When it's done, it'll also automatically collect the output you wanted back.

N.b., once a job is submitted, you cannot modify most of its properties (there are very good reasons for this).

GANGA/LHCb supports the following types of applications:

- Executable (binaries, scripts, *etc.*);
- Root (ROOT macros, PyROOT scripts);
- Gaudi-type applications (GaudiPython, Brunel, Moore, DaVinci, Panoptes, Gauss, Boole, Bender, Vetra).

GANGA/LHCb supports the following backends:

- Interactive (foreground on client node);
- Local (background on client node);
- Batch (LSF at CERN; SGE,PBS,Condor at other sites);
- Dirac (The Grid).

SPLITTING/MERGING

Users often want to run a large number of *similar* jobs. GANGA makes this easy.

GANGA/LHCb supports the following splitters:

- Input data (SplitByFiles, DiracSplitter);
- Gaudi-app (GaussSplitter, OptionsFileSplitter);
- General (GenericSplitter, ArgSplitter).



GANGA/LHCb supports the following mergers:

- TextMerger (text files);
- RootMerger (ROOT files);
- DSTMerger (DST files);
- General (SmartMerger, CustomMerger).



EXAMPLE JOB

To run DaVinci tutorial, in GANGA I'd simply do:

```
In[1]:j = Job()
In[2]:j.application = DaVinci(version='v26r3p2')
In[3]:j.application.optsfile =
    ['<path>/DaVinciTutorial_1.py', '<path>/Bs2JPhi.py']
In[4]:j.backend = Interactive()
In[5]:j.outputsandbox = ['DVHistos_1.root']
In[6]:j.submit()
```

To run on the Grid, we'd simply do `j.backend = Dirac()`. GANGA will automatically collect all of your modified files and send them w/ the job. Yes, it's really that easy.

EXAMPLE JOB

GANGA will tell you the status of the jobs – it'll update you whenever a job changes state, you can also check directly by doing `j.status`. Once the jobs are complete, GANGA will download the output automatically (and merge them if needed).

You can check the output of a job by doing, e.g.:

```
In[7]:j.peek()
```

```
total X
```

```
-rw-r--r-- 1 you z5 X Jan 5 10:00 DVHistos_1.root
```

```
lrwxr-xr-x 1 you z5 X Jan 5 10:00 stdout
```

```
⋮
```

or open a shell running ROOT w/ the file loaded by doing:

```
In[8]:j.peek('DVHistos_1.root')
```

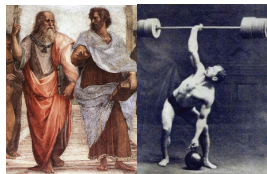
or specify the program you want to use:

```
In[8]:j.peek('stdout','less').
```

MORE GANGA LHCb FEATURES

GANGA doesn't just handle jobs, it also deals w/ data files & data sets:

- full support for logical & physical files including downloading, uploading, replicating, removing, obtaining metadata and replicas, etc.;
- `job.inputdata = browseBK()`



The screenshot shows the GANGA LHCb Desktop Environment. On the left is a file browser window with a tree view of files. On the right is a table with columns for 'Address', 'Name', 'Size', 'Created', and 'Last Modified'. The table contains many rows of data. A red circle highlights the 'Browse BK' button in the bottom right corner of the table window.

- bookkeeping queries can also be persisted in a BKQuery object and updated at any time w/o the need for the GUI or web interfaces.

OUTLINE

- 1 INTRODUCTION
- 2 DISTRIBUTED ANALYSIS @ LHC*b*
- 3 GANGA
- 4 ETC.
- 5 SUMMARY



Help is available for GANGA:

- Interactively in GANGA via the `help` function:
`In[9]:help(BKQuery)`
- Online via the GANGA manuals and GANGA/LHCb FAQ:
<http://ganga.web.cern.ch/ganga/user/index.php>
<https://twiki.cern.ch/twiki/bin/view/LHCb/GangaLHCbFAQ>
- Via the mailing list (lhcb-distributed-analysis@cern.ch).

For Python help, see <http://docs.python.org/tut/tut.html>

PERSISTENCE

```
ixplus256.cern.ch:64 (rlambert)
Rob Terminal
File Edit View Terminal Tabs Help

In [24]:jobs
Out[24]:
Registry Slice: jobs (14 objects)
-----
fqid | status | name | sj | app | v | be | backend.actualCE
-----
149 | new | Brunel QW Chrenkov | | Brunel | v35f7 | Local |
221 | completed | NTuple Loose Selection, BsBd events, CERN | | DaVinci | v24r5p1 | LSF |
224 | completed | DC06 selection, corrected, MC09 Bs and Bd data, CERN | 1 | DaVinci | v24r5p1 | LSF |
277 | failed | NTuple Preselection, Bd events, GRID | 35 | DaVinci | v24r5p1 | Dirac |
281 | failed | NTuple MC09 Selection, bb events, GRID | 107 | DaVinci | v24r5p1 | Dirac |
282 | failed | NTuple Preselection 2, bbmumu event, GRID | 134 | DaVinci | v24r5p1 | Dirac |
343 | failed | NTuple MC09 Selection, Bs events, GRID | 37 | DaVinci | v24r7p3 | Dirac |
344 | completed | NTuple MC09 Selection, Bd events, GRID | 35 | DaVinci | v24r7p3 | Dirac |
354 | completed | DST MC09 Selection, BsBd events, CERN | | DaVinci | v24r7p3 | LSF |
366 | completed | NTuple Presel, Bs events, GRID, CVS opts | 18 | DaVinci | v24r7p3 | Dirac |
367 | failed | NTuple Presel, bb events, GRID, CVS opts | 105 | DaVinci | v24r7p3 | Dirac |
368 | completed | NTuple Presel, Bd events, GRID, CVS opts | 18 | DaVinci | v24r7p3 | Dirac |
370 | failed | NTuple Presel, bbmumu events, GRID, CVS opts | 106 | DaVinci | v24r7p3 | Dirac |
371 | submitted | NTuple Sel bbmumu GRID more Truth | 106 | DaVinci | v25r1 | Dirac |

In [25]:box
Out[25]:
Registry Slice: box (5 objects)
-----
id | name | type | files | file0
-----
0 | Bs, Bd, CERN | LHCbDataset | 581 | castor:/castor/cern.ch/grid/lhcb/MC/MC09/DST/00005588/0000/00005588_00000252_1.dst
1 | bbmumu GRID | LHCbDataset | 1903 | /lhcb/MC/MC09/DST/00005609/0000/00005609_00000895_1.dst
2 | Bd GRID | LHCbDataset | 287 | /lhcb/MC/MC09/DST/00005588/0000/00005588_00000252_1.dst
3 | Bs GRID | LHCbDataset | 294 | /lhcb/MC/MC09/DST/00005580/0000/00005580_00000160_1.dst
4 | bb GRID | LHCbDataset | 2995 | /lhcb/MC/MC09/DST/00005113/0000/00005113_00000238_1.dst

In [26]:
```

DIRAC MONITORING

<https://lhcbweb.pic.es/DIRAC/LHCb-Production/lhcb/jobs/JobMonitor/display>

Jobs monitoring as lhcb@LHCb-Production

https://lhcbweb.pic.es/DIRAC/LHCb-Production/lhcb/jobs/JobMonitor/display

Jobs monitoring as lhcb@LHCb-Production

Selected setup: LHCb-Production

DIRAC SideBar

Selections

DIRAC Site: All

Status: All

Minor status: All

Owner: wrece

JobGroup: All

Date: YYYY-mm-dd

JobID:

Submit Reset

Global Sort

Selected Statistics

Global Statistics

JobId	Status	MinorStatus	ApplicationStatus	Site	JobName	LastUpdate [UTC]	LastSignOffLife [UT]	SubmissionTime [U]	Owner
922979	Failed	Maximum of resche	Unknown	LCG.GRIDKA.de	Ganga_ROOT_5.11	2008-12-05 23:19	2008-12-05 23:19	2008-12-05 10:23	wrece
920882	Failed	Uploading Job Outp	:/afs/cern.ch/user/lw	LCG.UKI-SCOTGR	Ganga_ROOT_5.11	2008-12-05 21:09	2008-12-05 21:09	2008-12-05 09:19	wrece
919804	Killed	Marked for terminat	unknown	ANY	Ganga_ROOT_5.11	2008-12-05 09:18	2008-12-05 09:18	2008-12-05 08:46	wrece
910243	Done	Execution Comple	/scratch/z5/wreco/	LCG.RAL.uk	Ganga_Gauss_v35	2008-12-05 07:23	2008-12-05 07:23	2008-12-04 13:10	wrece
910232	Done	Execution Comple	/scratch/z5/wreco/	LCG.RAL.uk	Ganga_Gauss_v35	2008-12-05 05:25	2008-12-05 05:25	2008-12-04 13:10	wrece
910221	Done	Execution Comple	/scratch/z5/wreco/	LCG.PIC.es	Ganga_Gauss_v35	2008-12-05 08:26	2008-12-05 08:26	2008-12-04 13:09	wrece
910209	Done	Execution Comple	/scratch/z5/wreco/	LCG.PIC.es	Ganga_Gauss_v35	2008-12-05 06:23	2008-12-05 06:23	2008-12-04 13:09	wrece
910199	Done	Execution Comple	/scratch/z5/wreco/	LCG.CNAF-T2.it	Ganga_Gauss_v35	2008-12-05 09:09	2008-12-05 09:09	2008-12-04 13:09	wrece
910188	Done	Execution Comple	/scratch/z5/wreco/	LCG.RAL.uk	Ganga_Gauss_v35	2008-12-05 04:49	2008-12-05 04:49	2008-12-04 13:08	wrece
910177	Done	Execution Comple	/scratch/z5/wreco/	LCG.GRIDKA.de	Ganga_Gauss_v35	2008-12-05 05:58	2008-12-05 05:58	2008-12-04 13:08	wrece
910166	Done	Execution Comple	/scratch/z5/wreco/	LCG.GRIDKA.de	Ganga_Gauss_v35	2008-12-04 23:24	2008-12-04 23:24	2008-12-04 13:08	wrece
910154	Done	Execution Comple	/scratch/z5/wreco/	LCG.GRIDKA.de	Ganga_Gauss_v35	2008-12-04 21:24	2008-12-04 21:24	2008-12-04 13:07	wrece
910145	Done	Execution Comple	/scratch/z5/wreco/	LCG.RAL.uk	Ganga_Gauss_v35	2008-12-05 05:31	2008-12-05 05:31	2008-12-04 13:07	wrece
910134	Done	Execution Comple	/scratch/z5/wreco/	LCG.PIC.es	Ganga_Gauss_v35	2008-12-05 08:16	2008-12-05 08:16	2008-12-04 13:07	wrece
910122	Done	Execution Comple	/scratch/z5/wreco/	LCG.PIC.es	Ganga_Gauss_v35	2008-12-05 08:19	2008-12-05 08:19	2008-12-04 13:07	wrece
910112	Done	Execution Comple	/scratch/z5/wreco/	LCG.GRIDKA.de	Ganga_Gauss_v35	2008-12-04 21:36	2008-12-04 21:36	2008-12-04 13:06	wrece
910099	Done	Execution Comple	/scratch/z5/wreco/	LCG.GRIDKA.de	Ganga_Gauss_v35	2008-12-04 21:36	2008-12-04 21:36	2008-12-04 13:06	wrece
910087	Done	Execution Comple	/scratch/z5/wreco/	LCG.GRIDKA.de	Ganga_Gauss_v35	2008-12-04 21:04	2008-12-04 21:04	2008-12-04 13:06	wrece
910076	Done	Execution Comple	/scratch/z5/wreco/	LCG.GRIDKA.de	Ganga_Gauss_v35	2008-12-05 07:42	2008-12-05 07:42	2008-12-04 13:05	wrece

Page 1 of 13

Items displaying per page: 25

Displaying 1 - 25 of 318

wrece@ lhcb ~ (/DC=ch/DC=cern/OU=Organic Units/OU=Users/CN=wreco/CN=667821/CN=Will Reece)

lhcbweb.pic.es Proxy: None

CORRECT USAGE OF COMPUTING RESOURCES

Testing/Debugging:



Full Running:



OUTLINE

- 1 INTRODUCTION
- 2 DISTRIBUTED ANALYSIS @ LHC*b*
- 3 GANGA
- 4 ETC.
- 5 SUMMARY**

SUMMARY

- The LCG provides LHC*b* users w/ a massive amount of CPU power and disk space.
- GANGA allows users to run jobs locally, on batch systems and on the Grid in a seamless way.
- GANGA is written in Python; its syntax is easy to understand.
- GANGA/LHC*b* provides a number of specific tools for running LHC*b* jobs wherever resources are available.
- Try getting started with the “hands on” GANGA/LHC*b* tutorial:
<https://twiki.cern.ch/twiki/bin/view/LHCb/GangaTutorial1>