GANGA Tutorial for LHC*b*

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

March 25^{th}, 2011

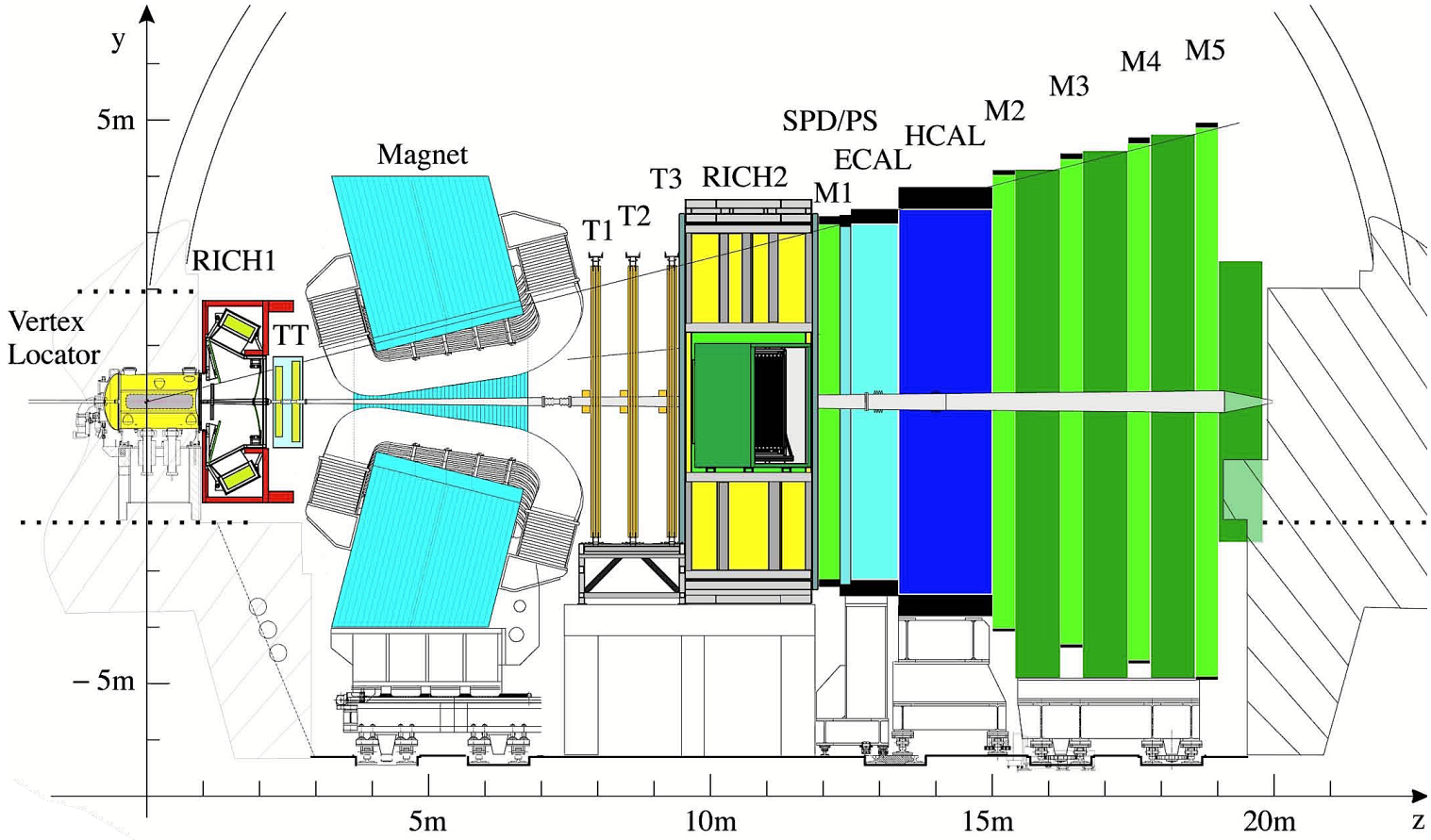
59^{th} LHCb Week

Outline

Outline

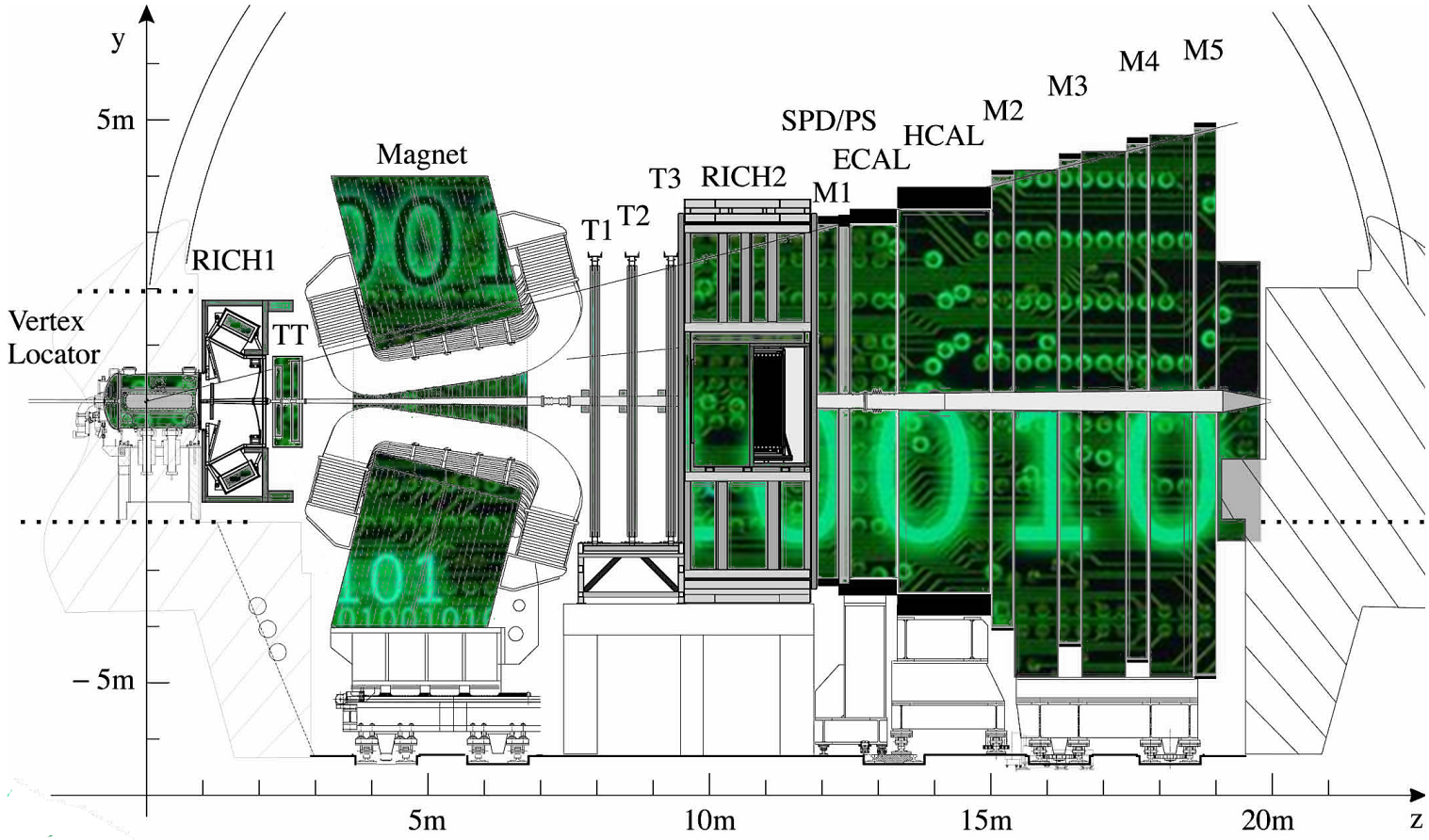
LHC*b* Data Taking

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



LHC*b* Data Taking

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

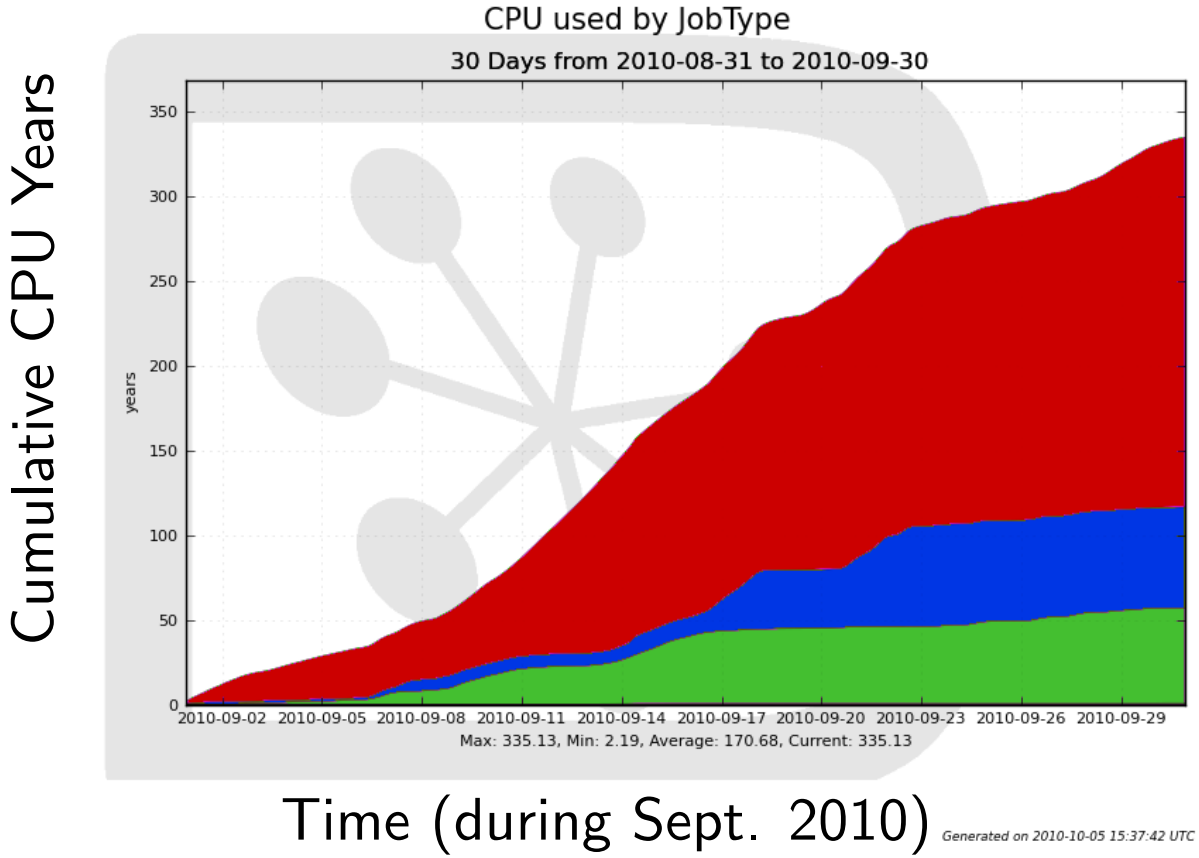


LHC*b* is a super bit factory (not to be confused w/ a Super B factory).

LHC*b* Computing Resource Usage

LHC*b* used over 300 CPU years (shared resources only) in Sept. 2010.

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



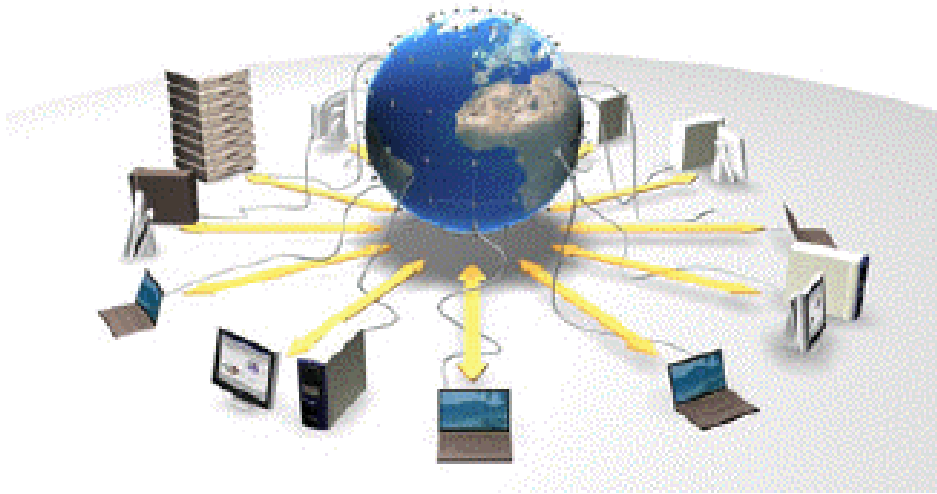
We've also used close to 700 TB of disk space in Sept. 2010. So, that's 10 CPU years and 35 TB every day. We need *The Grid*!

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 LHC*b* 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 `.globus` directory). By joining a *virtual orginization* (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 LHC*b* VO!

Outline

LHC*b* Division of Labor

GANGA is a user-friendly frontend that handles job definition and management for LHC*b* 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 LHC*b*. It does the *heavy lifting* for all DA in LHC*b*.

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

The DIRAC WMS/DMS

Distributed **I**nfrastructure w/ **R**emote **A**gent **C**ontrol

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;

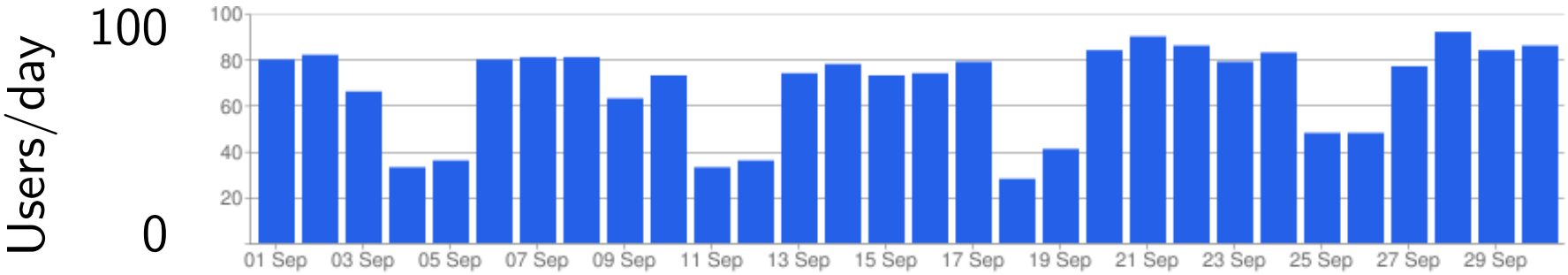
–NOT–

"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 LHC*b*

Almost 100% of LHC*b* grid users used GANGA in Sept. 2010



GANGA provides a *complete* analysis environment for LHC*b* and greatly simplifies the user experience (the topic of the rest of this talk). Thus, the vast majority of LHC*b* 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

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!

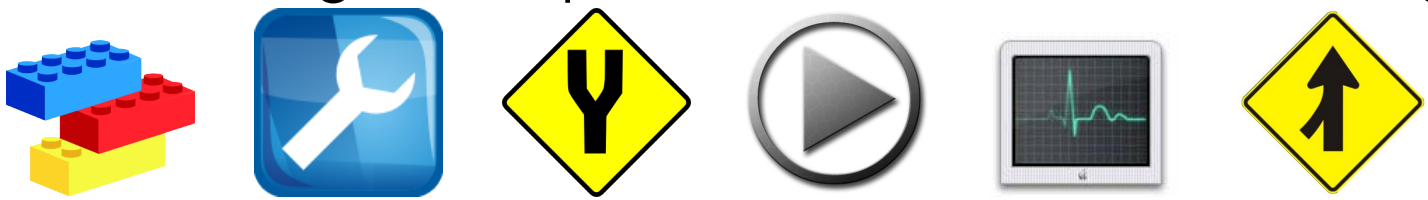
GANGA was developed to meet the needs of ATLAS & LHC*b* for a grid user interface and is now used by many other groups as well. Usage: 45% ATLAS, 45% LHC*b*, 10% other.

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 (`DSTMerger`);
- 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

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

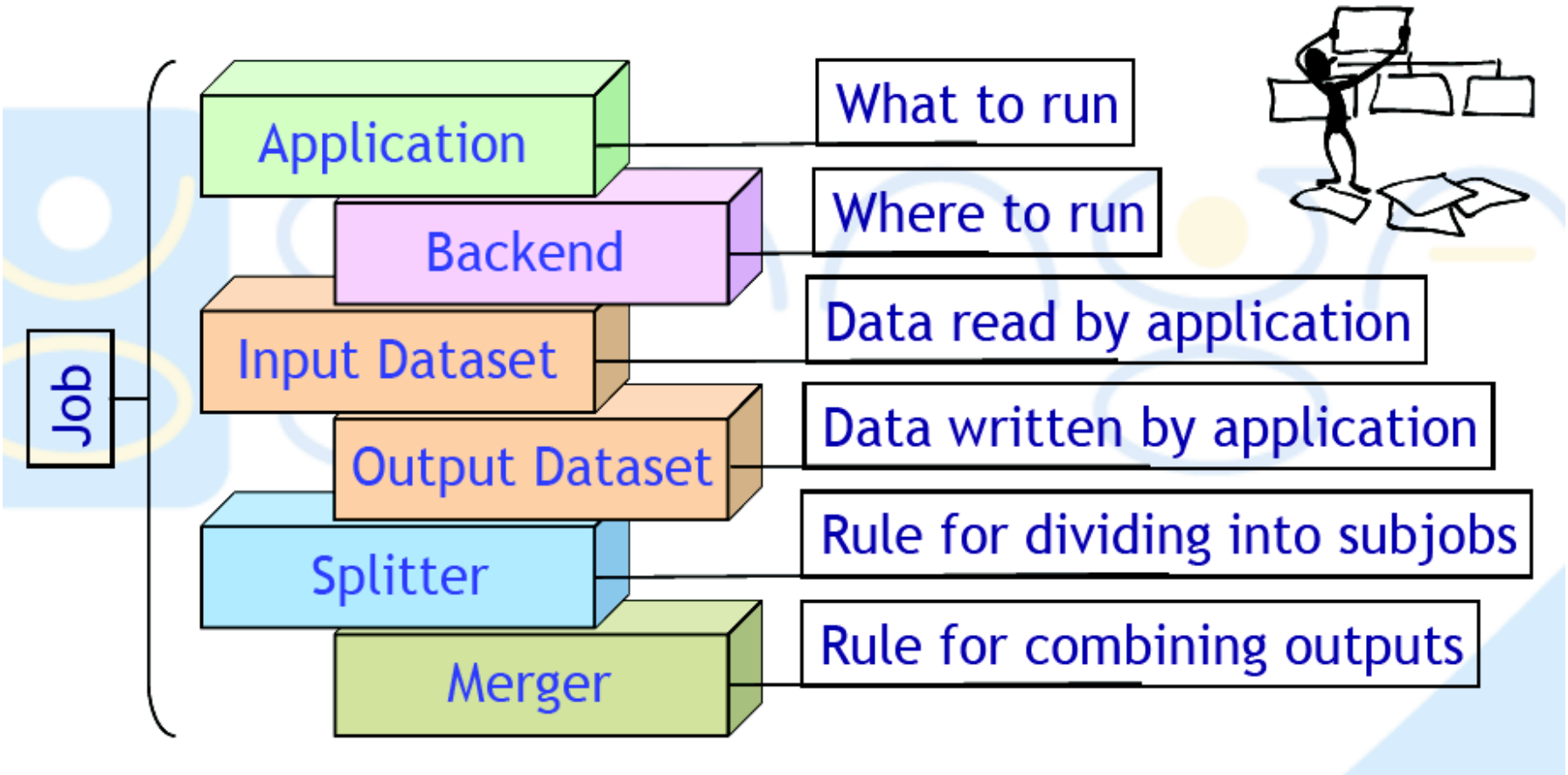
- Python syntax;
- Shell commands;
- TAB completion, scrolling thru your history, *etc.*

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).

Applications/Backends

GANGA/LHC*b* 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/LHC*b* 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/LHC*b* supports the following splitters:

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

GANGA/LHC*b* 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>/Bs2JPsiPhi.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:

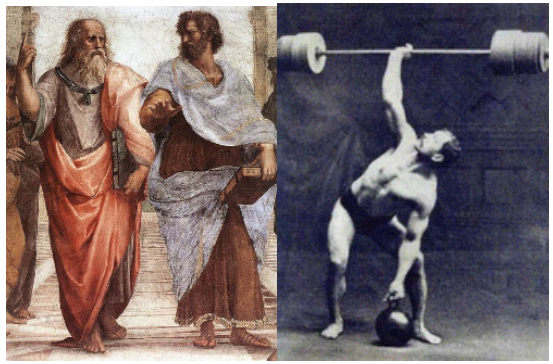
```
In[8]:j.peek('DVHistos_1.root')
```

or specify the program you want to use:

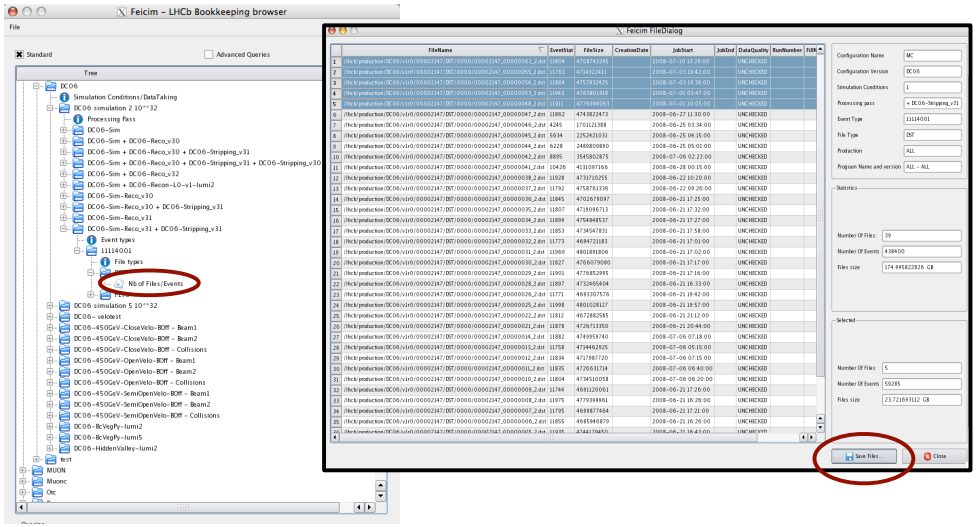
```
In[8]:j.peek('stdout','less').
```

More GANGA LHC*b* 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()`



- 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

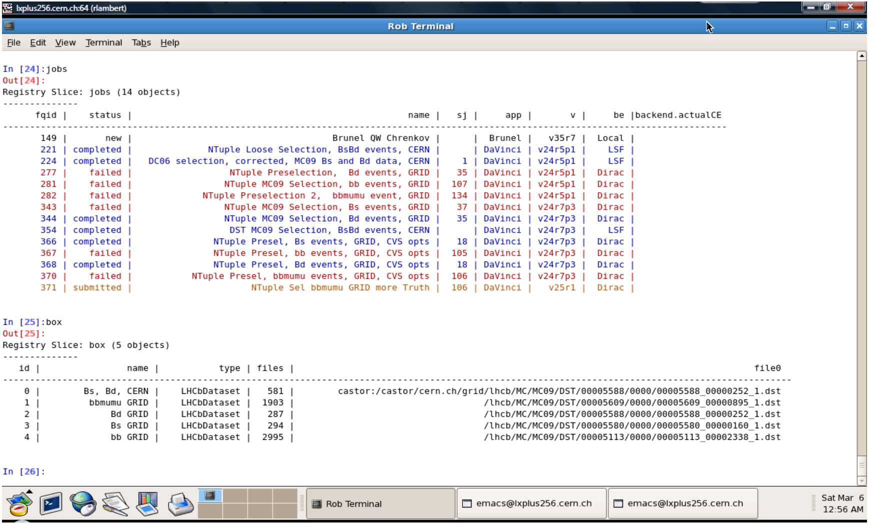
GANGA Help

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).

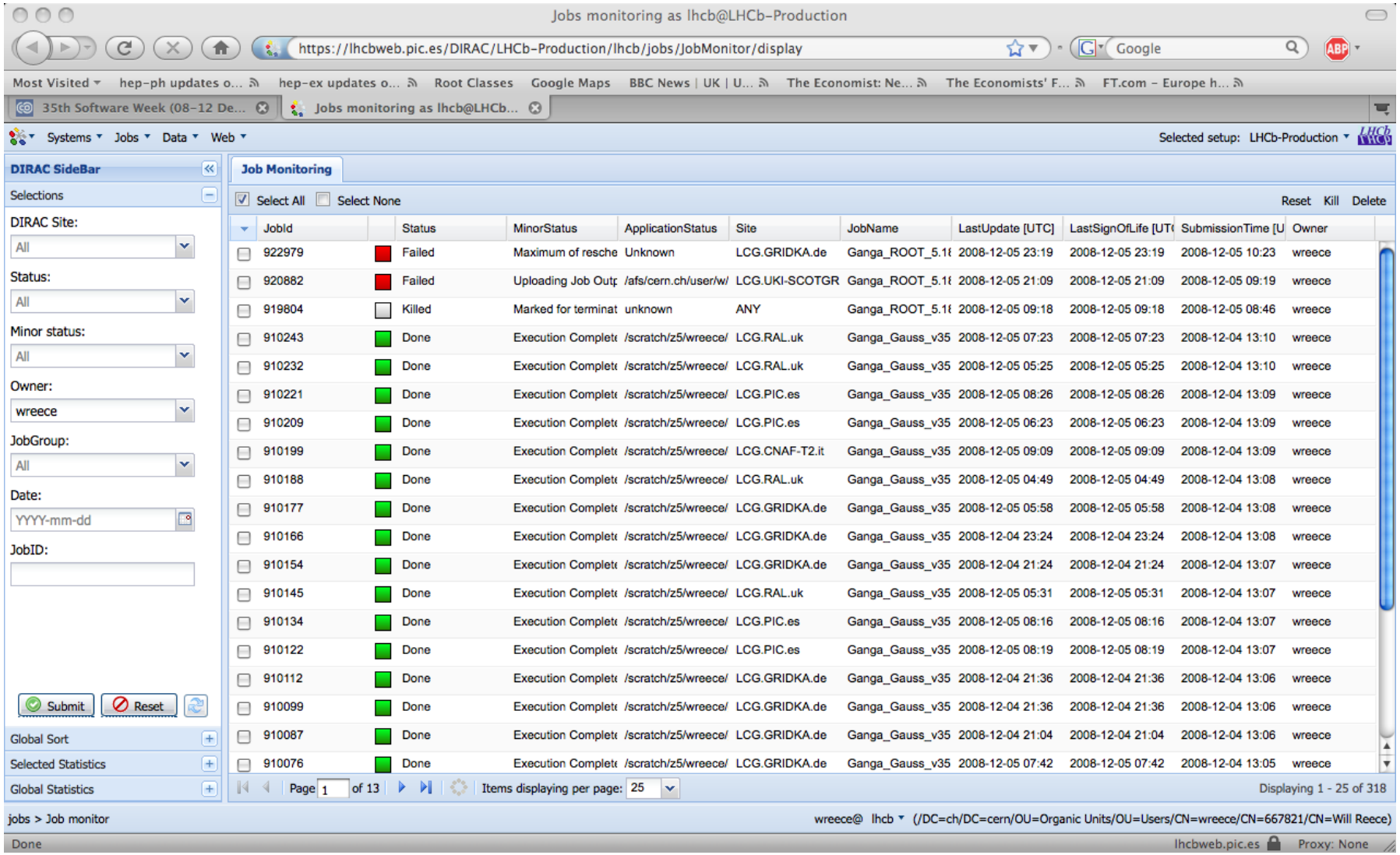
For Python help, see http://docs.python.org/tut/tut.html

Persistence



DIRAC Monitoring

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



Correct Usage of Computing Resources

Testing/Debuging:

Full Running:

Outline

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