Programmatic access to file syncing services

Cloud computing workshop

Patrick Owen
on behalf of the Ganga developer team

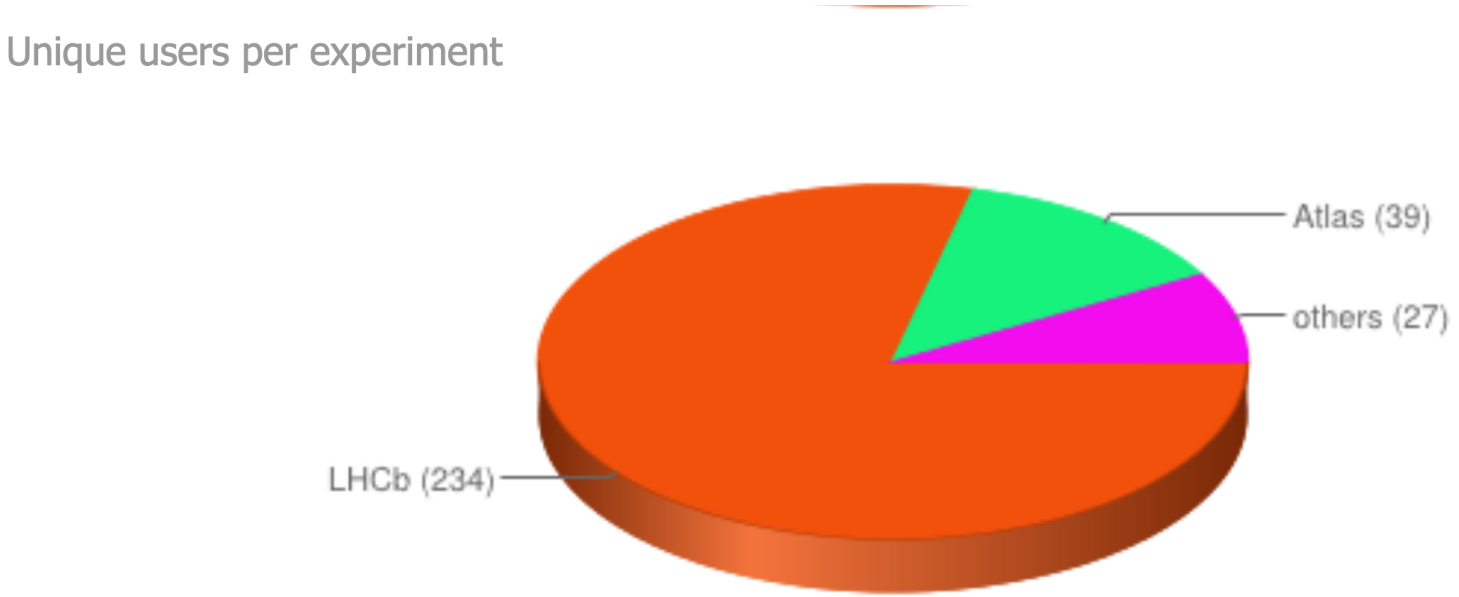
Imperial College London

Computing for LHC physicists

- Physicists at CERN often need to perform CPU/data intensive tasks.
  - Processing data from the LHC.
  - Running complicated analysis.
- These often involving using CPU years and/or pbytes of data.
- For these we often use the LHC Computing Grid (LCG), which is our distributed system of computing.
- Interacting with the LCG is not straightforward for the typical physicist.
  - We provide the software package Ganga to simplify interface.
  - Also try to help protect against expensive mistakes.

Ganga

- Ganga is a python API for distributed analysis.
- Primary use: submitting jobs to the LCG.
- Also used for infrastructure testing of ATLAS, CMS and LHCb (HammerCloud).
- Have about 300 users, most from the LHCb experiment.
  - Other users include ATLAS and SNO+.



Typical usage in Ganga

- Typical usage.
  - Run over a few hundred GB
  - Take a few CPU days.

```
In [11]:j=Job()

In [12]:j.application = Executable()

In [13]:j.outputfiles = [LocalFile("myfile.txt")]

In [14]:j.submit()
```

- j.ouputfiles attribute defines where user wants what and where.

Uploading and downloading files

- Interface simple and easily scalable.

```
In [6]:d = DiracFile('BdKsMuMu.root')


In [6]:d.put()
Ganga.GangaDirac.Lib.Files          : INFO     Uploading file /afs/cern.ch/user/p/powen/BdKsMuMu.root

In [7]:cd analysis
/afs/cern.ch/user/p/powen/analysis

In [8]:d.get()
Ganga.GangaDirac.Lib.Files          : INFO     Getting file /lhcb/user/p/powen/96800c6a-90a4-4dd0-813e-822a945b8fe9/BdKsMuMu.root

In [9]:d
Out[9]: DiracFile (
 namePattern = 'BdKsMuMu.root' ,
 guid = '8412E508-CEDD-9643-87DC-524560C37668' ,
 remoteDir = '96800c6a-90a4-4dd0-813e-822a945b8fe9' ,
 localDir = None ,
 lfn = '/lhcb/user/p/powen/96800c6a-90a4-4dd0-813e-822a945b8fe9/BdKsMuMu.root' ,
 failureReason = '' ,
 locations = [CERN-USER] ,
 compressed = False
 )
```

Current limitations

- Until last year, users could specify output to go:
  - Back to the local directory.
  - Left on user LCG storage.
  - Uploaded to separate mass storage system.
  - Previously only accessible through grid middleware.
- However, need to manually copy to local machine.
- Also, physicists often want to share very large files.

Case for cloud storage

- What one might want to achieve:
  - Ease of use for both the storage of the data and the retrieval.
  - Possibility to share large files safely and easily.
  - Integration with existing software solutions for cloud storage.
  - No new authentication mechanisms.
  - Secure (i.e. not transmitting and storing new tokens in plain text).

GoogleFile implementation

- Authentication set up via OAuth.

```
In [3]:j.outputfiles = [GoogleFile('myfile.txt')]
Go to the following link in your browser: https://accounts.google.com/o/oauth2/auth?scope=https%3A%2F%2Fwww.googleapis.com%2Fauth%2Fdrive.file&redirect_uri=urn%3Aietf%3Awg%3Aoauth%3A2.0%3Aoob&response_type=code&client_id=54459939297.apps.googleusercontent.com&access_type=offline
Enter verification code: 4/iPfYm_83eScDO3L23I2XhJu_BvKUJqCX0OrNyrdxEl8.8rkYXgWKjcEatIqXn3XwmZwD0sZGkwI
Ganga.GPIDev.Lib.File.GoogleFile    : INFO     Your GoogleDrive credentials have been stored in the file /afs/cern.ch/work/p/powen/private/gangadir/googlecreddata.pkl and are only readable by you. The file will give permission to modify files in your GoogleDrive. Permission can be revoked by going to "Manage Apps" in your GoogleDrive or by deleting the credentials through the deleteCredentials GoogleFile method.
```

- After job has finished, 'myfile.txt' uploaded to my Google Drive.
- Only 15GB available.
  - Helpful to look at small files locally when job finished.
  - Not so useful for sharing large datafiles.
  - Quickly clogs up my gmail quota ..
  - Requires separate authentication, yet another storage system ..

CERNBox

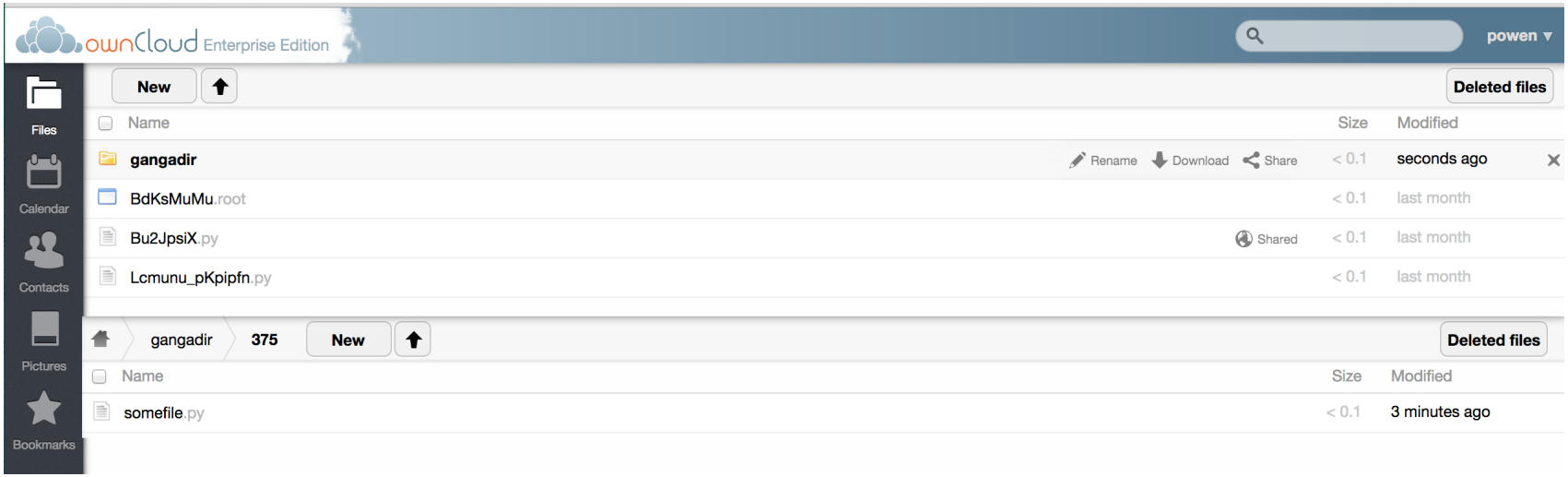
- The CERNBox server can solve these problems.
  - Larger space to start with (100GB).
  - Available for each CERN user.
  - Potential for syncing with larger experiment specific storage (EOS).
- All the user does is attach CERNBoxFile to output files attribute.

```
In [12]:j=Job()

In [13]:j.outputfiles = [CERNBoxFile('somefile.py')]

In [14]:j.submit()
Ganga.GPIDev.Lib.Job              : INFO     submitting job 375
Ganga.GPIDev.Lib.Job              : INFO     job 375 status changed to "submitting"
Ganga.GPIDev.Lib.File             : INFO     Preparing Executable application.
Ganga.GPIDev.Lib.File             : INFO     Created shared directory: conf-046e6329-5512-46bc-a7bd-600e37ae6a21
Ganga.GPIDev.Adapters             : INFO     submitting job 375 to Local backend
Ganga.GPIDev.Lib.Job              : INFO     job 375 status changed to "submitting"
Ganga.GPIDev.Lib.Job              : INFO     job 375 status changed to "submitted"
```

- Have prototype of CERNbox file, which interacts with server via WebDAV.
  - Can easily expand this to use other cloud storage systems.
- Currently just using basic authentication for testing.

CERNbox

- When job completes, makes folder inside gangadir and can sync to local machine.
- User has full control over path structure.
- Certain files can be placed in shared folders for collaboration.



Summary

- Ganga is a python API for doing distributed analysis in high energy physics.
- We have simple, scalable file interface for data management.
- Cloud computing will help simplify data analysis for physicists.
  - Physicists want to share large files easily.
  - They also like to work late at night perhaps with poor internet connection.
- Have implemented interface for Google Drive and CERNBox.
  - Cloud storage interface more easily allows other collaborations to join Ganga.
  - In early stages of development, not a huge amount of user feedback yet.