GridPilot – a graphical front-end to distributed compute systems

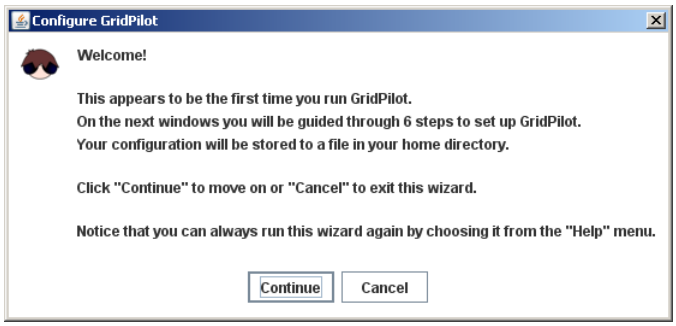
Introduction and quick-start

Frederik Orellana
Niels Bohr Institute, University of Copenhagen
December 2010

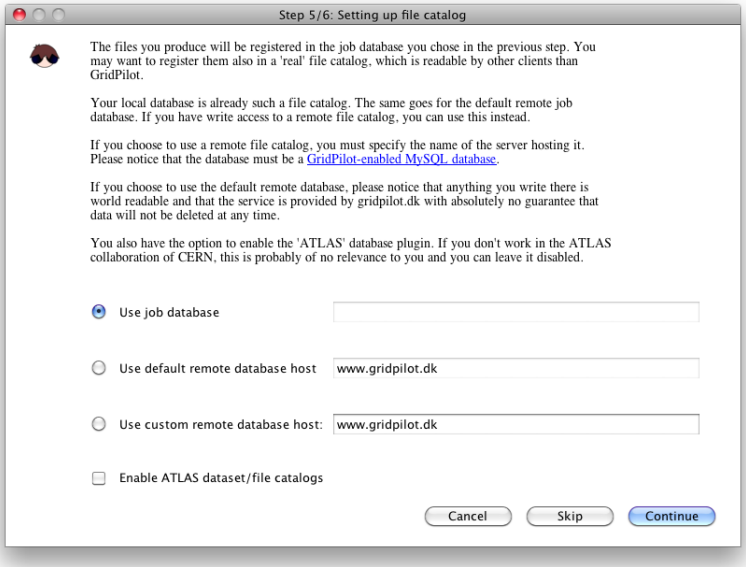
1 Introduction

GridPilot is a tool to control and keep track of your personal data production on grid and cloud resources. GridPilot was motivated by trying to simplify the large computing skills demanded of the physicsts in the ATLAS experiment at CERN, but developed into a general-purpose tool.

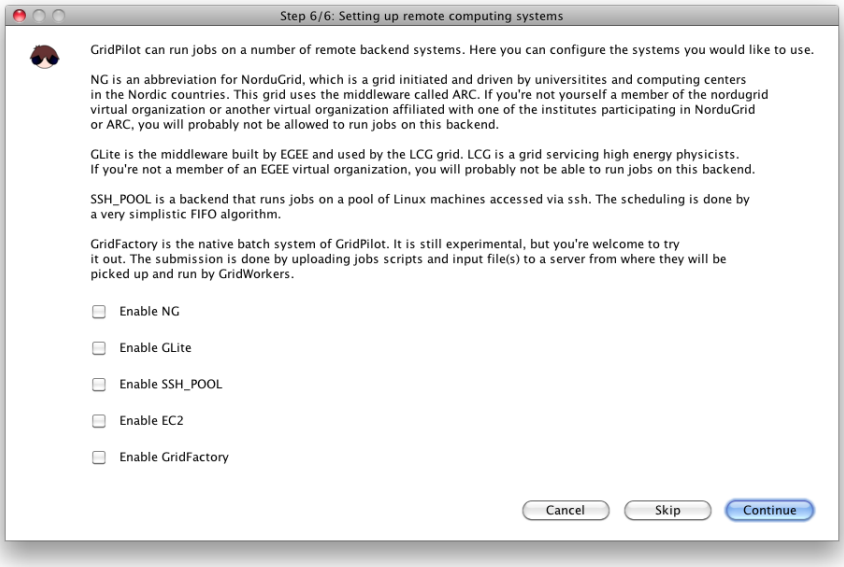
After downloading and installing, when you start GridPilot for the first time, you'll be guided through a setup wizard.



On most of the screens, you can simply click “OK” to accept the defaults. CERN/ATLAS users should pay special attention to screen 5 and remember to enable the ATLAS file catalog.



In general, attention need only be paid to the last screen (6), where computing back-ends are chosen.



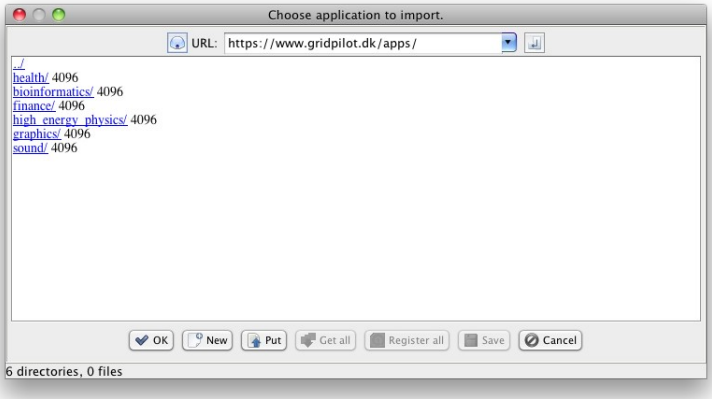
Notice that you can always configure GridPilot “manually” by editing your preferences or running the setup wizard again (“Help” → “Wizard: Configure GridPilot”).

The next section will help you get started by guiding you through a simple example.

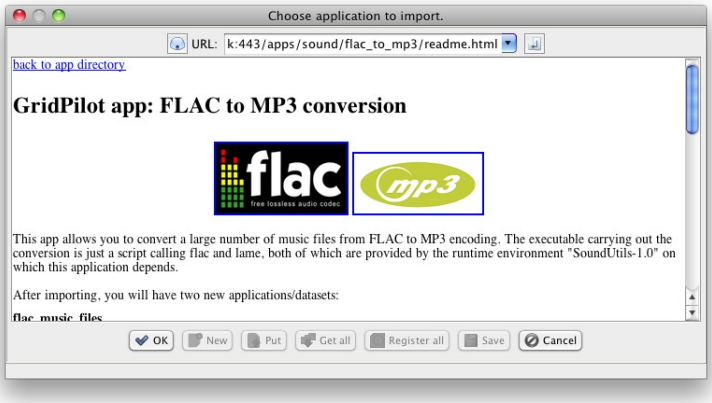
2 Converting 12 flac files to mp3 format

If you're running GridPilot for the first time, you'll be asked if you want to import an application from the GridPilot “app store”. Notice that you can always import applications by choosing “File” → “Import application(s)”.

In either case, you'll be presented with the GridPilot browser.

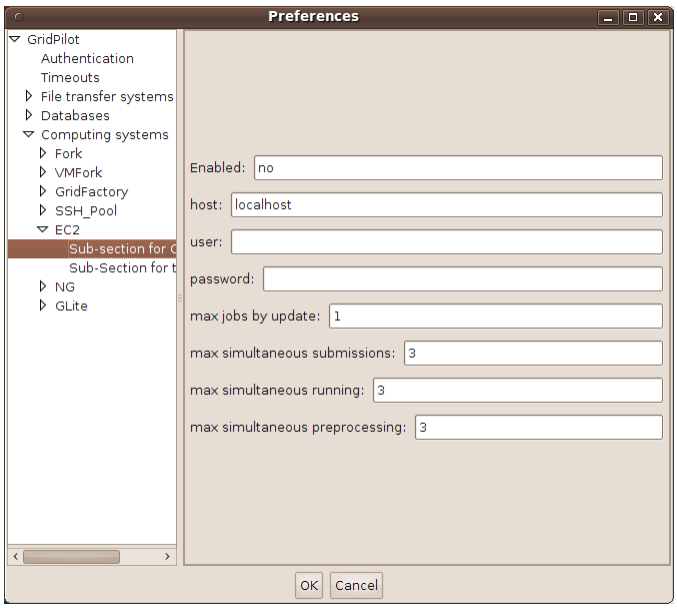


If you navigate to the directory “sound” and click on the directory “flac_to_mp3”, the browser will display the readme file of the application.

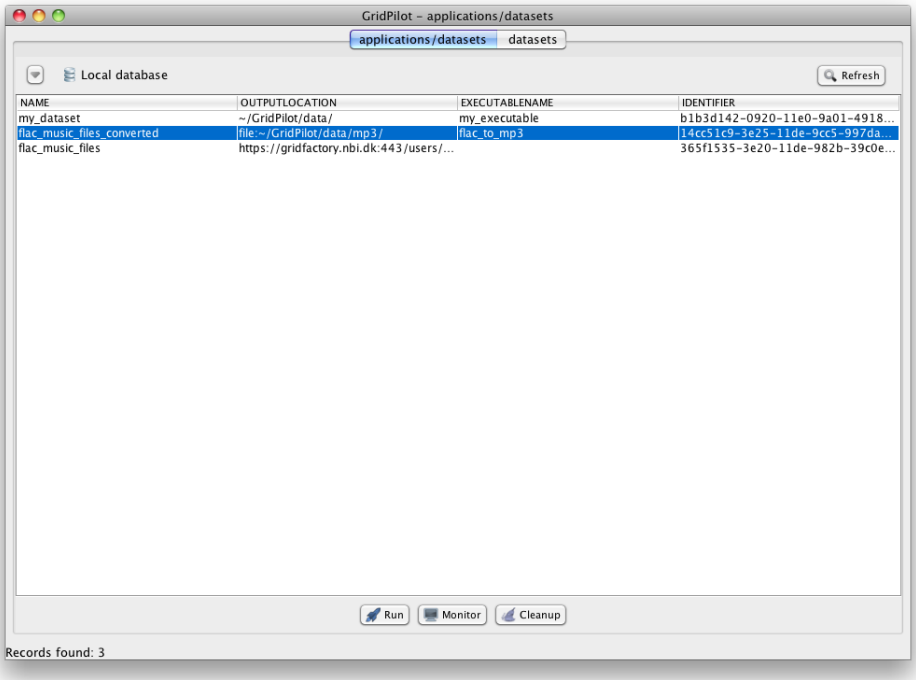


Read this file and click the “OK” button on the navigation bar at the bottom of the browser. If you want to read this file again, start a GridPilot browser from “View” → “New file browser” and use the URL history drop-down.

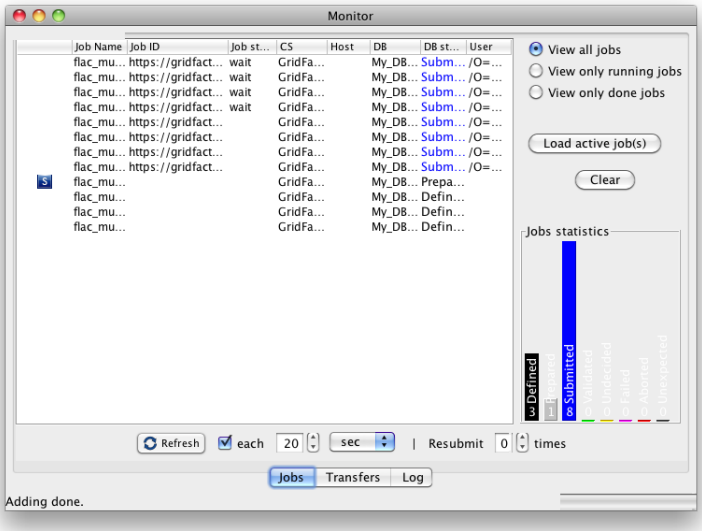
Now, in order to run this application, as you saw in the readme file, you must have either the VMFork, GridFactory or EC2 computing back-end enabled in your preferences. If you didn't configure one of them with the setup wizard, you can open the preferences and do this manually.



If you have at least one of these back-ends enabled, you can select the record “flac_music_files_converted” on the “Applications/Datasets” tab, click “Run” and choose a computing back-end to run on.



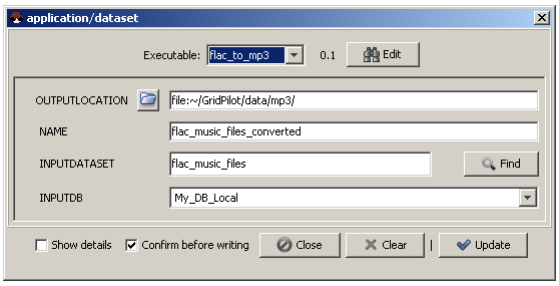
This will pop up a window for creating job definitions; here you can simply click “Create”. This will pop up another window asking if you want to add some files to this dataset. Simply click “Yes”, then “OK”. You will then populate the input dataset “flac_music_files” with 12 flac files in the public domain. After that a confirmation window will pop up – simply click “OK for all”. Then the GridPilot monitor window will open and you can follow as jobs are being submitted and run by the back-end system.



3 Data management

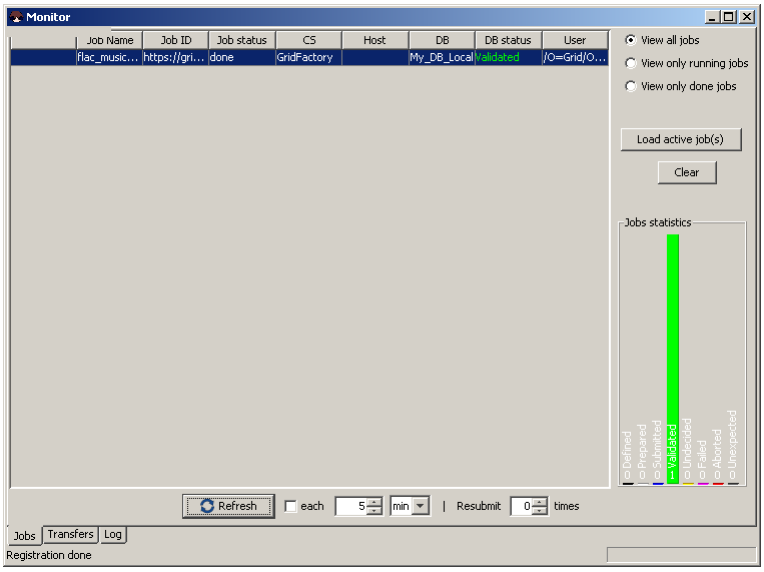
Once your jobs start to finish, you'll probably wonder where you can find the output files. To understand this, let's look a bit more in details at the application you just ran: On the main GridPilot window, right-click on the record “flac_music_files_converted” on the “Applications/Datasets” tab and choose “Edit”. Then a new window will pop up, allowing you to see and change all the fields of this record. You'll notice that “OUTPUTLOCATION” is set to “file:~/GridPilot/data/mp3/”. This means that the output files produced by this application will be put in the folder GridPilot/data/mp3 in your home directory. E.g. on Windows, this would typically be something like “C:\Documents and

Settings\Frederik Orellana\GridPilot\data\mp3". You can change this changing the text or by clicking on the little folder icon and navigate to a folder of your choice.



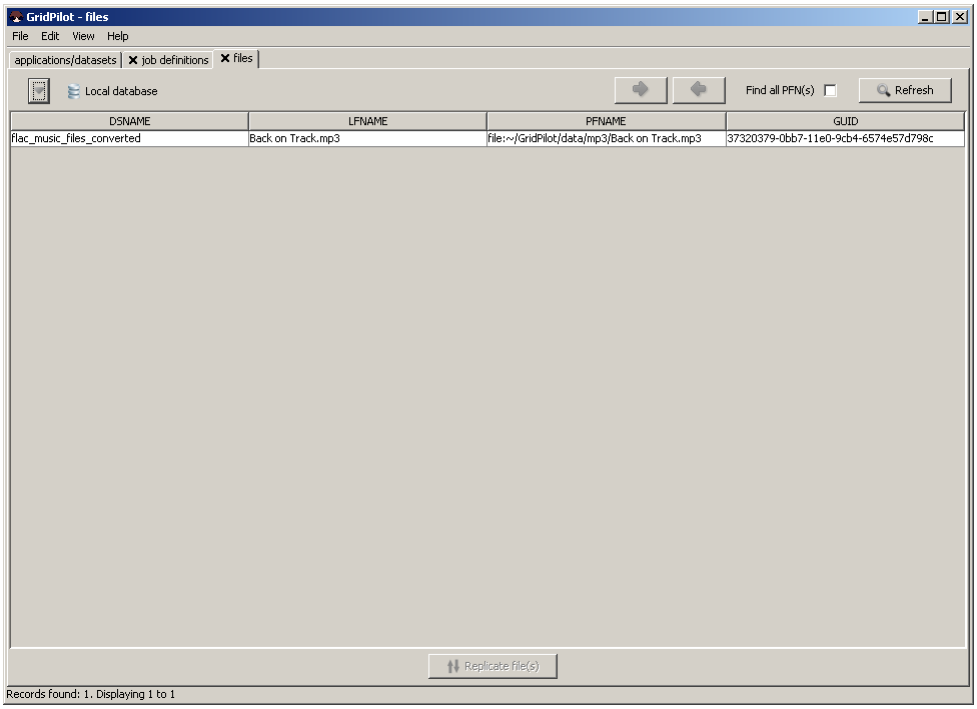
Notice that the folder you choose doesn't necessarily have to be on your own hard disk; i.e. the URL does not necessarily have to start with "file:". Other possibilities can be "https:", "sss:", "gsiftp:" and "srm:". Which of these will actually work, depends on which computing back-end you're using, which credentials you're using and which credentials are accepted by the file server in question. GridFactory, for example, can copy output files to https servers and does so using the credentials of the GridFactory server to which you submitted your jobs. NG (NorduGrid/ARC) and Glite (LCG/EGEE/EGI) can copy output files to https, gsiftp and srm servers and does so using the proxy credentials you've delegated to the server to which you submitted your jobs.

When GridPilot detects that a job has finished, it'll download and check that the stdout and stderr of the job.



If all is ok[1], the output file of the job is downloaded and registered. In the current case, registration means that a record is created in the "File" table of the local database. To inspect this record, on the main GridPilot window, right-click on the record "flac_music_files_converted" on the "Applications/Datasets" tab and choose "Show file(s)".

[1]Actually all that is checked is that stdout exists (it must, since the job wrapper script always outputs some text) and that stderr is empty. It is possible to define some patterns that are ignored in stderr in the preferences.



If you enable more database back-ends in your preferences, you can publish your files in other databases by simply copy-pasting records from one tab to another, but describing this in detail belongs in a more exhaustive document than the present.

4 Computing back-ends and software packages

GridPilot supports running computing jobs on a variety of back-end systems. In this section we'll look at 5 such systems. To enable each of these, in the preferences, click on “GridPilot” → “Computing systems” → [back-end system] → “Sub-section for GridPilot” and set “enabled” to “yes”. You should enable only those you need and for which you have valid credentials. Enabling more prolongs the startup time.

Your jobs may require software in order to run and may only run on a specific operating system. GridPilot keeps a table of available software packages and platforms available on each back-end system[2].

WARNING: GridPilot does currently not check that you have enough disk space to download virtual machines and software, so keep an eye on your available disk space.

4.1 VMFork

This back-end runs jobs inside a virtual machine running on your local computer. The virtual machine is downloaded from the software catalogue defined in “GridPilot” → “runtime catalog URLs”.

Anyone can use this back-end – as long as the machine on which GridPilot is running is running Linux or Windows and has enough free disk space (>30 GB) and RAM (>1 GB) available.

Notice that virtual machines can be quite large and downloading can take some time. Once downloaded, the virtual machine is cached. Therefore, the first job you run on a given virtual machine will take a long time to start.

4.2 GridFactory

This back-end runs jobs inside virtual machines running on remote computers – which are part of a GridFactory cluster. Again, the virtual machines are downloaded from the software catalogue defined

[2]If you first set “GridPilot” → “advanced mode” to “yes”, you may inspect this table by clicking “View” → “New tab with My_DB_Local” → “runtimeEnvironments”

in “GridPilot” → “runtime catalog URLs”, and also again, notice that the first job you run on a given virtual machine may take a long time to start.

By default, jobs are submitted to the central GridFactory server at www.gridfactory.org. This does not have any permanent resources attached, so whether or not your jobs start depends on whether or not some people are donating resources. To change server, click “GridPilot” → “Computing systems” → “GridFactory” → “Sub-section for this plug-in” and change “submission url”.

This back-end uses SSL certificate/key for authentication. The default location of these is two files, “usercert.pem” and “userkey.pem” in a folder “.globus” in your home directory. If you already have two such files, you should ask the administrator of the GridFactory cluster in question to allow your corresponding distinguished name to run jobs on his cluster. You can also temporarily move “.globus” somewhere else – in which case GridPilot will use default credentials that are allowed to run jobs on the central GridFactory server.

4.3 EC2

This back-end runs jobs inside virtual machines running on the Amazon Elastic Compute Cloud (EC2). The virtual machines are downloaded from the software catalogue defined in “GridPilot” → “Computing systems” → “EC2” → “Sub-section for this plug-in” → “runtime catalog URLs”.

To use this back-end, you must have two access identifiers that Amazon provide to anyone who has signed up at http://aws.amazon.com/. You must set them in the corresponding preferences entries.

4.4 NG

This back-end runs jobs on remote computers – which are part of NorduGrid and run the ARC middleware. Each cluster publishes the software it has installed and GridPilot uses this information to populate its “runtimeEnvironments” table for NG.

This back-end also uses SSL certificate/key for authentication. The default location of these is two files, “usercert.pem” and “userkey.pem” in a folder “.globus” in your home directory. To use this back-end, you must have such two files, they must be issued by a grid certification authority and your corresponding distinguished name must be allowed to run on one or more ARC clusters, i.e. you need to be member of a virtual organization recognized by NorduGrid.

You should fill in the field “GridPilot” → “Computing systems” → “NG” → “Sub-section for this plug-in” → “clusters” with the IP name of a cluster on which you are allowed to run jobs. If you leave this field empty, all known NorduGrid clusters will be queried – which will cause job submissions to be very slow.

4.5 GLite

This back-end runs jobs on remote computers – which are part of the EGEE/EGI or LCG grids and run the gLite middleware. Each cluster publishes the software it has installed and GridPilot uses this information to populate its “runtimeEnvironments” table for NG.

This back-end also uses SSL certificate/key for authentication. The default location of these is two files, “usercert.pem” and “userkey.pem” in a folder “.globus” in your home directory. To use this back-end, you must have such two files, they must be issued by a grid certification authority and your corresponding distinguished name must be allowed to run on one or more gLite clusters i.e. you need to be member of a virtual organization recognized by the grid in question.