Migration to Debian based worker nodes

This page details the changes incurred by our migration of the Research Grid worker nodes to a completely new software stack, which includes a new set of applications and middleware.

Please check it frequently for updates!

Contents

Brief introduction into the current state of the ResGrid cluster

Currently the worker nodes of the ResGrid cluster are split into two logical partitions, running separate software stacks:

- **The "old" RHEL6 based nodes**
  - Run RHEL6 (or RHEL7 on a very few special ones) operating system
  - Have their own application repository (accessible under /apps)
  - Are sub-partitioned in several SGE pools (for the purpose of scheduling):
    - "all" -- generic main pool
    - "interactive" -- dedicated to interactive work
    - "long" -- long jobs not subject to the 1week execution time limit
    - "large" -- large memory allocation
    - "RHEL7" -- run RHEL7 operating system
- **The "new" Debian based nodes**
  - Run Debian 10 (Buster) operating system
  - Have their own application repository (accessible under /apps -- **but different from the one above**) !!!
  - Are all part of a single SGE pool ("debian")
    - The plan is to reconfigure the SGE workload manager and implement a priority based scheduling system (ToDo)
    - *For the moment* a general 10days wall-clock time limit applied to all activities.

There is only one head node (researchgrid.gsb.columbia.edu) that provides access to *all* worker nodes. The destination distinction is managed by the **"Grid Wrapper"** -- a script that handles interactive requests as well as batch submittals, and which is transparently interposed in front of several "application invocations" in order to distribute the resulting workload over the ensemble of worker nodes. Depending on how the wrapper is invoked, the workload might end up in one or the other system partition, and furthermore, in one of the various SGE (scheduler) pools. See the **"Grid Wrapper"** subsection for more details.

The Grid Wrapper

The Grid Wrapper is a Perl script that insulates the user from the SGE workload manager's complexities, which also tries to make some educated choices regarding the best manner to process a certain workload.

The Grid Wrapper has multiple identities (!) as it can be invoked by several names. The purpose of this impersonation is to intercept calls to common applications such that the resulting workload is managed by the SGE scheduler, distributed over the set of worker nodes for better user response and increased system efficiency. Thus the "wrapper" function.

Let's see few examples:

Assume you're connected to the head node.

Typing the simple command: "matlab --grid_mem=4GB" :

- The command "matlab" is a link to the Wrapper and instead of launching MATLAB it calls the Wrapper
- The wrapper recognizes that it was called under the name of "matlab" and will:
    - Parse the command options, here only "--grid_mem=4GB" and recognize that the user needs 4GB of mem allocated to the task
    - Create a task request for the SGE scheduler using the proper SGE options to request 4GB or mem and run the real MATLAB application with all the passed arguments
    - Submit that request to SGE
- SGE will schedule and run that task on the least busy node

Any other commands that resembles an application name and is part of the list of calls displayed by the command "gridhelp", are applications similarly "wrapped" in these preparatory steps. Examples: R, STATA, SAS, etc.

A special case is running the wrapper to execute an arbitrary script, by calling "sge_run". The "--grid_*" options are SGE (scheduler) options and are processed in the same way. What comes after them in the list of arguments is the name of the script (or a system command) plus its own arguments. This latter part is the job "payload" and it's passed to the scheduler in the same manner as calling an application described above is. In fact the similarity goes in the other direction: an application wrap is just a special case of script/command scheduled execution, where the command to run is the application name that replaced the "sge_run" name.

THE GRID WRAPPER IS REDESIGNED AND CERTAIN COMMANDS WILL REQUIRED A SLIGHTLY CHANGED SYNTAX . Please check Towards a more rational Grid Wrapper for details.

Two (classes of) commands require new options:

```
stata  --grid_submit=batch  [other --grid_options]  -b do  <script>  <arguments>
R --grid_submit=batch  [other --grid_options]  CMD BATCH  <script>  <arguments>
```

All alternative invocations of R and STATA (see the next section), using **BATCH** mode, will require the additional "-b do" or "CMD BATCH" options. The "versioned" command forms that run on the old RHEL6 nodes are given until July 1st to add the options. The new Debian nodes will require it immediately.

Application Support

The table below contains the applications supported on each Grid "partition" (RHEL6 or Debian), their versions and the corresponding way to invoke them. (To maintain simplicity, where available we won't list the graphic interface alternative invocations -- the commands are similar, just prepend "x" to the listed command name) Reading the table is simple: e.g. typing "matlab" on the head node will run MATLAB R2019a on a Debian node, while typing "matlabr2017a" will run MATLAB R2017a on a RHEL6 node... Each command maps to a corresponding application on *one partition only*; to chose the Grid partition on which to run, pick the command alternative that activates the job on the desired partition.

As a general rule: **the most recent version of an application should be available on the Debian partition using the "version-less" command name**. (See "work in progress" note!)

Work in progress: all applications not listed below (but listed in the output of the "gridhelp" command) are available only on the RHEL6 partition. We're working on installing them on the Debian partition too, and will update this table as they become available.

Application	Command	Debian vers.	RHEL6 vers.
MATLAB	matlab	2019a	
	matlabr2007b		R2007b
	matlabr2016a		R2016a
	matlabr2017a		R2017a
R	R	3.6.0	
	Rold	3.5.2	
	R-3.0.2		3.0.2
	R-3.1.2		3.1.2
	R-3.2.2		3.2.2
	R-3.2.3		3.2.3
	R-3.4.0		3.4.0
	R-3.4.1		3.4.1
STATA	stata	STATA-mp 15	
	stata15		STATA 15
	stata15se		STATA-se 15
	stata15mp		STATA-mp 15
	stata14		STATA 14
	stata14se		STATA-se 14
	stata14mp		STATA-mp 14
Interactive session or script	sge_run	Interactive session or script	
	R6_sge_run		Interactive session or script
All other applic listed by "gridhelp"	<command>	[not avail]	<applic.vers.>

Notes

Retrieved from "http://gridserv03.gsb.columbia.edu/research/index.php?title=Migration_to_Debian_based_worker_nodes&oldid=5636"

This page was last edited on 23 May 2019, at 10:54.