# Guide to using the Weizmann ATLAS Grid cluster

31 May 2020

A guide to the policies and procedures in using the ATLAS Grid Cluster at the Weizmann Institute of Science

## General

Please follow the policies described in this document. Failure to do so will put you on our "Anti-social Users" list.

- Send all requests for help to: lcg-managers@weizmann.ac.il .
- When reporting a problem, provide all details needed to reproduce the issue. Furthermore, try to come with a **minimal** example demonstrating the problem, like the proverbial "Hello world" in the domain of programming.
- To get an account, you need to be in the Weizmann directory and have a Weizmann Unix computer account. You must abide by the same Acceptable Use Policy as required by the Institute's Computer Center.
- The cluster is intended to be used mostly in batch mode, i. e. you submit your program as a script to a scheduler, it is queued, and then run on an available "work node". Short test jobs and interactive analysis jobs may be run interactively on designated interactive computers; see below.
- The cluster is NOT your home base for browsing, email, etc. Use your laptop/desktop for these activities.
- There is no printer in the cluster. You must transfer files to print to your personal computer or a general-purpose WIS Unix server like clever or kesem.

## Security and login

The cluster is outside the Weizmann firewall and is protected by its own firewall.

- ssh access to the cluster is only from computers inside Weizmann, i.e. on the wired network or from the WIS Secure WiFi. To log in from home, Weizmann Hotspot WiFi, or abroad, connect to the Weizmann VPN (specifying **stvpn.weizmann.ac.il** as the VPN server) using your SecureID. Then ssh to wipp-ui1, wipp-an1 or wipp-an2.
- There is no telnet or ftp access to the cluster.

## Storage

- Linux does not like more than about 1000 files in a directory. Use subdirectories to organize your files. You and your jobs, and to some extent, your colleagues, will suffer if you have more than this number in one sub-directory.

- Do not use the following characters in Lustre file names: " ", "+","=". In general, stick to the alphanumeric subset of the ASCII table. This will make your life easier, e.g., when need arises to transfer data to/from a computer running another operating system. Spaces in filenames are especially notorious for causing troubles in various scripts; avoid them at all costs.

## Home directory

- Your home directory is only for high value files, for example program code that you have written, results that take a long time to compute, etc. Do not store temporary files here. This is the **only** file space that is backed up.
- You are limited to 30 GB and 50,000 files. This is for **all** the files on /srv01 that are owned by you, not just the ones in your home directory, i.e. the limits include your files in shared group directories, if there are any.
- Any file present at the time of the daily backup will be saved for **three months**, so **do not put temporary files here** because they will occupy space on the backup disk for 3 months.
- Avoid installing software in your home directory. You may request that software of general use be installed in the shared software area.
- Batch jobs should not write to your home directory. Several hundred batch jobs writing to home directories, which are all on a single RAID array, will overload the system and cause poor interactive response for all users logged on. See below for writing temporary files on the batch worker.

## Large data files – the Lustre file system

- Data files are stored on the Lustre file system. This is a high performance cluster file system capable of providing files from a single filename space to many work nodes at once. It is optimized for large files. Since the data volume is very large, currently about 1000 TB, there is no backup. Therefore, **do not keep source files or other files that are difficult to replace on Lustre**. More than once, files on Lustre have been lost.
- All computers in the cluster see the Lustre file system, /storage/…
- Each group has a Lustre subdirectory which has subdirectories for individual users and sometimes for shared data.
- Quotas (per user and/or per group) – both on the number of files and their total size – are enforced for the Lustre file system. Use "lquota" command to see the status of yours.
- The command, lfs ls -l /storage/… is faster than the Linux ls command.
    - If "ls" on Lustre is slow, remove any alias of ls with options like –color=tty etc, or just use "\ls".

## Temporary files for the duration of a job

The Torque batch system creates a per-job temporary directory on the batch worker's local hard disk. The env variable TMPDIR is set to the directory name. After the job completes and exits, Torque will recursively delete that temporary directory.

# *Batch system*

Torque (based on PBS, Portable Batch System) provides a management system for job processing. Jobs are submitted to job queues, and are then scheduled to run immediately or at some later time, depending on how busy the overall system is and on scheduling policies. Users submit the jobs, but Torque decides when a job starts, chooses which worker node the job runs on, makes sure the job doesn't overstay its requested time, and manages the return of output files to the job submitter.

You, the user, provide a script, consisting of commands to be processed. This might contain just a single command, which might be the name of another script, perhaps with some options or parameters, or the name of a pre-compiled binary file to be run. Or the script might contain a mixture of control statements and commands. The script is submitted using the qsub command.

- Many introductions to using PBS/Torque can be found via Google: "pbs torque introduction". Probably you should start with a job script from a colleague. Details specific to the Weizmann ATLAS cluster are elaborated below.

### *Job queues*

There are three queues for normal users:

- **N:** queue for normal jobs
- **S:** queue for short jobs
- **I:** queue for disk IO intensive jobs, i.e. jobs that read many files with little computation

Type "qstat -q" to see the maximum time and memory limits for each queue.

### *Local requirements*

- Please add the following to your batch script (turns off e-mail notifications)
  ```
  #PBS -m n
  ```
- Use the -o and -e options to redirect the stdout and stderr to a directory other than your home directory:
  ```
  qsub -o path_stdout  -e path_stderr ....
  ```

### *Specifying job memory requirements*

To request 1.5 GB, use: `qsub -l mem=1500mb`. You will also need to specify vmem accordingly, see here. Do not request excessive memory. If you do, you may block your own and others' jobs from running. If your request is modest, your job may run when others cannot.

### *Multi-core jobs*

If you have set up your job to run on, for example, four cores, add "`-l nodes=1:ppn=4`" to the `qsub` command, where the value of processors per node, "`ppn`" is the number of cores your job will use.

See more recommendations here and here, and here. By the way, feel free to browse the entire list archives.

For multi-node jobs, please contact lcg-managers@weizmann.ac.il.

### *Fair share*

The priority for starting a job in a queue depends on the "fair share" of a user and his/her group. Each group, including grid groups, is allocated a percentage of the CPU available in the cluster. Users in a group are allocated equal shares in the group's percentage. If a group has used more than its fair share, jobs of other groups will be given priority. If a user has used more than their fair share in his group, other users in that group will be given priority in the group. The usage history of the past 12 days is used by the scheduler to calculate priorities. More recent use is weighted higher than past use.

### *Preventing jobs from running on slower computers*

Each work node has one or more of the following attributes: `speed4`, `speed5`, `speed6`. They can be used to exclude slow nodes from running a job. For example, adding "`-l nodes=1:speed5`" to your `qsub` will exclude running on machines without the `speed5` attribute. The job will run on machines with `speed5` and `speed6`, but not on those with the `speed4` attribute. Note though that currently, the cluster is rather homogeneous (only ~30 nodes that max at speed4 and very few that reach speed6), so usually you do not need to use such a fine tuning.

## *Interactive processing*

- Interactive jobs are processes run from the command line, as opposed to batch.
- If your interactive process shows 90% or more CPU use with "`top`" for more than a few minutes, you must run it with `nice`. Add it to your script. You can change a running process with the `renice` command.
- Do not run heavy CPU jobs interactively for more than an hour or two; it slows down other users.

- It is forbidden to run interactively on work nodes, i.e. wipp-wn*.

### Analysis nodes

Two nodes, wipp-an1 and wipp-an2 (48 GB memory each), are designated for interactive analysis.

Note that Condensed Matter department members have priority on wipp-an3 and wipp-an4, which have exceptionally large memory for large matrix calculations.

### Matlab

All interactive nodes have Matlab licenses. Be sure to run Matlab with `nice`.

Matlab may also be run in batch. If you have set up your Matlab job to run on, for example, two cores, add "`-l nodes=1:ppn=2`" to the `qsub` command, where the value of "`ppn`" is the number of cores you have told Matlab to use. Matlab is configured to run in the single-core mode by default.

## File transfer

### Cataloged ATLAS datasets and files

Please use the RUCIO protocol to transfer data.

### Small files (< 500 MB) to/from external sites

If the remote site supports a gridftp server, run gridftp, which now supports ssh authentication as well as certificates, on wipp-ui1.
Avoid using `scp` or `sftp` to transfer large data sets to/from sites outside of Israel. The latency is too long for these tools to work well. Also, the file encryption is a heavy CPU burden.

### To transfer files to and from Weizmann

Use `scp`. `scp` must be run from a computer inside the institute (i.e. inside the SecureID firewall zone) or have the WIS VPN enabled (see above) with wipp-ui1 as the src/dst in the cluster.

## Tips

- You will probably find these utilities useful: `nohup`, `screen`.
- X-sessions: `ssh -X`
- Accessing remote files transparently (and more) from Linux, Windows, MacOS.
- There are many more options, including Putty, SSH Secure Shell Client, etc; a good X-window package for Windows is Xming.