Quarterly maintenance

Maintenance schedule 2020

In order to provide a stable service, regular maintenance periods are required to allow intrusive work on the clusters to be carried out.

There are 4 one day (exceptionally two days) downtimes per year for minor interventions and updates.

When	Notes
Q1 - March 30/31	2 days due to recabling
Q2 - June 29	
Q3 - Oct 26	1 day because of Electrical recabling
Q4 - TBC	

In addition there is an annual downtime week in January to allow for major work and software upgrades.

The next planned maintenance week is in January 2021.

The following sections give an overview of the changes carried out and details of any user visible effects that users of the DCSR clusters should be aware of.

March 2020

Jobs submitted before the maintenance period

Please note that the scheduler will not start jobs that are expected to finish after 7am on the 30th of March. This means that any long jobs submitted in the run up will be held in the queue even if there are free nodes so please take care to specify the shortest wall time possible.

For example, on Tuesday the 24th of March, if there are free nodes, a 5 day job will run but a 7 day job will remain waiting in the queue (state PD).

Whilst we will make every effort to maintain the state of the queue we cannot guarantee that your pending jobs will still be present after the maintenance.

Please be aware that after the maintenance not all nodes will be immediately available. The remainder will be brought online in the days following the downtime.

User Visible Changes

New Partition Structure

In order to simplify the management of the clusters the partition structure will be changed. The new partitions are:

- debug - 4 nodes in Wally to allow for quick tests with one job per user at any time
- wally - all nodes in the Wally sub-cluster
- axiom - all nodes in Axiom sub-cluster

This means that there are no longer partitions by wall time and all limits are imposed automatically by a job submit plugin and appropriate Quality of Service (QoS) policies.

The maximum run time remains 10 days. In order to request an allocation on Axiom that lasts for one week the required directives are:

```
#SBATCH --time 7-0
#SBATCH --partition axiom
```

HyperThreading turned off

HyperThreading is a CPU feature that allows two threads to share one execution core and can improve throughput in a number of typical enterprise computing scenarios. For HPC codes it generally degrades performance and makes it difficult to correctly and safely share nodes as well as to run multi-node MPI tasks. For this reason it will be disabled on all Axiom nodes and is already turned off for Wally.

The core count on Axiom will be reduced by 50% after this change so nodes that previously reported 64 cores will now report 32 and so on. Job scripts may need to be updated to reflect this change.

Default wall time of 15 minutes

The default run time for all jobs will be set to 15 minutes - this means that if you have not requested longer via an SBATCH directive then your job will be terminated after 15 minutes.

Topology aware scheduling

On Wally it will be possible to specify that you want all allocated nodes to be on the same Infiniband leaf switch. Whilst this may improve performance for communication intensive tasks, it can also lead to a significant increase in queuing time. It is also possible to specify how long you are prepared to wait for.

For example: requesting 8 nodes on the same switch and being prepared to wait 12 hours for this.

```
#SBATCH --nodes 8
#SBATCH --switches 1@12: 00: 00
```

Other Changes

SLURM 19.05.x and configuration changes

Update to SLURM 19.05.x and diverse changes to the SLURM configuration to improve performance and usability.

Infiniband network rebalancing

In order to increase robustness an extra IB switch will be added to both fabrics (Wally and Axiom) and certain nodes moved to the new leaf switch.

OS update to RedHat 7.7

General security and functionality updates. No user visible changes are expected.

Storage updates

Updates and maintenance on the BeeGFS /scratch file systems including new version and rebalancing.

June 2020

Jobs submitted before the maintenance period

Please note that the scheduler will not start jobs that are expected to finish after 8am on the 29th of June. This means that any long jobs submitted in the run up will be held in the queue even if there are free nodes so please take care to specify the shortest wall time possible.

For example, on Tuesday the 23rd of June, if there are free nodes, a 5 day job will run but a 7 day job will remain waiting in the queue (state PD).

Whilst we will make every effort to maintain the state of the queue we cannot guarantee that your pending jobs will still be present after the maintenance.

Please be aware that after the maintenance not all nodes will be immediately available. The remainder will be brought online in the days following the downtime.

User Visible Changes

R with multithreaded BLAS

R will be able to take advantage of multiple CPU cores by using multi-threaded linear algebra libraries (OpenBLAS or MKL).

In order to set the level of parallelism you can use the OMP_NUM_THREADS environment variable

Updated software stack

The new software environment will receive a minor update - the same applications will be available but sometimes with minor version changes. For example the version of Python is now 3.7.7 and R has moved to 3.6.3

In the case of problem please let us know! The previous stack is still available via the following command:

```
source /dcsrsoft/spack/bin/setup_old_dcsrsoft
```

Other Changes

Infiniband recabling

The recabling of the Infiniband networks will be completed

Ethernet recabling and reconfiguration

Improvements to the 10 Gb/s network

BeeGFS updates

Update to 7.1.5 and hopefully fewer bugs!

Oct 2020

Electrical recabling

The recabling of the Axiom and Wally racks has to be performed because of electrical security regulations

Q4 2020

Revision #27
Created 25 February 2020 07:34:31 by Ewan Roche
Updated 4 September 2020 08:20:19 by Roberto Fabbretti