Table of Contents

Stalled jobs

Reasons that a job can move into the Stalled state

Jobs can move into the stalled state if the watchdog process that is running alongside them loses contact with DIRAC central services. If contact is re-established then the job can continue executing. However, occasionally jobs can remain stalled for many days which indicates a more serious problem which the Grid shifters and experts have to deal with.

Problem with the job

- The job may be using up a lot of CPU and as such has become CPU-bound, meaning that the watchdog process cannot run as frequently to report on the job status.
- The job has been killed by the local batch system for using up too much resource (CPUTime, wallclock time, virtual memory) meaning that DIRAC no longer receives signals from the watchdog process.

Problem with DIRAC

- A central DIRAC service has gone down or has a problem meaning that the job cannot report its status. This will likely affect many jobs.

Problem with the site

- There has been some failure (of network or other infrastructure) which takes the site off of the Grid.

Debugging stalled jobs

Stalled jobs should be investigated by the shifters to determine the root cause of the problem.

If the job has just moved into the stalled state (check the last update time)

- The job may just be suffering from one of the "Problem with the job" reasons above and time should be given to see if the job comes back to life.
- In the meantime, the shifter should check if there are any other problems with the site or central DIRAC services, especially if many jobs are in this state.

If the job has moved into the stalled state many days ago

- In this case, there is no way that the job can recover (sites typically don't allow jobs to run for more than a few days).
- The shifter should check if there were problems with the site and/or DIRAC central services around the time of the last update of the job.
- Have a look at the job JDL:

```
$ dirac-wms-job-get-jdl 964529
```

- ◦ Is the CPUTime that the job requested is longer than that provided by the site? (NOTE: I'm not sure how these wallclock limits map onto the job requested CPUTime which is in LHCb-units)

```
$ dirac-admin-site-info LCG.RAL.uk
{'CE': 'lcgce05.gridpp.rl.ac.uk, lcgce04.gridpp.rl.ac.uk',
 'Coordinates': '-1.32:51.57',
 'Mail': 'lcg-support@gridpp.rl.ac.uk',
 'Name': 'RAL-LCG2',

$ dirac-admin-bdii-ce-state lcgce04.gridpp.rl.ac.uk | grep MaxWallClockTime
GlueCEPolicyMaxWallClockTime: 120
GlueCEPolicyMaxWallClockTime: 4320
GlueCEPolicyMaxWallClockTime: 4320
GlueCEPolicyMaxWallClockTime: 4320
GlueCEPolicyMaxWallClockTime: 4320
```

- ♦ If it is a user job then the user should be informed to change their requested CPUTime to something longer if the job hit the cpu limits of the site.
  ♦ If the CPUTime is OK, then the job may be using up too much memory (perhaps due to an application memory leak) and is being killed off by the batch system. The user should be contacted about this.

- ♦ Was the job a production job and did it upload a file before it moved into the stalled state? Is this file in the BK, LFC and on storage? Use the job JDL to see if there were any LFNs specified as outputdata then use these commands to check the file catalog and storage.

```
$ dirac-dms-lfn-metadata LFN
$ dirac-dms-lfn-replicas LFN
$ dirac-dms-pfn-accessURL PFN SE
```

- In the future there will be a client script to check if the LFN has been registered in the bookkeeping.

-- GreigCowan - 09 Dec 2008

This topic: LHCb > ProductionProceduresStalledJobs
Topic revision: r1 - 2008-12-09 - GreigCowan