Production shifter manual

S. Poss

v1.0

1 Introduction

This document decribes the way to monitor the status of the production system, with a list of links to check regularly.

The first part concerns the DIRAC part, the second the storage monitoring.

2 DIRAC's Production system

The monitoring occurs there:
https://volcd01.cern.ch/DIRAC/ILC-Production/ilc_prod/jobs/ProductionMonitor/display.

One needs to be registered in the *ilc_prod* group. Those currently allowed (at CERN) are Christian Grefe, André Sailer, Jan Strube, S. Poss. They are the people that can actually start, stop, clean productions. Viewing the page is not limited to those people. Granting more access is possible by diracAdmins and they are C. B. Lam, E. Hidle, S. Poss and A. Sailer.

The shifter should have multiple tabs open: one per production types. Set to auto refresh (see bottom of page) at once per hour at least. A close look should be put on the production of type *MCReconstruction_Overlay* as they are the most problematic ones. A special attention should be put on the evolution of the fail rate.

The shifter should have a look at the following twiki page that gives links to useful monitoring pages: https://twiki.cern.ch/twiki/bin/view/CLIC/DiracForAdmins

That page also gives useful links to plots produced by the accounting system of DIRAC, in particular the evolution of the final status rate (instantaneous and cumulative).

2.1 What to do in case of sudden increase of failure rate

1. **Don't panic!**[1]
2. Look at the production jobs in question from the Production Monitoring page by clicking on the faulty production → show jobs
3. Usually, there is a clear message in the Application Status column (in the Job Monitoring page, not the Prod Monitoring). See table 1 for quick reference.
4. Get one of the failed jobs' output by clicking on the job → Sandbox → Get output file(s).
5. Look into the std.out and the application.log file. In case the overlay does not work, there should be also a std.err file. Most errors are due to storage disks dying (massive failure) under the number of queries. In principle, this should not happen too often now as the Overlay files have been merged by bunches of 20, so each file is enough to cover 10 event (in the 3TeV case). No merging has been done for the 500GeV case, so some failure can be expected. This is not too likely as the files have been replicated at many places (CERN, IN2P3, RAL, Imperial College).
6. Some failures are harmless (then they should not be considered as failures, I know):
   - The overlay input automatically fails if it waits more than 5 hours before being able to find a empty slot to get the files. This limitation is not applied to jobs running at CERN, IN2P3-CC and UKI-LT2-IC-HEP.uk as there the files are obtained directly from the storage at those locations. This error implies a recreation of the job as the file is marked as unprocessed.
   - Sometimes the File Catalog times out because of too many connections. The error is then *Can't connect to dips://volcd01.cern.ch:9197/DataManagement/FileCatalog*. This does not matter: it usually occurs while registering the metadata at the end of a job, but it means the file is already uploaded. And if one job of the entire production finishes, then the metadata is properly set. This implies that the file is not reported as failed or processed. The DataRecovery agent is used for that (see later).

[1] Make sure you know where your towel is...

7. If no good reason for failure is found, stop the failing production (stop button on the web monitoring page) and call me!

Table 1: Most common errors that the job monitoring can show

Message	Significance
Failed to install software	Software tar ball is not reachable from the GRID. Normally all software tar balls should be replicated to at least 3 sites. Check replicas and replicate more if needed.
Failed InputSandbox Download	Most likely one of the steering files is only available from one storage element that is not available. Normally all files are replicated to at least 3 places. Same as for software: replicate as needed.
Failed InputData Resolution	Most likely the storage element at CERN is dead. Nothing to do but to stop all productions until situation is resolved: usually a few hours.
Failed to populate Log Dir	Means the logs could not be uploaded to LogSE (in our case volcd03).That's usually not so good because it means one of the log files was not produced. Can imply that one application failed to run. Getting the output file should show the problem. Mostly harmless.

Table 1: (continued)

Message	Significance
Output Data not found	As name suggest: one of the output file is not found locally: means one application did not produce it. Normally should not happen in production context as Output Files are named by convention automatically and the applications know what name the files should have. Can still happen if one application fails but returns a status 0 (like WHIZARD).
Overlay proc failed to get files locally	Looking at the logging info of the job gives more details. Jobs are recreated automatically. std.err file of the output files should contain the reason of the failure: usually happens when more than 20 overlay file could not be obtained.
No space left on device	Arrgh, that's not good: the machine hosting the services has no more disk space, rendering all interactions with the databases impossible. Solution: see section 2.3.
Any sort of exception in a Module	See C. B. Lam: he has the keys to the shop and knows how to handle the code. Call me.

2.2 What if no jobs seem to run at all for some time?

This can be because of multiple reasons:

1. The share at all sites has been used: very unlikely, but can happen in dedicated sites like CERN

2. Pilots fail always. This can be monitored from the Pilot monitor page. To understand, getting the pilot logging info of one of them can be done by clicking on it. Usually means putting a ticket on ggus: send me a mail. Need to ban the faulty site (see diracAdmin person).

3. Can be due to old pilot version having to be purged: when releasing a new ILCDIRAC version (bug fixes essentially), the existing pilots have to be finished before the new version is picked up. Unfortunately, there is no way to kill all pilots in one go (no pilot cleaning agent). In principle, the lifetime of a job that has nothing to do is a few hours max.

4. Matcher can be down for some reason: need to look at log files, see 2.3.

2.3 What if DIRAC services do not work or respond?

Example: the web portal is not responsive, or any dirac command returns an error.

This happens if the machine hosting the services is "down". That can be due to the fact that there is no space left on the device. Nothing can be done with prod manager rights or user rights. One needs direct acces to vobox. The people allowed to do that are S Poss, Bon Lam and Erik Hidle. They know what needs to be done (usually restarting the services is enough to free some space). If that does not work, contact me, and I'll act with a tough way.

It's also possible that a service is down for some reason. This has to be debugged by hand looking at the services log files, accessible from the VOBOX: direct access needed, see above people.

3 Storage monitoring

Essentially only one site to monitor: https://sls.cern.ch/sls/service.php?id=CASTORPUBLIC_ILCDATA.

Percentage free space should not go below 10% as then the garbage collector kicks in and the data is removed from the stager (not from Tape). That's a problem for the DST that have to remain on disk always, and the files currently being used have to be staged also. There is a stager service in DIRAC, so the staging requests are done by DIRAC when the files are needed, but that only works for prod jobs. So when the space becomes too low, one needs to remove the "old" files and the REC files. Removing REC files is more efficient in terms of space gain, but the list takes time to obtain.

Also issuing a stager_rm query is not straightforward: cannot pass too many files at once, so needs to be called "by hand".