Proposal for the Accounting of the DC'04 Production

Ricardo Graciani Díaz, Rubén Vizcaya Carrillo
Universidad de Barcelona
Juan Saborido Silva, Manuel Sánchez García
CERN

26 January 2004

Abstract

A proposal for the Accounting of the DC'04 production is presented that will allow to make reports of the produced data and the resources used presenting the result in different ways.

1. Introduction

The DC'04 production of LHCb will consist of few hundred thousand jobs running over about hundred days on few thousand WN. These Jobs will produced few hundred million events, occupying about 50 TB of storage, in the best case in one hundred thousand files.

One of the important aspects of the DC consist on the exact determination of the computing resources used, as well as the efficiency achieved on its used. For this purposed, a precise accounting system must be put in place that would later allow to produced accurate and detailed reports.

The aim of this proposal is to discuss the content of these accounting reports in order to guide the implementation of the accounting tool, making sure at the same time that the information necessary to produce the reports is made available and kept in a persistent way for its later analysis with the proposed Production System.

Along this proposal a **Job** will be defined as the complete process that runs on a WN. This job might consist of one or more **Step**s, build of modules around one of the different LHCb Gaudi Applications (Gauss, Boole, Brunel, DaVinci,…or a combination of them). These steps may have some input data, generated by a previous step in the job or by a different job, and some output data. A collection of jobs producing the same type data, what sometimes is called a dataset, form a **Production**. For accounting purposes, a Production is the smallest unit that can be queried.

There are two different aspects to be accounted, as mentioned in the abstract, first is the amount of data being produced (call it number of events, or files, …) and the second is the amount of resources used (actual CPU consumed) and how this resources are distributed.

In the present scenario, the necessary info is collected on a per job base by the DIRAC system, kept in the Monitoring DB, MDB, and at some later time, after the job execution has finished (successfully or not), it is copied to an Accounting DB, ADB, and removed from the MDB. What exactly triggers this transfer is still to be defined, but is not relevant for the purpose of this document. It is assumed that all information needed for the accounting is to be found at the ADB, and that it has been collected on the MDB by the Job Monitoring System.

The proposal is organised as follows, Section 2 describes the accounting of the produced data, Section 3 that of the resources used and Section 4 includes some miscellaneous remarks.

2. Accounting of Produced Data

This corresponds to reporting the amount of data produced in a given time interval by one or more Productions. The user must be able to select the start and end dates of the accounted period as well as a list of Productions to be included in the query. Queries with more than one production will produce a report for each of the Productions plus a report for the sum of all Productions. An option is to be included with a predefined list of official DC'04 Productions.

The following reports will be generated:

1. Plot of the incremental number of events produced as a function of the production day.

2. Plot of the incremental storage space used as a function of the production day.

3. Plot of the daily rate of produced events and used storage for the given date interval.

4. Summary Table of the number of submitted, successful and failed jobs, number of events produced, storage and CPU time used for the selected productions.

In order to efficiently produce these reports the following Job Parameters should be included defined for each job:

1. **Production ID**: ID of the production to which the job belongs.

2. **Submission Date**: date (and time) when the job is submitted to DIRAC system.

3. **Output Ready Date**: date (and time) when the output data files are make available to the system by update of the BKDB.

4. **Events**: actual number of successfully processed events. In the case of multi-step jobs the Production Manager should decide which of the steps (or sum of steps) is the relevant one for the accounting.

5. **Output Size**: actual storage space, in MB, of the produced files that are to be kept in the system. Same comment as above.

6. **Execution Time**: Total wall time needed to execute the job.

7. **CPU Time**: Actual execution time consumed by the job.

8. **Failed**: if jobs are rescheduled (or resubmitted) the total number of failures.

Failed jobs are supposed to report, if possible, the CPU and Execution Time, and no Events produced.

3. Accounting of Used Resources

This corresponds to reporting the resources used within a given time interval by one or more Productions. The user must be able to select the start and end dates of the accounted period. The query can be on a Production (or list of Productions) or on a Site (or list of sites).

3.1 By Production

The resources used by a Production or list of Productions are summarised. The accounted quantities are presented as total and mean values for each of the Job steps and for the complete Job. The following quantities are to be accounted:

1. CPU Time (in s) and ratio of CPU Time over Execution Time[1].

2. Input Data processed and Output Data produced (in MB).

3. Data Consumption Rate and Data Production Rate (in MB/s).

Some summary data for each step (and for the complete job) including number of submitted, successful and failed[2] steps (or jobs), total number of events produced, storage and CPU time used.

If a list of Productions is given, the results are presented separately for each one. Plots for 1, 2 and 3 are presented. All results are also presented as a table.

[1] These CPU time include no normalisation factor by the processing power of the WN.
[2] A Job or Step may be resubmitted after failure, therefore completed + failed may add more that submitted.

3.2 By Site

The resources used in a Site or list of Sites are summarised, including all the Productions predefined as official DC'04 (see above). The reports presented for each listed Site, plus an extra one for the sum of all the listed Sites. The accounted resources are:

1. Number of Jobs submitted, completed and failed[2].
2. Execution Time (in s) and ratio of CPU Time over Execution Time[1].
3. Normalised CPU Time (in s), taken as normalisation factor the test-benched time determined before the Production is launched.
4. Number of events produced (from the Events job parameter defined in Section 2).
5. Input Data processed and Output Data produced (in MB).

Some care is needed when defining production sites. In first approach they should correspond to DIRAC Agents but in the case of LCG production (over 50% is expected in this way) this might not be the case. Although LCG may provide job provenance tools this information must be collected by the corresponding Agent and properly filled in the MDB for each job and later transferred to the ADB.

Accounting information for non-official Productions will be available in the system, but since we are dealing with the accounting of DC'04, it does not seem appropriated to include them here.

3.3 Additional Job Parameters

For these accounting reports to be produced the following information has to be filled for each DC'04 Job:

1. **Steps**: Number of Steps.
2. **Step N Type**: High level description of the Step.
3. **Step N Events**: Actual number of processed events in Step N.
4. **Step N CPU Time**: CPU time for the execution of Step N.
5. **Step N Execution Time**: Wall time for the execution of Step N.
6. **Step N Input Size**: size (in MB) of the input data files processed in Step N.
7. **Step N Output Size**: size (in MB) of the output data files produced in Step N.
8. **Step N Failed**: if jobs are rescheduled, total number of times Step N has failed.
9. **Site**: Name of LHCb production Site where the job was run.
10. **CPU per Event**: Test-benched CPU Time per event for the Production.
11. **Error Message**: Description of the Error in case of Failure (it should include the Step Number).

In the above N runs from 1 to the total number of Steps defined in the Job. For failed jobs, even those rescheduled successfully, some statistics about the known reasons of failure will be needed. A meaningful **Error Message** parameter should be filled for each failing job, initialising the **Failed** parameter to 1. Should further failures occur, new messages are to be concatenated to the existing **Error Message** parameter and the Failed parameter is to be incremented by 1. The **Step N Failed** is also to be updated accordingly.

4. Miscellaneous

As mentioned in Section 1, the information needed for the accounting is collected by DIRAC Job Monitoring System on the MDB and later moved to the ADB. Since all this info is only needed once the job execution has finished it might be kept locally for a running job, retrieved by the Local Agent in the job output sandbox and finally the Agent does the update. This scheme avoids the need of outbound connectivity of the WN for this purpose.

The ADB will have a XML-RPC interface identical , at least in the first implementation, to that of the MDB. Both DBs have an identical structure.

5. Conclusions

This proposal is to be understood as a working document to be updated as new comments, suggestions, etc are presented.

Terminology

ADB	Accounting Data Base
DB	Data Base
DC	Data Challenge
MDB	Monitoring Data Base
WN	Working Node