A Study of ATLAS Grid Performance for Distributed Analysis

Sergey Panitkin\textsuperscript{1}, Valery Fine\textsuperscript{1}, Torre Wenaus\textsuperscript{1}
\textit{on behalf of the ATLAS Collaboration}

\textsuperscript{1}: Brookhaven National Laboratory, Upton, NY, USA

Abstract. In the past two years the ATLAS Collaboration at the LHC has collected a large volume of data and published a number of groundbreaking papers. The Grid-based ATLAS distributed computing infrastructure played a crucial role in enabling timely analysis of the data. We will present a study of the performance and usage of the ATLAS Grid as a platform for physics analysis in 2011. This includes studies of general properties as well as timing properties of user jobs (wait time, run time, etc). These studies are based on mining of data archived by the PanDA workload management system.

1. Introduction

The ATLAS experiment \cite{1} at the Large Hadron Collider (LHC) is designed to explore the fundamental properties of matter for the next decade. Since LHC start-up in 2009, the experiment has collected and distributed hundreds of petabytes of data worldwide to more than 100 computer centers. Thousands of physicists analyze tens of millions of events daily. The ATLAS Computing model \cite{2} is based on a Grid paradigm \cite{3}, with multilevel, hierarchically distributed computing and storage resources. ATLAS utilizes the PanDA workload management system \cite{4} (WMS) for distributed data processing and analysis.

2. Analysis Details

The PanDA server collects a variety of information about jobs running on the ATLAS Grid. About two years ago a summary statistics table was introduced in the PanDA database schema. Server side tools that allow to query this table were developed by the PanDA team. In this paper we present data about user analysis activities at all ATLAS analysis queues, that were collected from January 1, 2011 to December 31, 2011. The goal of the current project was to look at general trends and features in ATLAS user analyses activity on the Grid.

3. Analysis Results

3.1. General Statistics

Figure 1 shows the number of finished and failed analysis jobs per week, as a function of time. The total number of successfully finished user analysis jobs in 2011 was about 79 million. About 13 million user analysis jobs failed for various reasons during the same time period. The number of jobs per week was steadily growing over time and reached about 2.6 million jobs per week, for some weeks at the end of the year. The number of failures was approximately constant. The simple metrics shown in Figure 1 reveal an interesting sociological observation. The
distribution of finished jobs, reflecting the number of submitted Grid jobs, demonstrates several
spikes that seem to be correlated with preparations to major physics conferences (Moriond, EPS,
etc). One can also clearly see that during the Christmas vacation period the number of jobs
submitted to the grid dropped significantly. Perhaps this information may be useful in planning
allocation of Grid resources between analysis and production activities.

Figure 2 shows the error rate as a function of time. The error rate is defined as a ratio of
number failed jobs and total number of finished and failed jobs. It can be seen from Figure 2
that the error rate dropped significantly after April of 2011. Average error rate was 0.18 for
the whole year and 0.15 after April. The exact reason for this behavior is not clear and may be
connected several factors like improvement in ATLAS software and increase in reliability of grid
sites and central services.

3.2. Failure Types
Several hundred exit codes can be generated on error by various sources, from user code and
operating system to Panda pilot [5]. The PanDA WMS automatically tracks user job failure
types, they are classified and stored in an Oracle database. In our analysis these codes are
grouped into several broad categories that reflect major qualitative sources of user analyses
failures on the Grid. Figure 3 shows the distribution of these failure types as a function of time
in 2011. Distributed data management related failures are denoted as "nfaildat" in the Figure.
These were the dominant source of job failure at ≈4.2 millions failed jobs. The second ranked
source of job failures was due to so called "application failures" or failures related to ATLAS
software framework Athena and/or errors related to user code. They are labelled in Figure 3 as
"nfailapp" and led to a failure of ≈1.1 millions jobs. Time-out related failures, associated with
local data staging before and after the user job is running on a Grid worker node, are labeled
there as ntimeout. PanDA system related failures are labelled "nfailsys" in the Figure 3. These
two types of failures were approximately equal in damage, with ≈203.4 thousand jobs failed due
to "system" failures and ≈203.3 thousand failed due to time-out at stage-in or stage-out phase.
3.3. ATLAS Grid reliability for analysis jobs

It is interesting to separate "grid" and "user" related errors e.g. errors associated with Grid sites and Grid infrastructure (PanDA, data bases, catalogues, etc) and errors associated with "user code". Figure 4 shows such a metric, called here, for the lack of better term, "reliability". "Reliability" is defined here as one minus ratio of all "non-user" failure types and total number of failed and finished jobs. We note again that this metric is constructed in such a way that it
Figure 4. ATLAS Grid reliability for analysis jobs in 2011. See description in the text

takes into account only failures associated with Grid sites and services. In 2011 average Grid reliability for analysis jobs was 0.86. It can be seen in Figure 4 that the reliability slightly improved with time.

3.4. Grid Jobs Timing
From the point of view of scientists doing analysis on the Grid, this distributed computing infrastructure is just a tool that helps to achieve desired results. This tool should be not only robust but also should be able to help achieve results in a reasonable amount of time. We discussed some issues involved in robustness of the Grid in previous sections, in this section we will outline timing properties of the user jobs on the Grid. The most basic characteristics of the timing properties of the jobs submitted to the Grid are the wait and run times. We define the job wait time as a time interval between the moment a user submits a job to PanDA and the moment the pilot is assigned to get the jobs payload. Run time is defined as user code running time. Figure 5 shows wait time and run time distributions for user analysis jobs on the ATLAS Grid in 2011. It is clear from the Figure that the wait time is significantly larger than run time. Average job run time was about 100 Minutes. Average wait time was \( \approx 200 \) minutes, and was approximately constant, but with occasional peaks above 250 minutes. This almost constant behaviour is non-trivial, since number of jobs was growing dramatically during 2011 as it was shown in Figure 1. That is because more resources were allocated by the ATLAS Collaboration for analysis, in response to increased workload during 2011. The wait time dominates total average lifetime of the jobs on the Grid and determines analysis turn-around time. Ideally we want it to be as small as possible. Right now typical turn around time for the analysis jobs on the Grid is about 5 hours, which allows for more than one analysis iteration per day.

There are additional timing characteristics involved in the running of pilots and jobs on the ATLAS Grid. Before running actual user code the PanDA pilot stages input data to a worker node where the job will be running. This time is denoted as "tstagein" in Figure 6. Average stage-in time for analysis jobs was about 390 seconds. After user code is finished then the pilot stages out output data from the worker node to a persistent storage element. The corresponding time is denoted as "tstageout" in Figure 6. Average stage-out time was about 47 seconds. Average stage-out time is much shorter than the average stage-in time which is understandable since for analysis jobs typically the size of the input files is much larger than the size of the
output. When the pilot starts to interact with PanDA server one of the first operations that it performs is getting user job payload from the server. The time interval corresponding to this operation is denoted as "tgetjob" in Figure 6. On average this operation is short, compared to other pilot transactions, at about 10 seconds. One should also note that several ATLAS sites do not use data staging and provide direct read access from the storage element. For these sites stage-in and stage-out times will be close to zero. In our current analysis we did not attempt to separate sites with direct data access from data staging sites.

4. Conclusion
We presented a high-level overview of user analysis activities on the ATLAS Grid in 2011. 2011 was a very successful year of operation both for the LHC and for ATLAS distributed computing. There was a significant increase in number of user submitted jobs during that year. The number of failed user jobs was about constant over that period of time. The user jobs error rate decreased during April-May period and stayed at 15% thereafter. We also observed a significant decrease in application related failures during that time. The number of PanDA related failures stayed relatively small over the year. The largest source of error seems to be data management related. These errors affects all sites. The ATLAS collaboration is developing a new distributed data management system in order to address this problem. Grid reliability, in the analysis domain, improved with time from 80% to almost 90%. Average wait time for user jobs in 2011 was about 3 hours. Wait time can be as long as 4 hours at some sites. Average analysis job run time was 100 minutes. Data stage-in, stage-out can be several minutes on average and can give a non-negligible contribution to the job run time.

References
[1] ATLAS Collaboration, J. Inst. 3 (2008) S08003.
[2] Adams D, Barberis D, Bee C, Hawkings R, Jarp S, Jones1 R, Malon D, Poggioli L, Poulard, G, Quarrie D and Wenaus T 2005 The ATLAS computing model Tech. rep. CERN cERN- LHCC-2004-037/G-085
Figure 6. Times to stage data in and out and time to get job’s code. See description in the text.

[3] Foster I and Kesselman C (eds) 1998 The Grid: Blueprint for a New Computing Infrastructure (Morgan-Kaufmann)

[4] Maeno T., Overview of ATLAS PanDA Workload Management, Proc. of the 18th Int. Conf. on Computing in High Energy and Nuclear Physics (CHEP2010)

[5] Nilsson P., The ATLAS PanDA Pilot in Operation, in Proc. of the 18th Int. Conf. on Computing in High Energy and Nuclear Physics (CHEP2010)