Reliability Engineering Analysis of ATLAS Data Reprocessing Campaigns

A Vaniachine\textsuperscript{1}, D Golubkov\textsuperscript{2} and D Karpenko\textsuperscript{3}
on behalf of the ATLAS Collaboration

\textsuperscript{1} High Energy Physics Division, Argonne National Laboratory, 9700 South Cass Avenue, Argonne, IL 60439, United States of America

\textsuperscript{2} Experimental Physics Department, Institute for High Energy Physics, Protvino, 142281, Russia

\textsuperscript{3} Department of Physics, University of Oslo, P.b. 1048 Blindern, 0316 Oslo, Norway

E-mail: vanachine@anl.gov

Abstract. During three years of LHC data taking, the ATLAS collaboration completed three petascale data reprocessing campaigns on the Grid, with up to 2 PB of data being reprocessed every year. In reprocessing on the Grid, failures can occur for a variety of reasons, while Grid heterogeneity makes failures hard to diagnose and repair quickly. As a result, Big Data processing on the Grid must tolerate a continuous stream of failures, errors and faults. While ATLAS fault-tolerance mechanisms improve the reliability of Big Data processing in the Grid, their benefits come at costs and result in delays making the performance prediction difficult. Reliability Engineering provides a framework for fundamental understanding of the Big Data processing on the Grid, which is not a desirable enhancement but a necessary requirement. In ATLAS, cost monitoring and performance prediction became critical for the success of the reprocessing campaigns conducted in preparation for the major physics conferences. In addition, our Reliability Engineering approach supported continuous improvements in data reprocessing throughput during LHC data taking. The throughput doubled in 2011 vs. 2010 reprocessing, then quadrupled in 2012 vs. 2011 reprocessing. We present the Reliability Engineering analysis of ATLAS data reprocessing campaigns providing the foundation needed to scale up the Big Data processing technologies beyond the petascale.

1. Introduction

During the first run period of the Large Hadron Collider (LHC) [1] at CERN in 2010-2012, the ATLAS experiment [2] recorded 7.4 PB of “raw” $pp$ collisions data. Data reconstruction is a starting point for physics analysis. Following the prompt reconstruction, the ATLAS data are reprocessed, which allows reconstruction of the data with updated software and conditions/calibrations improving the quality of the reconstructed data for physics analysis. The large-scale data reprocessing campaigns are conducted on the Worldwide LHC Computing Grid (WLCG) [3]. For the success of the reprocessing campaigns conducted in preparation for the major physics conferences, a timely completion is critical. The WLCG computing centres provide tens of thousands of CPU cores and petabytes of disk storage for a faster reprocessing throughput.

During Run1, the ATLAS collaboration completed three petascale data reprocessing campaigns on the Grid, with up to 2 PB of “raw” data being reprocessed every year. Depending on conditions, it
takes $3\times10^6$ CPU-hours to reconstruct one petabyte of ATLAS “raw” data. Scheduled LHC upgrades will increase the data taking rates tenfold [4], which increases demands on computing resources. However, a tenfold increase in WLCG resources for LHC upgrade needs is not an option. A comprehensive end-to-end solution for the composition and execution of the reprocessing within given CPU and storage constraints is necessary to accommodate physics needs of the next LHC run. The ATLAS experiment needs to exercise due diligence in evolving its Computing Model to optimally use the required resources. Reliability Engineering provides the foundation needed to scale up ATLAS data reprocessing beyond the petascale.

2. Data Processing

The “raw” data from the ATLAS data acquisition system are processed to produce the reconstructed data for physics analysis. Events acquired within few minutes are collected in one “raw” data file. Files with events that are close in time (from the same “run”) are collected in one dataset. During reconstruction ATLAS applications are processing “raw” events data with sophisticated algorithms to identify and reconstruct physics objects such as charged particle tracks. Figure 1 shows the data processing flow of “raw” data and conditions/calibrations used in reconstruction operations. Since the data are comprised of independent events, massively parallel reconstruction applications process one event at a time. The “first-pass” processing of the “raw” event data at the ATLAS Tier-0 site provides promptly the data for quality assessment and analysis. Later, the quality of the reconstructed data is improved by optimizing further software algorithms and conditions/calibrations. For data processing with improved software and/or conditions and calibrations (reprocessing) we use WLCG computing resources.

![Figure 1](image_url)

Figure 1. Data from the ATLAS data acquisition system (top arrow) and conditions/calibrations are processed through “prompt” reconstruction at the Tier-0 site at CERN (top) and in reprocessing on the Grid (bottom).
2.1. Tasks vs. Jobs
In petascale data processing the task became a main unit of computation. A task is a collection of similar jobs from the same dataset (such as from one data-taking “run”) that could be executed in parallel. The task-based approach delivers “six sigma quality” performance for the petascale data processing campaigns with thousands of tasks [5]. The success of the task-based approach in ATLAS data processing resulted in a double exponential growth in the number of tasks requested for processing [6].

3. Analysis
In reprocessing on the Grid, failures can occur for a variety of reasons, while Grid heterogeneity makes failures hard to diagnose and repair quickly. As a result, ATLAS data reprocessing tolerates a continuous stream of failures, errors and faults. Transient job failures can be recovered by automatic re-tries. While our fault-tolerance mechanisms improve the reliability of data reprocessing, their benefits come at costs and result in delays making the performance prediction difficult. Designing fault tolerance strategies that minimize the duration of petascale data processing on the Grid is an active area of research. Reliability Engineering provides a framework for fundamental understanding of the data reprocessing on the Grid, which is not a desirable enhancement but a necessary requirement.

3.1. Cost of Recovery from Failures
Job re-tries avoids data loss at the expense of CPU-hours used by the failed jobs. Figure 2 shows that the distribution of tasks ranked by CPU-hours used to recover from transient failures is not uniform: most of CPU-hours required for recovery were used in a small fraction of tasks. Table 1 compares cost of failure recovery for three major ATLAS data reprocessing campaigns. In 2010 reprocessing, the CPU-hours used to recover from transient failures were 6% of the total core-hours used for reconstruction. In 2011 reprocessing, the CPU-hours used to recover from transient failures were reduced to 4% of the total CPU-hours used for the reconstruction. Figure 3 shows that most of the improvement came from reduction in failures in file transfers at the end of a job [8].

Figure 2. Tasks ranked by CPU time used to recover from transient failures.
The observed distribution approximately follows the behaviour of the Weibull distribution, which often models the process with a variety of failures having independent failure times (the “weakest link” model). Figure 4 shows similarity of the distributions of tasks vs. recovery costs for all three data reprocessing campaigns.

**Table 1.** Cost of recovery from transient failures for the reconstruction jobs.

| Reprocessing campaign | Input Data Volume (PB) | CPU Time Used for Reconstruction ($10^6$ h) | Fraction of CPU Time Used for Recovery (%) |
|-----------------------|------------------------|---------------------------------------------|------------------------------------------|
| 2010                  | 1                      | 2.6                                         | 6.0                                      |
| 2011                  | 1                      | 3.1                                         | 4.2                                      |
| 2012                  | 2                      | 14.6                                        | 5.6                                      |

**Figure 3.** Distribution of tasks vs. CPU-hours used to recover from transient job failures follows a multi-mode Weibull distribution [7].

**Figure 4.** Distribution of tasks/petabyte vs. CPU-hours used to recover from transient job failures follows log-normal behaviour.
As the latest campaign provided additional data for analysis, Figure 5 shows deviation of the empirical Cumulative Distribution Function (eCDF) from the straight line corresponding to the case of a simple Weibull distribution. Indeed, the symmetry of the Figure 4 distribution indicates a log-normal behaviour rather than a (skewed) Weibull distribution. Table 2 compares results of the maximum likelihood fits to the data using the log-normal (1) and Weibull (2) distributions with two parameters:

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(\frac{-(\ln x - \mu)^2}{2\sigma^2}\right)$$  \tag{1}$$

$$f(x) = \frac{k}{\lambda} \left(\frac{x}{\lambda}\right)^{k-1} \exp\left(-\left(\frac{x}{\lambda}\right)^k\right)$$  \tag{2}$$

The Kolmogorov–Smirnov (K-S) test values show that the log-normal fit describes the data better.

| Reprocessing campaign | $\mu$    | $\sigma$ | K-S test | $k$     | $\lambda$ | K-S test |
|-----------------------|----------|----------|----------|---------|-----------|----------|
| 2010                  | 3.29±0.09| 2.63±0.06| 0.07     | 0.41±0.01| 96.6±8.3  | 0.05     |
| 2011                  | 2.77±0.09| 2.50±0.07| 0.05     | 0.43±0.01| 54.5±4.9  | 0.10     |
| 2012                  | 2.92±0.07| 2.90±0.05| 0.04     | 0.35±0.01| 79.4±6.0  | 0.08     |

The log-normal behaviour of the distribution of tasks vs. CPU-hours used for failure recovery is unexpected. Although log-normal distributions are often observed in Reliability Engineering analysis of failures, such as mean-time-to-failure of the Grid hardware, these are expected when some multiplicative process is present, such as multiplicative degradation [9]. As we were unable to identify the multiplicative process describing our case where the CPU time used for recovery is added upon
each job retry, we compared our data with two necessary conditions under which the additive process may result in the log-normal distribution (when the number of variables is limited) [10]:

\[
\frac{|(\ln z^* - \ln z^*)^3|}{(\ln z^* - (\ln z^*)^2)^{1.5}} < \frac{|(z^* - (z^*)^3|}{(z^* - (z^*)^2)^{1.5}} \quad (3)
\]

\[
\frac{|(\ln z^* - (\ln z^*)^4) - 3(\ln z^* - (\ln z^*)^2)^2|}{(\ln z^* - (\ln z^*)^2)^2} < \frac{|(z^* - (z^*)^4) - 3(z^* - (z^*)^2)^2|}{(z^* - (z^*)^2)^2} \quad (4)
\]

Here \(z^*\) is the random variable (in our case it is the CPU-hours used to recover from transient job failures). Table 3 shows that conditions (1) and (2) are met in our case, since for all reprocessing campaigns the right-side expressions are much greater than the left-side expressions.

**Table 3. Conditions under which the additive process may result in the log-normal distribution.**

| Reprocessing campaign | Left-side of condition (3) | Right-side of condition (3) | Condition (3) | Left-side of condition (4) | Right-side of condition (4) | Condition (4) |
|-----------------------|---------------------------|-----------------------------|---------------|---------------------------|-----------------------------|---------------|
| 2010                  | 0.27                      | 7.90                        | True          | 0.23                      | 79.6                        | True          |
| 2011                  | 0.26                      | 7.17                        | True          | 0.31                      | 74.0                        | True          |
| 2012                  | 0.06                      | 10.1                        | True          | 0.46                      | 153.5                       | True          |

3.2. **Reprocessing Duration**

Since jobs generally complete or fail without producing side effects, the simple recovery mechanism involves automatic retry of failed jobs. Unfortunately, the automatic approach results in an unpredictable delay in task completion time dominated by repeated retries. Transient job failures and re-tries delay the reprocessing duration. Optimization of fault-tolerance techniques to speed up the completion of thousands of interconnected tasks on the Grid is an active area of Reliability Engineering research.

**Figure 6.** Number of jobs concurrently running on the WLCG clouds during ATLAS petascale data processing campaigns in November-December 2010, in August-September 2011 and in November 2012.
3.3. Results
During LHC data taking, Reliability Engineering analysis became critical for the success of the reprocessing campaigns conducted in preparation for the major physics conferences. To deliver results for the 2012 Moriond Conference, the processing of 2 PB of ATLAS data (that required twice more CPU-hours per petabyte) was completed in four weeks, achieving four-fold improvement in data processing power in comparison to the 2011 data processing (Figure 6). Earlier, optimization of reprocessing workflow and other improvements cut the delays and halved the duration of the petascale data processing on the Grid from almost two months in 2010 to less than four weeks in 2011 [5].

4. Conclusions
During three years of LHC data taking, the ATLAS collaboration completed three petascale data reprocessing campaigns on the Grid. Reliability Engineering analysis supported continuous improvements in data reprocessing throughput during LHC data taking. The throughput doubled in 2011 vs. 2010 reprocessing, then quadrupled in 2012 vs. 2011 reprocessing.

Scheduled LHC upgrades will increase the data taking rates tenfold, increasing demands on computing resources. A comprehensive end-to-end solution for the composition and execution of the reprocessing within given CPU and storage constraints is necessary to accommodate physics needs of future LHC runs. Reliability Engineering analysis provides the foundation needed to scale up ATLAS data reprocessing on the Grid beyond the petascale.

Acknowledgments
We wish to thank all our colleagues who contributed to ATLAS data reprocessing activities. This work was funded in part by the U. S. Department of Energy, Office of Science, High Energy Physics under Contract No. DE-AC02-06CH11357.

References
[1] http://lhc.web.cern.ch
[2] The ATLAS Collaboration, Aad G et al. 2008 The ATLAS experiment at the CERN Large Hadron Collider J. Inst. 3 S08003
[3] http://wlcg.web.cern.ch
[4] The ATLAS Collaboration 2012 Letter of Intent for the Phase-II Upgrade of the ATLAS Experiment, Scientific Committee Paper CERN-LHCC-2012-022; LHCC-I-023, http://cds.cern.ch/record/1502664
[5] Vaniachine A V for the ATLAS Collaboration 2011 ATLAS detector data processing on the Grid IEEE Nuclear Science Symposium and Medical Imaging Conference pp 104-107
[6] Golubkov D et al. 2012 ATLAS Grid Data Processing: system evolution and scalability J. Phys.: Conf. Ser. 396 032049
[7] Weibull W 1951 A Statistical Distribution Function of Wide Applicability ASME Journal of Applied Mechanics, pp 293-297
[8] Vaniachine A V on behalf of the ATLAS and CMS Collaborations 2013 Advancements in Big Data Processing in the ATLAS and CMS Experiments Preprint arXiv:1303.1950
[9] Kolmogorov A N 1941 On the log-normal distribution of particles sizes during break-up process Dokl. Akad. Nauk. SSSR 31 99-101
[10] Mouri H 2013 Lognormal distribution from a process that is not multiplicative but is additive Phys. Rev. E 88 042124