Simulation of the job processing performance at an ALICE Tier-2 site with MONARC

Č Zach¹, L Betev² and D Adamová¹ for the ALICE Collaboration

¹Nuclear Physics Institute of the AS CR Rez/Prague
²CERN - European Organization for Nuclear Research
adamova@ujf.cas.cz

Abstract. The MONARC (MOdels of Networked Analysis at Regional Centers) framework has been developed and designed with the aim to provide a tool for realistic simulations of large scale distributed computing systems, with a special focus on the Grid systems of the experiments at the CERN LHC. In this paper, we describe a usage of the MONARC framework and tools for a simulation of the job processing performance at an ALICE Tier-2 site.

1. Introduction
According to the ALICE Computing Model [1], the Tier-2 centers workload consists of simulation processing and end user analysis jobs. These two job categories have different requirements concerning the computing resources: simulation jobs demand long processing time (CPU time), while the analysis jobs usually require heavy Input/Output (I/O) load. This is reflected in the CPU/Wall time ratio, a quantity commonly used by the computing centres to track the job efficiency.

Since at each computing site there always is the demand to reach as high job processing efficiency as possible, a preliminary estimation of the influence of various configurations of the job processing environment is of a very high interest. In our work, we decided to investigate the influence of the following two factors:

1) the ratio of analysis to simulation jobs
2) the location of data processed by the jobs

We have performed a simulation of the effect of different conditions 1) and 2) on the processing performance of jobs at one of the ALICE Tier-2 sites: the computing farm Golias in Prague, Czech republic [2]. For our simulations, we have chosen the MONARC2 simulator [3] for its ability to simulate complex distributed systems, cf. [4] and the references there.

2. System setup for the simulation
Since two of us are working at the Prague site, we have chosen this site and a few close Tier-2s as the system to be simulated. As a warm-up and validation part of the work, we performed at first a simulation to reproduce a month-average performance of the Prague Storage Element, and namely the average traffic and number of opened sockets, which is the number of connected clients (the simulation was monitoring only active opened sockets).
In the second part of our work, we considered a system consisting of 6 ALICE Tier-2 sites from central Europe. The computing and storage resources of each site were parameterized using the data from the ALICE Grid monitoring framework MonALISA [5]. The topology of the model is sketched in Figure 1.

![Map of centers in the simulation. Speeds of links are given in Mbit/s.](image)

**Figure 1. Schema of the simulated system**

3. **MONARC2 simulator**

The presented simulations were performed using the software package developed within the project MONARC2 [3,4]. It is a discrete-event simulator which allows modeling of large distributed systems. The original intention of the project was the modeling and simulation of High Energy Physics (HEP) experiments' computing infrastructures.

MONARC2 is written in Java, with a built-in multi-thread support for concurrent processing. It offers an Interrupt-driven scheme for discrete event simulation, which is well suited to describe the stochastic arrival patterns and concurrent running tasks sharing resources, typical for the HEP computing projects.

MONARC2 structure consists basically of two parts. There is a core of the simulator (the packages: engine and network), which implements basic classes for creating specific objects, designed to be generic for any distributed system. Then there is the package monarc, which uses these tools to implement specific classes (like e.g. CPUUnit, DatabaseServer) and a set of classes for data visualization which allows monitoring of simulation results.
Figure 2. Job submission schema in AliEn

To use MONARC2 simulator, we had first to write configuration files specifying parameters of simulated systems. Then, extension/modification of some of the available classes was done (e.g. JobScheduler) and new ones were created, to implement specific components (e.g. AliEn Central Task Queue (TQ) [5]).

4. The job processing environment in ALICE

The ALICE Grid system includes 77 computing sites, out of which 70 are Tier-2s [6]. The disk storage part of the system is represented by 50 disk storage elements (SE) [7], the majority of them being operated by the xrootd/Scalla manager [8]. The whole system is coordinated by the ALICE Grid middleware AliEn [5]. The job submission procedure is orchestrated centrally from CERN with the help of the Central Task Queue (TQ), using the Job Agents (JA) "pull" functionality.

The schema of the job submission procedure is sketched in Figure 2. The steps of the job processing procedure were mapped into the simulator framework. A job in AliEn can go through multiple status types [5]. Most of them were implemented in the simulation and one was added (a JA-correlated job status "First_matched").

5. Validation of our simulation scheme

As mentioned in the section 2, we performed at first, as a validation study, a simulation to reproduce a month-average performance of the Prague SE, and namely the average outgoing traffic and number of opened active sockets.

The Prague site publishes 2540 CPUs but only part of them is available for ALICE jobs. During the simulated period, the average number of ALICE jobs processed at the Prague site was about 1000. The Prague SE represents a cluster of 3 storage servers with the total capacity of 60 TB. It is located at a department outside the farm Golas and is connected to the farm with a dedicated 1Gb/s link. Also, the cluster is accessible from the other ALICE Grid sites via the 2Gb/s public network, so there are combined I/O requests from Golas and external sites.
The cluster is operated by the pure xrootd storage manager. When a job requests processing of particular events included in different files stored at the SE, the data is streamed directly from the SE and only the requested events are read and processed on the fly. For this purpose, a number of sockets are opened, some of them inactive. In our simulations, we simulated only the active sockets.

The outgoing traffic from a Tier-2 SE is driven by the (user) analysis jobs, for which it is not straightforward to specify the profile, as the load tends to be chaotic. We used a parameterization of the run time and amount of processed data for analysis jobs, based on our examination of the Golias farm batch system logs and of output logs of analysis jobs accessible in AliEn.

For this simulation, we assumed that 25% of the running jobs in Prague were analysis jobs, which corresponds to the real operations during the simulated period. A plot showing the one-month (real) monitoring of the network traffic on the Prague SE is presented on Figure 3. Due to space constraints, we refer to our CHEP2010 presentation and to [9] for other plots.

Figure 3. Network traffic (OUT) on the Prague SE. 1 month, average 19 MB/s

Figure 4. Simulated network traffic on the Prague SE
The simulated network traffic on the Prague SE is shown on Figure 4. Unlike the real situation, there are no running jobs at the simulation take-off, and it takes some time to reach the average of 1000 jobs and the corresponding traffic on the SE. Similarly, there is a descent at the end of the simulation. From comparison of the simulated and actual plots (cf. also our CHEP2010 presentation and [9]) we concluded, that the simulation tool works according to our expectations.

6. The predictive simulation
In the second part of our work, we used the system setup described in the section 2. We performed simulations of the following quantities:
- the time profile of running jobs (in the state "Running-Saving")
- the job processing efficiency (CPU/Wall time)
- the data transfer speed

The simulations were performed for 4 different sets of conditions:
- jobs using only local data, 20% of jobs are analysis
- jobs using data stored anywhere,
- 20% of jobs are analysis
- 50% of jobs are analysis
- 75% of jobs are analysis

The first configuration was used as a reference and we adjusted the parameters of the model to get close to the real conditions at the Prague site, 20% of jobs being analysis is usually the case in Prague.

| AVERAGE                           | Local, 20% | Anywhere, 20% | Anywhere, 50% | Anywhere, 75% |
|-----------------------------------|------------|---------------|---------------|---------------|
| First_Matched $\rightarrow$ Assigned [minutes] | 1.46       | 0             | 0             | 0             |
| Running $\rightarrow$ Saving [hours]  | 1.84       | 2.16          | 3.52          | 5.88          |
| Efficiency                        | 69%        | 58%           | 47%           | 36%           |
| Transfer speed [MB/s]             | 0.40       | 0.42          | 0.99          | 1.49          |

Table 1. Summary of our simulation results

![Simulated job processing efficiency for different conditions](image)

Figure 5. Simulated job processing efficiency for different conditions

A summary of our results is given in Table 1. On Figure 5, we show the simulated job efficiency for the four simulation setups. Due to space constraints, we refer to our CHEP2010 presentation and to [9] for other plots.
7. Summary and Outlook

We have used the MONARC2 simulation tool to study the efficiency of job processing, the storage element performance and the data transfer speed at the ALICE Tier-2 site in Prague. We used the simulation of the Prague storage element performance as a validation study of the chosen simulator and then proceeded to a simulation with predictive results.

In the validation part, we reproduced a month-average performance of the Prague ALICE Storage Element. In the predictive simulation, we calculated the efficiency of job processing, the job processing time and the data transfer speed. Our simulation predicts, that using data stored at distant SEs and adding on the percentage of analysis jobs running at the site,
- the average efficiency of job processing will go down
- the average job processing time will grow several times
- the average transfer speed will grow.

Unlike the first two, the third observation is quite unexpected. The MONARC2 network simulation design does not provide a method to set an external load on a network entity. The network package is a part of the simulator core, for which the source code is not publicly available, so we tried to implement the feature by overriding methods for adding/removing messages to/from a network entity. However, this implementation cannot correctly treat the parts of network, which run messages bottlenecked elsewhere, giving other messages a speed boost.

As a summary, we conclude that the MONARC2 simulator is capable of simulating quite complex distributed systems. In our simulation, we used a rough parameterization of the simulated system and a simplified profile of the analysis jobs, mainly because of the time constraints. It would be interesting to perform the simulation using more complex parameterization of the simulated system and to simulate systems of more Tier-2 centers. This would need optimization of the simulator and high performance hardware, because already with our framework, the simulations are highly time consuming. Another alternative would be using the Grid for such simulations.

Acknowledgements

The work was supported by the MSMT CR contracts No. 1P04LA211 and LC 07048.

References

[1] ALICE Experiment Computing TDR, http://aliceinfo.cern.ch/Collaboration/Documents/TDR/Computing.html.
[2] Regional computing center for particle physics, http://www.particle.cz/farm/index.aspx?lang=en.
[3] Legrand I C, Newman H, Dobre C M, Stratan C 2003 MONARC Simulation Framework, ACAT'04, Tsukuba, Japan, http://monarc.cacr.caltech.edu:8081/www_monarc/monarc.htm.
[4] Dobre C M, Cristea V: A Simulation Model for Large Scale Distributed Systems, http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=4430426.
[5] Saiz P et al. 2003 AliEn - ALICE environment on the GRID Nucl. Instrum. Meth. A502 437 http://alien2.cern.ch/.
[6] ALICE Sites Performance, http://pcalimonitor.cern.ch/reports/index.jsp.
[7] ALICE Disk storage elements, http://pcalimonitor.cern.ch/stats?page=SE/table.
[8] XROOTD - Summary, https://savannah.cern.ch/projects/xrootd/; How to install XROOTD, http://alien2.cern.ch/index.php?option=com_content&view=article&id=33&Itemid=79#Xrootd_and_how_it_works.
[9] Zach Č 2011 Simulation of complex distributed computing systems in High Energy Physic Dipl. Thesis, Prague 2011 (work in progress).