Parallel n-tuples processing with PROOF on the Lecce Public Cluster

M.Bianco[1 2] E.Gorini[1,2], M.Primavera[1], F.Ricciardi[1], A.Ventura[1,2] and the ATLAS Collaboration [1]

[1]Istituto Nazionale di Fisica Nucleare, sezione di Lecce, Italy

[2]Dipartimento di Fisica, Università del Salento, Italy

1. Introduction

The final step of the analysis chain of a High Energy Physics experiment is often performed with a ROOT [2] n-tuple containing only the essential variables extracted from previous analysis stages. This part of a physics analysis is generally done locally at the institute and in an interactive way requiring fast response for the prosecution of the analysis. The size per event of these n-tuples can vary from hundred of bytes up to the hundred of kilobytes depending on the chosen event complexity. Given the huge rate of events collected at the LHC [3], the processing time of these n-tuples can be very large and depends on the complexity of the physics events under study. Also the final analysis has to be run for a significantly high number of times in order to tune cuts and to make histograms and graphs. All these characteristics can result in a very large processing time. It is therefore very important to have the possibility to fully exploit the local processing resources with the maximum efficiency using all the processor cores. To this purpose some tests have been done on the PROOF [4] architecture trying to use all the available cores of the on the Lecce Public Cluster.

2. Lecce Public Cluster

The Lecce Public Cluster is formed by a dozen of multi-core servers which are dedicated to many research activities also by other groups, ranging from Astroparticle Physics to Theoretical Physics. Most of the machines are 4 core server while few of them are newer 8 core machines. The processor CPU clocks are quite different, ranging from 2 GHz to more than 3 GHz for the newest ones. All the machines are daily and routinely used from the respective group owners and they are also part of the European GRID, running many batch jobs. All the tests done are obviously limited by the heavy use of these machines, but the average load of the servers can be considered constant since all the test have been repeated and performed in short time periods.

Data are stored on a 10 TB File Server which is connected through STORM (single interface) and are seen on each of the machines via NFS protocol.

3. PROOF on the Lecce Public Cluster

The Parallel ROOT Facility, PROOF, is an extension of ROOT [2] enabling interactive analysis of large sets of ROOT files in parallel on clusters of computers or multi-core machines. To work in PROOF the original ROOT macros need only to be embedded in a ROOT specific class (TSelector). The basic architecture should not put any implicit limitation on the number of computers that can be used in parallel and the system should be able to adapt itself to variations in the remote environment (changing load on the cluster nodes, network interruptions, etc.). The PROOF technology is also quite efficient in exploiting all the CPU's provided by multi-core processors. A dedicated version of PROOF, PROOF-Lite, provides a zero configuration solution to take full advantage of the additional cores available in today desktops or laptops. The PROOF cluster in Lecce has been configured assigning a single worker to each core of the Public Cluster for a total of up to 128 workers. This basic configuration of the PROOF cluster is very simple to implement and can be setup in very short time.

4. Test with SUSY n-tuples

Lecce ATLAS Group is involved in the inclusive Supersymmetry searches with presence of missing transverse energy, multi-jets and couple of opposite sign leptons in the final state [5]. We produce (on the GRID) custom ROOT n-tuples (together with the Milan and Pavia Groups) whose size per event is around 700 bytes for real data. The produced files have a variable size between tenth and few tenths of megabytes (average of 40 MB/file). The official n-tuples (called D3PD) have much bigger sizes and are much more difficult to copy store and analyze. Monte Carlo n-tuples have larger sizes (approximately 2-3 times larger than real data) depending on the physics channel (top-antitop, QCD processes, di-boson samples and so on) which is needed to simulate the various kind of background sources to the expected signal. We moved from the GRID to the Lecce NFS File Server about 2300 data files for a total of

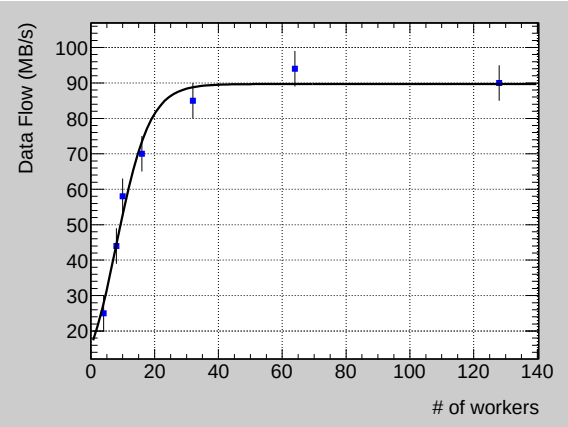

Figure 1. Analyzed data flow in MB/s versus the assigned number of workers in the PROOF pool

about 100 GB for the periods of data collection corresponding to 16 pb^{-1} acquired in 2010. Also other n-tuples have been produced for the study of trigger performance.

The analysis program is quite simple: it performs an event selection based on the variables of the ntuple and fills a few hundreds of histograms. The code can be run in the three standard modes, i.e. ROOT (single-core), PROOF-Lite (multi-core single machine) and PROOF (multi-core multi-machines) mode. Tests have been done in all modes for validation of the code, but we mainly exploited the PROOF mode.

In Figure 1 the analyzed data flow is shown versus the assigned number of workers defined in the PROOF pool. Each data point is an average of at least 3-4 tests and the error is the spread of these tests. A saturation is already observed at about 20 workers for this particular case where the data access speed is limited from the single Ethernet interface used on the NFS data server. No optimization of the network was performed and all the nodes of the pool were also used for other purposes so in this case the test did not reach the expected 120 MB/s limit which characterizes the Ethernet Gigabit interface. This limit has been in fact observed in many other performed tests but the stability to make an extensive test with this condition has never been reached. This test was performed during standard working time and was verified that load of the machines did not change too much during the test. The saturation at 90 MB/s corresponds to about 130000 processed Events/s.

This saturation at about 20 cores is due mainly to the network, being the analysis quite simple and then not taking too much computing time per event. For more complex analyses the network bottleneck should be less important and the use of many more cores should help to speed up the analysis.

At the moment the full data collected in 2010 (about 45 pb^{-1}) can be analyzed in less than a hour on the Lecce Public PROOF Cluster.

We bought and just received four 8-core machines (with 3 GB RAM/core) which we plan to dedicate only to this kind of tests (no GRID). Since each of these servers has a 500 GB internal hard drive, we will try to make a more complex PROOF cluster using a XROOTD [6] file server based on these internal disks which should optimize data transfers between the file server and the processing cores. This will allow to check if this configuration will solve the observed congestion problem and improve the cluster performance. The next step would be to try to couple XROOTD with the use of solid state internal disk arrays to further improve the performance of the cluster.

REFERENCES

1. ATLAS Collaboration is made of about 3000 Physicists coming from 170 Institutions of the following countries: Argentina, Armenia, Australia, Austria, Azerbaijan, Belarus, Brazil, Canada, Chile, China, Colombia, Czech Republic, Denmark, France, Georgia, Germany, Greece, Israel, Italy, Japan, Morocco, Netherlands, Norway, Poland, Portugal, Romania, Russia, Serbia, Slovakia, Slovenia, Spain, Sweden, Switzerland, Taiwan, Turkey, UK, USA.
2. R. Brun and F. Rademakers, ROOT: an object oriented data analysis framework, Nucl. Instrum. Meth. A 389 (1997) 81; see also http://root.cern.ch.
3. L. Evans and P. Bryant, LHC Machine, JINST 3 S08001 (2008).
4. M. Ballintijn et al., Parallel interactive data analysis with PROOF, Nucl. Instrum. Meth. A 559 (2006) 13
5. ATLAS Collaboration, Search for an excess of events with identical flavour lepton pairs and significant missing transverse momentum in $\sqrt{s}$ = 7 TeV proton-proton collisions at the ATLAS experiment, ATLAS internal note ATL-COM-PHYS-2011-149 (2011).
6. http://xrootd.slac.stanford.edu/