Distributed analysis with PROOF in ATLAS collaboration

S.Y. Panitkin[1], D. Benjamin[2], G. Carillo Montoya[3], K. Cranmer[4], M. Ernst[1], W. Guan[3], H. Ito[1], T. Maeno,[1] S. Majewski[1], B. Mellado[3], O. Rind[1], A. Shibata[4], F. Tarrade[1], T. Wenaus[1], N. Xu[3] and S. Ye[1]

[1]Brookhaven National Laboratory, Upton, NY 11973, USA
[2]Duke University, Durham, NC 27708, USA
[3]University of Wisconsin-Madison, Madison, WI 53706 USA
[4]New York University, New York, NY 10003, USA

Abstract. The Parallel ROOT Facility - PROOF is a distributed analysis system which allows to exploit inherent event level parallelism of high energy physics data. PROOF can be configured to work with centralized storage systems, but it is especially effective together with distributed local storage systems - like Xrootd, when data are distributed over computing nodes. It works efficiently on different types of hardware and scales well from a multi-core laptop to large computing farms. From that point of view it is well suited for both large central analysis facilities and Tier 3 type analysis farms. PROOF can be used in interactive or batch like regimes. The interactive regime allows the user to work with typically distributed data from the ROOT command prompt and get a real time feedback on analysis progress and intermediate results.

We will discuss our experience with PROOF in the context of ATLAS Collaboration distributed analysis. In particular we will discuss PROOF performance in various analysis scenarios and in multi-user, multi-session environments. We will also describe PROOF integration with the ATLAS distributed data management system and prospects of running PROOF on geographically distributed analysis farms.

1. Introduction

The Atlas Experiment [1] at the Large Hadron Collider (LHC) is supposed to start taking data in 2009. The experiment will record, reconstruct and analyze more than 10^9 events per year. Such an amount of data poses a serious computing challenge.

The Atlas Computing model [2] is based on a Grid paradigm [3], with multilevel, hierarchically distributed computing and storage resources. Raw data from the Atlas detector are received by the Tier 0 facility at CERN, where initial event reconstruction and calibration take place. Reconstruction jobs produce datasets in formats suitable for physics analysis and these datasets are distributed to Tier 1 facilities around the world. Tier 1 centers also store a fraction of raw data for archiving purposes and consequently, they can perform raw data reprocessing if necessary. Typically a regional Tier 1 facility has several smaller Tier 2 sites associated with it. The Tier 2 sites are primarily responsible for Monte-Carlo production and analysis. Tier 3 centers are envisaged primarily for end user analysis. In practice a Tier 3 can be part of an existing Tier 1 or Tier 2 facility.

Several streams of raw data selected by high level trigger will be delivered for reconstruction

Table 1. Parameters of some of the Atlas PROOF sites.

Site Location	Number of Cores	Disk Storage (TB)	Usage
BNL	112	45	analysis, SSD tests
Wisconsin-Madison	200	100	analysis, Condor-PROOF tests
Munich LMU/LRZ	40	40 (dCache)	analysis, performance tests

and storage at Tier 0, with subsequent distribution to Tier 1s. The average raw data event size is assumed to be around 1.6 MB. Reconstructed events will be stored in Event Summary Data (ESD) format with an expected event size of about 500 kB. Analysis Object Data (AOD) are derived from ESD, with a target event size of 100 kB. Both ESD and AOD are written to disk and distributed to various Atlas computing Tiers as POOL [4] ROOT [5] files. Typically, data in ESD and AOD formats are analyzed, using the Athena analysis framework [6], on the Grid. A special framework, Athena ROOT Access, was developed to facilitate ESD and AOD analysis in ROOT.
Finally, a format intended for "last step" analysis and histogramming, so called Derived Physics Data (DPD) is also defined. DPDs contain only objects and events needed for a particular type of physics analysis, thus reducing storage requirements and increasing processing speed. DPD files are written either as POOL/ROOT files or as pure ROOT files.
The addition of DPD in the computing model is due to acknowledgement of common practice by physicists of building sub-samples in a format suitable for direct analysis and display in ROOT [2].
The Parallel ROOT Facility - PROOF [7] is a distributed analysis system which allows to exploit inherent event level parallelism of high energy physics data. In its current implementation it is tightly integrated with Xrootd [8] and allows to fuse distributed computing and distributed storage for effective use of resources. It works efficiently on different types of hardware and scales well from a multi-core laptop to large computing farms. From that point of view it is well suited for both large central analysis facilities and Tier 3 type analysis farms. In this paper we'll attempt to describe PROOF related activities in the Atlas collaboration.

2. PROOF tests

Currently there are several Atlas groups worldwide that are interested in exploiting the PROOF technology for physics analysis and implemented PROOF farms of different scales. Most of them have the status of a test farm, but several sites were successfully used in day-to-day data analysis and Atlas wide computing exercises- like "Full Dress Analysis Rehearsals" - in 2008. Table 1 shows basic parameters of the largest Atlas PROOF sites operational at the time of writing.

2.1. PROOF performance and scalability studies

One important issue is the PROOF farm operation in a multi-group, multi-user environment. The PROOF resource scheduling mechanism ensures resource allocation to the jobs in order to optimize overall performance of the farm. That includes enforcement of usage policies, based on the description of priorities and quotas of different users and groups. PROOF implements resource scheduling at two levels. The first mechanism acts on every worker node. It controls the fraction of CPU resources, used by each job, according to the user priority. The second level is a central scheduler which determines the user priorities and assigns worker nodes to jobs.
PROOF performance and scalability with Atlas analyzes were extensively studied in different

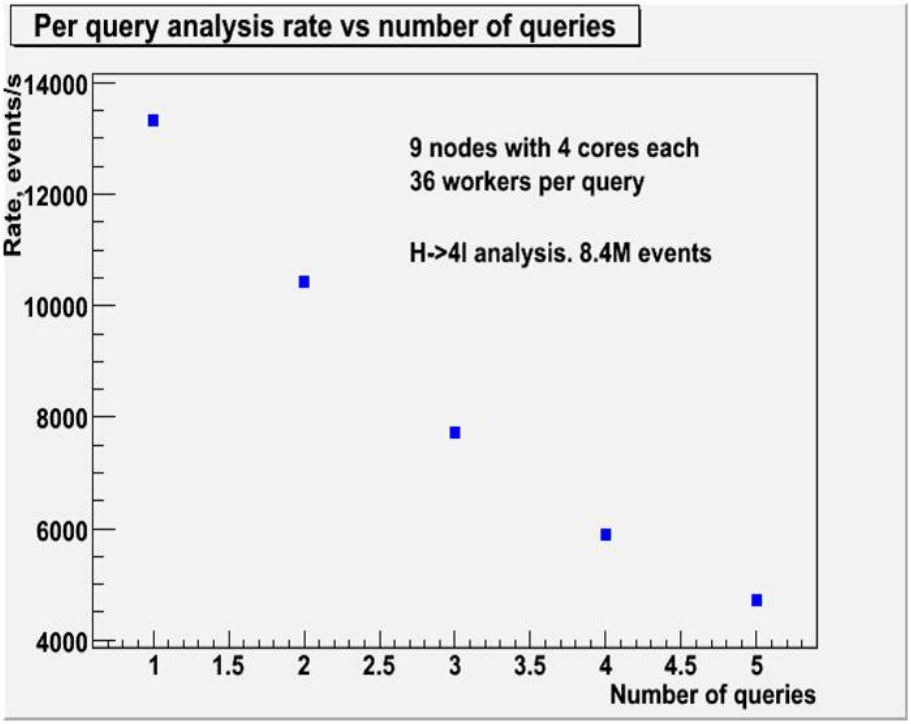

Figure 1. Analysis rate per query as a number of full occupancy queries. See description in the text

analysis scenarios. Since PROOF utilizes a local storage model and the analysis is performed on data located on the same node where the job is running, its performance scales up with the number of nodes quite well. One interesting characteristics is the PROOF farm performance in a multiuser, multi-session environment.

Figure 1 shows the analysis rate per query as a function of the number of simultaneously running full occupancy queries. In this context query means a ROOT macro executed on the farm, with several instances of the same analysis code running in parallel, processing different parts of a given dataset. Full occupancy query means that the query size is such that every available CPU core on the worker nodes is occupied with the analysis job. In that case full occupancy query was running with 36 workers.

This particular query is running a typical Atlas Higgs search analysis on a dataset in DPD format. It is both CPU and I/O intensive.

Since all queries in the scenario shown in Figure 1 have equal priorities, the decline in per-query analysis rate is a mere reflection of resource sharing. Each job gets a smaller allocation of resources when the number of simultaneously running jobs increases.

A more interesting measure of performance in this situation is the aggregated analysis rate of all jobs running on the PROOF farm. Figure 2 shows the aggregated analysis rate of the PROOF farm as a function of the number of full occupancy queries. One can see that the aggregated farm performance keep rising until the number of queries reaches four and then essentially saturates. The rise means that a single query of a given type can not utilize resources with 100% efficiency, probably due to a large fraction of time spend in I/O operations. The saturation means that, for a given load, available resources of the farm were exhausted at 4 running queries and it makes no sense to increase number of simultaneously running jobs any further. It is clear that the knowledge of the farm's performance "sweet spot" is important for optimal utilization of

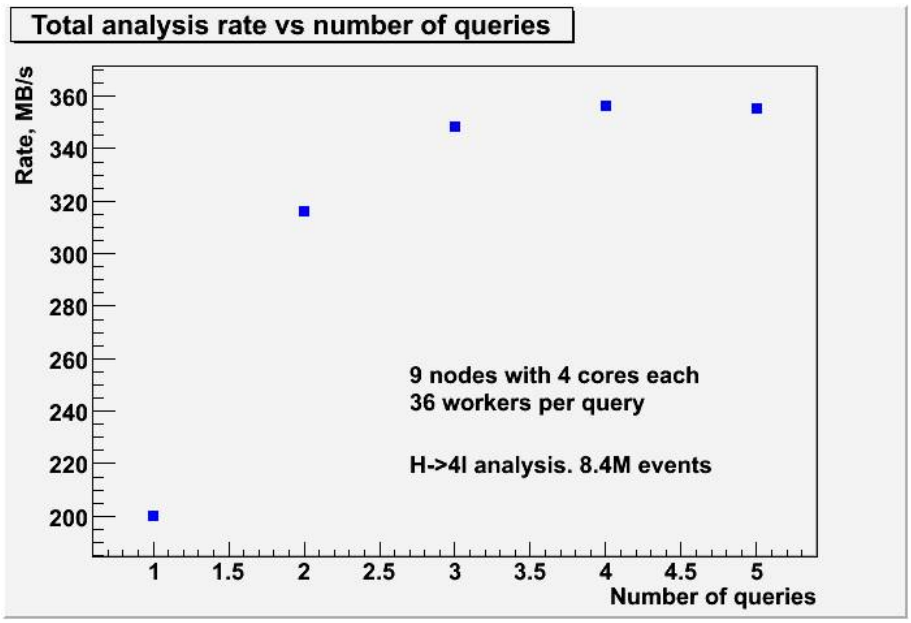

Figure 2. Aggregated analysis rate of a PROOF farm as a function of the number of full occupancy queries. See description in the text

resources and should be used for tuning load management parameters of the farm.

2.2. I/O performance

Typically analysis jobs require high input/output rates, with the input rate being more important. Since PROOF encourages local processing the performance of the local disk sub-system is a crucial parameter. We extensively studied different aspects of input-output performance of PROOF. In was found that a typical ROOT based Atlas analysis job requires 10 to 15 MB/s input rate. We also found that multicore machines with single hard disk drive (HDD) can support no more than two concurrent jobs near peak efficiency. The HDD performance deteriorates when the number of simultaneously running jobs increases beyond two, due to increased latencies associated with random disk access. We found out that HDD RAID arrays provide better performance but still can not keep up with full occupancy queries running on 8 core machines.

At the PROOF farm at BNL a study of Solid State Drives (SSD) was performed. It was found that SSD show an order of magnitude better performance than HDD and provides a good match for PROOF running on multicore hardware. Detailed descriptions of the results can be found here [9].

At the PROOF farm in Munich comparisons were made between storage elements based on Xrootd, dCache and Lustre. More details can be found in this volume [10].

2.3. Federated PROOF clusters

In principle PROOF and Xrootd architectures support federation of geographically distributed clusters. This is a very interesting capability for Tier 3 sites. In a view of very large data volumes expected at LHC the paramount question for a Tier 3 site is how to make a typically small and resource limited Tier 3 analysis facility useful? Is there enough space and CPU power to perform physics analysis in a reasonable time?

Proof cluster federation allows to pull together resources of several clusters and combine them

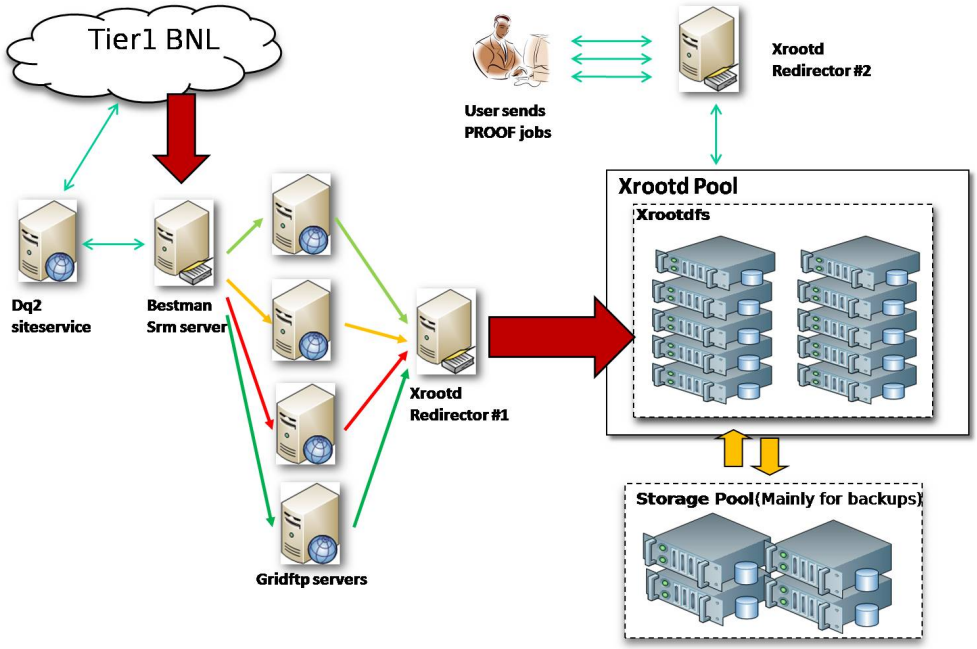

Figure 3. Data management scheme as implemented at the University of Wisconsin-Madison PROOF farm.

into a single analysis cluster. PROOF hides the complexity of the cluster topology and looks, for an end user, like a single analysis space of increased size. It's worth mentioning that such partnerships between federated sites do not need to be permanent and can be easily and quickly reconfigured, thus allowing flexible combination of various geographically distributed resources. First tests of federation of geographically distributed PROOF farms were performed by the Atlas collaboration. The PROOF farms at BNL and University of Wisconsin-Madison were successfully federated. In a view of stringent network security regime at BNL we used ssh tunnels for inter cluster communication. Since tunnels are needed only for Xrootd command channel communications and not for data movement this was not found to be a big limitation. Use of dual homed nodes for cluster federation will provide a better solution for linking clusters together in situations with strict site firewalling. Tests of performance characteristics of federated PROOF clusters, scalability and data management were not performed at this time and are clearly interesting and needed.

2.4. Data Management

One of the issues facing any PROOF installation is data management. As we mentioned before modern versions of PROOF rely on Xrootd for data discovery and delivery. By itself Xrootd is a "data serving" technology and includes a very limited set of tools for data management. Transferring large volume of data on the scale typically required for Atlas analyzes is not a trivial task. Chaotic nature of end user analysis also complicates the situation. In order to operate PROOF installation efficiently, one needs a variety of tools to copy data in and out of Xrootd space, a file catalog, tools for space management and disaster recovery. User level tools for data discovery are important as well.

Figure 3 schematically shows a PROOF farm data management system implemented at University of Wisconsin-Madison. It consists of several pieces providing necessary functionality. Data movement in and out the farm is done with the help of an Atlas-wide tool called DQ2 [11].

This layer manages data discovery and transfer from and to Atlas Grid sites, as well as central cataloguing services. The next layer is BeStMan [12] which provides a SRM interface to the Xrootd storage element. This approach allows not only to leverage tools developed for the Atlas Grid based distributed analysis in PROOF context, but also gives users a familiar set of tools for data discovery and management.
It is worth mentioning that detailed data management procedures and policies will be directly affected by analysis policies which will be developed for a particular site. Such analysis policies are currently under discussion.

2.5. PROOF farm monitoring
Monitoring is an essential part of any distributed storage and computing facility. One needs tools for understanding of various performance aspects, performance optimization, problem discovery and recovery, etc. Currently the farm at BNL employs a monitoring set up consisting of two components. For monitoring on hardware and operating system level we use the Ganglia system [13], that is standard framework at the Tier 1 facility at BNL. For monitoring of Xrootd the BNL farm utilizes a framework developed at SLAC and used in the BABAR experiment. It collects and displays, in real time, information about Xrootd files, clients and servers. Both monitoring frameworks feature convenient web based interfaces
The Monitoring system at The Wisconsin-Madison PROOF farm is based on the MonALISA [14] framework and is similar to the monitoring set up used by the ALICE experiment [15, 16] at CERN.

3. Conclusions
Currently several PROOF test farms are operational in Atlas. Significant experience with PROOF was gained. Several Atlas analysis scenarios were tested. Improved integration with the Atlas distributed data management was demonstrated. Working systems for farm management and monitoring were introduced. Federation of geographically distributed PROOF clusters was demonstrated. Several bugs in PROOF were discovered, reported to developers and fixed. Several Wiki pages are available for Atlas PROOF users with instructions and PROOF usage examples. Last but not least - several PROOF tutorials were given at Atlas Analysis Jamborees. In summary, PROOF is an attractive technology for Atlas, especially for Tier 3 centers. It provides a high performance platform for DPD analysis and complements Atlas distributed Analysis, in its present implementation, rather nicely.

References

[1] Armstrong W W *et al.* (ATLAS) 1994 Atlas: Technical proposal for a general-purpose p p experiment at the large hadron collider at cern cERN-LHCC-94-43

[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

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

[4] 2006 LHC persistency framework, pool of persistent objects for LHC. URL `http://pool.cern.ch/`

[5] 2008 The ROOT system home page. URL `http://root.cern.ch/`

[6] ATLAS Computing Group, ATLAS Computing Technical Design Report, CERN-LHCC-2005-022, 2005

[7] Ballintijn M *et al.* 2006 *Nucl. Instrum. Meth.* **A559** 13–16

[8] Dorigo A, Elmer P, Furano F and Hanushevsky A 2005 *TELE-INFO'05: Proceedings of the 4th WSEAS International Conference on Telecommunications and Informatics* (Stevens Point, Wisconsin, USA: World Scientific and Engineering Academy and Society (WSEAS)) pp 1–6 ISBN 960-8457-11-4

[9] Panitkin S *et al.* 2009 *Proccedings of International Conference on Computing in High Energy Physics (CHEP 09)*

[10] Calfayan P *et al.* 2009 *Proccedings of International Conference on Computing in High Energy Physics (CHEP 09)*

[11] Branco M *et al.* (ATLAS) 2008 *J. Phys. Conf. Ser.* **119** 062017

[12] 2009 Berkeley storage manager (bestman) URL `http://datagrid.lbl.gov/bestman/`

[13] Massie M 2004 *Parallel Computing* **30** 817–840 ISSN 01678191 URL `http://dx.doi.org/10.1016/j.parco.2004.04.001`

[14] Newman H B, Legrand I C, Galvez P, Voicu R and Cirstoiu C 2003 *Proceedings of International Conference on Computing in High Energy Physics and Nuclear Physics 2003* (La Jola, California) p 907

[15] Aamodt K *et al.* (ALICE) 2008 *JINST* **3** S08002

[16] Meoni M 2009 *ICCS '09: Proceedings of the 9th International Conference on Computational Science* (Berlin, Heidelberg: Springer-Verlag) pp 114–122 ISBN 978-3-642-01969-2