A Virtualized Tier-3g Facility Installation for WLCG Network of CERN

H. Saygın, S. Kuday, A. Alici

Istanbul Aydın University, Application and Research Center For Advanced Studies, 34295, Istanbul, Turkey

E-mail: hasansaygin@aydin.edu.tr, sinankuday@aydin.edu.tr, agah@aydin.edu.tr

I. Turk Cakir

Giresun University, Department of Energy Systems Eng., Giresun, Turkey

E-mail: iturkcak@cern.ch

Abstract. A Tier-3g Facility within the computing resources of Istanbul Aydin University has been planned and installed with the TR-ULAKBIM national Tier-2 center. The facility is intended to provide an upgraded data analysis infrastructure to CERN researchers who are the members in the recent nation-wide projects and international projects such as ATLAS and CMS experiments. The fundamental design of Tier-3g has been detailed in this work with an emphasis on technical implementations of the following parts: Virtualization of all nodes, VOMS usage for reaching fast experimental data in the WLCG network, batch cluster / multicore computing with HTCONDOR and PROOF systems, usage of grid proxies to access code libraries in AFS and CVMFS, dynamic disk space allocation and remote system mounting of EOS. We also present the interpretation of test results that was obtained by the simulation of typical analysis codes.

1. INTRODUCTION

Recently, cluster computer systems are based on computational operations have been widely used especially in disciplines like experimental physics, astronomy and medical sciences. The main need to handle the big data, that is obtained from the experiments and analyses, has led to the emergence of new technologies. Given the petabyte-scale data volumes produced by the Large Hadron Collider (LHC) experiments, a reliable high-speed network is a crucial requirement for an analysis facility. ATLAS (A Toroidal LHC Apparatus) [1] as a major experiment operates at the European Nuclear Research Center (CERN), explained the computing model a few years before the experimental data acquisition [2]. According to that plan, (i) transformation of the raw data received from the Tier-0 center to the distributed data by sharing among the Tier-1 and Tier-2 centers in the hierarchical structure, (ii) processing in CERN analysis facility giving researchers an active role in data management and (iii) reprocessing stages were predicted. The Worldwide LHC Computing Grid (WLCG), which was developed for this purpose, has successfully completed initial testing and initial data processing [3]. On the other hand, major institutes predicted that a network, performing central processing, would not be able to serve the user analysis as well as bulk processing tasks such as reconstruction and simulation. In this period, Tier-3g centers

Table 1. Virtualized server types and capabilities with minimum allocated sources.

No	Server Type	CPU	RAM	DISK
1	Head Node	8 Core	24 GB	2 TB
2	Interactive Node	8 Core	36 GB	6.5 TB
3	NFS Node	8 Core	24 GB	2 TB
4	Worker Node	8 Core	36 GB	6.5 TB
5,6	Other Nodes (LDAP, Squid ...etc.)	8 Core	24 GB	1 TB
	TOTAL	**40 Core**	**144 GB**	**18 TB**

have been developed which are set up for researchers, decentralized and included fully under the control of the institutes [4, 5]. These centers are part of the WLCG and can store the latest experimental data and carry out analyses in line submitting jobs to grid with the priorities of the owner institutes as indicated by the "-g "suffix. Researchers that are affiliated under the owner institute can develop and optimize their analysis software programs, using all of the technical possibilities of Tier-3g centers. Related studies [6] show that such facilities will gain importance for the high luminosity (HL) phase of LHC which is expected to continue to operate until 2023.

2. NATIONAL NETWORK & COMPUTING SOURCES

National Academic Network and Information Center (ULAKBIM), was established as an institute affiliated under the Scientific and Technological Research Council of Turkey (TUBITAK) in June 1996. The main aim of ULAKBIM is establishing and operating networks for research and education activities of Turkish universities and research institutions, provide links to domestic and international networks and support these networks by observing technological developments in the information technology era. In 2003, with more than 160 node points and 80 university centers, ULAKBIM network has been extended throughout the region and has become the member of European Research Network (GEANT). A National High Performance Computing Service (TRUBA) was introduced in the same year. TRUBA has recently reached an average monthly usage of 5M CPU/hour with 2 PB storage area, 1024 processors and 1130 user researchers In 2005, TR-Grid CA is accredited by EUGridPMA and became the main certification authority in the region.

Other computing clusters with a focus on research data analysis can be listed on-site of the developed universities nationwide. For this work, taking official permissions from the host university and ULAKBIM, all virtualized servers are provided with SL6 (Scientific Linux 6) on one physical cluster server of VMware Hypervisor 6.5 based on EMC SAN storages with specified capabilities as in Table 1.

3. DESIGN OUTLINES

3.1. Computing Elements (CE):

Basically, the Tier-3g system is a grid enabled computer cluster that hosts six or more different types of servers as shown in Figure 1. Each server including in the cluster is connected to a private network and have the VPN option to be connected from the public network (WAN). Briefly, head and worker nodes are the main elements of the batch system to process submitted jobs; NFS node that simultaneously lets GridFTP service to get data through WLCG, is also considered as the main storage; Interactive node lets users to create remote sessions to perform their analysis. All other servers are installed to perform the spesific services in their names (Ex: LDAP server for LDAP service) The number of interactive and worker servers in the system is designed to allow N servers to work together, as well as being proportional to the expected

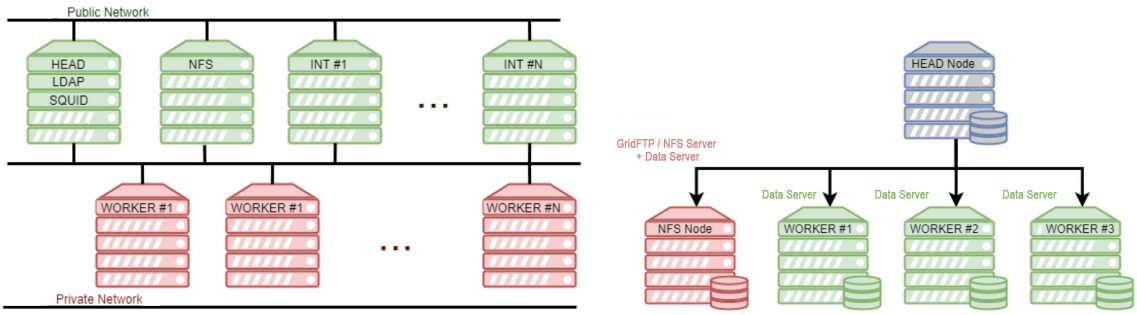

Figure 1. Initially designed network and server structure with N numbers of servers named as HEAD, INTERACTIVE, LDAP, SQUID, NFS and WORKER (left). Most recent structure of the cluster with data servers, merged GridFTP and NFS servers (right).

processing power capacity. One of the advantages of a fully virtualized system is the ability of dynamical provision CPU and memory resources. For the initial purposes; 40 CPU cores, 144 GB RAM and 18 TB disk space has been allocated in total and resources are sufficient to double the initial configurations according to user demands. The HTCondor batch processing system [11] is deployed on the cluster to manage user-jobs for the computing element as mentioned in the next section. The required console authorization for the Tier-3g system running over the command line is done from a single point through the LDAP (The Lightweight Directory Access Protocol) server. Even if the number of servers in the created system increases, the amount of time spent managing the server is saved because authorization is done from a single node. The SQUID Proxy server interfaces to the data that the servers in the Tier-3g system will download and keeps the necessary files in the cache, preventing the same data from being downloaded again. Lastly, NFS server has been installed to allow files to be shared between servers, and large files accessed without being held on the server.

3.2. Storage Elements (SE):

For DDM compatibility, we have chosen to use XRootD [12] as the storage element in the cluster since it has already been widely used in many tiers. Note that XRootD is a service implemented in ROOT framewok [13] that allows the user to remotely read data located at various destinations. In our setup, XRootD server daemons run on NFS, interactive and head servers and provide access to all data files in the systems mounted on NFS servers. However, XRootD is not the only method to get data for clients that are connected to interactive server. It is important to understand that there is an enormous difference between Tier-3s and Tier-1/2s comparing with their storage capabilities. Note that users that are often connected to Tier-3s perform analysis tasks and only need a portion of data with high processing capability while Tier-1/2s are centrally perform production and transfer tasks and need a massive amount of data. With developing technologies, users may no longer need to download data instead they may use remote mounting (e.g.: EOS) or remote analysing (e.g.: FAX) options. We have succesfully tested both of these options for our Tier-3g facility.

3.3. File System and Network Planning:

NFS server has been set to host scaled and distributed data managements such as XRootD and wide area network systems such as GridFTP. In Figure 2, the modified network structure after virtualizing NFS server is shown with the multiple services. Head server is positioned as the management center for file system and distributed stroge system. For required configurations of CVMFS (CernVM File System) and AFS [15] , head server set and authorized. Remote storage

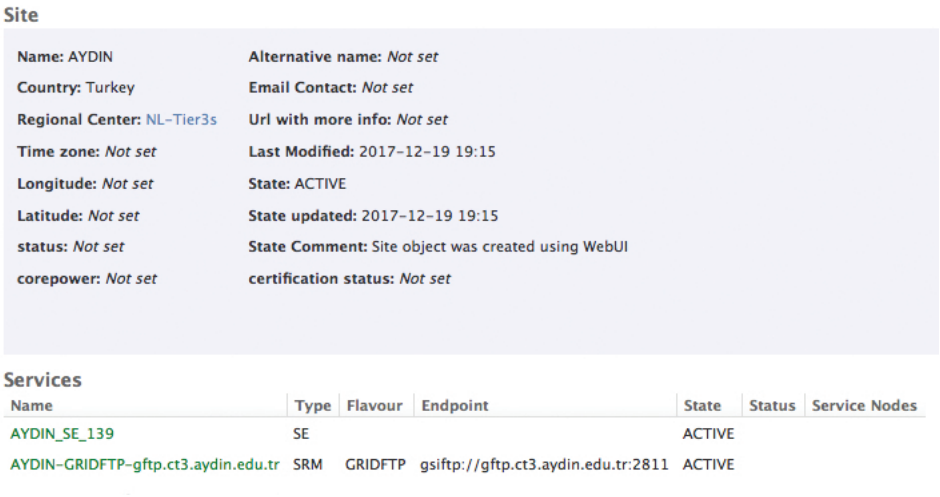

Figure 2. Registration information of SE and GridFTP in monitoring system of ATLAS

area of CERN utilized as EOS [14] is mounted via FUSE following the published IT descriptions at [16].

3.4. Batch System:
After server and network installations are completed for the system, it is required to choose and configure batch systems that allow users to run their analysis codes in the fastest and most efficient way. CERN documentation from SVN repositories [17] are taken into account configuring services such as firewall, local user settings, nfsv4 connection, LDP authorization, HTCondor configuration, CVMFS connection, Kerberos and needed related libraries. However, one can realise that the mentioned documentation has been prepared for a large scale tier with online physical servers. For virtual servers, admins have to write their own configurations manually. For HTCondor batch configuration, NFS server has been set and configured as the headnode and all worker servers are set with default configurations.

3.5. AGIS Registration:
To enable the grid option for our cluster, at the first stage, we have obtained a machine server certificate from the Turkish certificate authority - ULAKBIM. We have configured our virtualized GridFTP / NFS server such that XrootD service can reach properly and then allowed space-token has been created in the same server named as LOCALGROUPDISK. However, we haven't installed conveniently-used DPM services for disk storage management to avoid overloading on virtual systems. Note that new intention of ATLAS is to migrate from SRM protocol to GridFTP protocol as explained in Ref [18]. Although we did not deploy our system with DPM installation, that cause no errors or omissions for data transfers. At the last step of the registration, we have informed ATLAS with our deployments and technical capacity. As DDMendpoint name, we have prefered AYDIN_LOCALGROUPDISK under NL Cloud within the WLCG as in Figure 3.

4. DDM TRANSFERS & MONITORING
Tier-3 facility monitoring options are the same with using in ATLAS infrastructure for Tier-1s and Tier-2s. DDM dashboard monitoring and batch tools are avaliable for network&server log monitoring. However, since no production and group task jobs are accepted by the Tier-3 facility, only user subscriptions and admin transfer rules are visible with these tools. Considering the

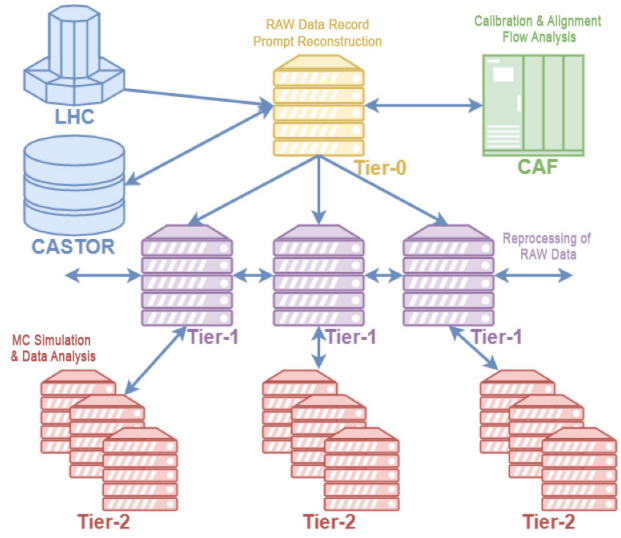

Figure 3. Data flow diagram of ATLAS

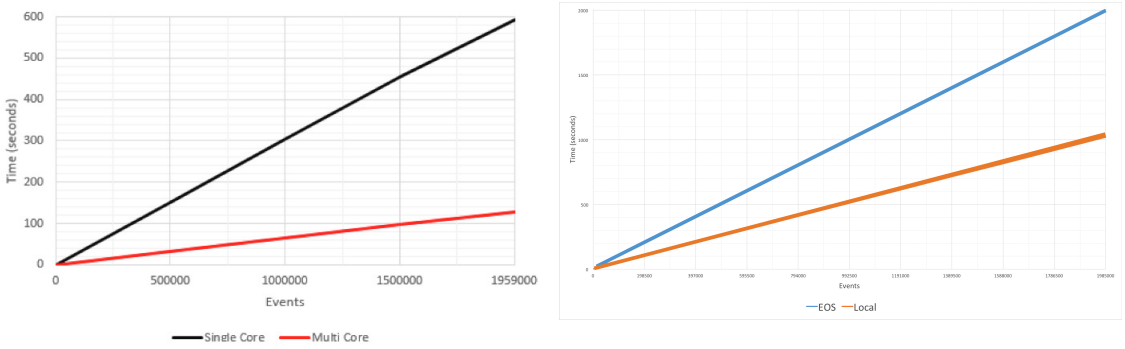

Figure 4. Basic Test results (left) and EOS vs. Local Test results (right)

data flow diagram in Figure 3 and taking into account the data structures in ATLAS, recently it is practical to accept DAOD data in to Tier-3 disks. Note that DAOD as the most compact form in derived data format, has approx. 0.01 of AOD (Analysis Object Data) format [20]. This corresponds to approx. %20 of RAW data of ATLAS which process data in a few PBs order in a year. DDM transfers are tested using Rucio tool [21] of ATLAS and succeded in the most cases. However, institute firewall restrictions caused "Connection timeout" failures that is solved after opening corresponding ports to SRM and GridFTP protocol in gridFTP server configuration.

5. PERFORMANCE RESULTS IN ANALYSIS

Soon after the Tier-3 installation process has finished, we have performed a several run tests to be able to see the quantitative evaluation of the cluster. Those tests can be categorized under the following titles: i.) Basic analysis tests ii.) Benchmark Event Loop Tests iii.) Remote storage (e.g: EOS) based analysis Tests.

For basic tests, we have prepared a simple multithreaded ROOT based analysis code that fills histograms using anti-kt jet algorithms [19] and kinematic variable calculations while accepting a simple event file as input. All input files are simulated for LHC runs similar to AOD data

format with different event sizes. As an example, one of the selected single core run worked over 1.959 million events for 593.74 seconds as shown in Figure 4(left). The same test applied for the multicore case with 48 CPUs using PROOF extention of Root and the results are obtained in apporaximately 125 seconds as shown in the same plot.

For event loop tests, we have used more sophisticated ROOT based analysis code similar to official ATLAS tools that can run on real data but filling only a few kinematic histograms as the output. Note that, official analysis code of ATLAS top group includes many subpackages such as systematic uncertainities, Pile-up Reweightings, Jet calibrations, ...etc. However we have used only fundamental packages as shown in the offline software tutorials [22].

Lastly, we have repeated the last performance test putting data inputs in EOS space area to remotely reach them over Rucio connection during the analysis runs. Results are shown as in Figure 4(right). All of the test results are taken online observing the network and storage performances.

6. CONCLUSION

In this paper, we aim to point the significant steps during the Tier-3 facility installation. We have observed that if targeted user capacity is manageably small, the virtualization of servers provides great advantanges since the system tends not to consume too much resources. However, for such a system, allocated worker hardwares can be low, as long as the users prefer running multi threaded analysis codes using PROOF. One crucial point is that users will need multi threaded codes if they choose to run it on a PROOF-farm instead of sending related jobs to the LHC grid. Therefore non-multi threaded analysis codes may be a big issue in the near future since they are getting voluminous and unmanageable. Another observation is that disk space problem of users is getting unimportant if the system allows them to mount remote disk spaces. Note that recent connection speeds allow users to perform remote analysis via mounted spaces without downloading data. In the near future, it may be expected to have new localizations for experimental data and user analysis by scientific collaborations.

References

[1] ATLAS Collaboration (Aad, G. et al.), "The ATLAS Experiment at the CERN Large Hadron Collider" - JINST 3 (2008) S08003

[2] Adams, D. Barberis, D. Bee, C. P. Hawkings, R. Jarp, S. Jones, R. Malon, D. Poggioli, L. Poulard, G. Quarrie, D. Wenaus, T. "The ATLAS Computing Model", Computing in High Energy Physics and Nuclear Physics 2004, Interlaken, Switzerland, 27 Sep - 1 Oct 2004, pp.1007 (CERN-2005-002)

[3] Shiers, J. "The Worldwide LHC Computing Grid", Computer Physics Communications, Volume 177, Issues 1-2, P 219 – 223 (2007)

[4] ATLAS Collaboration (Ueda, I. for the collaboration) , "ATLAS distributed computing operations in the first two years of data taking" - PoS ISGC2012 (2012) 013

[5] Gonzalez de la Hoz, S. et al. "Analysis Facility Infrastructure (Tier-3) for ATLAS Experiment", Eur. Phys. J. C. 54, 691 – 697 (2008).

[6] D. Adamova, M. Litmaath, "New strategies of the LHC experiments to meet the computing requirements of the HL-LHC era", PoS BORMIO **2017**, 053 (2017).

[7] Haupt, A., Kemp, Y. "The NAF: National Analysis Facility at DESY", Journal of Physics Conference Series, V 214, P5 (2010)

[8] Villapana, M. et al. "ATLAS Tier-3 within IFIC-Valencia analysis facility", ", Journal of Physics Conference Series, V 396, P4 (2012)

[9] G. Bell ; J. Gray ; A. Szalay "Petascale Computational Systems", Computer, Volume 39, Issue 1, Pages 110 – 112 (2006)

[10] Belov, S. et al. "VM-based infrastructure for simulating different cluster and storage solutions used on ATLAS Tier-3 sites" J. Phys.: Conf. Ser. 396 042036 (2012)

[11] Bockelman, B., et al.: "Commissioning the HTCondor-CE for the open science grid."J. Phys. Conf. Ser. 664, 062003 (2015)

[12] Gardner, R., Campana, S., Duckeck, G. et al.: "Data federation strategies for ATLAS using XRootD ", J. Phys. Conf. Ser. 513, 042049 (2014)

[13] Antcheva, M. Ballintijn, B. Bellenot, M. Biskup, R. Brun, N. Buncic, P. Canal,D. Casadei, O. Couet, V. Fine, L. Franco, G. Ganis, A. Gheata, D.G. Maline,M. Goto, J. Iwaszkiewicz, A. Kreshuk, D.M. Segura, R. Maunder, L. Moneta, A.Naumann, E. Offermann, V. Onuchin, S. Panacek, F. Rademakers, P. Russo, M.Tadel, "ROOT; A C++ framework for petabyte data storage, statistical analysis and visualization", Comput. Phys. Comm. 180 (12) (2009) 2499 - 2512
[14] Peters, A.J., Sindrilaru, E.A, Adde, G., "EOS as the present and future solution for datastorage at CERN", J. Phys. Conf. Ser. 664, 042042 (2015)
[15] Arsuaga-Rios, Maria; Heikkilae, Seppo S.; Duellmann, D.; et al. "Using S3 Cloud Storage with ROOT and CVMFS", J. Phys. Conf. Ser. 664, 022001 (2015)
[16] https://cern.service-now.com/service-portal/article.do?n=KB0003846
[17] http://svnweb.cern.ch/world/wsvn/atustier3
[18] https://twiki.cern.ch/twiki/bin/viewauth/AtlasComputing/SRM2GridftpMigration
[19] Atkin R. J. Phys. Conf. Ser., 645 (1), Article 012008, 10.1088/1742-6596/645/1/012008 (2015)
[20] Blomer J., "A quantitative review of data formats for HEP analyses"J. Phys. Conf. Ser., 1085, p. 032020 (2018)
[21] G Dimitrov et al. "Next generation database relational solutions forATLAS distributed computing "J. Phys.: Conf. Ser.513 042012 (2014)
[22] https://twiki.cern.ch/twiki/bin/viewauth/AtlasComputing/SoftwareTutorialxAODAnalysisInROOT