

7-8-2016

Deploying a CMS Tier-3 Computing Cluster with Grid-enabled Computing Infrastructure

Sean Stewart
sstew002@fiu.edu

DOI: 10.25148/etd.FIDC000759

Follow this and additional works at: <https://digitalcommons.fiu.edu/etd>

 Part of the [Physics Commons](#)

Recommended Citation

Stewart, Sean, "Deploying a CMS Tier-3 Computing Cluster with Grid-enabled Computing Infrastructure" (2016). *FIU Electronic Theses and Dissertations*. 2564.

<https://digitalcommons.fiu.edu/etd/2564>

This work is brought to you for free and open access by the University Graduate School at FIU Digital Commons. It has been accepted for inclusion in FIU Electronic Theses and Dissertations by an authorized administrator of FIU Digital Commons. For more information, please contact dcc@fiu.edu.

FLORIDA INTERNATIONAL UNIVERSITY

Miami, Florida

DEPLOYING A CMS TIER-3 COMPUTING CLUSTER WITH GRID-ENABLED
COMPUTING INFRASTRUCTURE

A thesis submitted in partial fulfillment of

the requirements for the degree of

MASTER OF SCIENCE

in

PHYSICS

by

Sean Stewart

2016

To: Dean Michael R. Heithaus
College of Arts, Sciences and Education

This thesis, written by Sean Stewart, and entitled Deploying a CMS Tier-3 Computing Cluster with Grid-enabled Computing Infrastructure, having been approved in respect to style and intellectual content, is referred to you for judgment.

We have read this thesis and recommend that it be approved.

Pete E.C. Markowitz

Lei Guo

Jorge L. Rodriguez, Major Professor

Date of Defense: July 8, 2016

The thesis of Sean Stewart is approved.

Dean Michael R. Heithaus
College of Arts, Sciences and Education

Andrés G. Gil
Vice President for Research and Economic Development
and Dean of the University Graduate School

Florida International University 2016

ACKNOWLEDGMENTS

I would like to thank my major professor, Dr. Jorge L. Rodriguez, for his endless support, patience, and guidance. His wealth of knowledge and experience, and his belief in me, were of invaluable importance to the completion of this project.

I would also like to thank Dr. Mengxing Cheng for his extensive contributions to my understanding of the workings of the cluster, and for dedicating so much effort and time to ensuring a successfully working cluster.

Finally, I would like to thank Dr. Pete E.C. Markowitz and Dr. Lei Guo for offering me their time and advice, and supporting me along the way.

ABSTRACT OF THE THESIS
DEPLOYING A CMS TIER-3 COMPUTING CLUSTER WITH GRID-ENABLED
COMPUTING INFRASTRUCTURE

by

Sean Stewart

Florida International University, 2016

Miami, Florida

Professor Jorge L. Rodriguez, Major Professor

The Large Hadron Collider (LHC), whose experiments include the Compact Muon Solenoid (CMS), produces over 30 million gigabytes of data annually, and implements a distributed computing architecture—a tiered hierarchy, from Tier-0 through Tier-3—in order to process and store all of this data. Out of all of the computing tiers, Tier-3 clusters allow scientists the most freedom and flexibility to perform their analyses of LHC data. Tier-3 clusters also provide local services such as login and storage services, provide a means to locally host and analyze LHC data, and allow both remote and local users to submit grid-based jobs. Using the Rocks cluster distribution software version 6.1.1, along with the Open Science Grid (OSG) roll version 3.2.35, a grid-enabled CMS Tier-3 computing cluster was deployed at Florida International University’s Modesto A. Maidique campus. Validation metric results from Ganglia, MyOSG, and CMS Dashboard verified a successful deployment.

TABLE OF CONTENTS

CHAPTER	PAGE
1. Introduction.....	1
1.1 Worldwide LHC Computing Grid	1
1.2 Computing Tiers	2
1.2.1 Tier-0.....	2
1.2.2 Tier-1.....	4
1.2.3 Tier-2.....	5
1.2.4 Tier-3.....	6
1.3 Compact Muon Solenoid Experiment.....	7
1.4 Rocks Cluster Distribution Software	9
2. Florida International University's CMS Tier-3 Computing Cluster.....	13
2.1 Project Aim	13
2.2 Frontend Node	13
2.3 Login Nodes.....	14
2.4 File Servers	15
2.5 Computing Element	16
2.5.1 Gatekeeper Node.....	16
2.5.2 Service Node.....	17
2.5.3 Compute Nodes.....	18
2.5.4 Job Flow in the Computing Element	19
2.6 Storage Element	20
2.6.1 Storage Resource Manager	20
2.6.2 Name Nodes.....	21
2.6.3 Data Nodes.....	22
2.6.4 Storage Flow in the Storage Element.....	23
2.7 Validation Software	24
3. Results.....	25
3.1 Ganglia.....	25
3.2 MyOSG.....	26
3.3 CMS Dashboard.....	28
4. Discussion.....	29
References.....	30

LIST OF FIGURES

FIGURE	PAGE
1-1: CMS data flow from Tier-0 through Tier-2	4
1-2: A Portion of the Rocks Kickstart Graph for FIU's Tier-3 cluster	11
1-3: List of the Rocks rolls installed on FIU's Tier-3 cluster	12
2-1: Schematic of job flow using glidein jobs	20
2-2: Schematic of the Hadoop Distributed File System architecture.....	23
3-1: Ganglia snapshot of load distribution, memory/CPU used, and network activity	25
3-2: Ganglia snapshot of the aggregated load on FIU's Tier-3	26
3-3: MyOSG displaying status of FIU's Tier-3	27
3-4: MyOSG listing the critical metrics for the CE and SE.....	27
3-5: CMS Dashboard metric result status	28

Chapter 1 Introduction

1.1 Worldwide LHC Computing Grid

The Large Hadron Collider (LHC), located in Switzerland, is the world's largest and most powerful particle accelerator. Each year, the LHC produces roughly 30 million gigabytes of data, made available to the over 8,000 physicists worldwide who are able to access and analyze this data in near real-time [1]. The ability for these physicists to process, analyze, and store LHC data from anywhere in the world is made possible by the Worldwide LHC Computing Grid (WLCG), often called the Grid.

The WLCG is a global computing infrastructure composed of two main grids, the European Grid Infrastructure (EGI) in Europe, and the Open Science Grid (OSG) located in the United States, as well as many associated regional and national grids [2]. Managed and operated by a collaboration between the four main LHC experiments (ALICE, ATLAS, CMS and LHCb), as well as the participating computing centers, the WLCG is the world's largest computing grid.

This global computing infrastructure allows multiple copies of data to be distributed and stored at different sites, which helps to provide physicists worldwide with a means to access data, regardless of their geographical location. The other main benefit of the WLCG is one of resource allocation. The computing resources—such as data storage capacity, processing power, sensors, and visualization tools—and the financial means needed to create and maintain those resources, are too great of an undertaking for a single institution to handle alone. Through the WLCG, self-sustaining computer centers worldwide are able to collectively pool their resources in order to achieve the effective outcome which would, if limited to a single institution, be impossible to obtain [2].

Additional benefits of the WLCG include data security, since data is replicated and stored across multiple sites, and around-the-clock monitoring and expert support as a result of computer centers existing across multiple time zones.

1.2 Computing Tiers

The WLCG links thousands of computers and storage systems located in over 140 computer centers across 41 countries [2]. These computing centers are arranged into four tiers, Tier-0 through Tier-3, which are configured to function as a single coherent system intended to store and analyze all of the LHC data. Each tier provides a specific set of services and resources which contributes to the overall production and stability of the WLCG. The collaboration of the computing centers with the WLCG is formalized through the negotiation and signing of a Memorandum of Understanding (MoU), with the exception of the Tier-3 centers, which have no formal agreement through the MoU [3]. This MoU describes the legal framework and agreements for the WLCG, as well as formalizes all WLCG partnerships. The services and responsibilities of each partnered computing center are described in the MoU.

1.2.1 Tier-0

The only computing center in the highest tier, Tier-0 (T0), is the European Organization for Nuclear Research (CERN) Data Centre (DC). The Data Centre and the LHC, where the collision data is created, are co-located at CERN. All of the LHC data first passes through the CERN Data Centre, with the Tier-0 being the first point of contact between the LHC and the Grid. CERN is responsible for collecting and safely storing the raw data, which consists of full event information that contains the digital readings from all of the experiments' detectors [4].

First, CERN accepts the raw data from the CMS Online Data Acquisition and Trigger Systems (TriDAS). Then, using the trigger information, this raw data is repacked—events from the unsorted online streams are sorted into physics streams of events with similar characteristics—into raw datasets. These repacked datasets are then archived to tape at the CERN Data Centre. Once these raw datasets have been archived on-site at CERN, additional copies are then distributed to the Tier-1 (T1) centers. This ensures that two copies of every raw dataset are stored—one at CERN and another at a Tier-1 center.

Next, after the calibration constants are obtained for each particular dataset, the raw datasets are set for the reconstruction phase, which aims to convert raw data into meaningful information for physicists by running the raw data through reconstruction algorithms. During this reconstruction period, the raw dataset is reconstructed as RECO (RECOstructed data) and AOD (Analysis Object Data) datasets. The RECO datasets contain very detailed reconstructed physics objects (tracks, vertices, jets, electrons, muons, etc.), as well as reconstructed hits/clusters. While RECO datasets can be used for analysis, they are simply too large for frequent use once a substantial data sample has been collected. The AOD datasets, on the other hand, are a relatively compact, convenient subset of the RECO datasets. While the AOD datasets provide physicists with the speed and flexibility necessary for analysis purposes, they lack the full event details and the fine resolution of the RECO datasets, and contain little-to-no hit/cluster information. However, these AODs still contain enough information about the event to be effectively used for most analyses [4].

Once the reconstruction phase is completed, each RECO dataset is distributed to the Tier-1 center where its corresponding raw dataset is stored. Finally, the full AOD datasets are distributed to all of the Tier-1 centers. The data flow from Tier-0 to Tier-1 is illustrated in Figure 1-1.

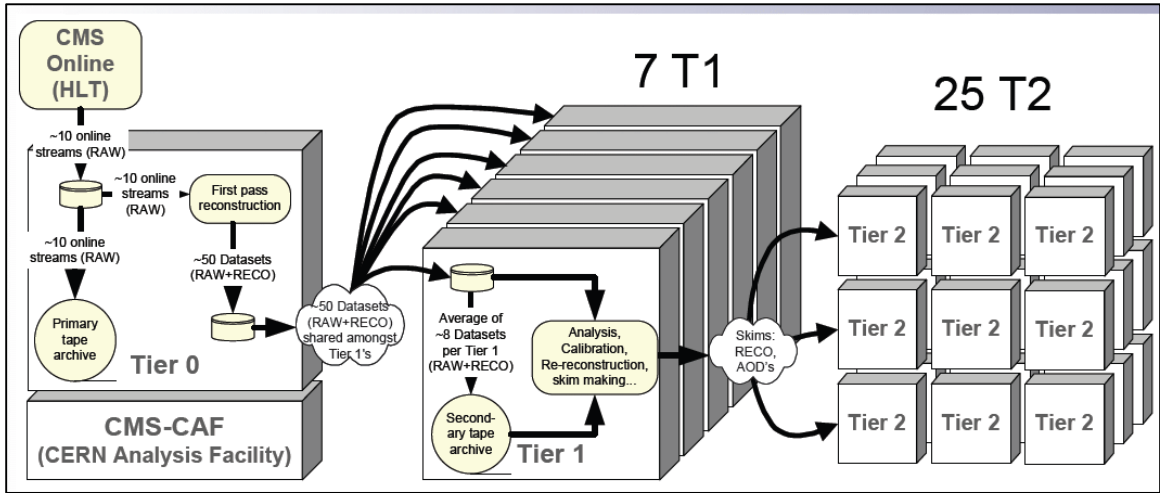


Figure 1-1: CMS data flow from Tier-0 through Tier-2. The number of T1 and T2 centers shown is outdated, but the data flow schematic remains accurate. Courtesy: [4].

1.2.2 Tier-1

The next tier of computing center, Tier-1 (T1), consists of 13 computer centers worldwide. These computing centers are typically national laboratories in collaborating countries, such as Fermi National Accelerator Laboratory (FNAL) for CMS, and Brookhaven National Laboratory (BNL) for ATLAS. These Tier-1's are connected to each other, and to CERN, through a dedicated high-bandwidth network called the LHC Optical Private Network (LHCOPN). The fiber-optic links of the LHCOPN, which work at 10Gbps, allow the Tier-1's and CERN to transfer large datasets to one another within a reasonable timeframe [5].

Since each of these Tier-1 centers is large enough to store LHC data, one of the main responsibilities of a Tier-1 center is to store a proportional amount—an amount agreed upon in the MoU—of the raw and reconstructed data sent from CERN, as well as a full copy of the AOD datasets. In addition to storing this data, each Tier-1 also securely stores a share of the Monte Carlo (MC) data created by the Tier-2 (T2) sites.

Since each Tier-1 stores a full copy of the AOD datasets, a subset of the raw and reconstructed datasets, and the MC data from Tier-2's, the Tier-1 sites are also responsible for distributing this data to the other associated Tier-2 sites.

In addition to being responsible for providing around-the-clock support for the Grid, Tier-1 centers also provide substantial CPU power for large-scale reprocessing operations. These operations include re-reconstruction, skimming, calibration, and AOD extraction (and the subsequent storage of this data output) [4].

1.2.3 Tier-2

The more numerous, smaller set of computing centers is the Tier-2 (T2), which sits below the Tier-1 centers in the computing hierarchy. Each Tier-2 is associated to a specific Tier-1 site. Currently, there are about 160 Tier-2 sites around the world. Typically, these sites are located at universities, such as the University of Florida, as well as other scientific institutes. Although smaller than the Tier-1's, the Tier-2 sites still have substantial CPU resources, and are able to store a sizeable amount of data.

The main responsibility for Tier2's is to produce Monte Carlo (MC) data [4]. This Monte Carlo data is a reconstruction of simulated events. Once the MC data is produced at a Tier-2, it's then distributed to its associated Tier-1 center, through which other Tier-2 centers can access it. The use of MC data, in conjunction with analysis performed on the

AOD data made available to the Tier-2's by their associated Tier-1 center, allows physicists to perform their research. Most physicists perform their analyses using the resources available at a Tier-2 computing center. However, these physicists have to compete for time and CPU resources with other users wanting to perform their own analyses, as well as the site itself, which has to allocate resources in order to fulfill its requirements, via the MoU, as a member of the WLCG.

1.2.4 Tier-3

The final tier in the hierarchy of LHC computing centers is the Tier-3 (T3). Although there are no strict requirements on the system's run time or data storage capabilities for Tier-3 sites—since there is no formal agreement between Tier-3s and the WLCG via the MoU—most Tier-3s are considered to be much smaller versions of Tier-2 sites, with resources at a few Tier-3s even being somewhat comparable to their Tier-2 counterparts [6]. Tier-3 sites have a wide range of architecture, which can span up to several local clusters. Typically, these Tier-3s are monitored and maintained by several individuals in a university department.

Tier-3 clusters allow physicists, typically those that are geographically close, and a local group of users, to locally host and analyze data from the LHC, and to submit grid-based jobs. Since Tier-3s are independent of the MoU, they are not bound to the same responsibilities as the Tier-1 and Tier-2 centers. This allows Tier-3 users to devote all of the Tier-3 center's CPU resources to their own studies. Ultimately, physicists using a Tier-3 do not face the same competition for time and resources as those at a Tier-2 site. Furthermore, Tier-3 centers provide physicists with more freedom, compared to a Tier-2

site, to test and debug the code they are using, as well as providing physicists the freedom and flexibility to run the analyses that they desire [6].

1.3 Compact Muon Solenoid Experiment

The Compact Muon Solenoid (CMS) experiment is one of the four main experiments that operate at the LHC at CERN [7]. The CMS detector is 21.6 meters long, 15m in diameter, and weighs about 14,000 tons. The CMS collaboration, which consists of over 3,500 scientists, engineers, and students from 193 institutes in 43 countries, built and operates the CMS detector which was created to examine proton-proton (and lead-lead) collisions at a center-of-mass energy of 14 TeV [8]. The CMS detector consists of several subsystems, which are designed to measure the energy and momentum of photons, electrons, muons, and other products of the proton collisions.

At the center of the detector is the interaction point. This is where the proton beams collide with one another. At the design luminosity, the detectors see an event rate of approximately 31.6 million collisions per second [8]. The trigger system reduces this rate to about 100 events per second to be sent for storage.

Surrounding the interaction point is a silicon-based tracker that can be used to reconstruct the paths of high-energy muons, electrons, and hadrons, as well as the tracks resulting from the decay of very short-lived particles, such as the bottom quark. The tracker can also be used to determine the momentum of charged particles by examining the curved path that these particles take, as a result of the presence of the magnetic field, as they pass through the silicone.

The next sub-system, which surrounds the silicon tracker, is the Electromagnetic Calorimeter (ECAL). The ECAL is a calorimeter made of lead tungstate (PbWO_4)

crystals and is designed to measure the energy of particles which interact primarily through the electromagnetic interaction, such as electrons and photons.

The ECAL itself is surrounded by a sampling calorimeter for hadrons called the Hadron Calorimeter (HCAL). The HCAL is composed of layers of brass with plastic scintillators. The main purpose of the HCAL is to measure the energy of hadrons, such as protons, neutrons, pions, and kaons.

Outside of the HCAL resides the large solenoid magnet which produces a 3.8 Tesla magnetic field. The bore of the magnetic coil is large enough to house the calorimetry system as well as the tracker. The magnet is 13m long and 6m in diameter, and is composed of refrigerated, superconducting niobium-titanium (NbTi) cables wound into a coil. The purpose of the magnet is to curve the path of charged particles in order to determine their momentum and charge.

The final components of the CMS detector, interspersed within the return yoke of the magnet, are the muon detectors. There are three muon detectors, the drift tube system (DT), the cathode strip chambers (CSC), and the resistive plate chambers (RPC). Together, these detectors can identify muons and measure their momenta.

The main goals of the CMS experiment are to explore physics at the TeV scale, study the properties of the recently found Higgs boson, look for evidence of physics beyond the Standard Model, such as supersymmetry or extra dimensions, as well as to study aspects of heavy-ion collisions [7]. The CMS experiment, just like the other LHC experiments, utilizes the WLCG's tiered-computing architecture in order to store and analyze data. CMS computing sites contain specific cluster setup and software in order to access CMS data and submit grid-based jobs. The CMS computing sites must allow for

the utilization of validation software, such as Site Availability Monitor (SAM) tests, which are site-specific functionality tests that determines whether a computing center has the CMS software properly installed, and is capable of receiving CMS jobs.

1.4 Rocks Cluster Distribution Software

The tool that was used to deploy Florida International University's (FIU) Tier-3 cluster was the Rocks Cluster Distribution (Rocks) software. Rocks is a Linux distribution intended for the implementation and management of high-performance computing clusters. Modern versions of Rocks are based off of CentOS, with a modified Anaconda installer that simplifies mass installation onto many computers [9]. Rocks was used as the cluster management software for the FIU Tier-3 cluster because of its scalability, its open-source source model, and because of its use among the CMS Tier-3 community, which provides the robust and rich support that is necessary for a viable commercial Linux distribution.

The Rocks software begins by completely installing the Operating System (OS) on a node. This is the basic management tool of the Rocks software [9]. As a result of the fact that the OS can be installed from scratch in a brief amount of time—and since once the installation is started, the process is largely automated—it often becomes faster to simply reinstall all of the nodes to a known configuration, rather than determine if a node was out of synch in the first place. Also, the short period of time for OS installation allows for the possibility of different, application-specific configurations to be installed on select nodes. This approach allows for great scalability, with Rocks clusters ranging from single-digit CPUs to over several thousand CPUs, like GridKa, the largest registered Rocks academic cluster, operated by the Karlsruhe Institute of Technology in Karlsruhe,

Germany [10]. Another important benefit to this approach is that this structure insures that any upgrade to a node will not interfere with actively running jobs.

Clusters may require an assembly of different node types, including compute/worker nodes, frontend nodes, login nodes, and file servers, among many other types. Each of these nodes takes on a specific role which requires a specialized set of software. One of the important aspects of Rocks is that it provides a mechanism through which customized distributions are produced that defines the complete set of software for a particular node. The software for several different node types are already defined in the base Rocks setup, but the option exists for users to define their own node types through editing existing Kickstart files, or creating entirely new ones.

Within a distribution, different node types are defined with a machine-specific Kickstart file, made from a Rocks Kickstart Graph. A Kickstart file is a text-based description of all the software packages and software configuration to be deployed on a node [9]. The Rocks Kickstart Graph is an XML-based tree structure used to define the Kickstart files. By using a graph, Rocks can efficiently define node types without duplicating shared components. Using this structure, many of the hardware differences can be abstracted, which allows the Kickstart process to auto-detect the correct hardware modules to load. A portion of the Rocks Kickstart Graph for Florida International University's CMS Tier-3 computing center is displayed in Figure 1-2.

The FIU Tier-3 computing cluster was deployed with the base rolls of the Rocks 6.1.1 (Sand Boa) software [11], along with several additional rolls which extend the base software, such as the Scientific Linux 6 roll (zfs-linux)—a Linux distribution roll developed by Fermilab and CERN—and the Open Science Grid (OSG) roll, which would

allow the cluster to join the Open Science Grid. The OSG roll that was used for FIU's Tier-3 cluster is OSG roll version 3.2.35 [12]. The OSG roll version 3.2.35 was created by Juan Eduardo Ramirez of the University of Puerto Rico. This roll was created in order to simplify the process of connecting a Tier-3 cluster to the Open Science Grid. A full list of the installed rolls is displayed in Figure 1-3.

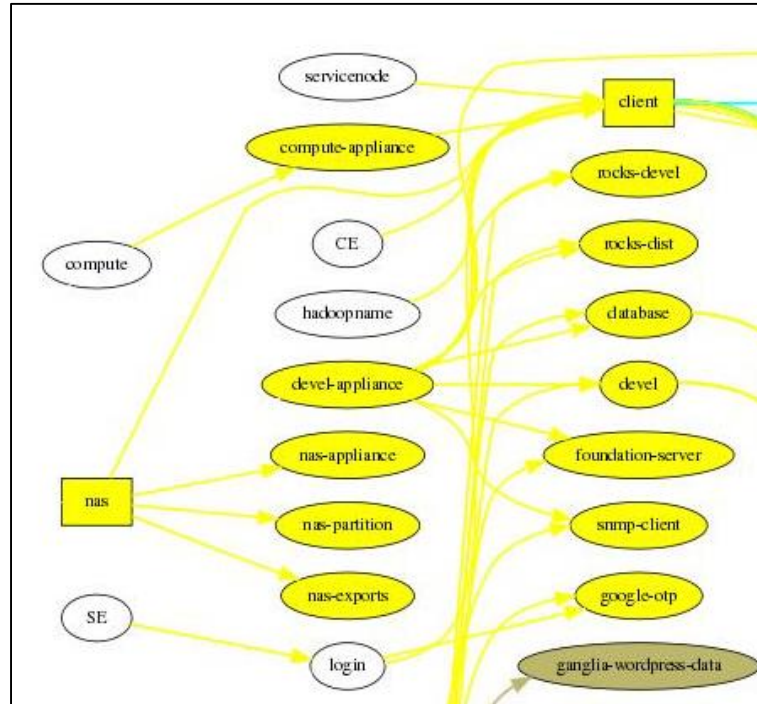











Figure 1-2: A portion of the Rocks Kickstart Graph for FIU's Tier-3 cluster. The white nodes are cluster-specific to FIU's Tier-3 cluster, created and defined through the Rocks Kickstart file. The other colored nodes represent different software located on the different nodes, as depicted by the connecting arrows.

Index of /roll-documentation

Name	Last modified	Size	Description
 Parent Directory			-
 area51/	26-Feb-2016 17:48		-
 base/	26-Feb-2016 17:48		-
 developers-guide/	26-Feb-2016 17:48		-
 ganglia/	26-Feb-2016 17:48		-
 osg/	26-Feb-2016 17:48		-
 perl/	26-Feb-2016 17:48		-
 python/	26-Feb-2016 17:48		-
 zfs-linux/	26-Feb-2016 17:48		-

Apache/2.2.15 (CentOS) Server at rocks6.hep.fiu.edu Port 80

Figure 1-3: List of Rocks rolls installed on FIU's Tier-3 cluster. The OSG and zfs-linux are the two additional rolls. The rest are provided in the base distribution.

Chapter 2 Florida International University's CMS Tier-3 Computing Cluster

2.1 Project Aim

The aim of this project was to deploy a CMS Tier-3 computing cluster with a grid-enabled computing infrastructure on Florida International University's Modesto A. Maidique campus. The cluster is officially referred to as T3_US_FIU, but will often be referred to as FIU's Tier-3 cluster. This cluster consists of 36 machines, with each type of machine performing a different task. The combination of these machines' tasks provides the cluster with grid-enabled functionality. As described in section 1.1, this allows both remote and local users to submit grid-based jobs to the cluster to perform their desired analyses. Additionally, the computing cluster provides local users with many local services including login and storage services.

The components of FIU's Tier-3 cluster that were deployed using Rocks, consists of 27 worker nodes, a frontend node, a computing element, a storage element, a service node, two login nodes, and servers running dedicated file services.

2.2 Frontend Node

The frontend node acts as a sort of central hub for the cluster, and is in charge of distributing the Operating System (OS) to the rest of the machines in the cluster, and also in charge of managing the software packages on the cluster. Additionally, the frontend also runs the global cluster services, such as the Domain Name Service (DNS), web services, cluster distribution, and a cluster-wide database in support of Rocks. Through the frontend, the other machines in the cluster are given their identities and roles. For example, when re-installing a node in the cluster, a PXE boot (Preboot eXecution Environment) can be initiated—wherein the node boots using only its network card—

through the frontend. Once the node is back up, it will have retained its identity in the cluster using information provided by the Rocks cluster management system that combines configuration instructions in XML files and parameterized cluster/node specifications storage in the Rocks MySQL database.

The frontend, which is called `Rocks6.hep.fiu.edu` on the FIU Tier-3 cluster, also allows administrators and users to monitor the cluster's status. For example, using the Ganglia roll provided in the base installation of Rocks, through the frontend, users can monitor the number of nodes and which ones are up and running, as well as monitor the CPU use, load, and recent memory used. Furthermore, Ganglia also allows users to view the job queue on the cluster and user metrics [13].

2.3 Login Nodes

There are also two login nodes in the FIU Tier-3 cluster called `quine.hep.fiu.edu` and `medianoche.hep.fiu.edu`. Each of these nodes is used by a different set of users; `Medianoche.hep.fiu.edu` is reserved for students at FIU that may or may not be involved with the FIU CMS group, while `Quine.hep.fiu.edu` is reserved for CMS users only. These login nodes act as user interfaces to the cluster, allowing users to access and utilize some of the features of the cluster. Most importantly, the login nodes allow for users to submit jobs locally and on the WLCG grid, as well as to check the status of submitted jobs. The login nodes also allow users to complete simple tasks, such as editing files—through the Emacs and vi editors, for example—and accessing fitting tools for plots (such as `Mn_Fit`), as well as the ROOT data analysis framework, a standalone version or one integrated with the CMS software. From the login nodes, users are able to access their

home directory located on the cluster's file servers, which contains all of their files and directories.

2.4 File Servers

The file servers for FIU's Tier-3 cluster, called `fs1.local`, have the primary purpose of providing a location for shared storage of computer files and directories. The FIU Tier-3's file servers operate as a Network File System (NFS), which is a distributed file system protocol that allows a server to share directories and files with clients over a network [14]. With NFS, users and programs can access files on remote systems as if they were stored locally. This is beneficial because it eliminates the need for users to have separate home directories on every network machine, which reduces redundancy. Instead, home directories are set up on the NFS server and are then mounted, on-demand, to machines throughout the network so that no matter what machine a user is accessing on the network, the user is still able to access their home directories. Additionally, NFS allows data that would otherwise be duplicated on each client, to be kept in a single location and accessed by users on the network. This reduces the amount of disk space being used on the local workstations.

The FIU Tier-3's file servers, `fs1.local`, consists of a single file server with sixteen 1 TB SATA hard drives. The process by which `fs1.local` stores data is through a RAID (Redundant Array of Independent Disks) process; namely RAID level 6, or RAID 6. RAID 6 consists of block-level striping, but also consists of double distributed parity [15]. For the case of distributed parity, the parity information is distributed among the drives. This means that, should one drive fail, no data is lost as long as all of the other drives remain functional. Double parity, on the other hand, provides fault tolerance for up

to two failed drives. This RAID 6 setup provides strong data redundancy, and therefore severely limits the occurrence of data loss.

2.5 Computing Element

The CE, along with the login nodes, the service node, and the worker nodes, comprise the computing element of the cluster. Together, these nodes fulfill the tasks that are needed in order to receive and process grid-based jobs.

2.5.1 Gatekeeper Node

The gatekeeper node, or CE, is the cluster node which acts as the grid gatekeeper. In FIU's Tier-3 cluster this node is called `fiupg.hep.fiu.edu`. The CE, which hosts the `globus-gatekeeper` service. The CE acts as the grid interface to the computing/processing elements. The gatekeeper authorizes and executes a grid service on behalf of a grid-user. Once the GUMS service, running on the service node, successfully maps a user, the information is sent to the CE. If the CE authorizes the user, the CE's gatekeeper runs a job manager to be operation which allows that user's job to be monitored and submitted to the cluster's batch processing system. The gatekeeper also provides monitoring of the jobs once they are running.

FIU's Tier-3 cluster utilizes the HTCondor software. This software is a batch processing system that manages the jobs for the computing element [16]. Once the gatekeeper authorizes a job, the `condor_negotiator` daemon running on the CE performs a matchmaking task whereby, after accessing a list of available worker nodes and pending job queues, the authorized job is then matched with the appropriate available worker node(s). This process is also responsible for switching the job to another machine, if necessary, as well as sharing available resources with other jobs in the queue.

2.5.2 Service Node

The service node, named `rocks.hep.fiu.edu` in FIU's Tier-3 cluster, hosts the Grid User Management System (GUMS) service, and the Frontier-squid service. GUMS is a Grid Identity Mapping Service [17], that authorizes and maps users from their global grid-identity to an appropriate local site Unix account and communicates this mapping to the gatekeeper.

In order for a user to submit grid-based jobs, the user must first obtain a grid certificate. These grid certificates are obtained through a Certification Authority (CA), such as OSG or CERN. If a user requests the issuance of a grid certificate and is deemed eligible to receive one by a certain Certification Authority, then that CA will create a Distinguished Name (DN) for the user. This DN is unique to the user and contains certain attributes such as the user's name, organization, city, state, etc.

This identity mapping is necessary when a site does not use grid credentials natively, but instead use a different mechanism to identify users, such as UNIX accounts. For example, if a user tries to submit a job to FIU's Tier-3 cluster, the user's site credentials, which can change depending on the machine being used, the cluster needs to verify that the user has the appropriate credentials (grid certificate) in order to do so. GUMS will then map the user's site credentials to their DN, and then return this mapping to the gatekeeper.

The service node also hosts the Frontier-Squid (Squid) software, which is an HTTP caching proxy software that is optimized for use with applications on the WLCG [18]. The Frontier system stores the conditions database for the CMS experiment. The conditions data is the complete parameters set that describes the CMS detector's

component's run parameters, the status of which change significantly over time. The conditions data also parameterize the status of the CMS software, which also changes significantly over time. Once installed, Squid caches, or stores, HTTP requests and retain copies of the data returned. This improves cluster response times by providing a shorter path between the cached copies and the requesting machine. Squid also reduces bandwidth consumption by caching frequently-requested data, since it allows multiple machines to access the same data, rather than each machine accessing its own copy of the conditions data.

2.5.3 Compute Nodes

The 27 worker nodes that are a part of the FIU Tier-3 cluster function as both the compute nodes and the data nodes for the cluster. For their function in the computing element, these worker nodes will be referred to as compute nodes. For the computing element of the FIU Tier-3 cluster, these compute nodes are the machines that perform the computations that need to be carried out for each job.

Each of the FIU Tier-3 compute nodes has two 1TB drives and is composed of 8 cores. Since, at maximum usage, each core is configured to run one job, a total of 8 jobs can be run simultaneously on each compute node. Therefore, a total of 216 jobs can be run simultaneously on FIU's Tier-3 cluster. For each compute node, the `condor_schedD` daemon is called upon when there is a queue of jobs waiting to be run. Once the compute node is available to begin processing a new job, the `condor_startD` daemon is run, which handles all of the details of starting and managing the job.

Each of these compute nodes has the `gLExec` program, which acts as a light-weight gatekeeper, installed on them [19]. `GLExec` is a tool used to address some security

flaws with the pilot/glidein job infrastructure, which is discussed more in the following section.

2.5.4 Job Flow in the Computing Element

First, when a user submits a job to the grid, the job is sent to a pool central manager node, for example the OSG pool. While the job sits in this pool, the pool central manager node contacts the Glidein Factory¹ which is used to locate available sites. The Glidein Factory then initiates a pilot, or glidein job, which acts as a shell of the user's job. This pilot job is then sent out to all sites on the grid that should be able to run the job. The first pilot job that starts is the one that gets the job [20].

Once these pilot jobs are initiated and sent out, they interact with the available cluster's CE, which hosts the cluster's gatekeeper. The gatekeeper then communicates with GUMS to authorize the execution of the pilot job. Once the pilot job is authorized by the gatekeeper, it is then passed on to the CE's HTCondor batch system. The CE's HTCondor batch system then locates available compute nodes on the cluster and submits the pilot job to these compute nodes. The compute nodes then pull the job from the pool central manager to the compute nodes using HTTP, and then invoking the gLExec tool. The job is then executed by the compute nodes. The pilot jobs will remain on the compute nodes for a short period of time, looking for another job that can be started on the site, before expiring.

¹ Here, the term Glidein Factory is used to denote the glideinWMS VO frontend, the glideinWMS collector node, and the glideinWMS Glidein Factory. The use of the Glidein Factory term in such a manner was made because the specifics of the glideinWMS inner workings are beyond the scope of this project.

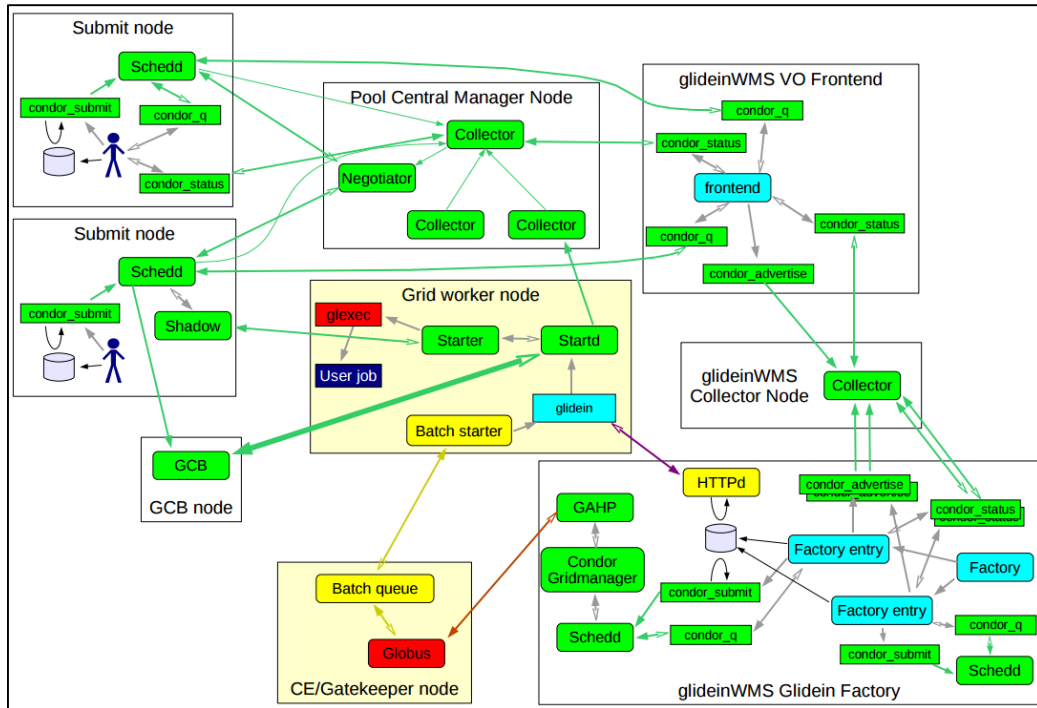


Figure 2-1: Schematic of job flow using glidein jobs. The job is submitted to the Pool Central Manager where it is sent to the Glidein Factory¹. From there, a pilot job is sent out to the CE/Gatekeeper node which passes the job on to the batch system once authorized. Once the worker node accepts the pilot job, the worker node then pulls in and executes the actual job. Courtesy: [21]

2.6 Storage Element

The SRM, in conjunction with the name nodes and the data nodes, make up the storage element of the cluster. These nodes all work together to handle the tasks necessary to transfer and store data from the grid.

2.6.1 Storage Resource Manager

The storage element node, or Storage Resource Manager (SRM), called `srm.hep.fiu.edu` in FIU's Tier-3 cluster, is the cluster node which hosts the Berkeley Storage Manager (BeStMan) software, and the GridFTP (Grid File Transfer Protocol). The function of the SRM, as identified by its namesake, is to manage storage requests

from the grid. The SRM is in charge of space management—allocating storage space and removing unneeded files. The SRM accomplishes these tasks, through the BeStMan software, by utilizing dynamic space reservation [22]. This means that, for example, as a job is being run, the SRM is ensuring that there is enough local disk space remaining in order to prevent a job failure, either by moving data to a location with more space, or by removing garbage files. This dynamic storage helps to avoid a loss of productivity within the cluster. The SRM also helps to unclog temporary storage systems—since the removal of files after they are used is not automated—by providing lifetime management of accessed files.

The SRM node utilizes GridFTP to transfer large data files between FIU’s Tier-3 cluster and other outside hosts. GridFTP is an extension of the universally-accepted FTP, and is used as a reliable, high-performance method of transferring very large files, such as the files associated with CMS data [23]. In order to facilitate the speedy transfer of such large data, the SRM machine in FIU’s Tier-3 has a 10 Gbps connection to the public network.

2.6.2 Name Nodes

Additionally, there are two name nodes in the cluster, with one being the primary name node and the other serving as the secondary name node. On FIU’s Tier-3 cluster, these machines are named h1.local. The secondary name node is co-located at the compute-0-0.local server. The secondary name node stores images of the primary name node at certain checkpoints and is used solely as a backup to restore the functionality of the primary in case of primary name node failure.

The name node acts as the centerpiece of the Hadoop Distributed File System (HDFS) [24]. Hadoop provides a distributed filesystem and a framework for the analysis and transformation of very large data sets [24]. Hadoop controls the data storage for the SE, much in the same way that the HTCondor controls the scheduling and running of jobs for the compute element.

The main role of the name node is to host the file system index. This file system index is essentially file meta data, or information about the file data. When a large data file is to be stored, Hadoop splits the data up into data blocks and stores these blocks across the data nodes. This process involves automated redundancy, since each data block is automatically replicated and stored across the data nodes as well.

For each data file, the name node keeps track of the number of data blocks, and also where in the cluster those data blocks are located. Furthermore, the name node is responsible for maintaining the directory tree of where files are located, as well as carrying out operations like opening, closing, and renaming files and directories. The issuing of commands to the data nodes, such as block creation, deletion, and replication are also handled by the name node.

2.6.3 Data Nodes

As described in section 2.5.3, the 27 worker nodes in FIU's Tier-3 cluster serve as both compute nodes and data nodes. Each worker node is partitioned so that a portion of their space is dedicated for use as a data node. As members of the SE, these data nodes are responsible for serving the read and write requests from the file system's clients. The data nodes work with the storage element to store data blocks, as well as perform the data block creation, deletion, and replication operations [24]. In addition to each 1.7 TB data

node, of which the entirety of one disk, and a portion of the other, are dedicated entirely to storage, there is also a 36 TB storage unit connected to the SRM that acts as a dedicated data node.

2.6.4 Storage Flow in the Storage Element

First, a request for data file transfer and/or storage is sent to the SRM. The SRM utilizes the BeStMan software to allocate enough space on the data nodes to fulfill the storage requirement. Then, through GridFTP, the data file is transferred to the site. Hadoop then splits the data file up into blocks, the blocks are replicated, and then sent to the data nodes. The name node then records the size, name, and location on the data node of each data block. The data nodes then store the data blocks, and perform any read/write function on the data. Then, periodically, the secondary name node will take images of the name node, in case of primary name node failure. This process is outlined in Figure 2-2.

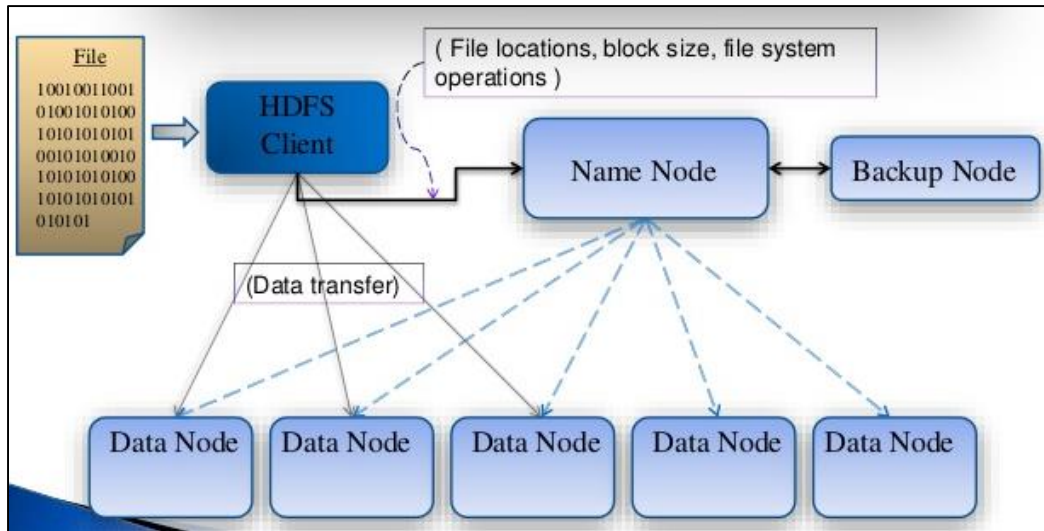


Figure 2-2: Schematic of the Hadoop Distributed File System architecture. The data file is split up into blocks and distributed to the data nodes by Hadoop. The file location and block size are sent to the primary name node, while the secondary name node takes periodic images of the primary name node. Courtesy: [25]

2.7 Validation Software

The Tier-3 cluster at FIU utilizes several different validation software in order to verify that certain resources and services of the cluster are working as intended. Two of the main instances of FIU's Tier-3 validation software are the Resource and Service Validation (RSV) software and the Gratia transfer probe.

Located on the FIU Tier-3 gatekeeper node, `fiupg.hep.fiu.edu`, the RSV software provides OSG sites a scalable and easy-to-maintain resource and service monitoring infrastructure [26]. Through RSV, administrators of the FIU Tier-3 cluster are able to run tests at scheduled intervals and view the results for validation. More importantly, the RSV software allows the OSG to run specific tests, called probes, to pick up resource information and WLCG interoperability information.

The Gratia transfer probe is the other main instance of validation software used by FIU's Tier-3 cluster. Like RSV, Gratia allows probes to be run on the cluster. These probes generate accounting information about the site, mainly file transfers, and allows, for example, the CMS Dashboard to record and store this information [27].

Chapter 3 Results

3.1 Ganglia

Figure 3-1 shows a snapshot of the Ganglia cluster monitoring software provided in the base installation of the Rocks cluster distribution software. This Ganglia snapshot shows a daily report of the load distribution, amount of memory/CPU used, and the network activity. Additionally, Figure 3-2 displays the Ganglia readout of the FIU Tier-3 aggregate load over the span of one hour. Figure 3-2 also lists all of the host machines that are installed on the cluster.

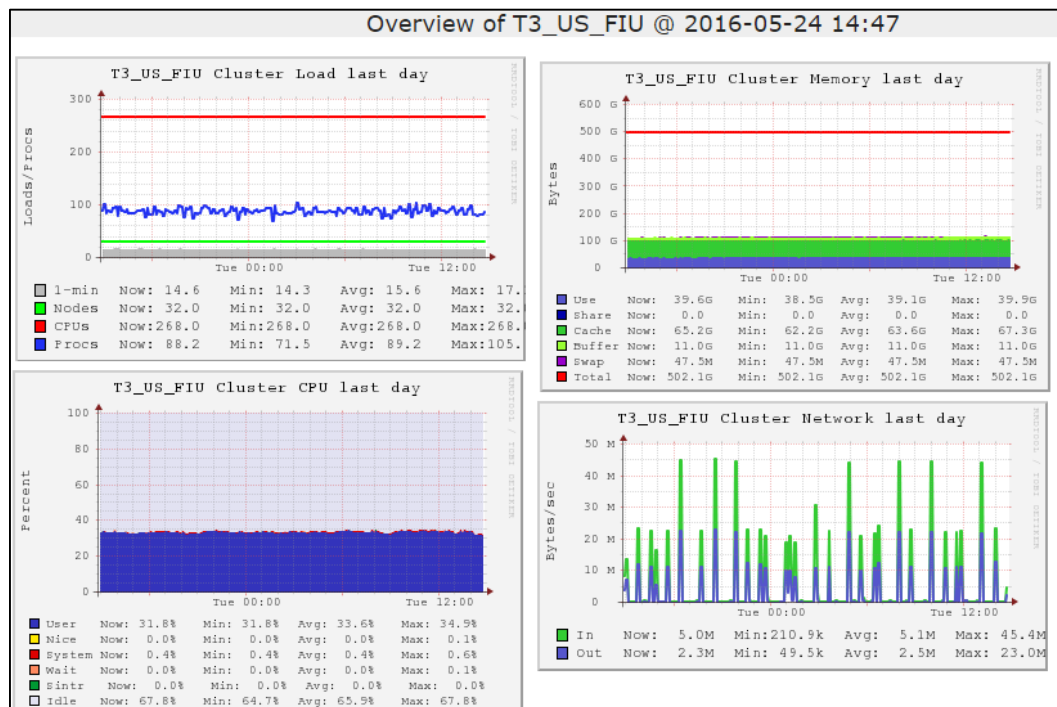


Figure 3-1: Ganglia snapshot of load distribution, memory and CPU used, and network activity for the day of May 24th, 2016

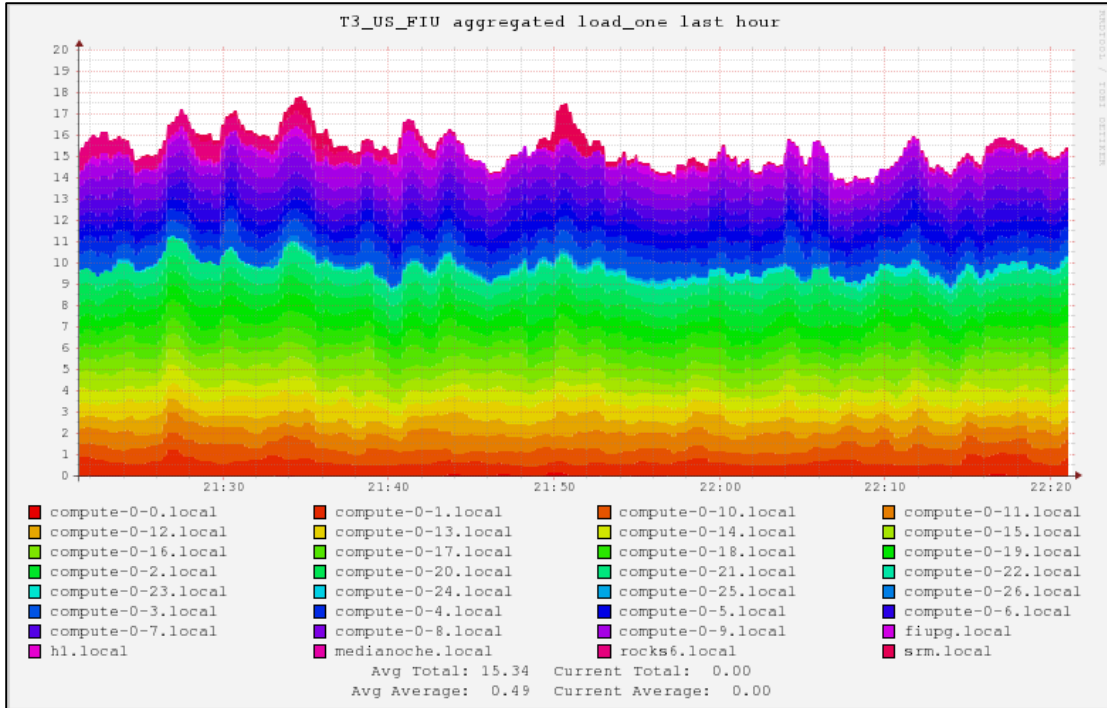


Figure 3-2: Ganglia snapshot of the aggregated load on the FIU Tier-3 cluster over the span of one hour from 21:20 EST to 22:20 EST on May 23rd, 2016.

3.2 MyOSG

Figure 3-3 shows a snapshot of the FIU Tier-3 cluster displayed on the MyOSG website, which presents information about OSG clusters [28]. The proper location of the FIU Tier-3 cluster is shown, as well as the display that both the CE and the SE (srm) are listed as having no issues. Additionally, another snapshot is taken from the MyOSG website and displayed in Figure 3-4. Figure 3-4 shows a list of the MyOSG critical metrics, all of which are listed as having no issues.

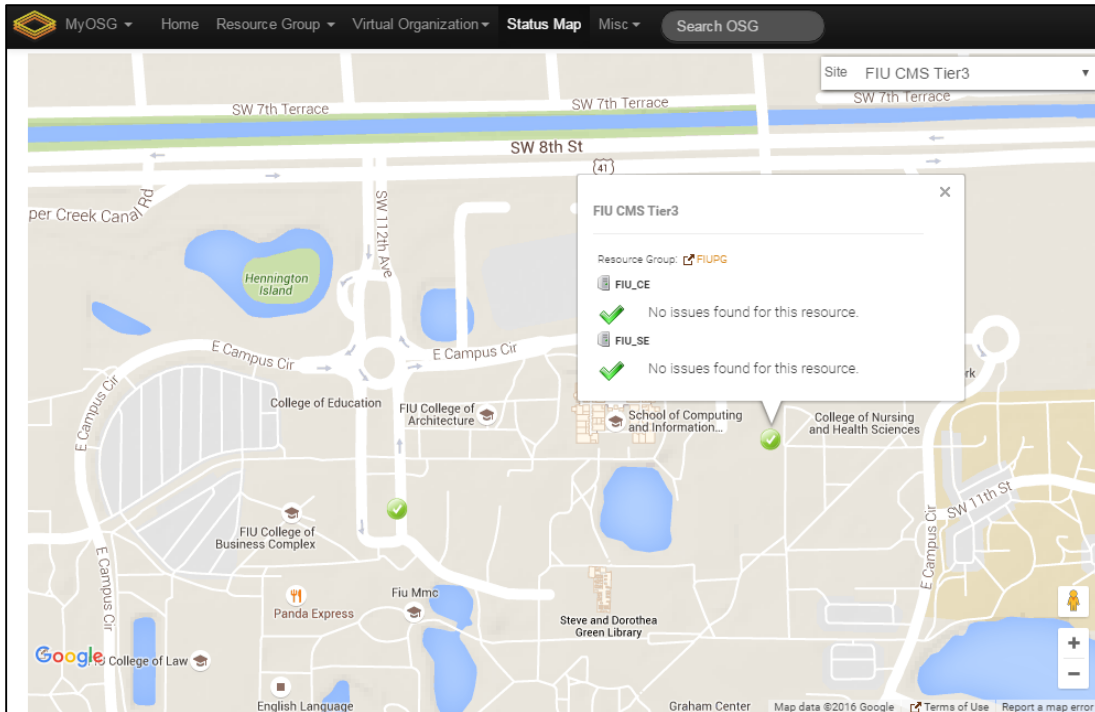


Figure 3-3: MyOSG snapshot displaying the location of FIU's Tier-3 cluster, as well as no issues being found on the CE or SE, which indicates a proper connection to the OSG. Courtesy: [28]

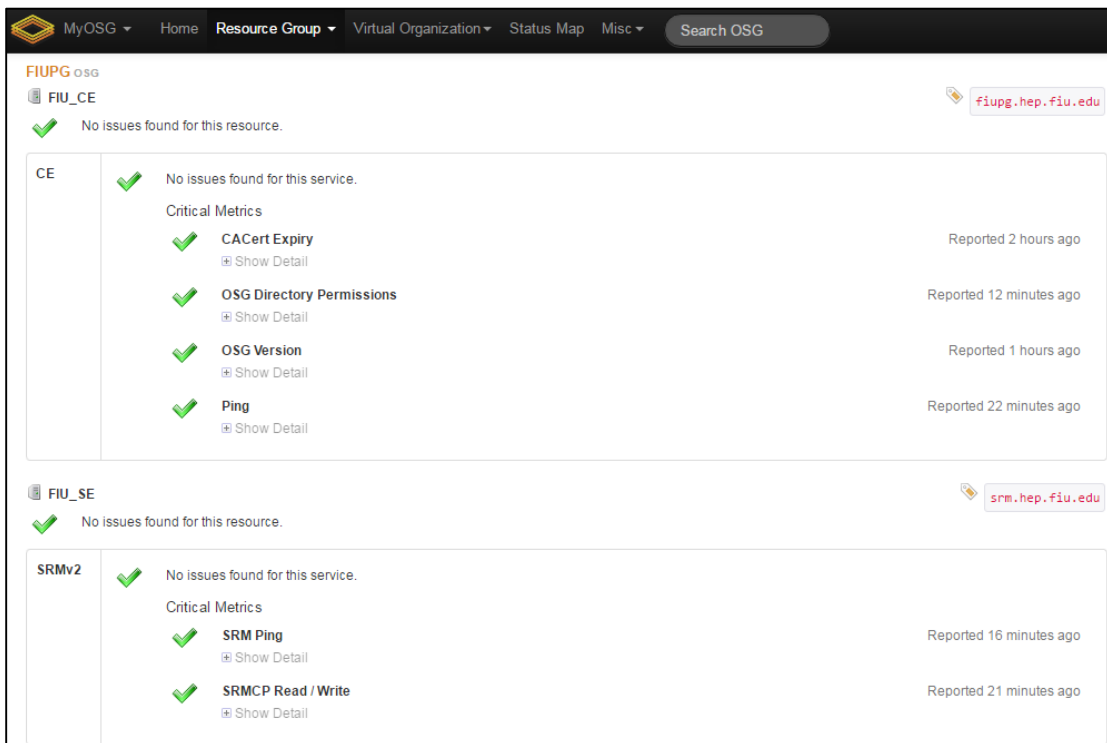


Figure 3-4: MyOSG snapshot listing the critical metrics for the CE and the SE. Having found no issues, FIU's Tier-3 cluster is properly connected to the OSG. Courtesy: [28]

3.3 CMS Dashboard

Figure 3-5 displays the CMS metric readout from the Site Availability Monitor (SAM) tests performed on FIU’s Tier-3 cluster, as displayed on the CMS Dashboard website, which displays information on CMS clusters [29].

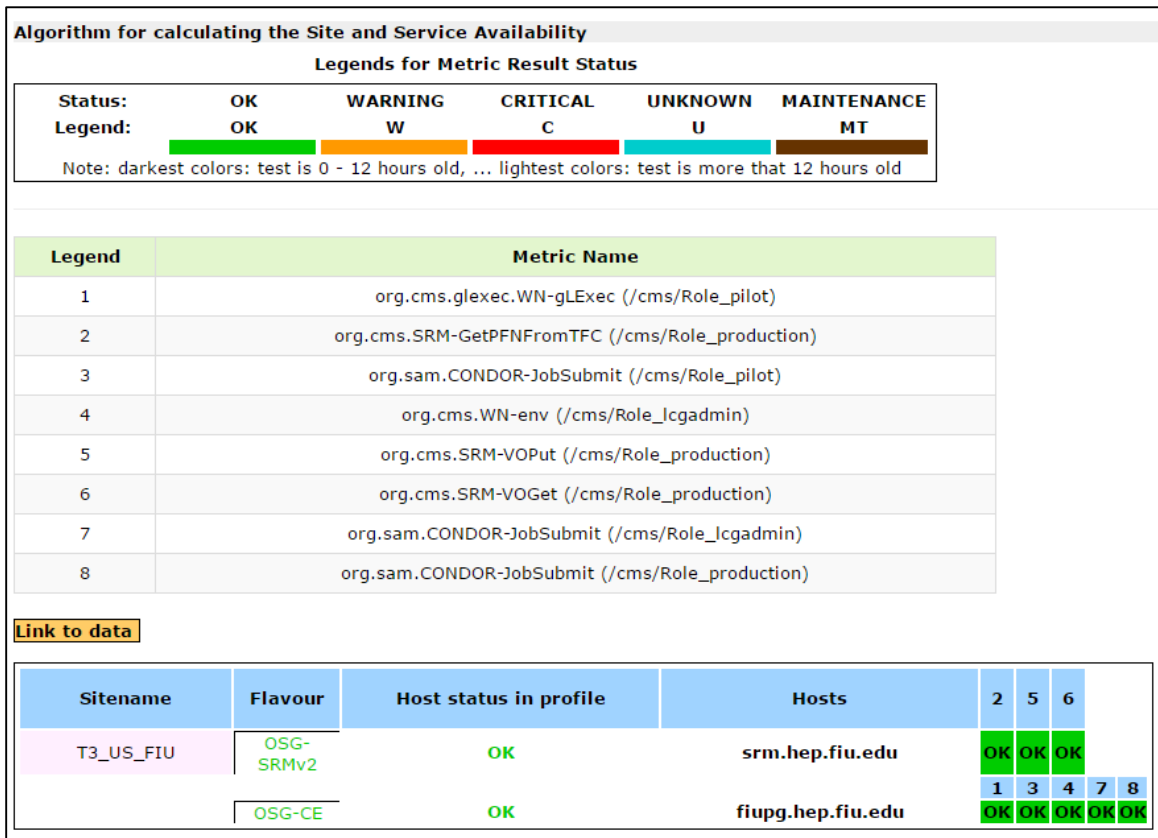


Figure 3-5: Snapshot from the CMS Dashboard website. The metric result status is displayed, as determined by the SAM tests. All metrics are listed as OK, which means the FIU Tier-3 cluster is deemed capable of accepting and running CMS jobs. Courtesy: [29].

Chapter 4 Discussion

Through the use of Rocks 6.1.1, all of the machines were given their identities and their roles in order to form FIU's Tier-3 cluster. The successful implementation of the FIU Tier-3 can be seen in Figure 3-1 and Figure 3-2, both of which were taken from the Ganglia cluster monitoring software. Together, these figures list all of the machines that are running, and showcase their communication and ability to work properly with one another in order to execute jobs.

The RSV and Gratia software, which send out probes for validation, were both used for the validation of the FIU Tier-3's successful connection to the Open Science Grid. As displayed in Figure 3-3 and Figure 3-4, both the computing and the storage resources of FIU's Tier-3 are properly connected to the Open Science Grid, with all of the metrics revealing no issues found. This confirms the grid-enabled status of FIU's Tier-3 cluster.

In order for a cluster to be authorized to receive CMS jobs, the cluster must meet certain CMS metrics, as determined by running Site Availability Monitor (SAM) tests. The results of these SAM tests are displayed in Figure 3-5, and indicate that both the storage and computing resources of FIU's Tier-3 cluster meet the criteria to run CMS jobs.

All of these results together show a successfully deployed CMS Tier-3 grid-enabled computing cluster.

References

- [1] “LHC: The Guide.” *CERN*. CERN, 2009. Web.
- [2] “Update of the Computing Models of the WLCG and the LHC Experiments.” *CERN*. CERN, 15 Apr. 2014.
- [3] “Memorandum of Understanding for Collaboration in the Deployment and Exploitation of the Worldwide LHC Computing Grid.” *CERN*. CERN, 28 Apr. 2015. Web.
- [4] “The CMS Computing Model.” *CERN*. CERN, 15 Dec. 2004.
- [5] “Proposed LHCOPN Operational Model.” *TWiki.CERN*. 24 Nov. 2010. Web. 5 May 2016. <<https://twiki.cern.ch/twiki/bin/view/LHCOPN/OperationalModel>>
- [6] “OSG Tier3 Twiki.” *TWiki.OSG*. 22 Feb. 2012. Web. 5 May 2016. <<https://twiki.opensciencegrid.org/bin/view/Tier3/WebBook>>
- [7] “The CMS experiment at the CERN LHC.” *CERN*. CERN, 14 Aug. 2008.
- [8] “Compact Muon Solenoid.” *CERN*. CERN, 13 May 2011. <<http://cms.web.cern.ch/content/cms-collaboration>>
- [9] “Rocks Cluster Distribution: User’s Guide for Rocks Version 4.3 Edition.” *Rocks*. Rocks, July 2007.
- [10] “GridKa – The German Tier-1 Centre for High Energy Physics Computing” Marten, Holger. 15 Jan. 2004.
- [11] “Rocks 6.1.1 (Sand Boa).” *Rocks*. 14 Apr. 2014 <http://www.rocksclusters.org/wordpress/?page_id=477>
- [12] “OSG Roll: User’s Guide version 3.3.25 Edition.” OSG, 15 Feb. 2016 <<http://rocks6.hep.fiu.edu/roll-documentation/osg/3.2.35>>
- [13] “Ganglia: User’s Guide version 5.0 Edition.” University of California, Apr. 2008.
- [14] “Network File System (NFS).” Ubuntu 16.04 Server Guide.
- [15] “RAID Basics.” Sun Microsystems. 2009. <https://docs.oracle.com/cd/E19168-01/817-3337-18/appa RAID_basic.html>
- [16] “HTCondor-CE Overview.” *TWiki.OSG*. 10 May 2016. Web. 12 May 2016.

- [17] “GUMS Install Guide.” *TWiki.OSG*. 10 Nov. 2015. Web. 7 May 2016.
- [18] “Frontier Squid Caching Proxy Installation Guide.” *TWiki.OSG*. 4 Dec. 2014. Web. 7 May 2016.
- [19] “Glexec Installation Guide.” *TWiki.OSG*. 14 Apr. 2016. Web. 6 May 2016.
- [20] “Job Submission Comparison.” *TWiki.OSG*. 6 May 2016. Web. 8 May 2016.
- [21] “Glidein WMS.” USCMS. Web. 7 May 2016.
<http://www.uscms.org/SoftwareComputing/Grid/WMS/glideinWMS/doc.v2_5/factory/design_glidein.html>
- [22] “Berkely Storage Manager (BeStMan).” Sim, Alex. 6 Oct. 2009.
- [23] “Overview of GridFTP in the Open Science Grid.” *TWiki.OSG*. 15 Feb. 2012. Web. 8 May 2016.
<<https://twiki.opensciencegrid.org/bin/view/Documentation/StorageGridFTP>>
- [24] “HDFS Architecture Guide.” The Apache Software Foundation. 4 Aug. 2013. Web.
<https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html>
- [25] “Hadoop Distributed File System.” Kulkarni, Anand L. 28 Aug, 2015.
- [26] “Installing, Configuring, Using, and Troubleshooting RSV.” *TWiki.OSG*. 13 May 2016. Web. 17 May 2016.
- [27] “Gratia Transfer Probe.” *TWiki.OSG*. 14 Oct. 2014. Web. 18 May 2016.
- [28] “MyOSG.” The Trustees of Indiana University. 2013. Web.
<<https://myosg.grid.iu.edu/about>>
- [29] “CMS Dashboard-SAM Visualizations.” CMS. 2016. Web.
<<http://wlcg-sam-cms.cern.ch/templates/ember/#/>>