



## Open Science Grid

Document Name	US LHC Baseline Services on OSG
Version	8
Date last updated	June 3rd, 2005
Authors	Rob Gardner, Frank Würthwein

**Abstract.** We summarize OSG requirements in support of a baseline set of services necessary for the US LHC experiments. We focus on deliverables for OSG 0.4 in late 2005, and identify areas where ATLAS and CMS expect to provide effort.

---

1	Introduction .....	4
1.1	Open Science Grid Philosophy and Implications to US LHC .....	4
1.2	US LHC Context.....	4
1.3	VO Specific “Edge Services” .....	5
1.4	Carrying Lessons Forward .....	6
1.5	Document Overview .....	6
2	Core Services that carry over from Grid3.....	6
3	Enhancements in OSG 0.2 infrastructure .....	7
4	Major Deficiencies and Concerns .....	9

4.1 Scalability and Robustness of the CE..... 9

4.2 No SRM based storage outside CMS T1,T2 and BNL..... 10

4.3 VO Policy and Authorization Infrastructure Lacking..... 10

    4.3.1 Authorization ..... 10

    4.3.2 VOMS Support Model..... 11

4.4 Accounting, Monitoring, and Information Services ..... 11

5 New Services in 2005 ..... 11

    5.1 Dataset Placement and Access ..... 11

    5.2 Late binding of Resources on the Compute Node ..... 12

    5.3 Dynamically deployed Edge Services..... 14

    5.4 Interoperability..... 15

6 Metrics and Goals ..... 16

7 Summary of OSG 0.4 Goals ..... 17

8 Services Beyond OSG 0.4 ..... 19

9 Performance Goals beyond OSG 0.4 ..... 20

V8	6/2/05	Added in info from apps. Meeting	FW
V6	5/31/05	Expanded sections 4, 5; fig.	RG
V5	5/25/05	Expanded sections 3,4,5 some	FW
V4	5/19/05	Post Workshop additions	RG
V3	5/2/05	Clean up	RG
V2	5/1/05	Add ATLAS text	RG
V1	4/29/05	Initial draft	FW



## 1 Introduction

The purpose of this document is to revisit and elaborate the set of baseline core services required by the US LHC program for running applications on OSG, and to provide these as recommendations and input to the OSG Technical Roadmap, being devised in the context of the OSG Architecture and Blueprint Activity. We focus on deliverables for OSG 0.4 in late 2005, and provide some vague outlook towards OSG 0.6 release and beyond.

### ***1.1 Open Science Grid Philosophy and Implications to US LHC***

The OSG is not a project with resources, central management, and a resource loaded schedule. It is a Consortium of Organizations that have come together to build a persistent national computing infrastructure on the premise that the collective is more than the sum of its pieces.

The “Open” in OSG implies a broad enough community of participants to make it unlikely that an infrastructure can ever be built via a centrally organized design, development, and deployment process that will lead to implementations of all services required by all participating communities. Instead, we have chosen a process that is based on principles like heterogeneity, incremental development, a strong focus on phased deployment from development to integration to production, and a focus on VO based, rather than OSG based service instantiations on top of a minimalist core OSG infrastructure.

Accordingly, OSG is committed to the principle that “VOs that require services beyond the baseline set should not encounter unnecessary deployment barriers for the same”, as expressed in the OSG blueprint in this fashion. We expect to rely on this basic OSG promise as we develop, integrate, and deploy the US LHC Baseline Services on OSG.

The Operations Model of the OSG also relies on a relatively thin central support and coordination structure provided by iGOC, plus significant support via participating VOs, called VO Support Centers. This model distributes the expected large workload for problem debugging in a hierarchical fashion, with first-lines of support coming in most cases from application domain specific expertise residing within the VOs.

### ***1.2 US LHC Context***

We describe here the driving factors for both CMS and ATLAS for the remainder of 2005.

For ATLAS the main issue for the remainder of 2005 is to develop the next generation production system which meets requirements for scalability and fault tolerance derived from lessons learned during the DC2 and Rome production exercises. This involves both the core infrastructure and other lower level services on which it depends. ATLAS would like to use OSG infrastructure and services where possible for reliable, managed access to storage elements, role based authorization for both storage and compute resources, and resource accounting/monitoring infrastructure. Three other key areas will be pursued by ATLAS during 2005: one, participation in LCG Service Challenge 3 which will focus on both throughput and service phases; second is the integration and deployment of a new ATLAS distributed data management system which may require deployment of site specific services and catalogs, and finally, the deployment of a distributed analysis service.

For CMS the driving factors revolve around enabling user data analysis on the CMS Tier-2 centers via the OSG infrastructure. Initially, CMS will limit its focus to its own sites for these services. However, from the CMS perspective it would be extremely beneficial if we could work out a subset of services, and a transition path that leads to CMS user data analysis first on ATLAS sites, and later more widely on OSG, via the OSG infrastructure. To accomplish its goals, CMS will need to overcome the same hurdles as ATLAS in terms of reliability, data movement and data access.

Both Atlas and CMS would like to benefit from common LHC service implementations on OSG as far as convenient, and deploy their own where this leads to successful operations more quickly. Over time we expect differences to decrease rather than increase.

### **1.3 VO Specific “Edge Services”**

Given the relative immaturity and fragility of many of the core middleware services in use, we envision a scenario in which resources will be partitioned into specific, VO-dedicated servers along side shared, open grid services used by many VOs in opportunistic fashion. A related issue is deployment of “foreign” VO-specific services on sites hosting common OSG infrastructure. This could be accomplished through scheduling of a gatekeeper-like resource via a grid job. OSG refers to such services as “dynamically deployed edge services”. We strongly encourage OSG to develop a framework that allows for such services.

### **1.4 Carrying Lessons Forward**

The Grid3 infrastructure that OSG extends has been exceedingly successful in satisfying the Monte Carlo production requirements for the US LHC program so far. However, to get ready for LHC data taking a number of major issues need to be addressed. Most of these issues revolve around providing a substrate of protected, robust, and fault tolerant services that will enable both user data analysis and production services on the OSG infrastructure. These reliability concerns result from extensive experience with GT2.x and GT3.x versions of GRAM and GridFTP during major production exercises in the last years. Site/facility architectural issues need to be revisited to address the Grid3 head node overloading problem, as has been well documented.

### **1.5 Document Overview**

For the remainder of this document we first describe the core services that carry over from Grid3, followed by a brief description of the new services, deployed as part of OSG this spring and summer. This is followed by a brief description of the major deficiencies that need to be addressed. We conclude with an outlook towards OSG 0.4 and beyond. We focus here on deliverables for OSG 0.4 in late 2005, and are significantly more vague about OSG 0.6 and beyond.

## **2 Core Services that carry over from Grid3**

The core services that carry over from Grid3 in the first deployment are listed briefly here.

- Basic grid interfaces to computing resources: Globus GRAM for access to local job schedulers and GridFTP for access to local storage.
- A group-based VO security infrastructure based on VOMS for authentication and access. Authorization (eg., priority or weighting factors configured for local schedulers) is left to the site administrator.
- Monitoring infrastructure which capture resource metrics (jobs executing, queued) at the VO level, and a site-level validation service which monitors and reports the health of the site periodically.
- An information catalog providing a summary of site attributes needed for client applications (the Grid3 \$DATA, \$APP, \$TMP, \$WN\_TMP variables).
- A shared file system and convention for it's use (group writeable data areas for temporary job directories and installed releases for applications).

- An operations infrastructure based on a combination of VO and iGOC support and coordination services.

### 3 Enhancements in OSG 0.2 infrastructure

As of writing, the current set of services deployed on the OSG Integration Testbed (ITB release 0.1.6) are the provisioning candidate for OSG 0.21. These described in the ITB twiki<sup>1</sup>. The main additions to the Grid3-type services are discussed briefly here. A large number of these are not yet in OSG 0.2.1 but are expected to appear within some 0.2.x release at least at some sites.

Privilege Project Components:<sup>2</sup>



- Use Privilege Project deliverables for management of mappings between grid users and local Unix accounts, and role-based authorization.
- Be able to create large pools of accounts that are meant for Jane Doe users with limited privileges. E.g., read but no write access to NFS spaces like \$APP, ...
- Be able to create a few (e.g. cmssoft, cmsprod, cmsdata and equivalent for Atlas) accounts that have special privileges, e.g. write access to NFS spaces like \$APP, ... . This will be used only by administrative users to install software that stays on the site persistent for all Jane Doe's.

Modifications to the SE

- SRM interfaces at those sites that have it
- Be able to do space reservation via SRM, at least for staging of job related input and output. By 2006 we'd like to be able to stage whole "blocks of files" (1-5 TB) into sites opportunistically for user data analysis. At present, it is not clear what level of space reservation will be available before OSG 0.4.
- Be able to put/get files from outside via SRM
- Be able to read files from inside via Posix-like interface.
- Be able to write files from inside via SRM.

---

<sup>1</sup> <http://osg.ivdgl.org/twiki/bin/view/Integration/ItbRel016>

<sup>2</sup> <http://osg-docdb.opensciencegrid.org/cgi-bin/ShowDocument?docid=134&version=0>

- SE equivalent of Privilege Project deliverables. This is not yet available in OSG 0.21 but is expected before the OSG 0.4 release.
- We expect that the majority of sites that do not provide SRM interfaces will nevertheless improve their CE/SE reliability by decoupling the CE and SE hardware. E.g. sites may deploy \$DATA plus gridFTP server on hardware separate from the CE such that the CE is not affected by SE overloads. In addition, we encourage sites to provide installation access to \$APP via a dedicated job queue in order to avoid CE overloads due to excessive fork queue usage.

#### Discovery Service

- Register ATLAS/CMS code versions that are installed, and how they are accessed from the compute nodes. Code versions are advertised via both the native discovery service mechanism as well as Generic Information Providers (GIP) 1.2. The latter is important for LCG interoperability.
- JobMon = a service that provides read-only access to CMS/Atlas sandbox environment for the user. It provides the grid equivalent of ls, ps, top, tail, etc.
- Applications need to be able to register reasonably simple exit messages via a sufficiently reliable mechanism at the site. These messages are then discoverable by the VO specific job bookkeeping and monitoring systems in an asynchronous fashion, implying that the messages persist in the discovery service for some reasonable amount of time. The discovery service thus functions like a buffered proxy.

It is likely that ATLAS and CMS will require separate service instantiations within the same Clarendon web services container. We are in the process of forming an OSG activity to co-ordinate integration and deployment of these services. We expect contributions from ATLAS, CMS, Globus, Griphyn, and PPDG.

Not all of these services are available in the OSG version presently deployed. Instead, we expect to role out many of these services over the next few months as incremental enhancements to the existing OSG core services. By doing so we will benefit greatly from the OSG commitment to heterogeneity and the well defined process for moving new releases from development through integration to deployment. In general, not all services will be available at all places.

## 4 Major Deficiencies and Concerns

### 4.1 Scalability and Robustness of the CE

Example incident that is completely unacceptable on many levels:

On Friday, April 29th, at 7:54am Soul Youssef sent an email out to [grid3-all@ivdgl.org](mailto:grid3-all@ivdgl.org) with the following content:

*“Would anyone submitting jobs to atlas.bu.edu in the ivdgl VO please stop now? We are under high load and may have to loose ~100x24 hours of batch production.”*

This kind of incidence is sufficiently common on Grid3 to make up a significant portion of the 23% job failure rate observed by Atlas during the Rome production exercise. The root cause of this kind of failure may be summarized as follows:

- CE and SE share the same piece of hardware. Overloads of either one cause trouble for both.
- Neither CE nor SE have any native mechanisms that would prevent resource overloads.
- Both CE and SE are shared resources across multiple VOs. Given that the VOs operate completely independent of each other on these shared resources, either one careless user or fluctuations in use can cause resource overloads on either CE or SE, and thus affect all VOs operating at the site.
- For the existing CE each pending job consumes significant CE resources. CE performance thus doesn't scale with either pending or running jobs, exacerbating the previous three problems. E.g. an OSG 0.2 CE on a 3.2GHz CPU with 2GB RAM is generally able to support no more than 150-200 jobs.

To address this, we expect four major changes for OSG 0.4:

- Replacing the traditional SE with an SE that implements traffic shaping and load balancing. In practice, we expect this to be implemented by deploying SEs with an SRM interface at either all OSG sites, or at least all CMS and ATLAS sites. As a side-effect, this will decouple SE and CE, as the SE will no longer reside on the same hardware as the CE.

- Upgrading the Globus gatekeeper to GT4. It is our understanding that the GT4 gatekeeper consumes (almost) no system resources per pending job because state is kept in files on disk rather than processes in memory.
- Deploy Condor-C such that the VO operates its own CE. This will help with scalability and make isolation of VOs from each other on the CE possible.
- Explore deployment of an “edge services framework” that possibly is based on virtual machine technology. In combination with Condor-C, this would allow us to isolate different VOs in different virtual machines on the same physical machine. We hope to use this to provide service guarantees on shared hardware resources at the site boundaries. This is discussed in more detail in Section 5.3 below.

#### **4.2 No SRM based storage outside CMS T1,T2 and BNL**

OSG 0.2 includes SRM based SE in form of SRM/dCache only. A support model for this implementation exists only for CMS Tier-2 centers as well as the Atlas and CMS Tier-1 centers at BNL and FNAL. At the very least we need an SRM implementation with a support model that includes Atlas Tier-2 centers, and preferably all sites on OSG. This is a crucial extension of OSG 0.2 because it has serious implications on SE and CE reliability as mentioned in Section 4.1.

#### **4.3 VO Policy and Authorization Infrastructure Lacking**

##### **4.3.1 Authorization**

Grid3 had very little in the way of an authorization infrastructure. Priorities for users were set using group accounts, where each member of the VO having identical priority for job queues as others in the VO. One needs, for example,

- Access to “fast” queues for data analysis requiring quick turn around.
- Priority access for software managers and debuggers of applications and middleware services.
- Other roles: for production managers and possibly specific physics working groups.
- General opportunistic use of resources within a VO.

There was no infrastructure to publish the settings for a given site’s policy for CPU usage.

There was no infrastructure at all for space allocation between users, nor policies in place for usage (expiration dates for cleanups).

#### **4.3.2 VOMS Support Model**

OSG security architecture relies on VOMS, an LCG product for which we need to understand the support model for use on OSG. Related support issues concern other components in the privilege model (GUMS, PRIMA, gPlasma, etc).

#### **4.4 Accounting, Monitoring, and Information Services**

As the OSG infrastructure is used ever more heavily by an expanding community, accounting and MIS become increasingly important. The LHC experiments continue to refine their requirements for accounting and have produced documents outlining minimal metrics. We encourage OSG accounting efforts to be closely aligned with developments in Europe in this area. This is particularly important as we expect to schedule ATLAS and CMS jobs globally. Appropriate and coordinated accounting is a political as well as technical requirement for success.

We observe that there is significant confusion today as to the requirements for these systems. We encourage OSG to start detailed discussions on requirements in this area as part of the OSG Blueprint activity to inform a program of work towards OSG 0.4. In addition, we encourage an effort to sort through the existing OSG 0.2 services in order to arrive at workable solutions for OSG 0.2 that require minimal effort.

## **5 New Services in 2005**

### **5.1 Dataset Placement and Access**

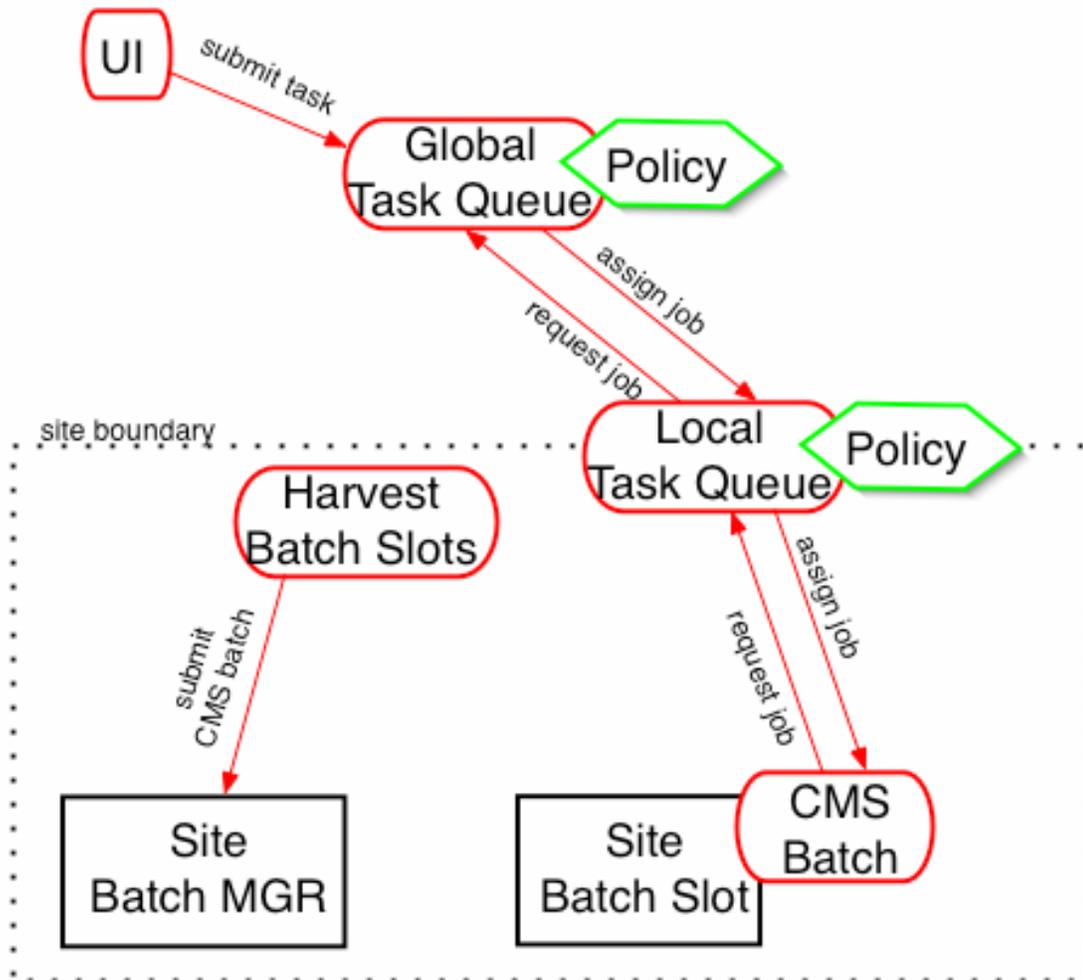
The following new services from CMS and ATLAS are moving towards using OSG for data analysis.

- a) Both ATLAS and CMS are developing or have developed new services for managed file transfers and dataset placement. CMS has developed the PhEDEx service for dataset placement into managed storage. The new ATLAS distributed data management service will be a follow-on to Don Quixote (DQ2). DQ2 will rely on an underlying transfer service, and is evaluating FTS from LCG.

- b) CMS local file catalogue (Pool based mysql); for ATLAS, a site catalog with a POOL file catalog interface is required (LFC and Globus RLS will be evaluated).

We expect that both ATLAS and CMS will focus on their own sites in 2005 in order to find solutions to their data analysis challenge in a well controlled environment. We then expect both experiments to deploy their lessons learned across a wider set of sites in OSG during 2006. We expect the “Edge Services framework” described in Section 5.3 to be crucial in expanding data transfer and access services for the LHC experiments across OSG.

## ***5.2 Late binding of Resources on the Compute Node***



**Figure 1** schematically describes the “late binding” architecture.

This is the “pull model” described in the blueprint document. Ideally, ATLAS and CMS would like to see the following characteristics:

- a) A global task queue that is global in the sense that resources in both LCG and OSG are submitted to at this level.
- b) A local task queue, batch slot harvesting, and VO specific batch infrastructure on the worker node (including its security architecture) that may be different between grids, and maybe even sites within a grid.
- c) Policy control points that allow the VO to configure their own resource utilization policies in a dynamic way, i.e. on a day by day basis according to experiment priorities.

- d) Harvesting of batch slots for the VO via a “glide-in” manager that is VO specific. This may not be possible at all OSG sites due to a sites security infrastructure. However, we expect that at least ATLAS and CMS sites within OSG will allow this functionality.

We expect to spend significant effort on developing a prototype infrastructure as a deployment candidate in time for OSG 0.4. This effort is coordinated with the EGEE WMS team in order to meet the interoperability requirements implied in a). It is an integral part of our strategy for overcoming the scalability and robustness challenges described in Section 4.1.

### **5.3 Dynamically deployed Edge Services**

The majority of OSG clusters deploy their compute nodes on private networks. Only CE and SE straddle the public/private network boundary. It would provide a lot of flexibility to the VOs if gateway resources could be scheduled in such a way that resources like CPU, disk, port ranges, etc. are guaranteed. The Privilege project infrastructure could then be used to implement roles that enable dynamic installation and upgrade of VO specific services. Operationally this may be done initially via multi-month lease of a piece of hardware to ATLAS at a CMS site, and vice versa. In the future, such leases would be much more dynamic, and accessible via grid access services to many VOs in OSG.

We propose formation of an “edge services activity” in OSG with the goal of developing an edge services framework based on virtual machine technology that will allow dynamic deployment of VO specific virtual machines as VO specific edge servers. Initial example services include Condor-C deployment at the site, including batch slot harvesting, as well as FroNtier DB cache based on squid technology. The layout of these services with those deployed in OSG are sketched in Figure 1. Future examples might include ATLAS and CMS specific data transfer agents, among others.

Timeline for this activity should be to have an initial set of services ready for deployment with OSG 0.4 in late 2005. We believe however that preliminary milestones along the way, demonstrating feasibility and exposing difficult aspects of the approach, should be identified and pursued early on.

At any rate, we expect both US ATLAS and US CMS will require partitioning of VO-dedicated CE and SE resources from shared, common OSG gateways until such services mature and become reliable.

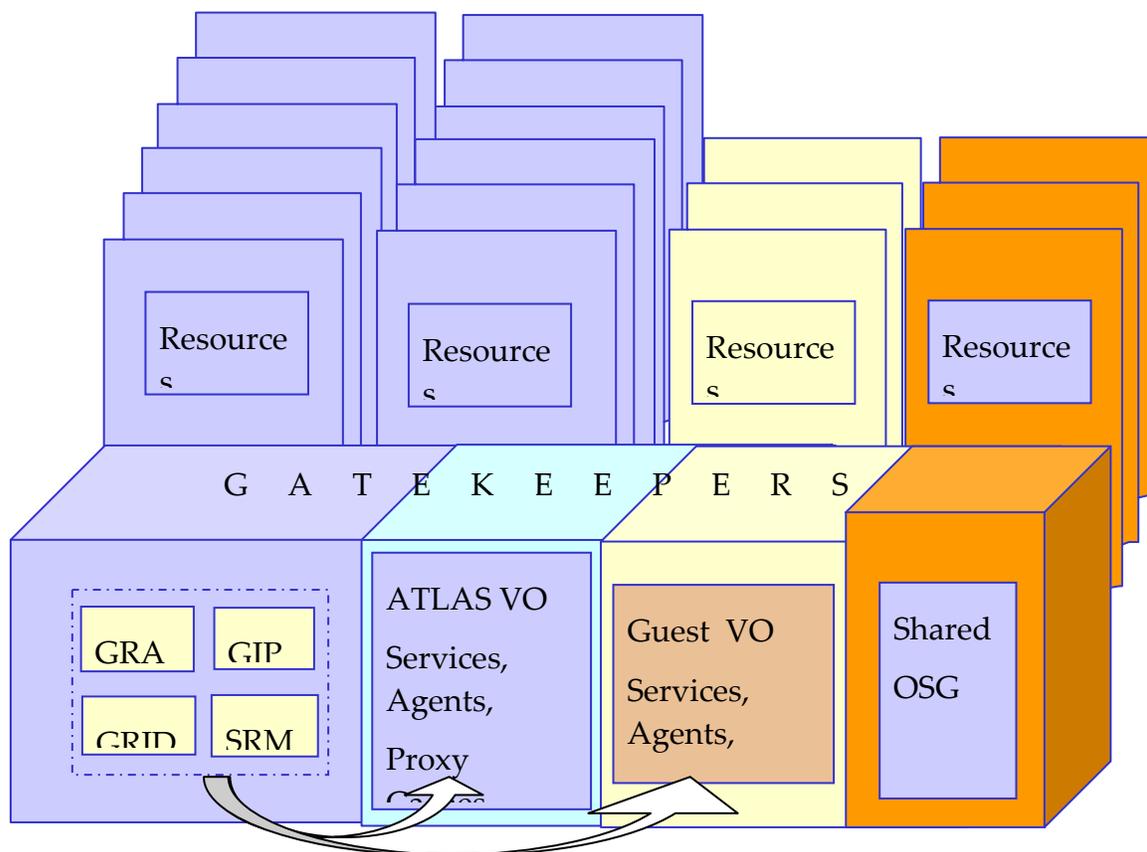


Figure 1 Site Architecture illustrating deployment of "Edge Services" for persistent, VO specific services and agents.

#### 5.4 Interoperability

ATLAS and CMS are interested in interoperability with both TeraGrid and resources deployed by the LCG project.

##### TeraGrid:

We expect development of OSG portal infrastructure to TeraGrid cluster resources. The TeraGrid as a whole will appear via such portals as one or more OSG sites. There shall be no additional requirements made on the ATLAS and CMS production job in order to run on the TeraGrid.

Initially, we expect to utilize only CPU power on the TeraGrid, with modest data import, e.g. a pile-up sample for CMS digitization (~50GB), and modest output, e.g. the output from a full simulation and reconstruction chain. At a later state,

we expect to be able to dynamically stage data into TeraGrid, using a standard OSG SE, i.e. an SRM interface, and then run both user analysis as well as data reconstruction applications.

We expect the initial goal to be met within the context of OSG 0.2. User accounts, authorization, software installation into common (community) areas on TG sites will need to be understood.

Given the well known scalability and robustness deficiencies of the OSG 0.2 CE/SE infrastructure it is obvious that large scale use of TeraGrid can only be achieved via a single OSG portal after these deficiencies are overcome.

### **Resources deployed by the LCG project:**

A standard OSG 0.2 CE includes GIP 1.2, and can thus present itself as an LCG-2 compatible site to the LCG-2 Resource Broker. We expect that this interoperability of LCG and OSG CE's will be maintained as LCG transitions to a CE based on gLite. Ideally, we'd like to see such interoperability accomplished in a number of ways:

- By having a Condor-C schedd at an OSG site receive jobs from a Condor-C schedd deployed as part of a gLite resource broker (RB). We are told that the gLite RB is compatible with GIP 1.2. Verification of this claim should be addressed soon after OSG 0.2 deployment within the interoperability activity.
- By publishing the minimal information about a site's attributes to a Condor ClassAd mechanism, making the site visible via Condor-G to LCG-2 executor (in the case of ATLAS).

## **6 Metrics and Goals**

At the OSG Applications workshop at SLAC on June 1-2nd we identified the following metrics and goals for OSG 0.2.

Metrics:

- (a) Efficiency of job execution. ATLAS and CMS commit themselves to determine the rate at which jobs submitted to OSG complete successfully.

We will share this information publicly at the granularity of site as well as all of OSG. Completing successfully shall be defined to exclude application failures. E.g. jobs that fail due to application core dumps, or central DB overloads are not counted in either nominator or numerator of the efficiency.

- (b) ATLAS and CMS commit themselves to provide information about the total amount of data produced, CPU cycles consumed, and data ingested by the applications on OSG.
- (c) ATLAS and CMS commit themselves to provide information about the total number of ATLAS and CMS users using OSG.

Goals:

- (a) ATLAS and CMS strive towards exceeding 90% efficiency as defined above for at least some sites on OSG within the 0.2 release series.
- (b) ATLAS and CMS strive towards an operational effort of less than 1 FTE for their production simulation efforts on OSG.

## 7 Summary of OSG 0.4 Goals

Throughout the document we have described a variety of services that we expect as part of the LHC baseline roadmap, and aren't available yet in OSG 0.2. In the present section we list those that we consider goals for OSG 0.4. We assume here that a first release candidate for OSG 0.4 will be available in November 2005, and ATLAS and CMS sites will have transitioned to an 0.4 production release by December 15th 2005.

- (a) Edge Services Framework. We expect that ATLAS and CMS will use the new Edge Services Framework to deploy at least one service each at each others sites.
- (b) Compute Element Enhancements:
  - a. We expect the web services version of GRAM in GT4 to be the default GRAM. We expect this to be transparent to all applications that presently use condor-G. This is a reliability and robustness goal.
  - b. We expect a quota policy for \$APP, as well as a deployment mechanism into \$APP that does not involve the fork queue. Both of these are reliability and robustness goals.

- c. We expect to support late binding WMS based on condor-c at ATLAS and CMS sites. This is mostly a functionality goal.

(c) Workload Management System

- a. We expect to deploy a workload management system that uses late binding of resources at ATLAS and CMS sites. We expect to do this in such a fashion that interoperability with LCG is preserved, assuming LCG deploys the gLite WMS on this time scale.
- b. We expect to deploy a “job inspection service”. A candidate for this exists in JobMon, a service developed in PPDG by CMS that utilizes the Clarens server technology that the OSG Discovery Service is based on. This service will provide read-only access to file and process space of the user sandbox while it is running.
- c. We expect to deploy a mechanism that guarantees reliable delivery of exit messages of jobs when they complete on a site’s batch slot. A candidate service will be designed and delivered via collaboration between ATLAS, CMS, and GADU utilizing DISUN, Griphyn and PPDG effort. We expect this mechanism to utilize the Clarens server technology that the OSG discovery service is based on.

(d) Storage Element

- a. We expect SRM based storage elements to be deployed at ATLAS and CMS sites that use the Privilege Project deliverables for SE.
- b. As a fallback for sites on OSG that do not have a viable support model for an SE with SRM interface, it would be worthwhile to explore transitioning to an SE that is based on RFT and gridFTP out of GT4. The gridFTP server would be deployed on hardware independent of the CE. RFT would provide queuing of data transfer requests. This could provide significant reliability improvements. ATLAS and CMS are not prepared to put significant integration effort into this solution but will deploy the service if shown to be useful by others.

(e) Monitoring, Information, and Accounting Services

- a. We expect to spend effort within the context of the Blueprint Activity to arrive at clear requirements and specifications in these

three areas. This will include recording and public presentation of the metrics described in Section 6.

- b. We expect to participate in an effort lead by tg-MIS that will provide a first implementation of the requirements in time for OSG 0.4 deployment.

(f) Performance goals

- a. We expect ATLAS and CMS operations to achieve efficiencies as defined above in excess of 90% on the majority of OSG sites. To reach this goal we are committed to provide education and outreach effort for sites that are falling short of this goal.
- b. We expect ATLAS and CMS to provide some education and outreach effort to other VOs in order to improve their operational efficiency on the OSG. On the flip side, we expect to restrict access to ATLAS and CMS sites for VOs that pose a risk to site reliability due to the way they structure their operations.
- c. We expect ATLAS and CMS operations to achieve efficiencies in excess of 95% for their operations on at least some sites.

## 8 Services Beyond OSG 0.4

There are a number of services that we expect to become available on OSG on time scales beyond the OSG 0.4 series of releases. We list them here without a clear indication of schedule or effort.

- Advanced space reservation. We expect ATLAS and CMS to replicate blocks of files of one to a few TB in size in order to exploit compute resources in an opportunistic fashion. E.g., replication might be triggered by a high watermark of pending jobs that require a given block of files. The VO specific WMS may then dynamically replicate the block of files to a site where it already has harvested significant CPU resources for the purpose of simulation. The late binding model would allow some degree of re-prioritization of CPU resources as data becomes available.
- Advanced networking services. The previous example of dynamic data replication based on load would require knowledge of network bandwidth as well as space reservation.

- Integration of interactive root based analysis into the WMS.

## 9 Performance Goals beyond OSG 0.4

Some performance goals need to scale with the total data volume of the LHC experiments. We discuss these goals here by describing the driving factors.

- a) Efficiency for job execution needs to scale with the size of the workloads submitted by ATLAS and CMS. As a guideline, we require that less than 3 re-submissions are required for completing all parts of even large workloads. For late 2007 we expect that large workloads may reach up to 1,000-10,000 jobs. We thus require that the failure rate to the third power is not much more than 1/10,000.
- b) Submission of  $O(1000)$  jobs should take no more than a fraction of a minute. It is expected that typical analyses will require  $O(1000)$  jobs. The user should not be exposed to latencies of scheduling jobs on the grid.
- c) The scheduler in the WMS needs to be fast enough to keep all resources busy that ATLAS and CMS have access to. A rough estimate of the number of batch slots in one of the two experiments worldwide may be obtained assuming 500 batch slots per site, and 100 sites, or  $O(1e5)$  batch slots. Assuming a typical runtime of a few hours, or  $O(1e4)$  seconds, this translates into a requirement of scheduling  $O(10)$  jobs per second.
- d) The overall grid system should be sufficiently easy to operate as not to require more than 1FTE to manage all of the ATLAS or CMS production simulation or reconstruction in the US. This will require ongoing improvements in operations as the data volume increases.
- e) We expect ATLAS and CMS to replicate blocks of files between sites based on number of pending jobs. This will require the ability to move multi-terabyte data volumes between OSG sites within hours. Network bandwidth as well as disk space and CPU availability will thus factor in to replication decisions. Reliable and sufficiently predictable data transfer at large volume is thus a likely performance goal beyond OSG 0.4.