OSG End-to-End Information Systems

Table of Contents

1 Document Change Log

Version	Date	Change Description	Prepared By
v1.0	March 1, 2010	First version of the Document	Anthony Tiradani
v1.0.1	March 12, 2010	Changes for feed back from March 1, 2010 meeting	Anthony Tiradani

2 Description of Activity

As the Generic Information Provider (GIP) infrastructure neared “feature complete” status, a USCMS effort was started to determine if there were any further requirements that USCMS would like to see fulfilled for the end-to-end Information Systems in the Open Science Grid (OSG). The effort concentrated on the Discovery Information Systems and ignored Monitoring and Accounting Information Systems. This effort was then expanded to understand the requests and requirements for Virtual Organizations (VOs) across OSG and included both Discovery and Monitoring Information Systems.

2.1 Activity Process

In May of 2009, a meeting was held within USCMS to discuss the current Discovery Information Systems and determine the future requests and/or requirements USCMS would have. It was determined that if there were going to be change requests, an understanding on how VO's across OSG use the Information Systems and what requirements and requests the VO's have for the Information Systems.

In December of 2009, a meeting was held to kick-off the effort to understand the requirements, requests, and future needs of the VOs participating in the OSG for the Information Systems. As a result of the December meeting, followup meetings were scheduled with individual VOs. The requirements and requests that were expressed in these meetings are recorded and summarized in Section 2.

2.2 Scope

2.2.1 Reliability

This report assumes that reliability is a built in concept. Reliability is defined as the service(s) is running, available, and responding to requests “all the time”. Metrics have been established and documented in a Service Level Agreement (SLA) with the OSG Grid Operations Center (GOC). The SLA for the BDII can be found at https://twiki.grid.iu.edu/bin/view/Operations/BDIIServiceLevelAgreement. An SLA does not currently exist for Resource Selection (ReSS) or OSG Match Maker (MM) as these services are not hosted or maintained by the OSG GOC.

2.2.2 Accuracy of Information

As detailed in Section 2, accuracy of the reported information is one of the concerns that has been expressed. In this context, accuracy is defined as being correct and timely.

2.2.3 Performance

There are performance metrics for some of the components of the Discovery Information System. At some point, metrics for all components of both the Discovery and Monitoring Information Systems should be defined and acceptable levels of performance should be set. This however is beyond the scope of this report.

2.2.4 Features

Besides reliability and accuracy, the requests and requirements listed in this report are features that go beyond the current capabilities of the Information System.

3 Summary of the Current Information System

The OSG has multiple sources of information. It offers MyOSG which contains site catalog information as well as providing an interface to monitoring information. The OSG Information Management System (OIM) and Resource & Service Validation (RSV) services feed MyOSG. The ReSS and BDII services are both fed by CEMon. CEMon calls the Generic Information Provider (GIP) to obtain the Discovery Information about a site. The GIP returns the Discovery Information in LDIF format that conforms to the Glue 1.3 schema. CEMon passes the LDIF directly to the BDII and transforms the LDIF into classads before passing it on to ReSS. OSG Matchmaker (MM) queries ReSS and submits probe jobs to the sites. The ReSS information is combined with the probe results to form OSG MM's final information store.

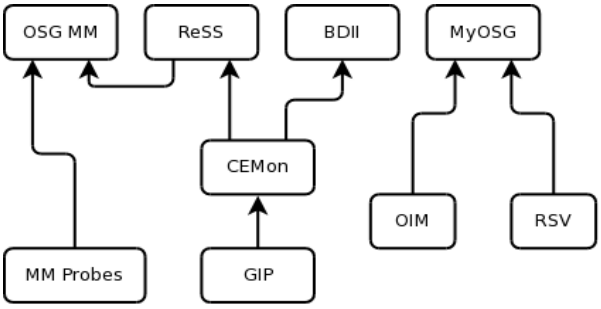

Illustration 1: Current Information System

4 VO Requirements & Requests

This section contains a summary of the requirements and requests that were gathered from VOs and service providers. The summary is split into three sub-sections. The first sub-section is a listing of the requirements that are currently met. Any changes to the Information Systems architecture must maintain current functionality of these requirements. The second sub-section details requests for improvement. The last sub-section details requirements and requests for features that are not currently met or not currently integrated coherently.

4.1 Requirements Currently Satisfied

The following components satisfy a current set of critical requirements for VOs to run jobs of the OSG. Any changes to architecture or information must not impact operation of these components.

4.1.1 BDII

The BDII is required for all OSG sites that wish to be inter-operable with the WLCG. The BDII information conforms to the Glue 1.3 schema. The OSG BDII acts as a top level BDII for the WLCG BDII's to query against. The CMS VO is the primary user of the BDII service. ATLAS uses the BDII for the FTS service and may be considering using BDII information in the future for their AGIS project.

4.1.2 Resource Selection Service (ReSS)

ReSS is used by several VOs to perform job matchmaking when determining where to submit jobs. LIGO and ENGAGE are two examples of VOs which make heavy use of ReSS.

4.1.3 OSG Matchmaker (MM)

OSG MM queries ReSS and performs additional information gathering tasks. It aggregates the information and provides a matchmaking service for various experiments that do not have formal VOs. Some formal VOs currently use this service as well and are completely satisfied with it.

4.2 Requests for Improvement

4.2.1 Single source of help

OSG needs a single source of help for the Information Systems. Right now we have multiple voices that are not all giving the same answers to questions. Some of this stems from an incomplete understanding of the entire Information System by the various sources of help.

4.2.2 Easier Querying

OSG needs to provide a method for easier querying of the Information System(s). VOs would like a single entry point for all information that integrates downtime information with data from the MyOSG service catalog, data from the Discovery Information system, and data from the Monitoring Information System.

An emphasis was put on having programmatic access to all the information. This means that the ability for a script or program to dynamically query the information systems without the need to manually build pre-canned queries is required. This will require an API for querying the information systems.

4.2.3 Consistency & Accuracy of Information

Consistency and accuracy of the information contained in ReSS, BDII, and OSG MM must be improved, tested, and verified. GIP generates the Discovery Information as raw LDIF. CEMon is responsible for reporting the information to ReSS and BDII. The raw LDIF is directly sent to the BDII Collector at the OSG GOC. CEMon transforms the LDIF to condor classads and sends the ClassAds to ReSS. Any inconsistencies encountered at this point are due either to the transform itself or the assumptions VOs make pertaining to the "meaning" of the data in ReSS. OSG MM queries ReSS and performs other tasks such as sending verification jobs to sites that are listed in ReSS. OSG MM then combines the results of the verification jobs with the ReSS data. One known inconsistency in the meaning of Information stored by OSG MM and the information generated by GIP is that OSGMM assumes that a cluster at a site is homogeneous. This isn't necessarily correct.

4.3 Requirements and Requests not Currently Satisfied

4.3.1 Software Reporting

Currently very little is known about the environment a job has prior to landing on a CE or a worker node. Some Workload Management (WM) systems require software knowledge prior to job submission. In other cases it is useful to know what software is available. Some VOs have specifically requested that GLOBUS_LOCATION is advertised. They have also requested the ability for sites to advertise specific software that is available on the CEs as well as the worker nodes. Additionally, a new requirement has been expressed to publish the environment a job will see upon execution.

4.3.2 Storage Reporting

The current schema for storage information is lacking several pieces of information. Information about opportunistic storage is completely missing from all Information Systems (BDII, ReSS, OSG MM, and MyOSG). At this point, a VO must contact each site individually to arrange for storage. The VOs would like for there to be opportunistic storage similar to opportunistic computing. Note: The OSG Storage group is currently working towards a solution for opportunistic storage.

Additionally, quota information is completely missing. A request has been made for used and total quota information per VO for both scratch directories and other forms of storage. Glue 1.3 does not provide any mechanism to report this and the Glue 2 implementation specifically specifies global values and omits quota information.

VOs have also expressed the need to be able to distinguish between pool accounts and individual accounts that a job will run as in order to determine the storage policies that a job will follow.

4.3.3 Verification

The issue of verifying whether the data seen in the Information Systems are correct and current has been brought up in nearly all the meetings with VOs and service providers. A mechanism is needed to verify that the reporting resource and the services that it is reporting is/are in the Information System(s) and are consistent across all the system components (MyOSG/OIM, RSV, BDII/ReSS). Additionally, the information being advertised must somehow be verified as correct.

Dynamic discovery of resource specifics was a request made as a way to help verify the information about resources that are advertised. Specifically, subclusters should be automatically detected and published. This would remove the need to manually update subcluster information everytime a purchase or major change is made to the computing cluster.

ENGAGE, LIGO, and FermiGrid all currently submit active probes to every site in OSG to perform verification by VO. In other words, they check to see what VO information is advertised by a site, then they attempt to use the information. So, for example, if Site A advertises that LIGO can write to /mnt/scratch/ligo, then the probes check to make sure that the ligo directory is either mounted or is accessible and that the correct directory permissions are set.

4.3.4 Schema Extensions

A subset of the requirements and requests fall under the category of extending the current schema with custom fields and sections. Two specific examples are for subclusters and per job information.

The batch system eviction and preemption policies are not advertised in Glue 1.3. These policies should be advertised at a subcluster level. Since the policies and machine configurations can differ between subclusters, total and free job slots should also be advertised at this level. An additional piece of information that some VOs have expressed interest in is the local batch system priorities at a subcluster level.

Some sites restrict the amount of resources a single job can acquire and use. Critical information that needs to be advertised per job are amount of memory a job can use, the number of cores that a job can access, and the amount of scratch space available to the job.

4.3.5 Information Integration

Currently to get a complete picture of a site, a VO must query two sources of information, at a minimum. VOs would like to be able to perform one query and get back Discovery information, Monitoring Information, and MyOSG Information. The data formats mentioned range from classads to xml. Additionally, VOs would like to have the ability to programmatically and automatically update or correct information in MyOSG.

There are two other requests for integration. The first is for native FTS reporting. The two VOs that would like to see native FTS reporting are CMS and ATLAS. There is a provider for

GIP that copies Discovery Information from FTS, modifies it slightly to fit into the Glue 1.3 schema, and forwards it to the BDII. The second is for better RSV integration. Currently, there is a human readable display in MyOSG for the current results and a history graph, but better integration into the full information system is desired.

4.3.6 Interoperability

The WLCG is moving towards Glue 2.0 as the next schema for Discovery Information systems. As such in order to maintain interoperability, the OSG will need to extend it's current Discovery Information system to handle Glue 2.0.

Interoperability with TeraGrid's Information System has been requested by an OSG group to help facilitate sending overflow jobs to TeraGrid.

4.4 VO Concerns

Two concerns were introduced by VOs as they presented their requirements and requests. The first is in reference to ongoing development effort. Are the Discovery Information systems going to have continued development effort (unlike VORS/MYOSG that were deliverables and therefore had a development deadline)? The second concern asks if the OSG MM will be integrated into the overall OSG information system so that it has an SLA with 24-hour maintenance and support.

5 Non-VO Required Changes

Condor 7.5.1 is starting the path to migrate from “old classads” to “new classads”. ReSS will eventually need to be updated to handle the new syntax. The second required change is the migration to the Glue 2.0 schema for WLCG interoperability.

6 Use Cases

Current use cases for the three largest stakeholders in OSG are listed below to help understand the impact of the recommendations.

6.1 ATLAS

For the most part ATLAS (Illustration 2) has chosen to not use the Discovery Information systems from either WLCG or OSG. ATLAS is in the process of developing “A Grid Information System” (AGIS). AGIS takes data from manual input, WLCG SAM tests, and MyOSG. The ATLAS workload management system, PanDa, takes legacy information from “Tiers-of-ATLAS” and current information from AGIS. FTS is the only component that uses the BDII. As long as the Glue 1.3 BDII is maintained for FTS, any changes to the OSG Information System will have minimal impact on current ATLAS operations. However, there is talk that ATLAS is considering using the BDII as well to feed information into PanDa as well.

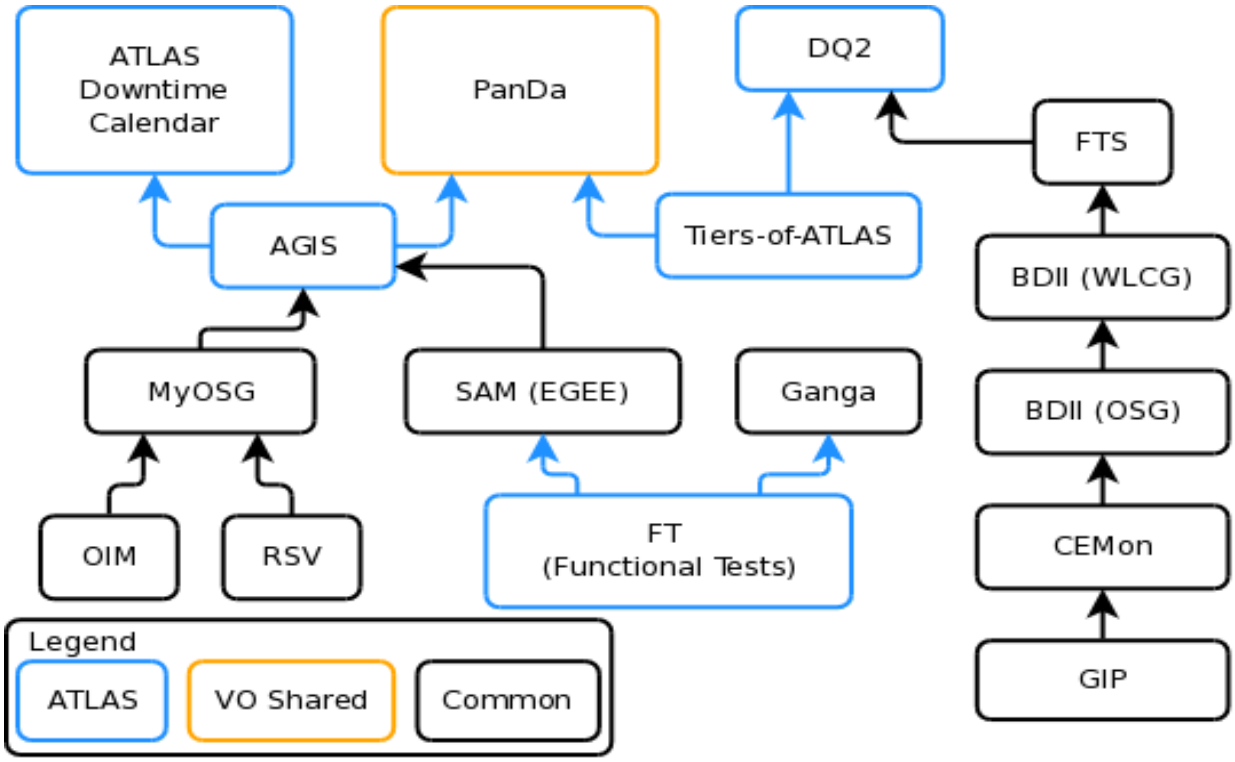

Illustration 2: ATLAS Use Case

6.2 CMS

CMS (Illustration 3) makes use of the BDII for Discovery information and OIM/RSV/MyOSG for downtime and monitoring reporting to the WLCG. However, CMS only uses the OSG BDII indirectly. The WLCG BDII's query the OSG BDII to get OSG's Discovery Information. All CMS applications query the WLCG BDII's. WLCG interoperability must be maintained for CMS. This includes the upcoming Glue 2.0 schema change.

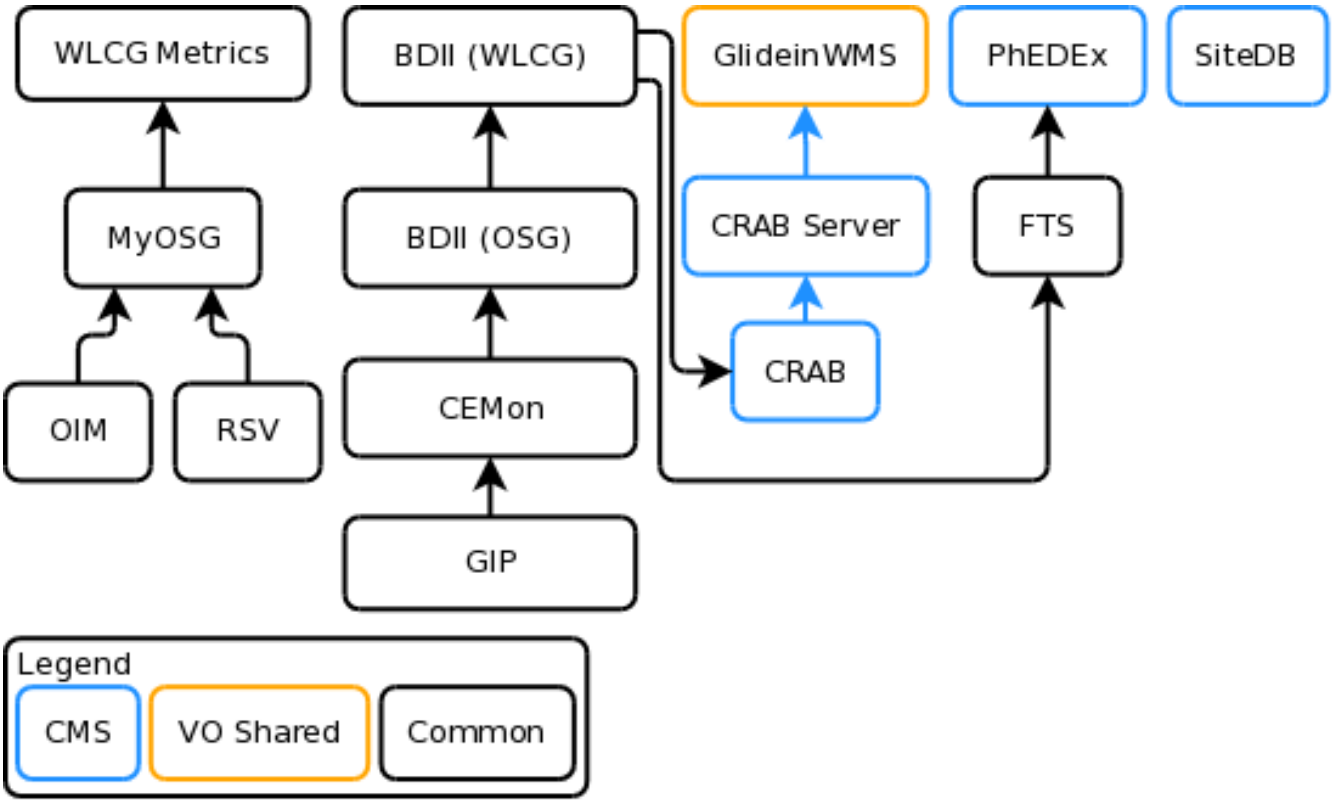

Illustration 3: CMS Use Case

6.3 LIGO

LIGO (Illustration 4) uses the Pegasus Workload Management tool. They pull information from the OSG Matchmaker to determine the sites that LIGO jobs will be submitted to. OSG MM queries ReSS for all sites in OSG. It then sends probe jobs to the sites that verify and extend the information that is reported to ReSS.

LIGO does not use the BDII for Discovery information. However, LIGO would like to be able to use MyOSG to load downtime information as well as some static information. This is shown by the dotted blue line.

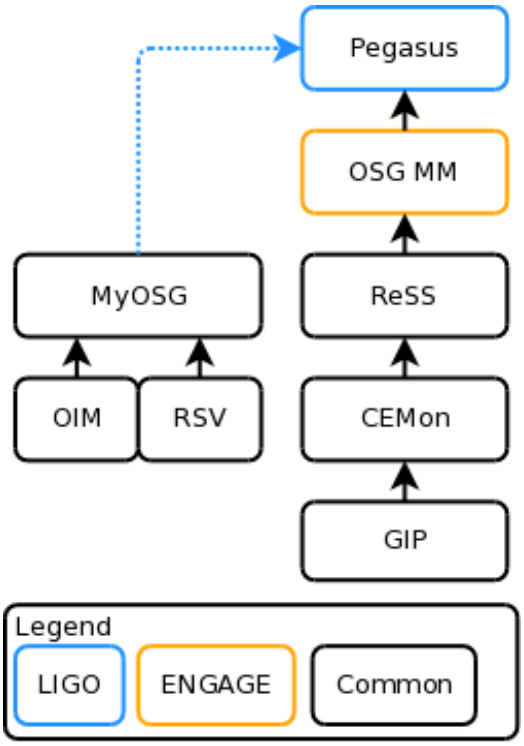

Illustration 4: LIGO Use Case

7 Recommendation Proposals

There are three proposals that utilize existing OSG components. The three listed here cover a minimalistic case, roughly maintaining the status quo, an incremental approach that leads toward the next proposal, and the most effort intensive case that will probably require a satellite proposal. Additionally, a brief look at implementing TeraGrid's Information System is included as an option.

7.1 OSG Minimal Effort / Status Quo

This is the absolute bare minimum change necessary to keep compatibility with the WLCG. None of the new feature requests would be satisfied by this recommendation. Of the three main stakeholders in OSG, CMS would be the only one to make use of the change. However, it would still take time for CMS to migrate their applications to make use of Glue 2.0 Discovery Information.

ATLAS would not be affected since they do not use the BDII for much. ATLAS would have to put all their requests and requirements into AGIS which would effectively replace most of OSG's Information

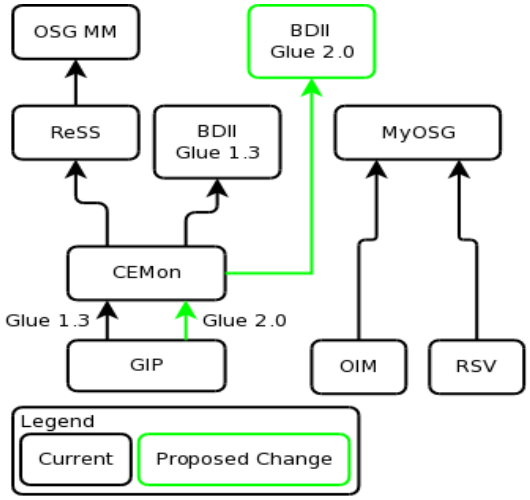

Illustration 5: Minimal Effort

System.

VOs that have needs that are not met in the current Information System will not benefit from this change either. All requirements and requests will have to be fulfilled by the VOs themselves or possibly a collaboration with ENGAGE to extend OSG MM to handle the needs and requests.

7.2 OSG Incremental Approach

This recommendation is suggested to cover the case where more than the minimal effort and/or more than the status quo is desired but funding cannot be arranged for dedicated effort to develop recommendation 5.3.

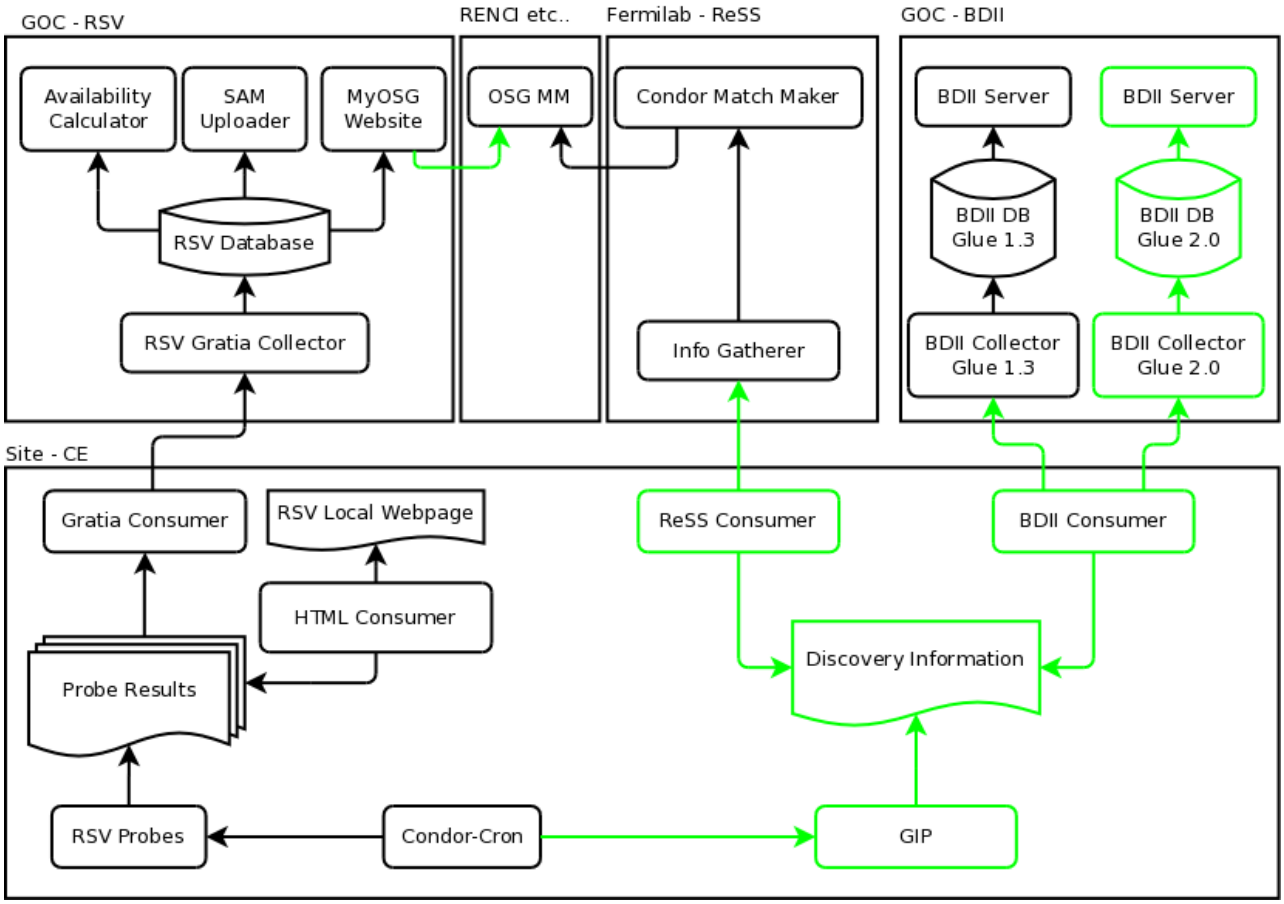

Illustration 6: Incremental Approach

Instead of being launched by CEMon, GIP is now launched by Condor-Cron within the RSV infrastructure. (Please note: This proposal is to use the existing architecture from RSV, not to combine the databases or consumers.) GIP is modified to output Discovery Information conforming to an OSG generic schema. New RSV Consumers would be written to translate the Discovery Information into the required formats. In the near term this would be a ReSS Consumer that would translate to the “Old ClassAd” format and the BDII Consumer which would translate to Glue 1.3 LDIF. Eventually, the BDII Consumer would need to be extended,

or a new Consumer written to translate the Discovery Information into Glue 2.0 format. Likewise, either the ReSS consumer would need to be extended or a new on written to translate the Discovery Information to the “New ClassAd” format. These Consumers would pass the translated information to the appropriate collectors.

The ability for schema extensions can be added to the ReSS Consumers and, to a limited extent, the BDII Consumers. This means that most extensions will exist in ReSS and OSG MM and not in the BDII. Additional work is needed to ensure that the ReSS schema is consistent with BDII schema and that they both are consistent with the understood “meanings” of the data.

PROS:

1) OSG removes one service from the CE package that it must support.
    1. CEMon is removed
    2. The RSV infrastructure fulfills the role that CEMon once served
    3. Reduces OSG's exposure to changes from a development team that may not consider OSG a priority (Note: To date, the VO services project has had a good working relationship with the external development team.)
2) This recommendation will meets some of the feature requests and requirements by allowing for extensions to the Glue 1.3 schema in the Discovery Information system for ReSS and OSG MM.
3) This recommendation gives the option for future enhancements. Potentially, VOs and service providers could be encouraged to contribute desired extensions.
4) This recommendation is required for the recommendation in section 5.3.

CONS:

1) Integration of MyOSG and RSV data with the Discovery Information is not accomplished with this recommendation.
2) Additional testing will be required for the components that replace CEMon.
3) Verification of all types of information is also not address with this recommendation.
4) Development effort will need to be allocated to this solution.

7.3 OSG Dedicated Development Effort

This proposed recommendation builds upon the incremental approach detailed in section 7.2. It adds a translation or transform layer to handle multiple query formats. OSG already has ReSS and BDII for Discovery Information (OSG MM queries ReSS). The basic idea for this proposal is to make it relatively simple for OSG to add different query formats in the future. For example, adding XML and JSON query formats would be as simple as adding a “plugin” to the architecture.

The Discovery Information Systems will report information in a generic OSG specific

"schema". Using a generic schema provides the ability to report data not currently allowed or covered by the Glue 1.3 schema. In addition, because of the transformation layer, extensions can be added to the schema without breaking existing services. Listed below are two ideas on how to implement this proposed recommendation.

7.3.1 Implementation A

As in the Incremental Approach (section 7.2), GIP runs within the RSV infrastructure. Condor-Cron launches GIP which writes out the Discovery Information in a format that conforms to a generic OSG specific schema. This information is placed in a flat file in the

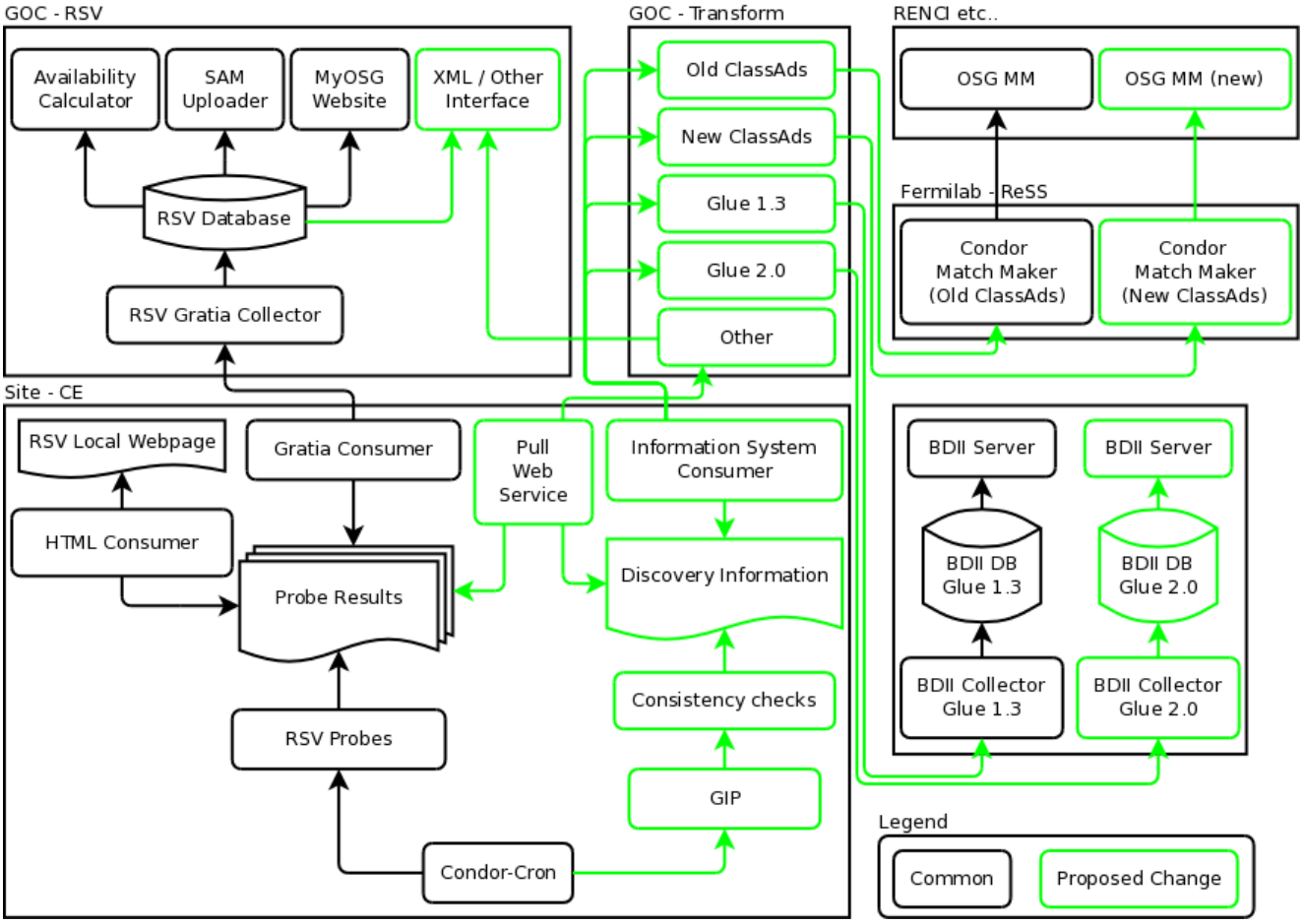

Illustration 7: Dedicated Effort (a)

same manner that the RSV probe results are. An Information System Consumer is developed and added to the RSV infrastructure. This new consumer reads the resulting Discovery Information and the results from the RSV Probes to integrate the two information sources. The consumer pushes the integrated information to the new transformation layer. The transformation layer transforms the data from the generic schema to the specific schema of the destination query service (Glue 1.3, Glue 2.0, ClassAds, etc). This particular implementation adds a web service to the CE that answers queries for other formats. Pictured is an XML format query system. XML based queries would Additional formats such

as JSON can be added to the web service. The actual “user” query point would exist at the OSG GOC in order to integrate MyOSG information such as down times.

PROS:

1. The Glue Schema limitations are removed
2. Query Services, such as ReSS and BDII become plugins to the overall Information System. This allows for greater flexibility for future formats. Adding a new query format does not require any changes to the reporting mechanisms
3. Consistency checks are built into this model
4. Future interoperability efforts can be undertaken without impacting existing Query Services, since all Query Services are plugins to the overall Information System.
5. Provides the ability to meet most if not all the VO and Service Provider requests

CONS:

1. Introduces a critical service layer (the transformation layer) that will be hosted by the OSG GOC.
2. This is a complete redesign of the Information System. As such it will require extensive testing and validation.
3. Significant dedicated effort will be required
4. Instead of reducing the number of query services (ReSS, BDII) that OSG utilizes, this approach maintains them and provides a path forward to add more (XML, JSON, etc).
5. A web service is added to the CE to provide the “pull” functionality that the XML query interfaced will require. Note, however, that CEMon already has this functionality enabled on OSG CE's.

7.3.2 Implementation B

Implementation B is essentially the same as Implementation A except that a central data store is introduced. This removes the need for the web service on the CE and by extension, potentially reducing the load on the CE. The data store becomes another critical service that the OSG GOC would also be responsible for.

PROS:

1. All Discovery Information queries get their results from the same data store.
2. The web service in Implementation A is eliminated
3. Historical data could be stored

CONS:

1. The central data store becomes a critical service. If it is down, then all of OSG

potentially is down. This can be mitigated using High-Availability technologies and techniques, but it is still a consideration.

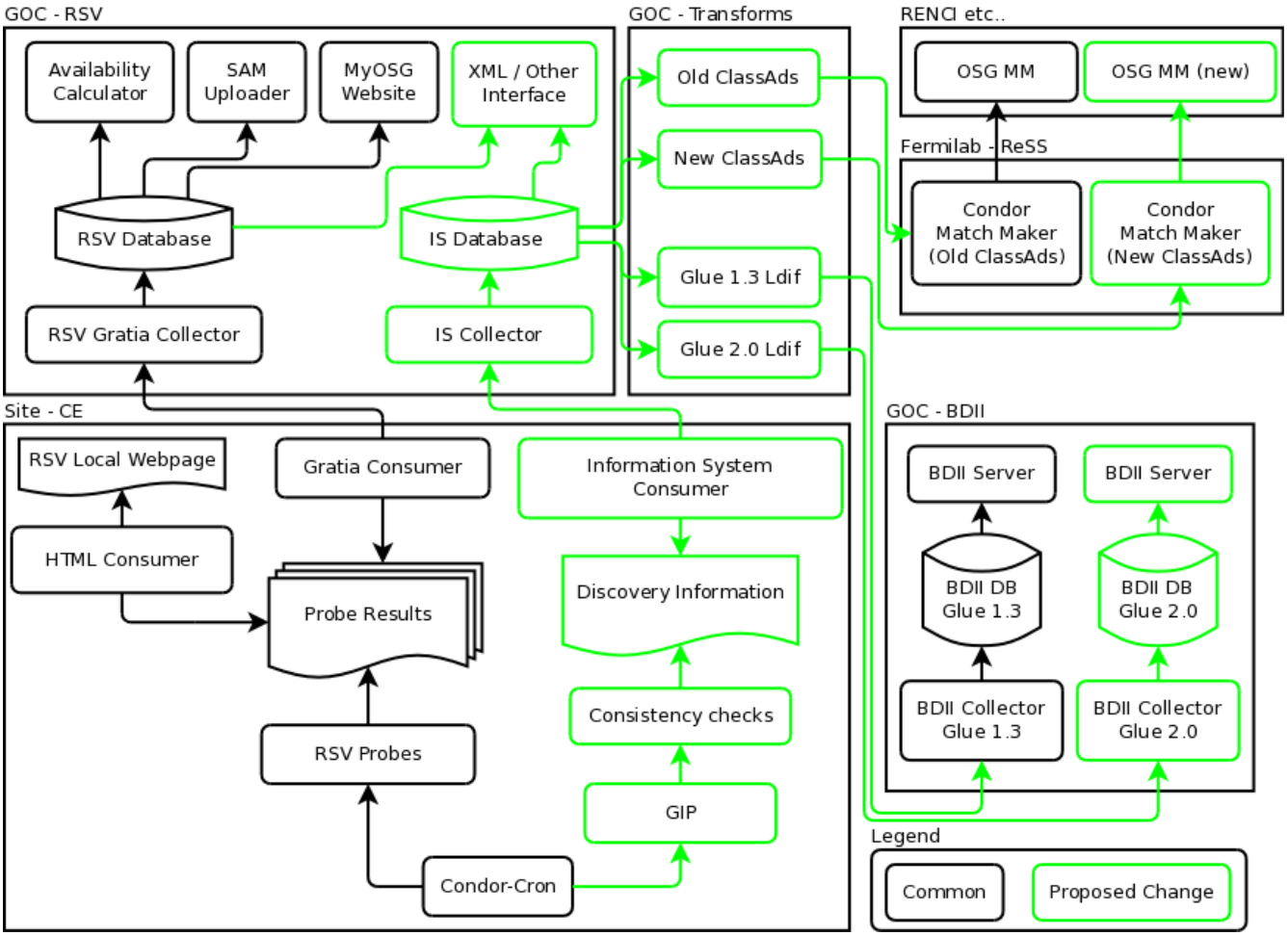

Illustration 8: Dedicated Development Effort (b)

7.4 Implement TeraGrid's Information System

Another approach that should be considered is to adopt the Information System from another Grid, such as TeraGrid. TeraGrid has already done some work on abstracting their Information Services to be an independent component that can be adopted by other Grids. They are currently working with other grids to bring the abstracted Information System online.

TeraGrid Information Services offers some functionality that currently does not exist within the OSG Information System. For example, most information is available in JSON and XML format via a REST interface out of the box. It has the Capability Kit Registry which provides the ability for sites to publish software, services, and/or any other “capability”. Each Kit uses its own schema, eliminating the limitations of the Glue Schema. Already built into the TeraGrid information Services is a subset of the Glue 2.0 schema in XML format.

However, there are missing components that OSG would have to develop to ensure current interoperability levels and to ensure that information that is currently being reported isn't lost.

For example, very little, if anything is published about storage in the TeraGrid Information Services. Most likely OSG would develop a new Capability Kit for storage. This does have the advantage of allowing the storage group to define their own schema. Additionally, there is no LDIF support at all for the TeraGrid Information Services. OSG would have to write a transformation plugin to obtain the Glue 1.3 LDIF and in the future, the Glue 2 LDIF. OSG would still need to maintain the BDII as an addition to the TeraGrid Information Services.

One additional point that must be considered is the question of what Globus plans to do with MDS4. The TeraGrid Information Services backend is based on MDS4. Globus has essentially given up on GRAM4 due to a host of factors including the unmaintainability of the Java libraries that Globus 4 was built on. What will happen to MDS 4? It will do no good to adopt a technology that won't be supported going forward.

PROS:

1. The core services are developed by TeraGrid, not OSG
2. Barriers to interoperability between OSG and TeraGrid are partially removed. Jobs should be able to be matched, and information about both grids can be freely exchanged.

CONS:

1. Implementation would implement a completely new system, requiring testing and verification of all components
2. Out of the box, the TeraGrid Information Services do not provide all of the functionality that OSG requires. For example, storage information is missing and there is no BDII interface. These components will have to be developed by OSG.
3. Functionality to report VO support on a resource does not exist in the TeraGrid Information Services.
4. It is unclear how many of the VO requests will be satisfied by the TeraGrid Information Services without additional development effort.