The benefits and challenges of sharing glidein factory operations across nine time zones between OSG and CMS

by

I Sfiligoi[1], J M Dost[1], M Zvada[2], I Butenas[3], B Holzman[4], F Wuerthwein[1], P Kreuzer[5], S W Teige[6], R Quick[6], J M Hernández[7], J Flix[7,8]

[1]University of California San Diego, La Jolla, CA 92093, USA
[2]Karlsruher Institut für Technologie, 76021 Karlsruhe, Germany
[3]Vilniaus universitetas, LT-01513 Vilnius, Lithuania
[4]Fermilab, Batavia, IL 60510, USA
[5]RWTH Aachen University, III. Physikalisches Institut A, 52074 Aachen, Germany
[6]Indiana University, Bloomington, IN 47405, USA
[7]CIEMAT, 28040 Madrid, Spain
[8]Port d'Informació Científica, E-08193 Bellaterra, Spain

Many VOs have adopted the pilot-based WMS paradigm, also known as “overlay infrastructure”.

In this paradigm, resources across multiple administrative domains are aggregated into VO specific overlay pools, or “virtual clusters” (VC). Each VO has full control over its own VC, and can thus easily implement priorities between the final users.

A broadly adopted pilot WMS is **glideinWMS**.

One characteristic of glideinWMS is the clear separation between

- the VO-facing layer implementing the provisioning logic (**VO Frontend**)
- the Grid-interfacing layer responsible for the actual provisioning (**Glidein Factory)**

This clear division allows the VOs to keep the control of their provisioning policies, while outsourcing the operations of the resource provisioning service to a dedicated team of Grid experts. This results in a clear division of labor, with the former supporting the domain scientists and their applications, and the latter working closely with IT professionals that physically manage and control the resources in each administrative domain.

The Glidein Factory operations is mostly independent of the served VO, allowing for instances serving multiple VOs

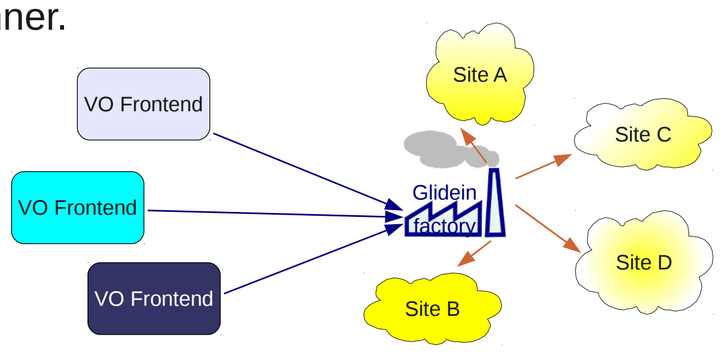
and thus reducing the total cost of ownership (TCO) through economies of scale.

Resource provisioning is clearly separated from resource usage, with the former managed by dedicated IT personnel.

Standard users are thus never exposed to the complexities of Grid infrastructure and perceive the overlay pool as just any other compute cluster.

Logically, a Glidein Factory instance is just a slave. It receives orders from one or more VO Frontend instances and acts accordingly, with each VO Frontend treated in an independent manner.

The main added value of a Glidein Factory is to insulate the served VO Frontends from the details of resource provisioning.

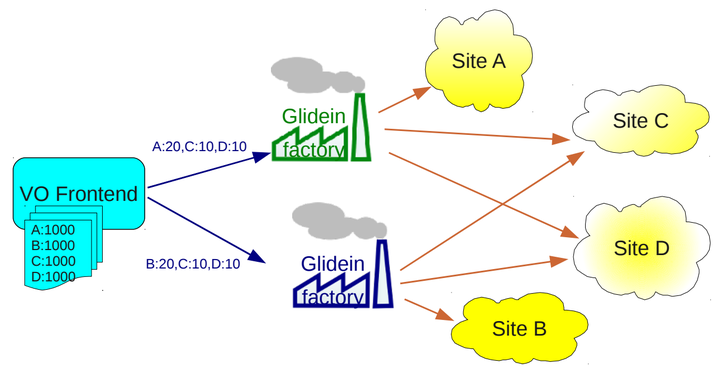


The protocol between VO Frontend and Glidein Factory is based on the **principle of constant pressure**. When a VO Frontend needs a large number of additional resources, it does not ask for all of them in well defined chunks; instead, it **asks for a stream of resource provisioning pilots.**

(with the understanding that the VO Frontend will tell the Glidein Factory when to stop provisioning more)

A nice property of this approach is **the possibility to request multiple streams from the same resource provider**, e.g. the **use of multiple Glidein Factories**.

(without having to guess how effective these requests will be)



CMS and OSG jointly operate four Glidein Factory instances.

The most obvious benefit of having multiple Glidein Factories is **providing redundancy** to the glideinWMS ecosystem, thus eliminating the single point of failure.

In case one Glidein Factory instance stops working, **others pick up the load** and the VOs hardly ever notice it.

This benefit extends both to unscheduled and scheduled downtimes, adding the nice side effect of making maintenance of the services a relatively transparent process for users.

A related benefit is **scalability**. By partitioning the glidein requests over multiple Glidein Factory instances the total number of supported provisioned resources **scales linearly with the number of instances**.

If we were to hit scalability limits in our deployments, we could easily overcome them by instantiating yet another instance, either at one of the four existing data centers, or anywhere else.

In order to achieve full benefits from multiple Glidein Factories, each instance used by a VO must be configured the same way. **Full synchronization is however not possible.**

(differences at the various instances both of available hardware and deployment policies)

Cannot sync byte by byte, due to differences in

Network setup | **File paths** | **Supported sites** | **SW version**

Have developed a **tool for selective cloning** of attributes from one factory to another, allowing for semi-automated propagation of changes. (Only occasional manual adjustments needed)

Remaining issues

- Major software version changes (typically backward but nor forward compatible)
- Human error

OSG and CMS operate the four factories with a single common team.

Most operational issues stem from Grid-related problems and are very similar for all the Glidein Factory instances;
solving a problem in one instance thus very often solves the problem for all instances.

The Grid landscape is composed of hundreds of sites, so **change is inevitable**. The Glidein Factory operators must keep up with it.

(We observe that each week at least one new service is added and one old deprecated)

This includes both **noticing** the change, as well as making sure that it is both **legitimate** and, if a new service is added, **properly configured**.

We have developed tools that help us reduce the human effort needed in doing this, but it is far from being fully automated.

The Glidein Factory operators are also responsible for **monitoring** the success rates of such requests, and act if too many are failing.

The **main challenge is the sheer number** of provisioning requests flowing through the system.

We have developed tools to filter out the logs of well behaving glideins, and some that flag the logs of the obviously broken ones. However, this still leaves a substantial number of logs that require some human parsing.

Each of our instances serves about 50k glideins/day

Operators span five locations and nine time zones.

The four Glidein Factories are located at CERN, Fermilab, Indiana University and UCSD, and one operator is at KIT.

By operating in shifts, we provide **up to 17 hours** of support a day with each operator working only his regular business hours.

The actual coverage is of course not complete; e.g. the hardware problems are dealt with by the local people only. However, the redundant nature of glideinWMS mitigates this problem significantly, i.e. as long as at least one Glidein Factory is fully functional the served VOs do not suffer, so the perceived coverage is indeed complete.

The increased head count also leads to the establishment of a **collective memory**. This allows for an easier handling of both personal needs of the various operators, such as vacation and sick days, as well as personnel turnover.

This is especially important for CMS, since due to various reasons the CMS operators at CERN can be hired for at most two years; we indeed already had a change of operator recently.

The high turnover expectation lead also to **better documentation** of both the glideinWMS architecture and the actual operational procedures.
UCSD has recently hosted two glideinWMS related workshops, the material of which was used to train both active operators and new hires.

As an added bonus, the same training material has also been used as educational material for undergraduate students who occasionally collaborate with us.

One major problem of having the team physically distributed across multiple continents is **communication**. Given there are up to nine hours of time zone difference between various operators, there is effectively **no overlap in business hours** between them. This means that most communication happens through a **bulletin board-like medium**, where the operators of one shift leave notes to the next one.

We are still experimenting with various tools. Details about which products are used not presented.

Text-based, e-communication is of course not ideal, so occasional **off-hour communication is still required**.

One regularly scheduled occurrence is a **weekly meeting** in which the operators in Europe work an hour later and the operators in California start an hour earlier. This is useful **mostly for the establishment of social ties**.

(although sometimes it is also an invaluable forum where operational questions can be addressed)

For rare major problems it is up to the group coordinator to work off-hours and indirectly bridge the gap between the various team members.

The other problem are **local responsibilities**. Most operators work only part time on the Glidein Factory operations and spend the rest of the time dealing with other, local responsibilities.

Local activities often prioritized over the global Glidein Factory operations.

Both due to the closer physical proximity of local leaders and the perceived redundancy of operators due to a relatively large operator pool.

Work unable to be performed by one operator can **occasionally be compensated** for by the others in the group. However, it is **not sustainable for extended periods of time**.

So we are making an effort to closely monitor the situation and, if needed, apply any necessary mean to correct it.

The overall experience has been very positive, and we look forward on continuing on this path.