SLM and SDM Challenges in Federated Infrastructures

Matti Heikkurinen[1,2] and Owen Appleton[1,2]

[1] Emergence Tech. Limited
[2] gSLM Project
{matti,owen}@emergence-tech.com
www.emergence-tech.com
www.gslm.eu

Abstract. Federation of computing resources imposes challenges in service management not seen in simple customer-supplier relationships. Federation is common in e-Infrastructure and growing in clouds through the growth of hybrid and multiclouds. Relationships in federated environments are complex at present, and must be simplified to allow structured service management to be improved. Input can be taken from commercial service management techniques such as ITIL and ISO/IEC20000 but special features of federated environments, such as complications in inducement and enforcement must be considered.

Keywords: Cloud, multicloud, e-Infrastructure, Grid, HPC, Service Level Management, Service Delivery Management, Service Level Agreement.

1 Setting the Scene

Online and distributed services are now essentially endemic in the IT sector, so much so that their nature is often ignored by end users of systems such as Gmail or various online storage services like Dropbox. Both the academic and industrial/commercial sectors are likewise awash with distributed services. In the commercial sector there is a long history of network based IT services, but recently Cloud computing has been the focus of attention. Clouds are online services based on large numbers of virtual machines that can be used to provide compute, storage and other services.

In the academic sector the umbrella term for distributed services in Europe is e-Infrastructure (cyberinfrastructure in the USA). This term includes a wide range of "facilities, resources and collaboration tools" that allow for "computing, connectivity, storage and instrumentation"[1]. This is a stack of services, from pan-European network services such as GÉANT at the bottom of the stack up through an infrastructure layer of high performance or high throughput computing systems up to service layers that are exposed to users. High throughput computing, also known as Grid computing, involves the coupling of geographically distributed heterogeneous, generally commodity resources into a system similar to a computing cluster. In comparison, high performance computing (HPC) involves tightly coupled and geographically local high specification machines, so-called supercomputers. HPC does also work in a distributed manner, through connections between supercomputers, but it is a different approach to distribution.

M. Alexander et al. (Eds.): Euro-Par 2011 Workshops, Part I, LNCS 7155, pp. 64–72, 2012.

In all these cases, there is an element of dynamic federation. These systems are designed for users with variable needs that for a range of reasons choose not to work with fixed local resources.

2 The Nature of Federated Infrastructures

2.1 The ICT Management Challenge

There is no way around the need for management of IT services. Provision of infrastructure services (i.e. large-scale, general purpose IT services) will always involve an important management component. Correct architectural design, fault-tolerant soft- and hardware and other technical solutions can improve the reliability of the infrastructure and convince users to migrate their applications.

However, management processes and organizational approaches are key in ensuring that Quality of Service (QoS) can be maintained in situations that automatic tools cannot cope with. Such situations are inevitable in complex systems reliant on variable networks and used by fallible users.

These challenges have been addressed by the commercial and public sector in the last few decades, and a large body of knowledge addressing this management challenge has been collected under the disciplines of Service Level Management (SLM) and Service Delivery Management (SDM). They collect, analyze and derive sets of processes that are used to define, agree, monitor and manage the provision of a huge range of services. These provide a structured approach to the service provision lifecycle, from offering service catalogues to negotiating enforceable legal agreements, coping with service failures and closing agreements. While SLM-SDM has legal elements and can specify technical metrics, it also pays attention to reaching common understandings and other human elements in the service domain. These themselves are necessary for legal reasons, as for instance in most systems contracts must represent a meeting of the minds to be legally enforceable.

In the commercial IT service sector, SLM-SDM is a mature area, a component of IT Service Management (ITSM). Several internationally accepted systems for SLM-SDM exist, notably ITIL, ISO/IEC 20000, eTOM and COBIT. However, these systems have not been used in the e-Infrastructure domain beyond small-scale pilots focusing on the activities within a single site. In fact, even when addressing complex, cross-organizational value networks, these SLM/SDM frameworks have always assumed a fairly static contractual model that can be broken down into bilateral "producer-consumer" agreements that have been negotiated well in advance. Dynamically federated e-Infrastructures, such as Grids, challenge this model and necessitate adaptation of many of the tools and procedures on the management level.

In clouds, SDM-SLM has been considered, as one would expect from commercial organizations already familiar with he concepts, but the implemented service level agreements (SLAs) for instance, are very weak by any normal IT service management standard. SDM-SLM also seems to ignore the need to federate, or so called Hybrid and Multi-cloud infrastructures. It appears that cloud providers assume (or hope) that any 'federation' you need can be accomplished within their service. However, such vendor lock-in will not suit all customers by any means.

In both clouds and academic e-Infrastructure, participants have tried to shortcut the management challenge through automation. For instance, in the Grid field automated SLA negotiation has long been the most commonly seen effort in the SLM-SDM area, but has ignored the more fundamental questions of agreement and negotiation. Rather one-size fits all SLAs have been attempted with only limited success.

Automation is an important step, but one might summarize the service management meta-process as understand, communicate, manage, monitor and then finally automate. Grids and some other e-Infrastructure have tried to jump to the end of this process, and hence found they have an automated system that lacks SLAs anyone will agree to.

2.2 Why Federation Is Necessary?

There are a number of drivers for federation of resources, through cloud, Grid, HPC or any other system. These include requirements for high capacity resources, mitigating risk through vendor independence, simplification of remote service accessibility or perhaps increased service redundancy. At heart most of these can be expressed in economic terms. High capacity resources might be needed transiently (peak demand) but might be unaffordable to provision permanently and locally. Equally vendor lock-in might present a dependence on an outside firm that presents a major financial risk.

Politics can also play a part. The original motivation for development and deployment of Grids came from the need for very large scale resources, that were either economically infeasible in one location, or if not, were politically impossible to select a single location for.

In the cloud domain, there are some basic questions on federation. Clouds are often presented as federating individual compute or storage units - conventionally virtual machines – into a consistent whole. But Clouds often aggregate the resources instead of federating them. While a cloud brings together many virtual machines their ability to act as a seamless distributed whole is limited. In comparison, Grids show lower reliability but more completely unify resources Grids, despite their issues can also scale redundantly as they are based on open standards and heterogeneous resources. Failure of a single site, or even the sites in a whole country is not terminal for a user. Partly due to their experimental and academic pedigree, Grids can cope better than Clouds with local failures and the diversity of software, equipment and groups involved provides security. For example, recent failures in the Amazon cloud services took down many major websites. Industry commentators have since noted that many firms rely on single cloud providers such as Amazon, despite never allowing it with local resources, where a backup system for key infrastructure was long a given [2].

The multi-cloud approach recognises and avoids these risks by spreading services across multiple providers. Hybrid clouds are similar, mixing internal and external cloud resources. Both emphasise provision of vendor-neutral interfaces to various Infrastructure/Platform as a Service (IaaS/PaaS) services.

Whatever their differences, the e-Infrastructure and cloud visions both contain – implicitly or explicitly – the ideas of easy, instant access to remote resources that for whatever reason, exceed the capacities and capabilities of a single organisation. At the same time, both Grid and Cloud domains have matured to the level where solving the interoperability issues between different solutions is seen as a key challenge: benefiting from the economies of scale is only possibly if the resources are – in one way or the other – put into a single pool.

2.3 Examples of Federated Infrastructures

For the moment, the largest federated e-Infrastructures are most likely the European HPC and Grid services. For example, EGI (the European Grid Initiative) acts as an umbrella organisation for an infrastructure consisting of almost 350 resource providers from 57 different countries with hundreds of thousands of CPU cores[3]. The federated supercomputing infrastructure PRACE consists of 21 European supercomputing centres that share resources to solve high-end HPC challenges[4].

In the cloud domain, the main providers are companies such as Amazon, Rackspace, Salesforce, Google and Microsoft's Azure. In the academic space, many installations are based on Eucalyptus (an open source private cloud solution compatible with Amazon's offerings) or make use of OpenNebula (a toolkit for heterogeneous cloud resources).

The multicloud space, however, remains smaller. RightScale provides a multicloud management and monitoring system supporting services including Amazon cloud services, Rackspace and the open source and Eucalyptus based clouds. Open source option Scalr offers most of the same functionality, and there are several other open source multicloud projects in the academic space, but they are not yet widely adopted. While currently a relatively small sector, it seems likely that as the hype on clouds dies down and the limitations of single vendor solutions become apparent, multiclouds will grow rapidly, and likely match if not surpass the size of current e-Infrastructure services.

2.4 The State of Federated SLM

We consider the current state of federated SLM rather weak. In the e-Infrastructure domain, the services' academic origins meant that systems were developed based on informal sharing, without financial customer-provider relationships. Imposing such models early on would have likely stifled innovation as academic and academic organisations are generally wary of engaging in financial relationships but very open to less formal collaborative ones. While at the network level, SLM is quite thorough, in the Grid layers, SLM is generally limited to very weak SLAs which are not easily enforceable.

The cloud does feature SLAs, but the service guarantees are rather trivial. The Amazon EC2 SLA [5] has a maximum penalty of 10% of a customer bill for periods where annual availability falls below 99.95%. Furthermore, the 10% is not a refund, but credit against future Amazon EC2 purchases. Rackspace LoadBalancer service offers 5% of fees paid for each hour the SLA is not met, up to a total of 100% of fees paid. However, again this is paid as credit toward future purchases.

Multiclouds seem to offer initially just collections of the SLAs of the individual services. This is essentially aggregation rather than federation. While it may be sufficient for simply replicating a single system, running a service with interdependent components means that a failure of one Cloud can take down the whole service but compensation is based on the value of the failed part.

For the multicloud or e-Infrastructures to be broadly successful, they will require SLM-SDM that inspires confidence in their client base. In the e-Infrastructure domain, the European Commission has funded the gSLM project [6] to bring approaches from the commercial SDM-SLM arena to the area.

2.5 The SLM Problem

We have described and demonstrated the need for an approach to SLM in federated environments, here we try and consider how a model can be formed. A first step is to consider that while experience, inspiration and expertise from commercial SDM-SLM can be extremely useful, one cannot simply drop SDM-SLM into a multicloud or e-Infrastructure. Dynamically federated systems have specific challenges that must be addressed to implement effective service management.

In e-Infrastructure, multiple organisations contribute to a single, large-scale service. These organisations are connected through relationships that began informally though they now start to seek formal connection. Jumping from no formal agreements to highly formal agreements would not be well accepted by the participants, who appear generally nervous about codifying service levels, even if they are at a level that is easily achievable. There are also complex issues about delegated responsibility. If a single party fails, causing a service provided by a large number of participants to fail, customers must be compensated, but neither penalizing one organisation for the total value of the service, not penalizing all organisations for the failure of one are fair or likely to be accepted. SLAs, the most common form of SLM attempted in the e-Infrastructure sphere, also tend to assume availabilities not seen in Grids and some other e-Infrastructure. Due to the complexity and heterogeneity of the system, availability or success rates around of 65% are not uncommon, which makes many commercial approaches hard to implement.

In the cloud, the product is standardised in a way e-Infrastructure generally isn't. They also have real SLAs that apply to all users, but while SLAs tend to promise good availability, they are weak on penalties for non-performance. They can also decouple user level value from the unit of sale. While the individual value of a single virtual machine may be low to the provider versus the whole pool, it might be high to the user. A complex service requiring many virtual machines running together might be rendered ineffective by a single failure, in a way that say, Grids would recognise and cloud providers would not. Such a failure might be within the bounds of acceptable failure on Cloud SLAs, and is one reason that clouds have not simple replaced e-Infrastructures. This would be a case for single clouds showing features of aggregation rather than federation, and currently multi clouds (apparently more federated) would show the same problem unless the multicloud providers provide their own SLA and accepts the risk themselves.

2.6 Modelling Federated SLM Relationships

In trying to bring a new model to the federated computing landscape, there are a number of steps that can be taken to start the process. The first of these is to map the relationships and risks in each service type.

Plain clouds show a simple relationship model (customer-provider), but for reasons previously discussed, single providers may not serve all needs. In the multicloud, we have several options for how relationships can be structured. Figure 1 shows the relationships in two types of potential multicloud situation.

Customer 1 pays directly two cloud providers and the multicloud provider to control and monitor their multiple cloud resources. While the aggregated multicloud

provider may guarantee their service (as in their dashboard availability etc) they make no promises about the services they aggregate. Thus their service can be relatively low cost, as they assume no risk. From a customer point of view this multicloud scenario works best when using different cloud providers per purpose, e.g. storage from one and compute from another.

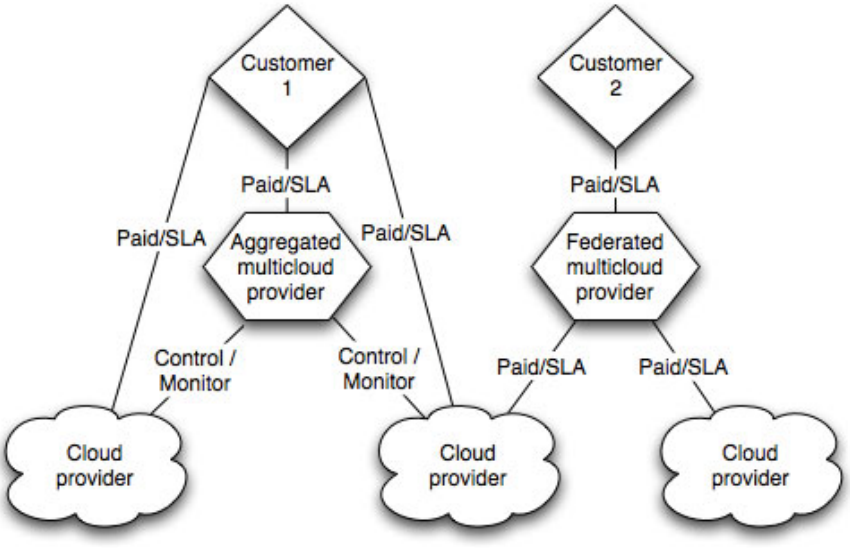

Fig. 1. Two kinds of multicloud relationship maps

Customer 2 deals only with the federated multicloud provider, who provides an abstraction layer that hides the various clouds used to provision their service. They then have SLAs with the cloud providers themselves. In this case the customer deals with a single interface that federates resources such that failure of the overall service will be managed by the multicloud providers (perhaps by redeploying from a failing cloud provider to a working one). Clearly taking on this risk will incur costs, but the costs will provide the user with security not present in the aggregated model.

In e-Infrastructures, particularly Grid infrastructures, reducing the number of relationships to the logical minimum is a challenge complicated by the number of actors. Figure 2 illustrates this relationships challenge.

An ideal situation would be a strictly hierarchical chain where users through various intermediaries accesses resources, this is however complicated by the existence of Grid Infrastructures (essentially federators) of different scales. Countries operate federating bodies for resources, but equally some subject areas or other groups may operate as federators. Both deal with resource owners at so-called sites. There are then secondary federators that collect primary ones into larger bodies. On the customer side, user groups come together as virtual organisations.

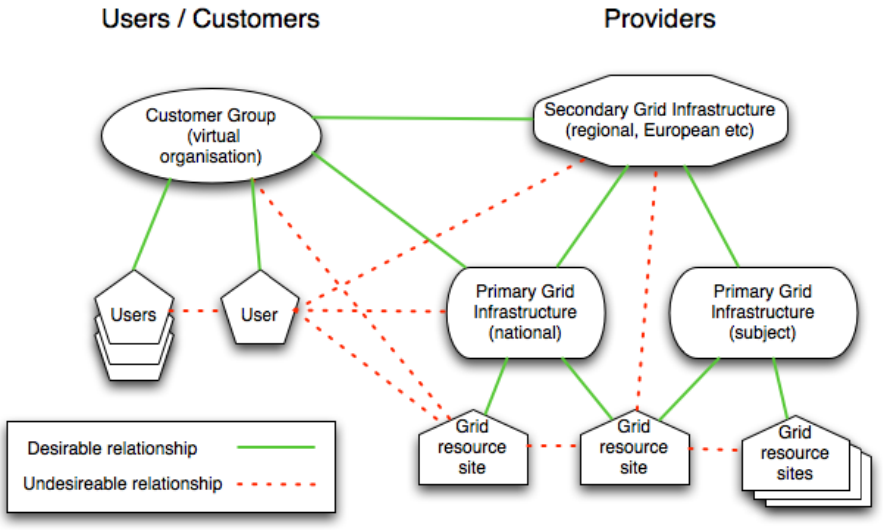

Fig. 2. Relationships in Grid type e-Infrastructure

At present almost every possible relationship occurs, whether desirable or not - including the unnecessary relationships of Figure 2. Reduction to a single chain would let the system be considered, in SLA terms, as an SLA between a user group and a federating Grid infrastructure, which would then have operational Level Agreements (OLAs) or Underpinning Contracts (UCs) with other federators and/or individual sites. The current situation is more complex in several regards. First, many relationships are captured only in Memoranda of Understanding (MoUs) or at best extremely weak SLAs. Second, many relationships that undermine a clear schema for responsibility exist, such as individual users having connections to resource owners, bypassing the federation of both users and resources. Third, the schema shown in Figure 2 ignores the underlying network layer, where each site (and potentially user) will have a relationship with a network infrastructure to support it, though network SDM-SLM is considerable more mature.

In both the cloud and Grid type e-Infrastructure cases we might say the ideal model is the one with the fewest relationships possible. These relationships should be those that can resemble to the greatest extent a conventional commercial customer-provider relationship, where the provider should assume responsibility for all downstream provision issues, even if this imposes risk that must be shared or represented in the contract with the user.

2.7 Enforcement, Inducement and Penalisation

Once a working schema for relationships is derived (and implemented) then agreements must be structured that are not only acceptable, but sufficiently enforceable. Currently cloud agreements appear technically enforceable, but sanctions may not be sufficient to justify major investments in reliability. On the other hand, e-Infrastructures codify very

low levels of service and recognise the impact of failure, but offer no path for enforcement or penalisation on failure. The cloud domain may at the first glance seem to be in a better state, but the SLM-SDM as implemented refers almost exclusively to single providers. The Multi-cloud, with its hopefully federated resources currently offers no meaningful SDM-SLM at all.

On both sides a balance must be struck between guarantees strong enough to tempt a user to adopt a system but not so draconian as to discourage providers. In the cloud this should be simple, since at least all participants are legal entities and agreements between participants are formal. The issue will be to induce single cloud providers to recognise that they must provide SDM-SLM that is compatible with their services being federated. Users must then recognise that the federators assume a risk on their behalf, for which they must be compensated. A similar situation has been seen in software licensing, where all distributed computing systems require rethinking of license models. It is unreasonable for a user that uses 10 copies of an analysis package all month but then uses 10,000 copies on a cloud one day to pay for 10,000 licenses. Equally it is unreasonable for cloud providers to be able to ignore that users require their services to be aggregated in the SDM-SLM they offer.

For Grids the challenges stem from the lack of economic basis to many agreements. One simply cannot impose financial exchanges and penalties in an academic system that has operated on a model of sharing and informal agreement overnight. The multinational nature of Grids and other distributed e-Infrastructure also means that disputes are often between organisations that do not have formal bilateral agreements, and are often in different countries. For instance a site in one country failing may cause a service failure dealing with a federator in a second country, causing many sites federated by that body to lose business, time or custom. At present there is little recourse for those that suffer from failure without being responsible for it. The only sanctions available tend to be embarrassment and loss of professional reputation. Anything beyond this generally involves exclusion from the service, which is a ‘nuclear option’ that becomes immediately a political problem. This means that small failures are not punished, and there is little inducement to perform well beyond avoiding catastrophic, politically unpleasant failures.

One option, short of instituting financial relationships, is some sort of quanta of credit, whether it is for service delivery or for service quality. Even though such a ‘currency’ might be of no economic value, by quantifying it, small changes in service quality could be easily demonstrated. By engaging competition to perform against the yardstick of ones service credit balance, e-Infrastructure could engage organisations in improving their service in the same way social networks encourage participants to make ever more connections through metrics relating to social activity.

3 Conclusions and Future Work

This paper offers only the first steps in a model for federated SDM-SLM, by examining two kinds of system that show (or will show) federation. They clearly have some common issues though many individual ones. In both cases lessons can be learned from commercial SDM-SLM, though systems such as ITIL cannot be wholesale imported. New approaches in areas such as inducement and enforcement must be

considered. These issues will be considered by the gSLM project [5] in the next year and will be presented at future events. The gSLM project also intends to produce a maturity model for SDM-SLM in Grid infrastructures. The concept being to show Grid providers and users a path for evolutionary improvement of service management, such that they can select appropriate SLM-SDM measures as infrastructures grow and develop. This will also form part of a strategic roadmap to be released in last 2012 on the larger issues of SLM-SDM in Grids.

The authors of this paper and colleagues from the gSLM project will also seek to collaborate with members of the cloud community on common issues of federated SDM-SLM. This has already begun through participation at the IEEE BDIM2011 workshop [7] on Business Driven IT Management where the issues were discussed with SDM-SLM experts from the commercial sector, including major IT service providers. It will continue through collaboration with projects such as mOSAIC, and results will be released through the gSLM web site.

Acknowledgements. This paper was produced in part through the gSLM project, which is co-funded by the European Commission under contract number 261547.

References

1. Definition of e-Infrastructure taken from the European Commission website, http://cordis.europa.eu/fp7/ict/e-infrastructure/ (accessed 04.07 2011)
2. http://www.infoworld.com/d/cloud-computing/the-failure-behind-the-amazon-outage-isnt-just-amazons-107?page=0,1 (accessed 03.07.2011)
3. http://www.egi.eu/infrastructure/Figures_and_utilisation/ (accessed 01.07.2011)
4. http://www.prace-ri.eu/ (accessed 29.06.2011)
5. http://aws.amazon.com/ec2-sla/ (accessed 04.07.2011)
6. http://www.gslm.eu (accessed 30.06.2011)
7. BDIM 2011 formed part of the 12th IFIP/IEEE International Symposium on Integrated Network Management (IM 2011), http://www.ieee-im.org/, http://www.bdim2011.org/Workshop/Welcome.html (accessed 01.07.2011)