Adaptive Service Level Management for Grids

Adrian Li Mow Ching[†], Dr Lionel Sacks[†], Paul McKee[‡]

† University College London, ‡ BT

Abstract: Recent advances in Grid computing have lead to real world deployments of grid implementations in the eScience and commercial domains. With the increasing demands on these resources, the role of Service Level Agreements (SLA) becomes hugely important, as an SLA is the means used to define the terms of usage and obligations for the relevant parties. However, the current methods used to manage grid resources are not capable of supporting SLAs. The system has to be robust and be capable of adapting to changing resource demands from continuous SLA requests. In this paper we present our adaptive approach to SLA management for Grid systems based on the adaptive mechanisms of cognitive control in the human brain and an ontological approach to SLA decomposition.

1. Introduction

Recent advances in Grid computing have lead to real world deployments of grid implementations in the eScience and commercial domains [1][2]. With the increasing demands on these resources, the role of Service Level Agreements (SLA) becomes hugely important, as an SLA is the means used to define the terms of usage and obligations for the relevant parties. However, the current methods used to manage grid resources are not capable of supporting SLAs. Current approaches to sharing grid resources are based on hard partitions set by system administrators and batch schedulers that place grid jobs in a queue while providing a single level of service.

A more intelligent form of resource management is required where by the process of establishing an SLA and allocating resources to the jobs associated with that SLA is done in an automated manner. Furthermore, the system has to be robust and be capable of adapting to changing resource demands from continuous SLA requests.. In this paper we present our adaptive approach to SLA management for Grid systems based on the adaptive mechanisms of cognitive control in the human brain and an ontological approach to SLA decomposition.

The work presented in this paper is part of the SOGRM project, an EPSRC funded collaboration between University College London and BT that is investigating Self-Organised approaches to Grid Resource Management. This project aims to develop techniques for distributed resource management for Grid networks as a contribution to the development of a network-wide service architecture.

2. Service Level Agreements

An SLA is a contractually binding agreement between an operator / broker and a user that specifies the terms of resource usage. It is composed of service level objectives that define targets the operator much achieve for the service provided. It also describes how these are monitored and what measurements are used to represent the desired level or service. SLAs have to be validated and monitored though out their lifetime to ensure the provision of the defined service levels and to enforce any penalties that might be incurred. As well as defining objectives such as start and end times, availability, important parts of an SLA are the performance parameters that indicate the performance level a user can expect from the service. These are expressed in application related parameters and vary according to the type of service described in the SLA.

In a realistic grid environment, there will be a continual flow of incoming SLA requests. Such a busy environment requires an automated management system that can cope with the continual flow of SLAs, and the ability to adapt resources to the new demands placed on the system. The partitioning of resources becomes a dynamic problem, where resources are constantly being reallocated.

In a commercial context, it is desirable to provide more than one level of service as there will always be users who are willing to pay more for stricter guarantees for service level objectives and those who willingly pay less for lower levels of service. We define this classification as our QoS mechanism, where a user is guaranteed that a percentage of jobs submitted will complete within the desired

completion time. A commercial operator of grid resources also has high level business objectives that dictate the characteristics of the system. Examples of such objectives are allowing low priority jobs to run on high priority resources so that resources are loaded for the shortest time possible in anticipation of new incoming SLA requests, or allowing the reallocation of resources where resources from low priority jobs that don't have strict deadlines are used for high priority jobs thus allowing the operator to accept more requests.

SLAs are expressed in natural language with the performance targets defined as easily understood parameters that are associated with a particular grid service or application. The user need not be aware of the implementation of the operator and how the performance targets are achieved. In order to support the high level objectives, the system must be able to translate these in to measurable quantities that can be used by a resource management system to accurately determine the amount of resources that need to be committed.

2. Related work

Various aspects of SLA management within grid computing have been addressed. The specification and monitoring of commercial SLAs was approached in [3]. A language was described for unambiguous and precise specification of SLAs, and a monitoring engine that aggregates measurements from multiple sites. The SNAP protocol [4][4] was developed for negotiating SLAs and coordinating resource management in distributed systems. It presents a model for managing the process of negotiating access to, and the use of resources through the definition of a framework within which reservation, acquisition and task submission can be expressed for any resource in a uniform fashion. In [5][5] the authors identify commercial grids argue that the unique requirements in SLAs necessitate the use of a dynamic offload infrastructure. Finally, the approaches envisaged to be adopted in short term in the context production-level GRID management such as pan-European project EGEE (Enabling GRIDs for eScience in Europe) [6] are based on simple fairness criteria such as the balance between the number of resources a VO contributes to the common pool of resources and the resources it consumes.

3. Adaptive architecture based on cognitive control

Our SLA management architecture is based on the architecture of cognitive control in the prefrontal cortex as described in [7]. Cognitive control is defined as the ability to act or think in accordance with internal goals. The architecture proposed is based on a modular model in which the lateral prefrontal cortex is organised as a hierarchy or representations originating from the pre-motor cortex and processing distinct signals involved in controlling the selection of appropriate stimulus-response associations. Control processes are separated in to two distinct categories of, processes that operate with respect to perceptual context and processes that operate with respect to the temporal episode in which it is occurring.

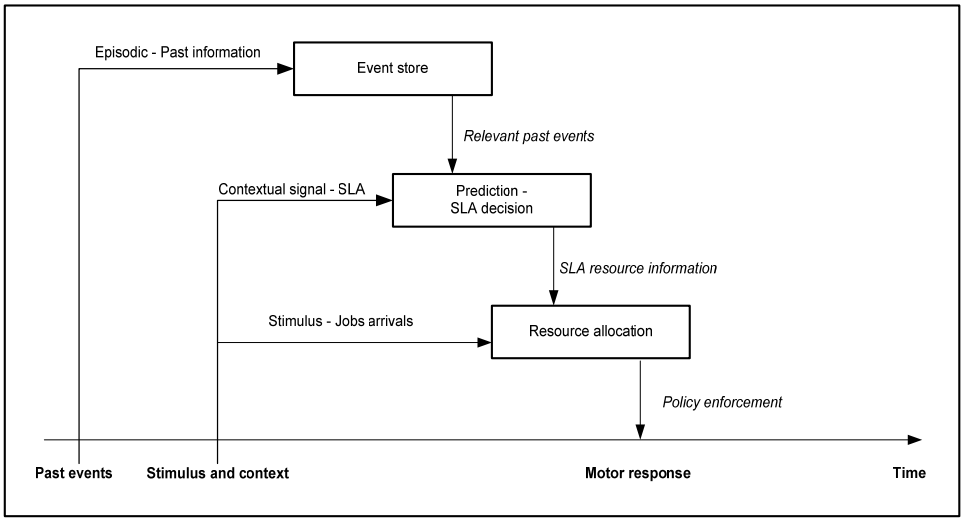

Figure 1 – Architecture based on cognitive control in the prefrontal cortex

The functional model is represented as three nested levels of processing. First, sensory control selects appropriate motor actions. Second, contextual control selects pre-motor representations according to external contextual signals that accompany stimuli, and thirdly episodic control, that is involved in selecting set of stimulus-response actions evoked in the same context and according to events that previously occurred. We applied this architecture to our scenario as illustrated in figure 2. In our system cognitive control is the ability to make decisions for accepting incoming SLA request and performing the necessary actions such as reallocating resources, based on the internal goals of our system which are the top level business objectives. In the first level of processing, the stimuli in our system are the jobs that are constantly being executing on resources and the associated motor responses are the actions taken by the node level components to allocate resources in accordance to the job requirements. The contextual signals in the second level are represented by SLA request, where the context is the requirements of a job and the related bounds defined in the SLA. The third level of processing, the episodic information, corresponds to historical data that is fed in to the lower levels of processing. Using past utilisation patterns, future resource usage levels are predicted based on both current levels of resource utilisation and resource requirements from new incoming SLA requests. This feedback loop provides the system with a self-organised functionality that will adjust resource allocation to current jobs if it is unable to fulfil existing agreements and conversely, it will accept more SLAs if jobs are completing in shorter time that predicted or if a user is not utilising all of the resource that have been committed. Using this model we can modulate lower level resource control through the decision based on higher level objectives.

In the human brain, the more frequently an action is performed in response to a particular stimuli, the quicker the brain is able to provide the correct information to sensory nodes telling it to repeat the same action. For example, if for the first time a small very hot object is thrown at a person, the first response of the brain would be to tell the hand and arm to reach out and catch the object. The heat stimuli would then produce the control and motor signals that tell the hand to drop the object as it's too hot. The next time the object is thrown at the person, the same control and motor signals telling the hand to catch the object would be invoked, but then the episodic control signals would remember from past experiences not to touch the object and would send the corresponding motor signals that would tell the hand to withdraw from its motion. In subsequent occurrences the person would know from past events not to touch the hot object and as such the relevant control and motor signals would be generated that makes the person move away from the hot object. Being able to retrieve and process episodic information more quickly, allows the correct control and motor signals to be selected in response to particular events. We took this in to consideration by maintaining a record of the decisions made and corresponding action performed when an SLA arrived and was accepted. In the future when an SLA of similar characteristics arrives, the system can know what actions to take based on previous responses and on the outcome on the system of those actions.

4. Resource requirements

Grid jobs have widely varying resource usage characteristics. For example, the execution of a job that calculates a genome sequence utilises the CPU, memory and disk resources, whereas a data collection job requires disk and network capacities. Furthermore, the performance of one task on a set of resources differs according to the type of resources in the grid system. While common benchmarking such as MIPS or FLOPS gives us an indication of the capability of a system, these are often simplistic and are not representative of the performance over a wide range of scenarios. They also give no indication how much time it will take to execute a particular job. To capture this non-deterministic behaviour we used an ontological approach as an inference mechanism that enables the understanding of resource commitments required by each SLA.

An ontology [8] is a conceptual vocabulary that that captures the relationships amongst entities within a knowledge domain. The relationships between entities are specified such that the semantics of the domain are richly expressed. We defined the terms in our ontology based on the perspective of being able to measure resource commitments.

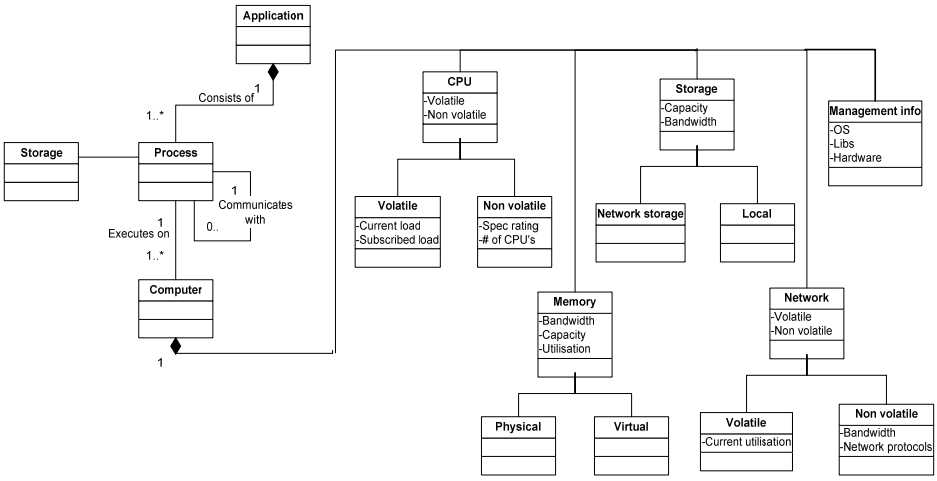

Figure 2 - Application ontology

Figure 2 illustrates our information model. An application can be decomposed in to one or more processes that executes on more or more CPU. An application also has requirements for storage, memory and network bandwidth. In order to ensure accurate data within the information model, each time an application completes its execution and each time an SLA expires, the estimated resource usage was compared to the actual usage of resources.

5. Prototype implementation model

We developed a model using the Repast simulation framework. Repast is a discrete-event based simulation toolkit, predominantly for the design of agent based simulations. SLA requests were generated using Poisson arrival rates and XML as the syntax language. Each SLA is decomposed to a set of resource requirement that is measured as a form of resource commitments. These requirements are input to a prediction model that would estimate if there are enough resources to fulfil the SLA. Predicting the capability of the system is performed continuously with information about past events fed back in to the ontology inference to improve accuracy.

6. Conclusions

Service level management is a crucial component for the development and deployment of future grid services. To accommodate future usage of resources, grid operators require a system that can handle SLA requests and adapt resources to changing demands. Our approach uses a cascade model based on the architecture of cognitive control in the prefrontal cortex with feedback loops that relay past information back in to the system. Using this model we can modulate lower level resource control through the decisions based on higher level objectives and improve the accuracy by feeding information about past events that have occurred in the same context back in to the information model.

References

[1] The UK e-science grid. http://www.escience-grid.org.uk/

[2] IBM Grid Computing. http://www-1.ibm.com/grid/

[3] Sahai A, Graupner S, Machiraju V, van Moorsel A. “Specifying and Monitoring Commercial Grids through SLA”, CCGrid, May 2003

[4] Czajkowski, K., Foster, I., Kesselman, C., Sander, V., Tuecke, S.: SNAP: A Protocol for Negotiation of Service Level Agreements and Coordinated Resource Management in Distributed Systems, submission to Job Scheduling Strategies for Parallel Processing Conference (JSSPP), April 2002

[5] Leff, A., Rayfield, J., Dias, D., “Service-Level Agreements and Commercial Grids” pp. 44-50 IEE Internet Computing July/August 2003

[6] EGEE, http://public.eu-egee.org

[7] Koechlin E., Ody, C., Kouneiher F., The Architecture of Cognitive Control in the Human Prefrontal Cortex. www.sciencemag.org, vol 302 November 2003

[8] T.R Gruber. A translation approach to portable ontology specifications. Knowledge acquisition, 5(2): 199-200, 1993