GFD-I.6
Category: Informational

U. Schwiegelshohn
University Dortmund
R. Yahyapour
University Dortmund
December 2001

Attributes for Communication between Scheduling Instances

Status of this Draft

This draft provides information for the grid scheduling community. Distribution of this memo is unlimited.

Copyright Notice

Abstract

This document describes a set of attributes of a lower level scheduling instance – as found locally on systems – that can be used by a higher level scheduling instance – as found in an Grid environment to interact with remote local scheduling systems. This set of terms provides directions for implementers of new schedulers that are used in Computational Grids.

Contents

1. Introduction

A Computational Grid typically consists of a variety of different resources with different owners. Usually those resources are not exclusively dedicated to Grid usage. For instance, a computer may temporarily be removed from the grid to solely work on a local problem. Many of those single Grid resources use local management systems that often include local scheduling instances.

In general, sites can freely participate in Grid computing by offering resources provided that certain conditions like security requirements are met. The interaction between those grid resources during the execution of a job requires a scheduling layer that uses a different scheduling paradigm than that of local or centralized schedulers. While local scheduling usually involves a single scheduling instance that has access to all system information, grid scheduling requires interaction to remote sites and their local scheduling systems. This suggests the use of several scheduling layers for the grid. Although the details of this scheduling architecture have not yet been decided, it is clear that those layers need to exchange information.

In this context, a distinction is made between ***lower level scheduling instances*** for the local scheduling of resources and ***higher level scheduling instances***, that are used for interaction in coordinated scheduling in a Computational Grid. Both scheduling instances need to work efficiently together in order to make best use of the Grid resources. At this moment it is not clear whether the various scheduling instances in a grid will form a static hierarchy. Therefore, in this document we use the expression *higher level scheduler* to denote a scheduler that is actually querying another scheduler for possible allocations for a query. This other scheduler is called a *lower level scheduler*. Hence, the expression "higher" results from the direction a query takes during the process of scheduling and does not necessarily indicate a fixed hierarchy of schedulers. That is, a higher level scheduler at one moment may as well be a lower level scheduler next during another scheduler communication. Those lower level schedulers may be part of queuing systems, like PBS (see http://www.openpbs.org), LSF (see http://www.platform.com/products/lsf), LoadLeveler (see http://www-1.ibm.com/servers/eserver/pseries/software/sp/loadleveler), NQS (see http://www.cray.com/products/software/nqe), or they may already include scheduling features as those provided by GARA in Globus (see http://www.mcs.anl.gov/qos).

The purpose of this document is the definition of attributes that describe those available features of such a lower level scheduling instance that can be exploited by a higher level scheduling instance. These attributes facilitate the interaction and cooperation between the different levels of a

Grid scheduling system as there may be different local schedulers with different features. Note that the document only focuses on the attributes for scheduler communication without addressing mechanisms for this communication.

Also these attributes do not describe the structure and syntax for the resource description, like, for instance, the minimum number of processors or the amount of memory requested. This information can be accessed through the Grid information service. While this is also an important part of the grid infrastructure it should be specified in a different document, see GWD-GIS-005. Instead only the features offered by the lower level scheduling instance to the higher level scheduling instance are addressed in this document. Consequently, there are no attributes in this document to determine the set of resources a lower level scheduler is responsible for. Further, although this document does not define the interface between the different scheduling layers in detail, it can be seen as the first step in this direction.

Note, that the features described by the attributes are neither an obligation nor a limitation for the design of a lower level scheduling system. For instance, some features are not relevant for certain resource types. Therefore, they are not considered in the corresponding local schedulers.

Further, the presence of an attribute indicates that the specified feature is available on the lower level scheduling instance and can be used by a higher level scheduling instance. On the other hand if such an attribute is not present for a lower level scheduling instance then the higher level scheduling instance can assume that it cannot use the specified feature on this lower level scheduling instance.

Finally, this list is based on scheduling concepts used or discussed today. Future developments may provide additional features of scheduling instances that will require corresponding adaptation of the scheduling attributes. Therefore, this list should serve as a starting point. As a next step following this informational document, there may be a discussion of existing local schedulers with respect to the availability of these attributes and a prototype implementation of a higher level scheduler that can use those attributes.

2. Typical Scenario

In a typical example, the higher level scheduling instance (a grid job scheduler) coordinates the scheduling for multi-site application or helps to select the best resources for a job among different possible resource offers. Typically, this scheduler itself has no direct control over resources. Therefore, it needs to communicate with and appropriately trigger lower level scheduling instances. Those lower level schedulers either control resources directly or have some kind of access to their local resources. However, note that the concept is not restricted to two levels of scheduling instances. Therefore, the lower level scheduling instance can be a local scheduler for a single resource or it may be a scheduling system that manages several resources. If more than two levels of scheduling instances are used then a scheduling instance in the middle needs to collect attributes from possibly several lower level scheduling instances, combine them and provide those combined attributes to a higher level scheduling instance. In this document we do not address the method of combining attributes, as there are several possible solutions to this problem, like a logical AND or attributes associated with a list of resources. The type of the resources handled by lower level schedulers is not restricted to computing resources but may also include other resources as, for example, some network bandwidth that is controlled by a bandwidth broker.

In this document we will use the term allocation for assignments of resources to a request. An allocation is tentative until it is finally executed, that is, until resources are actually consumed. The schedule maintained by some scheduling instance gives information on the planned or guaranteed allocations. Also note that any guarantee of an allocation only refers to the guaranteed assignment of resources, but does not guarantee the completion of a job.

As a simple example assume a Computational Grid that includes the following resources from different institutions:

- Several clusters of workstations
- a visualization cave
- a special database
- a network with bandwidth brokerage that connects the other resources.

On request of an application a Grid job scheduler tries to find a combined allocation that includes a set of 5 workstations, a visualization cave during one stage and the access to a database at another stage. To this end, the Grid job scheduler has to interact with the local management instances of the various resources to find suitable allocations. In order to do its job the Grid job scheduler needs not only information about the available resources but also about the available features of the corresponding local schedulers.

For instance, the Grid scheduler wants a guaranteed completion time of allocation for one stage in order to make an advance reservation for the visualization cave in the next stage. If such a guarantee cannot be provided by any of the corresponding lower level scheduling instances, the Grid scheduler prefers an exclusive allocation to limit the influence of other independent jobs. Finally, if such a feature is also not available, the Grid scheduler may reserve additional resources if it can preempt the job at one stage and migrate it to a less heavily used set of resources, if necessary.

We suggest that this information about the features of lower level scheduling instances can be provided in form of a list of attributes.

3. Attributes of allocation properties

These attributes are useful for a higher level scheduling instance to determine timing, lengths and reliability of allocations.

3.1 **Revocation of an allocation**

This attribute indicates that the local management may revoke an already made allocation while the allocation is still tentative. This may even happen although the higher level scheduling instance has fulfilled all its requirements to keep the allocation. For instance, the resources of the allocation may be withdrawn from Grid usage or may be used to execute another job with a higher priority. Therefore, this attribute indicates that the allocation is not guaranteed. Note that this attribute is independent of any process that must be executed by the user system or the higher level scheduling system to prevent deallocation according to the deallocation policy of a lower level scheduling system.

3.2 **Guaranteed completion time of allocations**

This attribute indicates that the local management guarantees the completion time of an allocation. This does not necessarily mean that the actual allocation is known but only that the system guarantees that the requested resource allocation will be executed before a given deadline.

3.3 **Guaranteed number of attempts to complete a job**

This attribute indicates that the local management guarantees that a specified number of attempts are made to complete a job. However, the completion time is not specified. For instance, such an attribute is useful for the transfer of data over a network.

3.4 **Allocations run-to-completion**

This attribute indicates that the local management will not preempt, stop or halt an application after it has been started. Once started the allocation will stay active on the given resources until the end of the requested time or the completion of the job.

3.5 **Exclusive allocations**

This attribute indicates that the allocation runs exclusively on the provided set of resources. The resources are not time-shared and the executed allocation is not affected by the execution and resource consumption of another allocation running concurrently.

3.6 **Malleable allocations**

This attribute indicates that the resource-set of an application can change the resource set during execution of the allocation, that is, the local management supports the addition or removal of resources from applications during run-time. This modification of the allocation is not controlled by the higher level scheduling instance.

Option: Moldable allocations

This option indicates that the local management can only increase the resource set of an allocation during run-time. In contrast to malleable allocations resources are not taken from the application.

4. Attributes of available information

These attributes allow the higher level scheduling instances to determine which information it may obtain from a lower level scheduling instance and how reliable this information is.

4.1 Access to the tentative schedule

This attribute indicates that the local management returns on request the complete current schedule for present and future allocations. This information can be helpful for the grid job scheduler to determine suitable timeslots that can be used, for instance, for co-scheduling in multi-site computing. There are many alternative options possible for this attribute if the local management does not return the complete schedule as the access to the schedule may be limited due to system policies or restrictions of the local management.

Option: Projected start time of a specified allocation

This option indicates that only the projected start time of a specified allocation is provided, if the higher level scheduling instance has the required access rights.

Option: Partial information on the current schedule.

This option indicates that the lower level scheduling instance provides its complete information on the current schedule. However, as there may be other access to the resources that is not controlled by the lower level scheduling instance, this information may be only partial with respect to the resources.

4.2 Exclusive Control

This attribute indicates that the lower level scheduler is in exclusive control of the resources and no allocations can be scheduled on those resources without using this lower level scheduling instance. This information is useful to determine the reliability of scheduling information.

4.3 Event Notification

This attribute indicates that the lower level scheduling instance supports event subscription. A higher level scheduling instance may subscribe to specific events and use this information for monitoring or rescheduling of allocations. Note that this document neither specifies the event types nor gives a definition of interfaces to query supported event types. This attribute only indicates that such a mechanism is available.

5. Attributes for manipulating the allocation execution

These attributes indicate supported scheduling functionalities of the lower level scheduling instance that can be used by a grid job scheduler to directly modify a running allocation.

5.1 Preemption

This attribute indicates that the local management allows the temporary preemption of an allocation by a higher level scheduling instance. In this case the corresponding application is stopped but remains resident on the allocated resources and can be resumed later. This preemption is not synonymous with the preemption in a multi-tasking system that typically happens in the time range of milliseconds. It only indicates that the local management offers the ability to remotely initiate a preemption of another allocation to e.g. temporary free resources for other usage or to synchronize two allocations on different resources. Also note that this kind of preemption does not necessarily require checkpointing. For instance, if machines of the resource set go down, it may not be possible to resume a preempted allocation.

5.2 Checkpointing

This attribute indicates that the local management supports the checkpointing of a job. A file of a checkpointed job is generated that allows a later continuation from that point. The checkpoint file may also be migratable to other resources but this is not required.

5.3 Migration

This attribute indicates that the local management supports the migration of an application or part of an application from one resource set or subset to another set. This way an application can be stopped at one location and the corresponding data are packed such that the application can be moved to another location and be restarted there. Note that this migration process is initialized and controlled by the higher level scheduling instance. The attribute does not include migration of allocations within the domain of the lower level scheduling instances that are not influenced by the higher level scheduling instance, see malleable allocations. Also, the migration attribute does not necessarily require the presence of the checkpointing attribute.

5.4 Restart

This attribute indicates that the local management supports the receiving and the restart of a stopped and packaged application from another resource set.

Option: Checkpoint restart

This option indicates that a restart is only possible from a checkpoint file, that is, the system does not support migration on the fly.

6. Attributes for requesting resources

These attributes indicate functionality of the lower level scheduling instance that is useful for a higher level scheduling instance to determine which information must be provided when requesting resources and which answers to expect.

6.1 Allocation offers

This attribute indicates that the local management supports the generation of potential resource allocations for a request. For instance, if several resources are capable to fulfill a request, a Grid job scheduler can first query those systems for the allocation and afterwards make its decision to accept the allocation.

Options: Single/multiple offers for a request.

This option indicates that the local management may provide several offers for a request with possible overlapping allocations. For instance, a Grid job scheduler may use this feature for multi-site application where corresponding allocations must be found on different sites.

6.2 Allocation cost/objective information

This attribute indicates that the local management system can return objective or cost information for an allocation. In case of several allocation offers, a Grid job scheduler can, for instance, use this information for the evaluation. The cost/quote for a specified allocation usually relates to the policy that is applied by the lower level scheduling instance. This represents the scheduling objective of the owner of the resource.

6.3 Advance reservation

This attribute indicates that advance reservation is supported according to the proposed advance reservation protocol.

6.4 Requirement for providing maximum allocation length in advance

This attribute indicates that the local management requires an allocation length to be given in advance. Historically, resource requests often have been submitted without additional information on the amount of time that the resources will be used. These programs have been started and run until completion. Current scheduling algorithms as, for example, the backfilling algorithm, require additional information on the maximum allocation length.

6.5 Deallocation policy

This attribute indicates that some kind of deallocation policy for pending allocation applies. Some systems pose requirements that must be met to keep the allocation valid. An example for such a policy is the requirement that an allocation must be repeatedly confirmed until the execution of the allocation. These policies must be further specified to allow the higher level scheduler to keep the allocations.

6.6 Remote co-Scheduling

This attribute indicates that the local management allows co-scheduling where the actual resource allocation and schedule is generated by a higher scheduling instance. This includes the generation/cancellation of allocations on the local schedule by the higher level scheduler. For instance, a Grid job scheduler can use this property to quickly co-allocate resources at different sites in order to fulfill a more complex job request.

6.7 Consideration of job dependencies

This attribute indicates that the lower level scheduler takes dependencies between allocations into account if they are provided by the higher level scheduling instance. For instance, in case of a complex job request the lower level-scheduling will not start an allocation if the completion of another allocation is required and still pending.

7. Security Considerations

Security issues are not discussed in this document. The scheduling scenario described here assumes that security is handled at the point of job authorization/execution on a particular resource. However, it is acknowledged that communication between schedulers should involve appropriate authentication and acknowledgement for protection against service disruption attacks.

8. Authors' Address

Uwe Schwiegelshohn
Uwe.Schwiegelshohn@udo.edu
Computer Engineering Institute, University Dortmund

Ramin Yahyapour
Ramin.Yahyapour@uni-dortmund.de
Computer Engineering Institute, University Dortmund

9. Copyright Notice

This document and the information contained herein is provided on an "AS IS" basis and THE GLOBAL GRID FORUM DISCLAIMS ALL WARRANTIES, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO ANY WARRANTY THAT THE USE OF THE INFORMATION HEREIN WILL NOT INFRINGE ANY RIGHTS OR ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

10. Intellectual Property Statement

The GGF takes no position regarding the validity or scope of any intellectual property or other rights that might be claimed to pertain to the implementation or use of the technology described in this document or the extent to which any license under such rights might or might not be available; neither does it represent that it has made any effort to identify any such rights. Copies of claims of rights made available for publication and any assurances of licenses to be made available, or the result of an attempt made to obtain a general license or permission for the use of such proprietary rights by implementers or users of this specification can be obtained from the GGF Secretariat.

The GGF invites any interested party to bring to its attention any copyrights, patents or patent applications, or other proprietary rights which may cover technology that may be required to practice this recommendation. Please address the information to the GGF Executive Director.