A Framework for Providing Hard Delay Guarantees in the EGEE

P. Kokkinos (1), K. Koumantaros (2) and E. A. Varvarigos (1)

1: Research Academic Computer Technology Institute (RACTI), Patras, Greece

2: Greek Research and Technology Network (GRNET), Athens, Greece

Introduction

Future Grid Networks should be able to provide Quality of Service (QoS) guarantees, in order to support real world commercial applications and complex scientific simulations and computations.

In this work we present a QoS framework that provides:

a) hard delay guarantees to Guaranteed Service (GS) users.

b) fair sharing of resources among Best Effort (BE) users.

By the term "user" we do not necessarily mean an individual user, but also a Virtual Organization (VO), or a single application, using the Grid infrastructure.

Description of the Framework

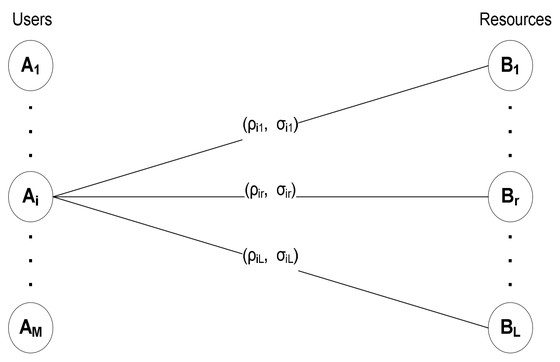
We consider two kind of users: GS and BE users, who generate tasks of GS or BE type, respectively. Also there are various types of resources based on the types of tasks they serve (GS or BE or both) and on the priority they give to each type.

Guaranteed Service (GS) users

The GS users are leaky bucket constrained, and so they follow a (ρ, σ) constrained task generation pattern, which is agreed separately with each resource during a registration phase. Each resource has a Weighted Fair Queuing (WFQ) local scheduler.

ρ_{ir} is the long term workload generation rate, that GS user i will submit to resource r.

σ_{ir} is the maximum size of tasks (burstiness) that GS user i will ever send, in a Weighted Fair Queuing very short time interval, to resource r.



A GS user i registers to a resource r, if conditions (1) and (2) hold.

$$r_{ir} \le g_{ir}(t) = \frac{C_r \cdot w_{ir}}{\sum_{k=1}^{N_r(t)+1} w_{kr}} \quad (1) \qquad J_{ir}^{max} \le J_r^{max} \quad (2)$$

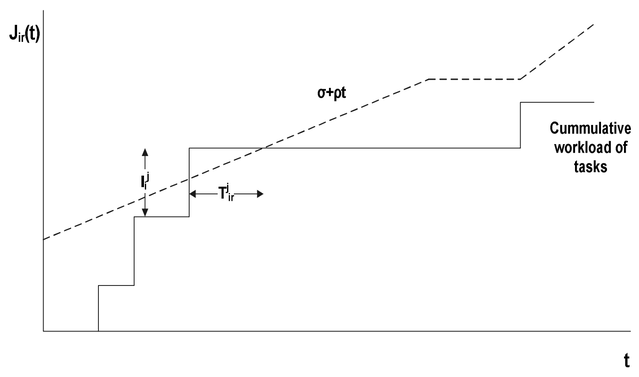
C_r is the computing capacity of resource r, $N_r(t)$ is the number of GS users already registered to resource r at time t, and w_{ir} is the weight of the GS user i for using the resource r, J_{ir}^{max} the maximum task workload user i will ever send to resource r and J_r^{max} the resource's maximum acceptable task workload.

A GS user i must not invalidate the (ρ, σ) constraints, agreed with resource r. So condition (3) must hold.

$$J_{ir}(t) < s_{ir} + r_{ir} \cdot t, \quad \forall t > 0 \qquad (3)$$

$J_{ir}(t)$ is the total computational workload submitted by GS user i to resource r in the interval $[0, t]$.

If a GS task j invalidates (3), then the GS user must locally withhold this task for a time period, denoted by T_{ir}^j, until (3) becomes valid again.



A resource r provides hard delay guarantees to a GS user i, if condition (1) and (3) are valid. The delay bound B_{ir}^j given for a task j is equal to (4).

$$B_{ir}^j \le T_{ir}^j + d_{ir}^j + \frac{s_{ir}}{g_{ir}} + \frac{J_{ir}^{max}}{g_{ir}} + \frac{J_r^{max}}{C_r} \qquad (4)$$

In (4) we also consider queueing delay T_{ir}^j and communication delay d_{ir}^j.

In order for a task j, of GS user i, to be scheduled to resource r, the conditions (5) and (6) must hold.

$$I_i^j \le J_{ir}^{max} \quad (5) \qquad D_i^j \le B_{ir}^j \quad (6)$$

I_i^j the workload and D_i^j the deadline of task j, belonging to GS user i.

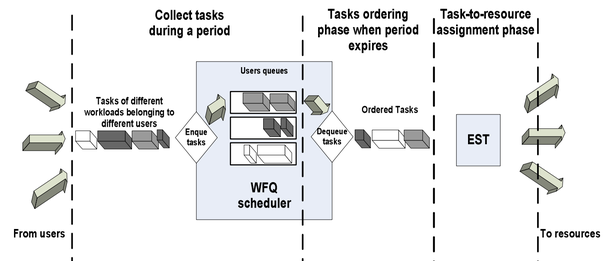
Best Effort (BE) users

For BE users, the framework describes a fair scheduling procedure that provides fairness among users instead of fairness among tasks.

The notion of user fairness is more appropriate for Grids, since the main entities in Grids are not the tasks but the users creating them.

The user fair algorithm, called WFQ/EST consists of two phases: task-ordering and task-to-resource assignment.

User fairness is mainly achieved in the first, task-ordering phase, where a Weighted Fair Queuing (WFQ) scheduler is used. In the second any task-to-resource assignment algorithm can be used.



Resources Types

We consider the following resource types:

a) **GS**, handles only GS users.

b) **BE,** handles only BE users.

c) **GS_BE_EQUA**L, handles equal GS and BE users.

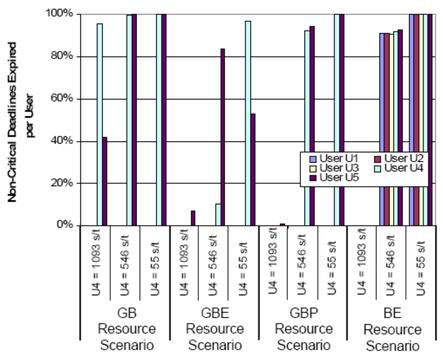
d) **GS_BE_PRIORITY**, handles GS users with higher priority than the BE users.

Simulation Results

We implemented the centralized version of the proposed QoS framework in the GridSim simulator.

In order to obtain realistic simulation parameters, we used the results of a Grid profiling study, of the task inter-arrival times, queue waiting times, task execution times, and data sizes exchanged at the kallisto.hellasgrid.gr cluster (part of the EGEE infrastructure).

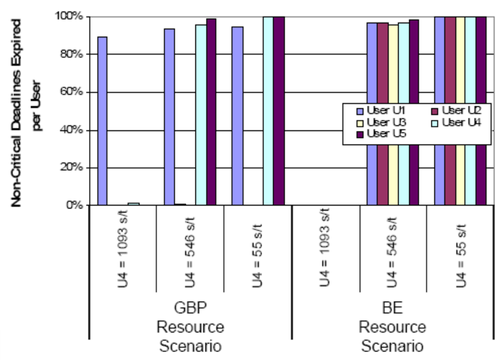
In our simulations we used 3 GS and 2 BE users, 3 resources, and a meta-scheduler. The GS users correspond to the Atlas, Magic and Dteam VOs, while the resources CPU capacities are based on that of the kallisto's cluster.

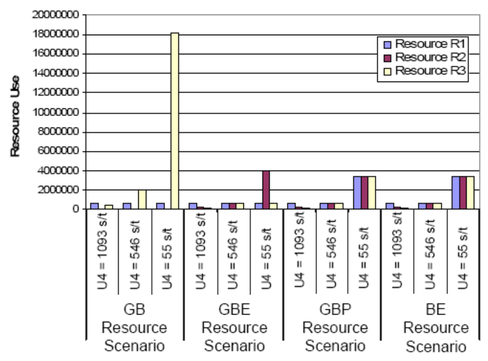


The per user percentage of the number of tasks that miss their non-critical deadlines, for various resource scenarios and task inter-arrival times (in secs/task) of BE user U4.

We observe that in all cases the GS users (U1, U2, U3) do not miss any of their deadlines, verifying that our framework succeeds in providing hard delay guarantees to GS users.

As long as the GS users respect their (ρ, σ) constraints, even with small deviation, our QoS framework succeeds in providing them with hard delay guarantees. Even when the GS users violate their (ρ, σ) constraints, our framework still benefits them.



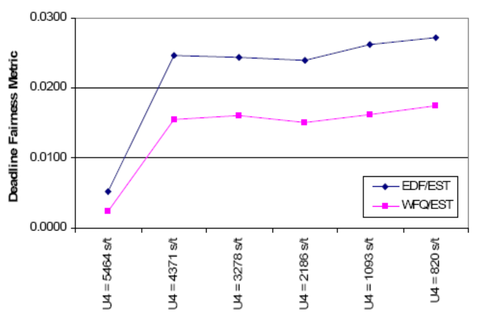


The resource use, for various resource scenarios and task inter-arrival times of BE user U4.

Resource R3 is utilized more in the GB resource scenario, since it handles exclusively BE tasks.

The deadline fairness metric: $\frac{1}{N} \cdot \sum \frac{\left| D_i^j - \max(Y_i^j, D_i^j) \right|}{D_i^j}$

The proposed user fair scheduling algorithm, provides fairness among BE users, while also improving task delay performance.



The framework in the EGEE

The proposed QoS framework can be used in the EGEE. The GS users can be Virtual Organizations (VO) or different User Interface (UI) machines, using the EGEE.

The Computing Elements (CE), capable of serving the GS users should be defined during their installation. These CE will publish to the Information Service of EGEE (Berkely Database Information Index - BDII) not only data regarding their current load but also information needed for the operation of the proposed QoS framework.

When a user wishes hard delay guarantees, then he can ask such a service from the Workload Management System (WMS). The WMS will be responsible for the registration of the user to GS resources and for the scheduling of his tasks to registered resources, capable of executing them before their deadline expires.