Workload balancing model of tasks with deadlines and other QoS requirements, in service-oriented Grid computing environments.

Ignacio Blanquer, Vicente Hernández,
Damià Segrelles, Erik Torres

Universidad Politécnica de Valencia
(UPVLC)

iblanque@dsic.upv.es

Summary

- Introduction.
- Objectives.
- A Use Case.
- Workload Balancing Model.
- Conclusions and Future Works.

Introduction: Quality of Service

- Quality of service is the ability to provide different priority to different applications, users, or data flows, or to guarantee a certain level of performance.
  - Def. from Wikipedia.

- In Computing, it can be seen as success on fulfilling a previou agreed set of requirements fo

Introduction: Performance

- Performance can mean many different things to many different people:
  - Whereas to service providers, performance is a matter of allocating resources for executing as many simultaneous users’ requests as possible, without affecting the perception that users have about the behaviour of a service.
    - Closer to the concept of Efficiency
  - For users, a good understand of performance could be whether or not a service can meet a deadline.
    - Closer to the concept of Response Time.
- A flexible Workload Balancing mechanism,

Introduction: Performance vs. Reliability

- Although concerned with the Response Time, most of the time, the users simply need to get a result before a date
  - The target is not the minimum time but to complete the work earlier than the agreed date.
    - "I have my presentation on Friday, but it is good enough for me to have the results of the test Wednesday the latest".
- Performance is important but reliability is even more important.
  - A service delivering results far before the deadline in 75% of the cases would not be better than a service delivering results on the deadline

Objectives

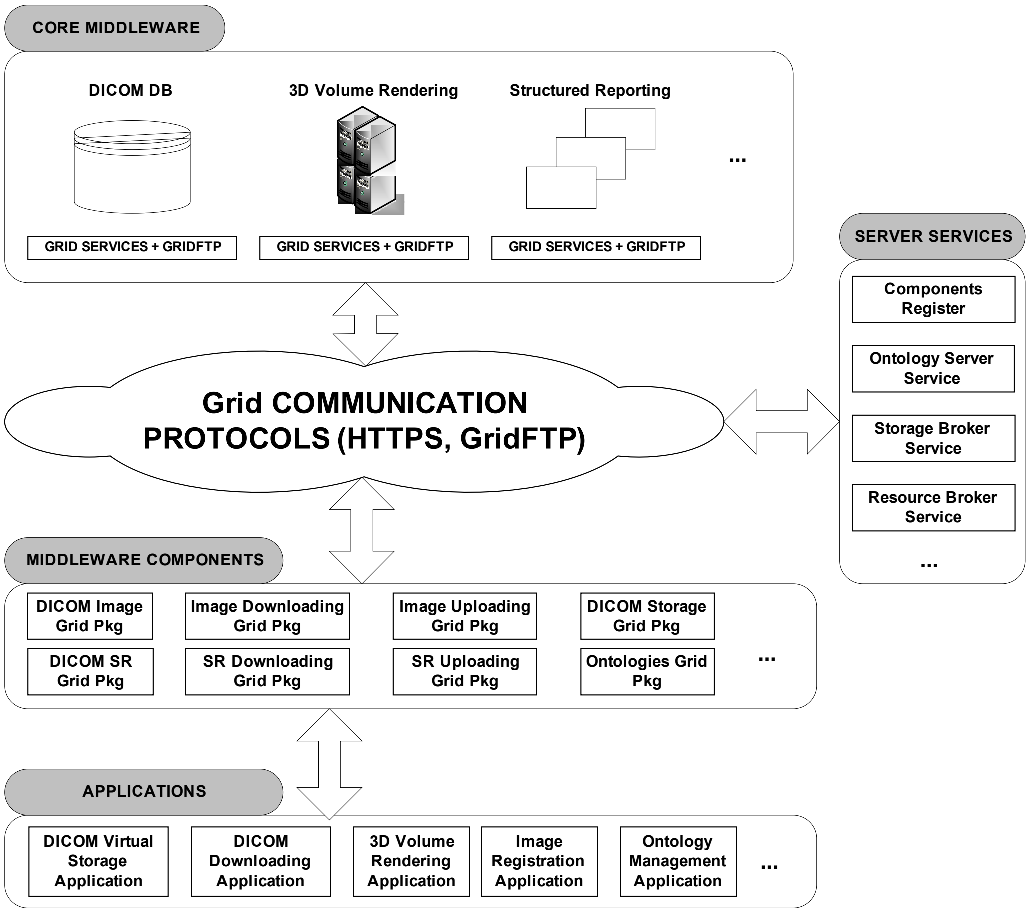
- This work is focused on integrating a new model for workload balancing into service-oriented Grid computing environments, with the aim of ensuring that tasks fulfil their deadlines and other QoS requirements,
  - Availability, Response Time, Throughput, Security, ...
- While improving the efficiency of the resource utilization.

QoS Workload Balancing Model: Concept

- The QoS model proposed is not simply a Load Balancing Mechanism
  - It is P2P and oriented to O(G)SA architectures.
  - It tries to foresee the evolution of the system to avoid overloading.
  - It focus on providing a reasonable degree of predictability for the performance of services.
  - It negotiates with resources the reliability on achieving the SLA.
  - It covers the complete problem of monitoring the resources, matching the requirements and selecting the resources.
  - It is not a simple priority-based schema, but a multi-

TRENCADIS Use Case

- TRENCADIS is a software architecture developed by our group for managing DICOM Objects in OGSA-based Grid environments.



CORE MIDDLEWARE

DICOM DB — GRID SERVICES + GRIDFTP

3D Volume Rendering — GRID SERVICES + GRIDFTP

Structured Reporting — GRID SERVICES + GRIDFTP

...

Grid COMMUNICATION PROTOCOLS (HTTPS, GridFTP)

SERVER SERVICES

- Components Register
- Ontology Server Service
- Storage Broker Service
- Resource Broker Service
- ...

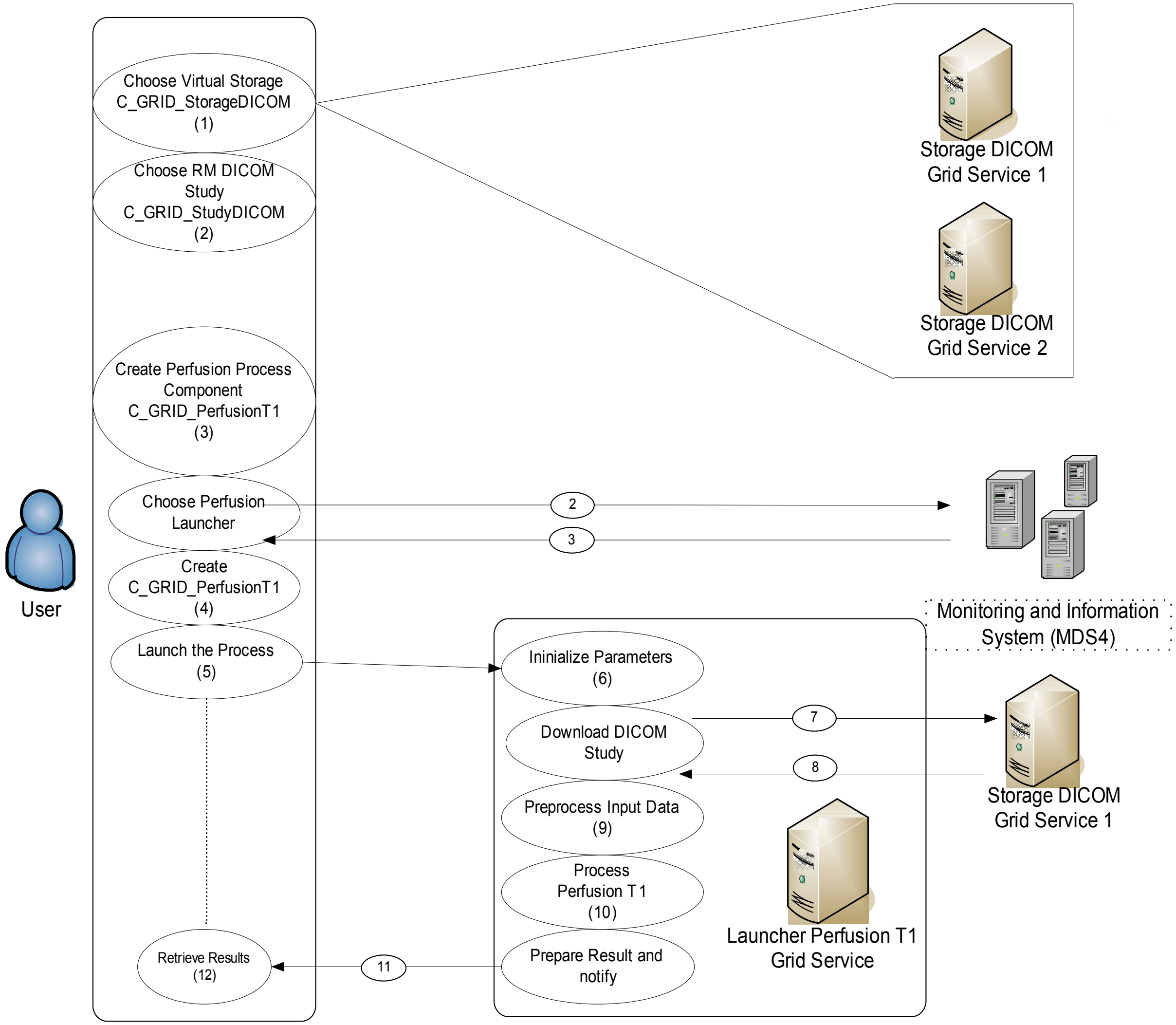
MIDDLEWARE COMPONENTS

- DICOM Image Grid Pkg
- DICOM SR Grid Pkg
- Image Downloading Grid Pkg
- SR Downloading Grid Pkg
- Image Uploading Grid Pkg
- SR Uploading Grid Pkg
- DICOM Storage Grid Pkg
- Ontologies Grid Pkg
- ...

APPLICATIONS

- DICOM Virtual Storage Application
- DICOM Downloading Application
- 3D Volume Rendering Application
- Image Registration Application
- Ontology Management Application
- ...

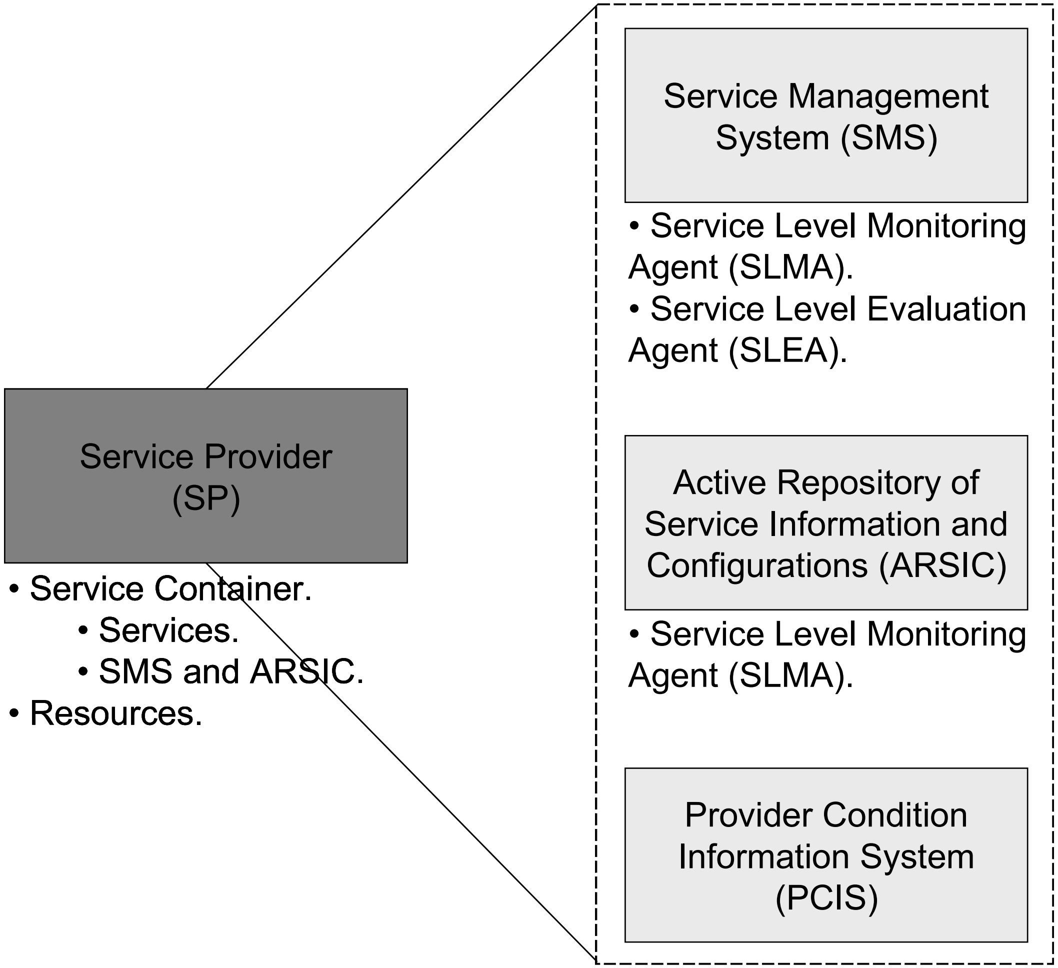
Example: Calculation of parametric images for contrast-enhanced Perfusion T1 in TRENCADIS


Choose Virtual Storage C_GRID_StorageDICOM (1)
Choose RM DICOM Study C_GRID_StudyDICOM (2)
Create Perfusion Process Component C_GRID_PerfusionT1 (3)
Choose Perfusion Launcher
Create C_GRID_PerfusionT1 (4)
Launch the Process (5)
Retrieve Results (12)
User
Storage DICOM Grid Service 1
Storage DICOM Grid Service 2
2
3
Monitoring and Information System (MDS4)
Inicialize Parameters (6)
Download DICOM Study
7
8
Preprocess Input Data (9)
Process Perfusion T1 (10)
Prepare Result and notify
11
Launcher Perfusion T1 Grid Service
Storage DICOM Grid Service 1

Workload Balancing Model: Requirements

- QoS publishing and matching: The major problem in using QoS for workload balancing.
  - Requirement: Specifying, mapping and monitoring QoS.
    - A set of measurable QoS indicators that can be used by clients to indicate their QoS requirements.
    - A Service Level Indicators (SLI) specification, which is used to describe and evaluate the service levels. It is a measure for an entity and period.
    - A Service Container Health Indicator specification, which describe the workload of the service containers at any time.
  - Requirement: Allocating entities and negotiating the compliance of the agreed service levels.
    - An algorithm for workload distribution, which can be used to allocate a service (or a set of services) for

Workload Balancing Model: Stack of Components



Service Provider (SP)

- Service Container.
  - Services.
  - SMS and ARSIC.
- Resources.

Service Management System (SMS)

- Service Level Monitoring Agent (SLMA).
- Service Level Evaluation Agent (SLEA).

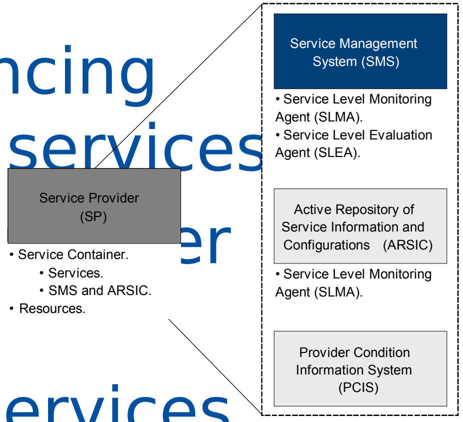
Active Repository of Service Information and Configurations (ARSIC)

- Service Level Monitoring Agent (SLMA).

Provider Condition Information System (PCIS)

Workload Balancing Model: Service Management System (SMS)

- The services use the SMS as the entry point to the stack, subscribing their QoS indicators to the SMS.
  - After that, the services produce SLI (Service Level Indicators) values that are monitored by the SMS.
- The main role of the SMS is to manage the workload.
  - The SMS provides several load balancing mechanisms, which can be used by services
  - The SMS is deployed in the same co[illegible]er as the service.
  - The SMS have a global view of the services



Service Management System (SMS)

- Service Level Monitoring Agent (SLMA).
- Service Level Evaluation Agent (SLEA).

Active Repository of Service Information and Configurations (ARSIC)

- Service Level Monitoring Agent (SLMA).

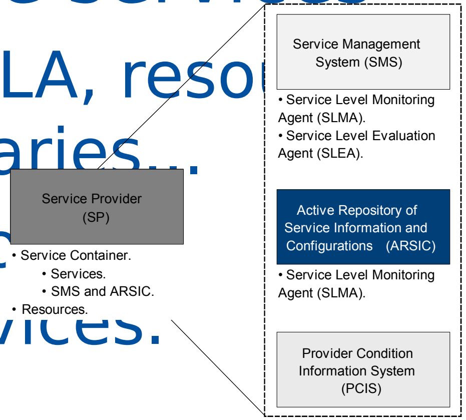
Provider Condition Information System (PCIS)

Service Provider (SP)

- Service Container.
  - Services.
  - SMS and ARSIC.
- Resources.

Workload Balancing Model: Active Repository of Service Information and Configurations (ARSIC)

- Keeps historical record of SLI, and updates the SMS when a service changes its status
  - e.g. when a service reports that can not complain with a QoS requirement that was fulfilled before.
  - It provides the linking among the different SMS.
- Additionally, it provides a repository of configuration issues related to the services
  - QoS requirements of the services, SLA, reso configuration, sw packages and libraries….
  - This enables the architecture with fc dynamically reconfiguring the services.


Service Provider (SP)

- Service Container.
  - Services.
  - SMS and ARSIC.
- Resources.

Service Management System (SMS)

- Service Level Monitoring Agent (SLMA).
- Service Level Evaluation Agent (SLEA).

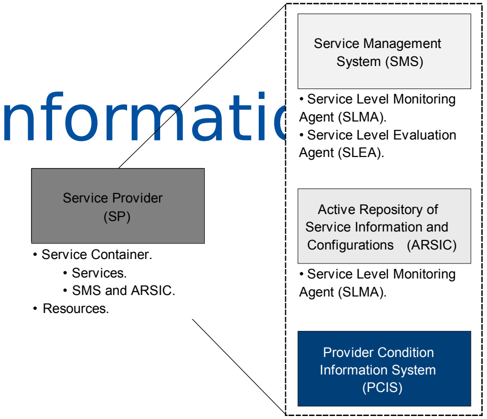
Active Repository of Service Information and Configurations (ARSIC)

- Service Level Monitoring Agent (SLMA).

Provider Condition Information System (PCIS)

Workload Balancing Model: Provider Condition Information System (PCIS)

- Keeps record of the “health” status of the service containers.
  - Understanding “health” as the values for the data used in the SLI.
  - It has the instantaneous value for those indicators (with lower latency than the SMS).
- It provides the SMS with additional information to the SLI to predict the predisposing of a service to deliver a service level.
- The PCIS provides the instantaneous informatio penalise the most selected resources.


Service Provider (SP)
• Service Container.
• Services.
• SMS and ARSIC.
• Resources.
Service Management System (SMS)
• Service Level Monitoring Agent (SLMA).
• Service Level Evaluation Agent (SLEA).
Active Repository of Service Information and Configurations (ARSIC)
• Service Level Monitoring Agent (SLMA).
Provider Condition Information System (PCIS)

Workload Balancing Model: Workload Mechanism

- Resources are classified dynamically into three groups
  - Services strongly predisposed to meet the QoS (and mostly occupied).
  - Services predisposed to meet the QoS (generally free).
  - Services predisposed to violate the QoS (mostly free).
- The SMS takes the decision of the rightmost cluster of resources according to the SLIs the job requirements and the “health” status of the resource

Conclusions

- In this work we have introduced a workload balancing model for optimally allocating Grid services, in order to meet requesters' deadlines.
- The model consists of a set of components which work together for delivering predictable service levels, especially predictable execution times.
- This is a general-purpose model, valid for different kinds of services and

Future Works

- Future efforts must be done in order to extend the use of the model to more complex scenarios.
- Complementary studies are now in progress for BLAST in Grid (BiG), an application developed in the context of the EGEE (Enabling Grids for e-Science) project.
- Our interest in studying this use case is driven by the necessity of introducing

Contact

Universidad Politécnica de Valencia
Camino de Vera s/n
46022 Valencia, Spain
Tel: +34-963879743
Fax. +34-963877274
E-mail: iblanque@dsic.upv.es, vhernand@dsic.upv.es, dquilis@itaca.upv.es, etorres@itaca.upv.es