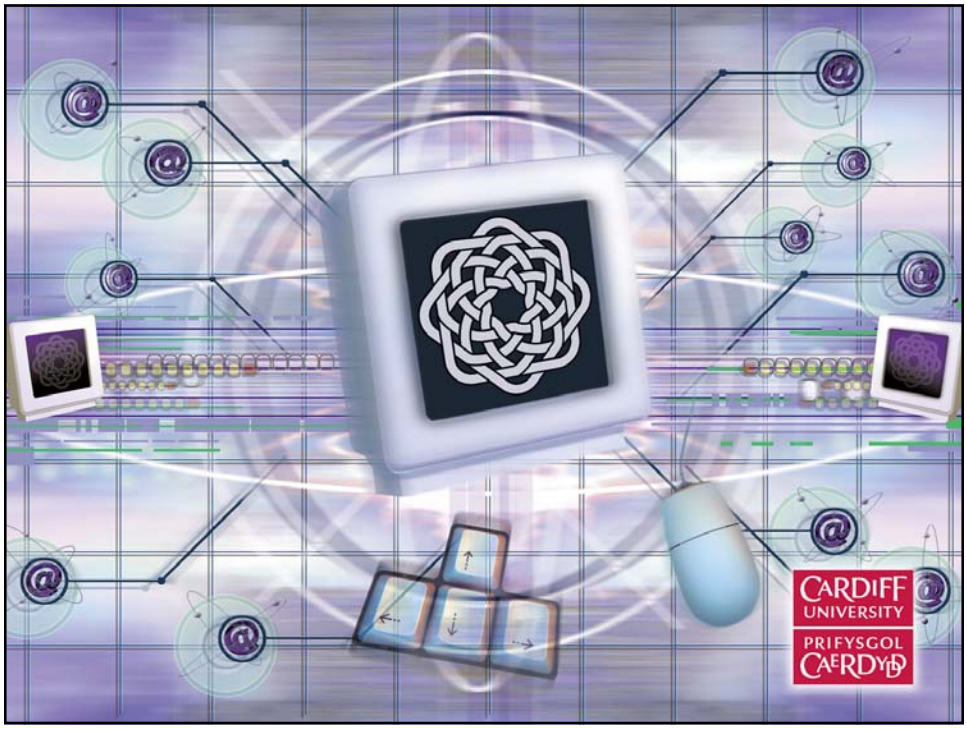


Welsh e-Science Centre

Canolfan e-Wyddoniaeth Cymru

Regional Grid Centre

QoS Adaptation in Service-Oriented Grids

by

Rashid Al-Ali

http://www.cs.cf.ac.uk/User/Rashid/

Rashid@cs.cf.ac.uk

Telcordia™ Technologies

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

Outline

- QoS Management
  and why it is needed in Grid Computing
- Objectives
  of QoS Model and G-QoSM architecture
- Adaptation Strategies
- Implementation Status
- Conclusion

QoS Management

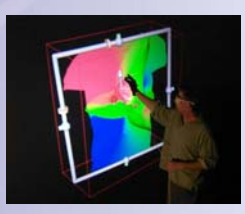
- QoS management:
  - Covers a range of activities - from resource specification, selection and allocation, through to resource release.
- A QoS system should:
  - Specify QoS requirements
  - Map QoS requirements to resource capability
  - Negotiate QoS with resource owners
  - Establish contracts/SLAs with clients
  - Reserve and allocate resources
  - Monitor parameters associated with QoS sessions
  - Adapt to varying resource quality characteristics
  - Terminate QoS sessions

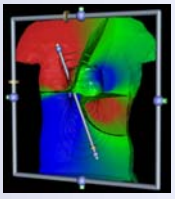
When is QoS needed?

- Interactive sessions
  - Computation steering (control parameters & data exchange)
  - Interactive visualization (visualization & simulation services)
- Response within a limited time span
- Co-scheduling or co-location support

- **Application QoS**
- User perception, response time, appl. security, etc.
- **Middleware QoS**
- Comp., Memory and Storage
- **Network QoS**
- BW, Packet loss, Delay, Jitter

From SCIRun, University of Utah





Types of QoS Services and Adaptation in G-QoSM

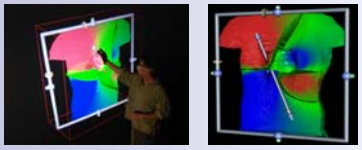
- Guaranteed QoS
  - constraints to be exactly observed
  - SLA is precisely/exactly defined
  - adaptation algorithm/optimization heuristics
- Controlled-load QoS
  - some constraints may be observed
  - Range-oriented SLA
  - optimization heuristics
- Best-effort QoS
  - any resources will do
  - no adaptation support

Why is Adaptation Required?

- In environments with shared resources, e.g. Grids:
  - Resources get congested or even fail !!
    - May lead to undesirable results !!
- Protection is needed for users with contracts (SLAs)
- Providing more capacity than is immediately needed is important – to cater for demand surges or provide support in disaster situations
  - However, spare capacity must be efficiently utilized !!

Our adaptation approach aims to solve this problem

From SCIRun, University of Utah



Literature Survey on Adaptation

- In the context of multimedia applications (Hafid et al.):
  - Based on a user profile, QoS manager considers system offers
  - QoS manager selects an optimal offer called (user offer)
  - During playback of MM document, if network/server gets congested the QoS manager considers alternative (system offer)
  - If resources are reserved for the new offer, then the QoS manager automatically changes to the new system offer, i.e. adaptation
- In the context of Grid computing (Foster et al.):
  - Designed adaptive control system
    - *Actuators* that permit online control
    - *Sensors* that permit monitoring of resource allocation
    - *A decision procedure* that allows entities to respond to sensor information by invoking actuators (adaptive)
  - Example: a loss rate sensor might acquire information from an edge router. Then a decision procedure adapts the network reservation based on the information obtained from the sensor using Globus/GARA Create/Modify primitives.

G-QoSM

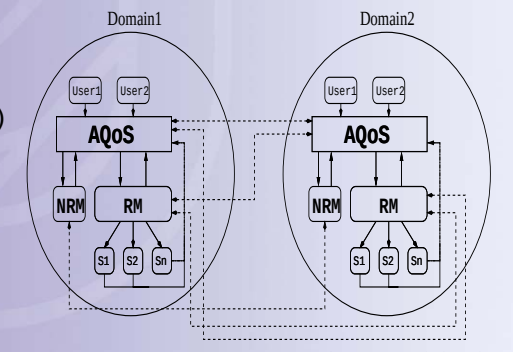
Providing a middleware (G-QoSM) for QoS management in service-oriented Grids, which builds on OGSA

- Supporting service & resource discovery using QoS properties
  - Extended UDDIe
  - Globus MDSv3.0 (currently looking at Service Data Elements to encode QoS properties)
- Providing QoS guarantees and establishing SLAs
- Providing QoS management
  - Monitoring resource unitization
  - SLA conformance
  - Adaptation strategies

In G-QoSM, *"QoS provides assurance on quantitative characteristics (e.g. Network & Compute QoS) and qualitative characteristics (e.g. service reliability) for execution of a Grid service."*

Main Components of the G-QoSM

- AQoS
  - Interacts: users, RMs
  - SLA Negotiation
  - Adaptation
- Resource Managers (RMs)
  - Globus/GARA - DSRT: reserve and bind services
  - UDDIe: registry service to publish & discover
  - NRM (Diffserv BB) adm. contrl, configures routers to provide QoS, negotiates SLAs at network level



Adaptation Scenarios

In G-QoSM: adaptation is the activity by the AQoS broker, in terms of resource allocation; (1) to compensate for QoS degradation or (2) optimize resource utilization

- New Service Request
  - When there are insufficient resources
  - Adaptation is used to free resources
- Service Termination
  - When a service successfully completes execution
  - Adaptation used to upgrade QoS of other current services:
    - Resources will be given to services which had their QoS reduced as a result of a new service request
    - Services that are not currently receiving 'best' QoS
- QoS Degradation
  - When QoS falls below the specified QoS in the SLA
  - Adaptation used to restore degraded QoS

Adaptation Solutions

- **Resource Allocation Optimization Heuristics**
  - Optimize resource utilization, by increasing the number of requests admitted
- **Adaptation Algorithm**
  - Reserve adaptive capacity to be used during congestion or resource failure
    - Resources never under-utilized (dynamic property)
    - A minimum resource capacity (left to the systems administrator) is reserved for 'best effort' – users with no SLA can also use resources

Resource Allocation Optimization

- Aim: optimize resource utilization while maximizing cost.
- If we define a set: $QoS = \{ a_1, a_2, ..., a_n \}$, where each a_i represents a different parameter of interest such as CPU, network bandwidth, etc.
- Then we could compare sets: $QoS_x = \{ a_{1x}, ..., a_{nx} \}$ and $QoS_y = \{a_{1y}, ..., a_{ny} \}$
- QoS values specified in SLA as: $a_y \leq a_i \leq a_x$ Or $a_i=\{ x, y, z \}$
- Assume the existence of a cost function: $Cost(a_i) = c_i * a_i$
- Then: $Service\_ Cost\ (QoS) = \sum_{i=1}^{n} (ci * ai)$
- The proposed heuristic: $Total\ Cost = max \sum_{i=1}^{n} (Service\_Cost(QoSi))$
  n is the total number of active services

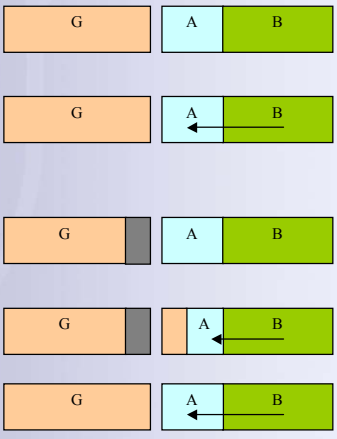
The QoS broker implements this heuristic
by varying the resource quality selection, based on the agreed-on SLA,
aiming to optimize resource utilization and maximize overall cost.

This heuristic could be mapped to the Generalized Assignment Problem (GAP)

Basic idea of the Adaptive Algorithm

Aim: compensation for QoS degradation for 'guaranteed' class only

- Assume capacity$_{Total}$: $C= C_G + C_A + C_B$
- *'best effort' can uses the adaptive capacity, as long as its not used by the 'guaranteed'*
- *When QoS degrades for 'guaranteed'*
- *Then adaptive is utilized to compensate for the degradation*
- *'best effort' can still utilize the remaining capacity of the adaptive, as long as its not used by the 'guaranteed'*
- *When the congested capacity is restored, the adaptive capacity can be used entirely by the 'best effort'*



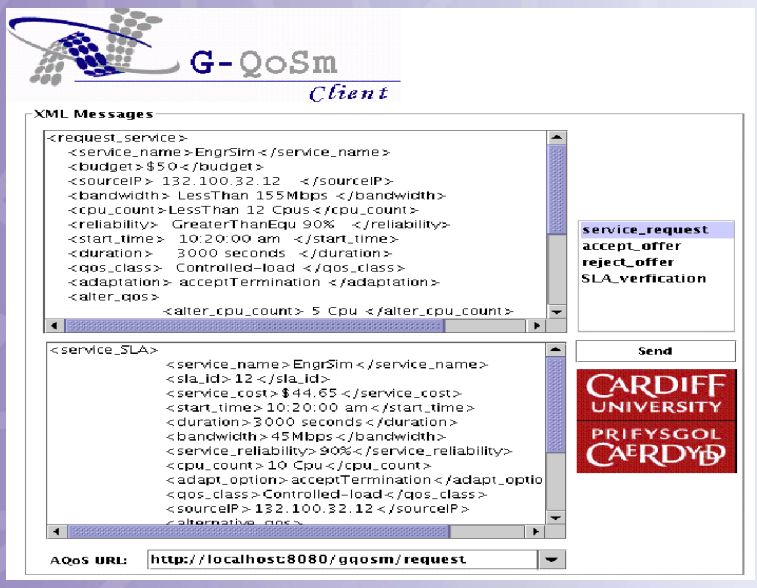
o Before invoking the adaptive function:
  o Ensuring that the request at time (t) $\leq$ the agreed upon in the SLA
  o Ensuring that the total capacities within all SLAs at time (t) $\leq C_G$

Adaptation Algorithm

- Aim: compensation for QoS degradation for 'guaranteed' class only
- Assume a total capacity: $C = C_G + C_A + C_B$
  where G, A & B denotes 'guaranteed,' 'adaptive' and 'best effort'
- ***Adapt***(c(u,t), g(u))
  Net capacity $N_G(t) = C_G(t) - \sum_{u \in G} g(u)$
  If $N_G(t) < 0$; (guaranteed can't be honoured)
  Then ADD ($\sum_{u \in G} g(u)$ - $C_G(t)$) from A to G
  ADD ($C_A(t)$ – [$\sum_{u \in G} g(u)$ - $C_G(t)$]) from A to B
- Before invoking the adaptive function:
  – Ensuring that the request at time (t) ≤ g(u) --- (SLA)
  – Ensuring that $\sum_{u \in G} g(u) \leq C_G$

Implementation

- The implementation test-bed is based on:
  – Linux RedHat 7.2
  – Globus toolkit v2.0
  – J2SDK v1.4.0
  – UDDIe
  – Apache tomcat application server
  – GARA/DSRT
  – NRM (Diffserv Bandwidth Broker)



A client interface screenshot, showing the client entered a 'service_request' and the AQoS broker replied with a service offer

A screenshot, showing activities undertaken by the AQoS broker

Conclusion

- The need for service agreements to enforce QoS
- As more Grid providers come on-line, the role of adaptation becomes more important
- Adaptation is also useful to enable commercial use of Grid services
  - Whereby a service provider can offer services in different classes
  - Learn from existing experience of multimedia services

To address these challenges, we propose the G-QoSM:

- Service discovery using QoS properties
- guarantees QoS at the 3 levels and establishing SLAs
- Support for QoS adaptation
- Implementation using GARA/DSRT, NRM/Diffserv BB, and UDDIe
- This summer – aim to integrate with Java CoG kit and a particular scientific Grid application domain at Argonne National Laboratory