HAL
open science

An Advisory QoS Service for GRID Based Learning Environments

Martin Bateman, Collin Allison, Ross Nicoll

To cite this version:

Martin Bateman, Collin Allison, Ross Nicoll. An Advisory QoS Service for GRID Based Learning Environments. 4th International LEGE-WG Workshop Towards a European Learning Grid Infrastructure Progressing with a European Learning Grid, April 27-28, 2004, Stuttgart, Germany, 2004, Stuttgart, Germany. pp.5. hal-00190338

HAL Id: hal-00190338

https://telearn.hal.science/hal-00190338

Submitted on 23 Nov 2007

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

An Advisory QoS Service for GRID Based Learning Environments

Martin Bateman, Colin Allison, Alan Ruddle, Ross Nicoll
School of Computer Science
University of St Andrews
{mb, ca, alanr, jrn}@dcs.st-and.ac.uk

This paper argues that the assumption of end-to-end QoS provision for use in Grid based applications is unsafe. We present an architecture which provides an adaptive QoS service for GRID applications which are prepared to adapt rather than assume the best and then suffer the worst. This is particularly relevant to Internet based GRIDs. An example application from the field of collaborative learning environments is given, featuring adaptive QoS for Real Time Protocol (RTP) based audio and visual conferencing.

1. QoS AND THE GRID

There is growing evidence of an assumption relating to the Grid, e.g as in [1], that Quality of Service (QoS) provision, specifically resource reservation, will be universally available. This assumption is currently unsafe as resource reservation is only available on a limited basis in various experimental networks, such as[2].

There is the argument that resource reservation will eventually be available on the Internet and therefore available for use within Grid based applications. However this means that the majority of the routers which make up the Internet would have to be upgraded in order for the wide deployment of resource reservation. This seems unlikely since there is currently no consensus on the form that QoS provision should take and also the monetary cost of such an upgrade programme. Secondly, there is some evidence [3] that resource reservation is only effective for certain network traffic patterns. As the future traffic patterns of the Internet are very difficult to predict, (who, in 1998 would have said that 60% [4] of the traffic on the Internet would be made up of peer-to-peer rather than web traffic), it is doubtful that the investment needed to upgrade all the routers to be QoS aware will be available. As the second biggest user of Internet bandwidth is World Wide Web traffic there is little need for QoS provision for normal web browsing.

Since resource reservation is not a viable option for QoS management and that, within systems that impose specific QoS requirements, it is often the case that what they are actually looking for is consistency in the available network resources rather than absolute QoS guarantees, we propose that a method which is able to predict the likely network path characteristics is a pragmatic alternative to reservation of network resources.

We present an architecture which is able to provide predictions on network load over given paths, as well as an example application which is able to use this architecture to better provide QoS for its' users.

The example we give is the introducation of telepresence into a collaboratibe learning environment called Finesse [5]. This particular application is interesting, as there are QoS requirements of both web-based and real-time traffic to take into account.

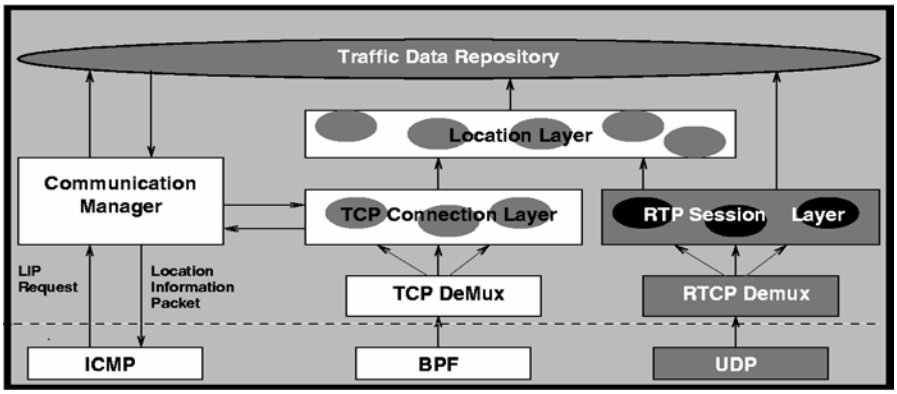
2. ADAPTIVE QoS FOR RTP-BASED SERVICE

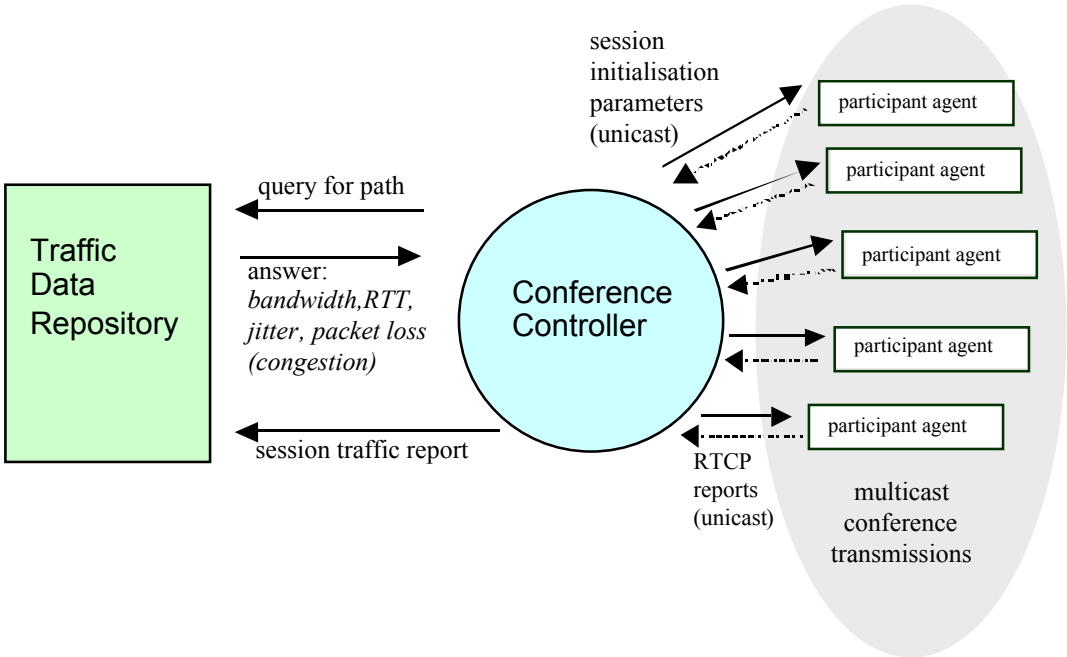
The general benefit of an adaptive system is that in the scenario where bandwidth is plentiful, the system can make full use of the available resources, thereby ensuring a high QoS for the participants. On the other hand, when bandwidth is seriously constrained, a lower QoS, that does not waste the constrained resource or act unfairly in relation to competing traffic, should be specified. At what point in time should an adaptive system react to changes in infrastructure QoS? Research [6] suggests that *consistency of quality* is often more important to users than the actual quality of service achieved. For example, if during the lifetime of a video session, the achievable frame rate

varies between two bounds, it is arguably better to transmit at the lower bound throughout the session rather than try to dynamically optimise the quality.

To achieve consistency it is necessary to know at the start of a session what the network conditions are likely to be for the duration. Previously, it has been envisaged that this problem could be addressed by combining admission control with resource reservation [7] to provide a guaranteed QoS for the duration of a session. Assuming (realistically) that such facilities are unavailable, what can be achieved in their absence?

The approach described here is premised on the assumption that the QoS available on routes from identifiable access points to other participants will not change dramatically from session to session, although they may evolve. This extends earlier work on a TCP Location Information Server (LIS)[8] (See Figure 1). Firstly, statistics are collected about the conditions experienced by all the connections to a conference, and stored in a common repository. These are analysed with a view to making predictions at the start of subsequent conferences about the conditions that are likely to prevail during their lifetimes.





The predictions are used to initialise relevant parameters, so that a QoS appropriate to the network connections can be specified at the start of, and maintained for the duration of the conference. This can be achieved without expert user intervention, which is an important consideration where the participants in a conference are not expert users.

This adaptive approach also allows for improvements in network infrastructure to automatically result in higher quality connection parameters being selected at conference startup. For example, if a participant who normally connects from home upgrades from a modem to ADSL then this will be reflected in improved QoS statistics and higher quality parameters will be used when initialising the codec for that connection in future.

The monitoring of RTP traffic by an LIS has a number of benefits. It widens the base of traffic from which congestion information is derived thereby increasing the proportion of paths for which such information can be made available to TCP sources. This will increase the accuracy of some predictions, and consequently the fairness of network resource sharing, leading to a better understanding of the potential reductions in latency experienced by users. By also allowing multimedia applications to make the best use of available resources the QoS experienced by users of applications that use RTP can be improved. Finally, by tailoring resource utilisation more closely to availability, congestion can be reduced.

3. A CONFERENCE CONTROLLER ARCHITECTURE

The Conference Control Architecture (CCA) shown in Fig.2 is an example of an adaptive application that can utilise feedback from the Traffic Data Repository described above. The CCA in turn provides a service to a learning environment, enabling the dynamic constitution of an RTP-based conference that is contextualised for each participant in the QoS timeliness dimension.

3.1. Conference Controller

At the initiation of the conference the CC queries the TDR for information on past traffic conditions on the network paths to be used. From this, a prediction is made on how the network path will behave during the current session. A policy, which has embedded knowledge of the physiological needs of audio, video and the end-user, is used to suggest the configuration of the audio/video components for use in the session.

For the duration of the conference, the CC monitors the entry and exit of participants and receives reports from PAs concerning the network conditions. In the case of a strong mismatch between expected and experienced conditions, it may be necessary to adjust bandwidth allocation and other parameters. When this happens, update instructions are sent to each PA, using the Lightweight Reliable Multicast Protocol. At the end of the conference, the CC generates a report to the TDR of the network conditions and traffic characteristics between each of the participants. This report allows the repository to determine the accuracy of its predictions and to update the data held on the utilised paths.

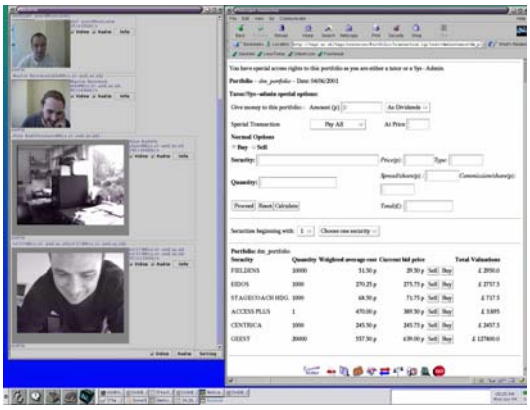


Figure 3: A CCA Enhanced Learning Environment

3.2. A CCA Enhanced Learning Environment

The components of the CCA have been implemented as a proof-of-concept system to enhance Finesse. The Participant Agents are Java Media Framework-based applets.

This makes it possible to allocate conference sessions to groups, in the same way that they are allocated Portfolios and Notebooks. Let us consider an acceptable quality video connection: an H.263 video codec set to 176 x 144 display size and 24 frames/s requires 120Kb/s; an audio connection using LPC encoding at 8 Khz mono requires 5.6Kb/s. per second. Given these parameters a conference would require a reasonable total of 125.6Kb/s per sender. This is slightly less than the bandwidth available over an ISDN-2 home circuit (128Kb/s) and well within the 512Kb/s promised by ADSL.

Fig. 3 shows the Finesse environment (the transaction page is showing) expanded to include an allocated conference resource. The desktop machines taking part were equipped with a mixture of relatively inexpensive frame grabbers and cameras. The QoS for this environment is unusual in that the share data for the shared objects (portfolios) has to be both timely *and* reliable. However, the delay sensitivity for share data can be specified in terms of minutes rather than seconds, or fractions of seconds. The participants in the test sessions were able to chat and use visual cues from each other's telepresence to talk about portfolio maintenance decisions.

3.3. Multicast Issues

Unfortunately, there is extremely limited support for multicast outside of the research environment. The MBone, which accounted for most of the interconnected multicast sites, has effectively ceased to exist. IPv6, includes native support for multicast, however many people, at least in the USA, are now suggesting that moving over to IPv6 is not worthwhile [9] [10], leaving the protocol with an uncertain future.

Such problems can be worked around by implementing multicast at the application instead of network level. There are tradeoffs for this, as dedicated routers typically have higher uplink bandwidth than peers downlink of them, and are better placed to duplicate multicast packets efficiently. Certainly, dedicated routers can never have less bandwidth or worse positioning than a peer which is downlink of them if packets are routed optimally.

For application level multicast to work well, the Conference Controller must take into account network conditions between peers, and decide on how packets will be routed. Other applications may also negatively affect performance. Obviously this adds a significant degree of further complexity to the policy used to determine codec configuration.

However there is an advantage to application level multicast; while with network level multicast all peers are sent exactly the same media streams, with application level the peers with more bandwidth can receive higher bit rate streams, and then transcode them down for sending to peers with less bandwidth. This allows for a better QoS than might be possible otherwise, for conferences with a significant difference in the peers' connection speeds.

4. CONCLUSIONS AND FUTURE WORK

This paper has argued that the assumption of QoS provision in Grid based applications which are in general use on the Internet is an unsafe one. With limited support for end-to-end QoS it is better to use measures which predict the likely QoS provision that and adapt an application to it.

Within this scope we have detailed how to provide a QoS aware service to a Grid conferencing service so that it can be adaptive, and to that extent, contextualized for each participant.

We have successfully integrated a prototype of this system into an existing learning environment (Finesse). This preliminary study has shown the feasibility and promise of core mechanisms for achieving adaptive, dynamically constituted conference sessions. Work now needs to be done on making this QoS advisory service more automated and available for use in any Grid-based collaborative learning environment.

REFERENCES

[1] Al-Ali, R. J., Rana, O. F. and Walker, D. W. (2002) G-QoSM: Grid Service Discovery Using QoS Properties in *Computing and Informatics Journal, Special Issue on Grid Computing*, vol. 21, pp. 363-382

[2] Teitelbaum, Hares, B. S., Dunn, L., Neilson, R., Narayan, R. and Reichmeyer, F. (1999) Internet2 qbone: Building a Testbed for Differentiated Services. *IEEE Network*, vol. 13, pp. 8-16

[3] Breslau, L. and Shenker, S. (1998) Best-effort versus reservations: a simple comparative analysis. Presented at ACM SIGCOMM'98, Vancouver, British Columbia, Canada

[4] Briere, D. and C. Bacco, C. (2003) Peer-to-peer traffic - friend or foe in *NetworkWorldFusion*

[5] Michaelson, R., Helliar, C. V., Power, D. M. and Allison, C. (2003) Group Work and the Web: FINESSE and TAGS. *Learning and Teaching for Business: case studies of successful innovation*, pp. 122-137

[6] Bouch, A and Sasse, M. A. (2000) The Case for Predictable Network Service presented at MMCN'2000, San Jose 2000

[7] Braden, B., Zhang, D. Estrin, S. and Jamin, S. (1996) Resource ReSerVation Protocol (RSVP) -Version 1 Functional Specification. IETF Working Document, IETF IETF

[8] Ruddle, A. (2000) A Location Information Server for the Internet presented at International Conference on Communications and Computer Networks, Las Vegas.

[9] Charney, B. (2003) U.S. shrugs off world's address shortage in *CNET News.com*

[10] Garfinkel, S. (2004) Internet 6.0 in *Technology Review*