



DataGrid

PROPOSAL FOR JOINT COLLABORATIVE INVESTIGATIONS BETWEEN DATAGRID WP7 AND DANTE

WP07: NETWORK SERVICES

Document identifier:	DataGrid-WP7-TR7-2.0304.0.6
Date:	29/04/2002
Work package:	WP07: Network Services
Partner(s):	PPARC, NIKHEF, CERN, INFN
Lead Partner:	CNRS
Document status	DRAFT
Author(s):	T. Ferrari, R. Hughes-Jones, P. Primet
Contact:	R. Hughes-Jones @man.ac.uk
File:	DataGrid-WP7-TR7-2.0304.0.6.doc

Abstract:

This document describes a collaboration between WP7 of the EU DataGrid project and Dante to study the operation of a number of network transport technologies over the GEANT core.



CONTENT

1. INTRODUCTION	3
1.1. EXECUTIVE SUMMARY	3
1.2. TIME SCALES	3
1.3. THE ROLES OF DANTE AND WP7	3
1.4. APPLICATION AREA	3
2. INTRODUCTION TO THE PROPOSAL	4
3. DISCUSSION OF THE TESTS	6
3.1. TCP PROTOCOL EXPERIMENTS	6
3.1.1. Latency and Packet Jitter	6
3.1.2. Throughput	6
3.1.3. Packet loss frequency	7
3.1.4. Packet loss pattern	7
3.2. DEMONSTRATION OF SUSTAINED TCP TRANSFERS	7
3.3. TESTS AND PILOTING OF QOS SERVICES	7
3.3.1. Tests with IP Premium Service on the Core Network	7
3.3.2. Tests with Scavenger Service on the Core Network	8
3.3.3. End to End tests with IP Premium and BioMedical applications	8
3.3.4. Potential Future Investigations	8
3.4. TEST OF RELIABLE MULTICAST PROTOCOLS	9
4. REQUIREMENTS	10
5. CONCLUSIONS	11

1. INTRODUCTION

1.1. EXECUTIVE SUMMARY

We propose a collaboration between WP7 of the EU DataGrid project and Dante to study the operation of a number of network transport technologies over the GEANT core. We propose experimental activities in the following areas:

- Analysis of TCP performance when run over a wide area high-speed network infrastructure
- Demonstrate high throughput data flows for times similar to that will be used in moving bulk data in the Grid environment.
- Pilot the IP Premium service including making end to end tests with biological use case applications.
- Test and assist with the definition of a Scavenger or “less then best efforts” service.
- Explore the use of reliable multicast for file replication.

We propose the use of end-user equipment located in the GEANT PoPs to run as traffic generators, and analysis and monitoring points for the test traffic.

We also propose to share the network performance analysis and measurements collected by the WP7 DataGrid network monitoring work, including the data from WP7 RIPE boxes installed at the edge of several EU DataGrid sites.

1.2. TIME SCALES

Technical discussions with Dante have estimated that the tests would require equipment in the Dante PoPs for about 4 months after installation. However, as detailed in Section 3, testing would not be continuous and would be done in full collaboration with Dante. It is also important that the data is analysed and understood as the tests proceed, thus allowing time for clarification or repetition.

1.3. THE ROLES OF DANTE AND WP7

The Dante engineers and WP7 members would be the researchers in this collaboration. We propose the following general roles:

- Installation and de-commissioning of the WP7 equipment at the PoPs would involve both WP7 and Dante engineers.
- The Dante engineers would perform operational logistics. This might include:
 - Agreeing the timing of various tests.
 - Ensuring the correct test equipment was connected as agreed.
 - Checking the configuration of the PoP routers.
- The tests and initial analysis of the data would be performed by WP7
- Interpretation of the data would be done jointly, as would the production and review of papers and reports.

1.4. APPLICATION AREA

We believe that this collaboration will provide useful experience and benefits to Dante, the European NRNs, the EU DataGrid project and will also be directly applicable to other Grid projects.

2. INTRODUCTION TO THE PROPOSAL

This document details a test plan for a collaboration between the Networking Workpackage (WP7) of the EU DataGrid project and Dante to explore the use of a number of transport technologies over the Geant core. Our proposal is within the framework of the Collaboration Agreement that has been established between the EU DataGrid project and DANTE at INET2001 on June 6th 2001.

Technical discussions of this proposal have already taken place, and the exchange of technical specifications, working documents and reports have promoted effective cooperation between Dante and EU DataGrid project. Representatives of DANTE have attended the WP7 regular meetings since the end of 2001.

Many of the middleware components of distributed systems, like authentication, database replication and the exchange of jobs and input/output data, require reliable transmission among grid nodes. The grid nodes involved include the computing elements, storage elements, and the information servers. The most widely deployed transmission protocol is TCP. One of the purposes of the DataGrid project is to implement a working distributed Grid computing system spanning Europe and serving a large pool of applications in a range of different scientific areas, like high energy physics, earth observation and biology. The DataGrid grid testbed will be also interconnected with remote grid platforms in the US, so the performance of applications over long-distance links is of primary importance.

According to our estimation, the initial deployment of grids and user applications, for particle physics alone, will produce a very large volume of data, with at least 200 Mbps of continuous and/or intermittent traffic. For this reason, the optimisation of TCP both in terms of configuration and stack implementation is particularly important to maximize its performance and efficiency of compute and network resource utilization.

The DataGRID project will generate large amounts of traffic with different quality of service and performance requirements. Concurrently with the DataGRID project, GEANT and the NRENs plan to set up different QoS mechanisms in their networks in order to provide differentiated services to the users. We propose to provide and deploy a Biomedical use case GRID application that requires Quality of Service in order to provide interactive access and feature matching of Medical images with those held in distributed Biomedical databases. This offers the opportunity to collaborate in the definition, evaluation and the tuning of advanced future QoS services.

Middleware components and some Grid applications need to replicate data simultaneously on a large set of different sites and all these transfers have to be perfectly reliable. Candidate applications of reliable multicast transport include large database replication, information exchange between information indexes and resource information repositories, and data and code distribution. To build a reliable multicast service in a Grid environment, one may benefit from the IP multicast routing service provided by the network infrastructure which uses the unreliable UDP protocol. The DataGRID project is investigating and evaluating a reliable transport protocol designed for multicast, called Reliable Multicast Protocol (RMP) that aims to recover all packet losses at the multiple receiver sides.

As well as the host based network monitoring installed by WP7 on its test sites (Amsterdam, Manchester, Daresbury, RAL, UCL, Lyon, CERN, Bologna), WP7 has installed five RIPE boxes in the main DataGRID testbed sites (Geneva CERN, Lyon In2p3, Amsterdam NIKHEF, PPARC

Daresbury, INFN Bologna). These 1-way measurements are stored and processed by RIPEnc. All of these network monitoring data provide us with a good view of the end to end performances of the network links used in the DataGRID testbed. Sharing this monitoring data and that gathered in the GEANT PoPs would help both DataGrid and Dante, but it would also allow the NRENs to have a better understanding of the behaviour of different network paths, when injecting large amount of traffic. In particular, the loss patterns and inter-loss delay distribution analysis would be of great importance for understanding the transport protocol dynamics. Having simultaneous end to end and Dante backbone performances results may also be a good way to understand where and when quality of service mechanisms are really necessary.

WP7 would like to perform the following tests on the production infrastructure:

- Analysis of TCP performance when run over a wide area high-speed network infrastructure.
- Demonstration of high throughput data flows for times similar to that to be used in moving bulk data in the Grid environment.
- Pilot the IP Premium service including making end to end tests with biological use case applications.
- Test the performance and behaviour of a Scavenger or “less than best efforts” service.
- Explore the use of reliable multicast for file replication.

We believe that this collaborative work would benefit Dante and the NRN community in the following areas:

- Better understanding of how high speed infrastructures react to load.
- Help Dante with definition of services for TCP based applications.
- Tuning of the network infrastructure for better handling of best effort bursty traffic.
- Exploration of what QoS provisions might be needed for a less than best efforts service and the effect of operating this service on the core network.
- Understanding of queue behaviour in over provisioned networks.
- Integration and sharing of information from the existing Geant monitoring and the WP7 end to end monitoring to provide better understanding of the network performance.
- Feedback on the use of multicast for innovative applications.
- Provide additional information on the benefits of the IP Premium service.

3. DISCUSSION OF THE TESTS

In this section we present details of the procedures required for each of the tests discussed in the previous sections. Estimations of the time and network load required are also given. It is clear that the work will be done in collaboration with the Dante engineers and at times suitable to the production operation of Geant.

3.1. TCP PROTOCOL EXPERIMENTS

The understanding of TCP dynamics is particularly important to verify the performance and the limits of the transport protocol in a production environment characterized by high-capacity links and very low or null packet loss probability.

In this section we present the list of goals of the proposed testing activity. To analyse TCP performance when running at high speeds we will inject stream of test traffic between the equipment at pairs of Dante PoPs and measure the following metrics:

Latency, variation of the latency (also known as packet jitter), throughput, packet loss, packet loss distribution.

Given that we will be using 1 Gigabit Ethernet NICs in the test equipment, we estimate that the load on the network for the throughput tests would be ~ 1 Gigabit between two PoPs, and that individual throughput tests would last between 1 and 10 minutes. In the case of multiple aggregate streams, the overall traffic load would be a small percentage of the link capacity present in the core of the network.

3.1.1. Latency and Packet Jitter

We will measure the round trip times as a function of the message size sent using a simple request-response application protocol. Measurements will be made using UDP/IP TCP/IP and ICMP ping. Typically 1000 measurements will be made for about 400 message sizes. The tests should take about 2-3 minutes and we expect the impact on the network to be very light. This traffic will be best efforts.

3.1.2. Throughput

We are interested in verifying the ability to maximize resource utilization when the TCP application runs on a truly high-speed wide-area infrastructure. Having traffic generators located at PoPs gives the possibility to avoid bottlenecks, which could potentially be present, when traffic goes through local and regional networks. It also gives the possibility to test TCP in presence of real-life background traffic, which is not possible in the laboratory environment.

In particular, we would like to test the impact of standard congestion control and congestion avoidance algorithms when run on high-speed links. Round trip time is also a key factor since it reduces the promptness in reacting to congestion notification.

We propose to use one of our PCs in the PoP to record the packet headers from the test traffic to provide a time sequence analysis of the TCP streams.

Tests will also be made using controlled streams of UDP packets. This complementary information will assist in the understanding of the TCP network and stack performance.

3.1.3. Packet loss frequency

TCP monitoring tools will be based on Web 100 and tcptrace. We would like to analyse the potential impact of well-tuned TCP applications when running on an over-provisioned network, where by “over-provisioned” we mean a packet-loss free network with average link utilization well below the link capacity. These measurements would be made using the throughput traffic described above.

As shown in the literature, aggregation of well-tuned TCP streams can potentially generate a highly bursty TCP aggregate: we would like to verify the capability of TCP application to cause long and short-term congestion. This test would be performed by generating multiple TCP streams between each pair of PoPs.

This test could be of interest for GEANT in order to quantify the impact on the existing infrastructure of bulk data transfers based on well-tuned TCP applications and consequently to evaluate the need of active queue management techniques to proactively prevent short and long-term TCP congestion.

3.1.4. Packet loss pattern

In case of non-null packet loss the analysis of the packet loss pattern is important to check if packets are lost randomly and individually or rather in groups of continuous packets. The former case is an indication of temporary congestion produced by a large aggregation of traffic, while the latter is a symptom of low tolerance of network devices to burstiness produced by individual streams.

Packet loss frequencies and patterns will also be measured using controlled streams of UDP packets. This complementary information will assist in the analysis of the TCP behaviour.

3.2. DEMONSTRATION OF SUSTAINED TCP TRANSFERS

The test equipment installed at the PoPs provides the opportunity to demonstrate high bandwidth, high throughput transfers operating for long periods of time. Together with the monitoring already provided by Dante, these tests would investigate the impact on the core network components, as well as providing evidence to potential uses that such transfers are feasible today. It is envisaged that these tests would initially be tried between pairs of PoPs for 1 hour, then 3, 10, and finally 24 hours. The aim would be to demonstrate transfer rates approaching 1 Gbit/s. This is ~ 10% of the current core load and is not expected to have significant impact production traffic. Tests would be performed at appropriate times in collaboration with Dante engineers.

3.3. TESTS AND PILOTING OF QOS SERVICES

3.3.1. Tests with IP Premium Service on the Core Network

This area of activity could be used to investigate and pilot the QoS services being offered on GEANT. Latency and Throughput tests similar to those described in section 3.1 would be made from PoP to PoP across the Geant Core. In this case data streams suitable both for IP Premium and background traffic would be generated. The tests would measure and demonstrate the preference given to the IP Premium traffic. Two sets of tests will be done: (a) pre-marked packets could be offered to the core routers to test re-marking and (b) the Geant routers could police and mark the packets.

3.3.2. Tests with Scavenger Service on the Core Network

There is also considerable interest within the Grid community, in using Scavenger or “less than best-efforts” flows for continuous bulk data transfers. Data streams of test traffic will be generated and marked with the agreed “less than best-efforts” code point. These will be sent from PoP to PoP over the Dante Core network. The Latency and Throughput of the received traffic will be measured during the transmission as a function of time. During the Scavenger flow, further data streams will be introduced some sent as “best efforts” and others marked with the IP Premium code point. The achieved throughput of each type of traffic will be measured. The tests would measure and demonstrate the preference given to the “best efforts” and IP Premium traffic. Appropriate QoS configurations for the Scavenger and IP Premium services would need to be made on the PoP routers.

3.3.3. End to End tests with IP Premium and BioMedical applications

This will require the implementation or transport of IP Premium traffic over the access links and the NRN networks.

An instrumented medical application will be deployed on several sites around Europe. Different services, like Premium Service and Best Efforts, will be used to carry the data flows; the flows will be monitored and the performance of the selected service will be measured. Different scenarios will be examined for end to end Quality of Service.

The proposed application has been developed within WP10 (Biomedical applications) of the DataGrid project in close collaboration with WP7. It is a content-based query database application that requires many images to be transferred and processed. The images size is about 20Mbytes and the rate can be between 1 to 10 images per second, depending the number of processing sites and storage sites. To offer a better response time, processing and communication phases will be performed in parallel. As the application is interactive, the usage of priority traffic like that of the Premium Service may offer great benefit.

The ultimate goal is to deploy the application on the DataGRID testbed sites. For example, the end user application specialists would be located in the IN2P3 centre in Lyon where they will initiate the tests; and the image databases and processing sites would be located at remote DataGRID testbed sites. In this case, Lyon will be directly connected to the RENATER Lyon-Paris link and the use of end-to-end QoS will require the different NRNs to be operating the IP Premium Service.

Different test scenarios will be used depending on which NRNs and campus networks support QoS at a given time. As an interim step, while the NRNs prepare the QoS services, the image databases and processing sites could initially be at the test stations located in the GEANT PoPs.

3.3.4. Potential Future Investigations

In addition to the goals listed above, we are also interested in running tests with active queuing mechanisms, like WRED, and congestion avoidance mechanisms, such as ECN deployed. This work would investigate the interaction between the TCP flow control and congestion control mechanisms and the active queuing techniques. These measurements would show how the network queuing and control techniques might influence the throughput achieved.

For these tests, the PoP routers would need to be configured with active queue management techniques.

3.4. TEST OF RELIABLE MULTICAST PROTOCOLS

The understanding of the behaviour of Reliable Multicast Protocol in a well-provisioned production network is very important, for the design and performance optimisation of such specific transport services. The main aim of the experiments we propose is to analyse in detail the requirements of this challenging potential application of the IP Multicast service. The overall multicast traffic load would be a small percentage of the multicast capacity present in the core of the network.

In the framework of the technical collaboration with the DANTE consortium, we propose to conduct experiments with the Tree-based Reliable Multicast Protocol (TRAM) using the standard IP Multicast service offered by GEANT. TRAM is the protocol that has been selected for detailed examination within WP7 of the DataGRID project as it is adapted to bulk data replication. TRAM is an open-source protocol realised by SUN Microsystems. It has been tested in our local testbed and different simulations have been conducted to study its scalability. The RMP protocol performance depends mostly on aggregate loss rate and loss patterns.

These tests would require the current basic multicast service using Sparse Mode PIM provided by Dante.

These tests may be extended to end to end tests when IP multicast services are available across the NRENs.

Packet loss analysis :

We propose to first characterize the IP Multicast service from the aggregate loss probability point of view. For this, we will use the mUDPmon tool, an extension of the UDPmon that has been developed within the WP7.

Comparative analysis

We will deploy the TRAM protocol on the test stations in order to measure its behavior with different information sizes and to compare the results with our simulations and local experiments.

4. REQUIREMENTS

The following list summarizes the list of requirements in order to achieve the goals stated above.

1. Availability of two or more Linux PCs in three PoP locations. The PCs will be rack-mounting to save space.
2. The PCs should be connected to the PoP router by a Gigabit Ethernet line card. In order to avoid interference with Dante's production monitoring, WP7 suggests installation of a dedicated Gigabit Ethernet blade in the Juniper router providing connectivity for the test equipment. WP7 is in discussion with Juniper for the provision of these blades.
3. The test PCs will be connected to the Juniper router via a suitable Gigabit Ethernet switch, provided by WP7, to allow detailed monitoring of the TCP flows.
4. The PoPs should be selected to maximise the RTT between the test equipment.
5. WP7 will loan the Linux systems they will have kernel version 2.4.16 in order to support the Web 100 patch, which provides monitoring of the TCP stack performance.
6. The test workstations should be accessible remotely using ssh.
7. Software needs to be installed on the test PCs to produce and monitor the test traffic. The list of packages of interest includes: mgen, rude/crude, ping, netperf, iperf, UDPmon, tcpdump, tcptrace and Web100.
8. Router monitoring. The support of MIBs in the PoP routers for traffic levels, queue and packet loss monitoring allows both the testers and the GEANT engineers to analyse and understand the measurements. It will also allow evaluation of the degree of overprovisioning and the robustness of the infrastructure.
9. The IP Premium alpha tests will use the DSCPs defined by Dante. The end to end medical tests will require IP Premium support in the Geant Core and the participating NRNs.
10. The reliable multicast tests are end-to-end and will require multicast routing to be enabled in the Geant Core and the participating NRNs. It is expected that Dante would provide the current basic multicast service using Sparse Mode PIM.
11. Technical discussions with Dante have estimated that the tests would require access to the Dante PoPs for about 4 months after installation. However, as detailed in Section 3, testing would not be continuous and would be done in full collaboration with Dante.

5. CONCLUSIONS

We have proposed a set of experimental activities focused on the TCP and RMP transport protocols which require the availability of traffic generators located in three of the GEANT PoPs.

We have also proposed to share the network performance measurements collected by the WP7 DataGrid network monitoring work, including the data from WP7 RIPE boxes installed at the edge of several EU DataGrid sites. These data could help the NREN and DANTE engineers in a better understanding of the end to end behaviour of the network. In return, we would appreciate suitable access the GEANT RIPE databases in order to conduct finer analysis of our end to end measurements and to improve our high performance transfer tools.

We will also provide the DANTE consortium with a use case application that will generate traffic requiring different quality of service levels.

We will explore the use of and the configuration needed for a Scavenger or “less than best efforts” service.

We think that the proposed activities could be of great interest for both the Grid community Dante and the NRNs. In the former case the results will provide important information about middleware and end-user application performance when running on a high-speed network environment. For Dante and the NRNs, the experiments will provide information about the performance and robustness of the GEANT infrastructure under high TCP load and, in addition to this, may help with the introduction of advanced TCP-oriented queue management techniques and differentiated services like Premium and Scavenger Services.

Reliable multicast is a challenging service for the future, we think that such collaborative experiments will help provide better configuration and tuning of the IP Multicast service, not only for traditional VC application but also for new Grid transport services.