

14.09.06

Deliverable DS3.6.2: Report of the PERT Group



Deliverable DS3.6.2

Contractual Date:	31/08/06
Actual Date:	14/09/06
Contract Number:	511082
Instrument type:	Integrated Infrastructure Initiative (I3)
Activity:	SA3
Work Item:	Wi6
Nature of Deliverable:	R
Dissemination Level	PU
Lead Partner	DANTE
Document Code	GN2-06-096v5

Authors: T. Rodwell (DANTE)

Abstract

The report begins with a review of the origins of the Performance Enhancement and Response Team (PERT), which was included in the GN2 project plan on the strength of a successful pilot PERT conducted in the last year of the GÉANT project. The PERT roles are then described, and distinguishes between the key members of the PERT, namely Duty Case Managers, Special Case Managers, and Subject Matter Experts. The PERT staff rely on two key systems to help them manage and investigate cases – the PERT Knowledge Base contains important information for both PERT staff and end-users alike and the PERT Ticket System (PTS) is a bespoke, web-based application that enables the tracking and control of open PERT cases. Appendix A gives an overview of the GN2 PERT cases submitted to date, whilst Appendix B is comprehensive overview of the Logistical Session Layer device that can be used to circumvent the issues related to long distance TCP connections. The conclusion of the report is that the PERT should extend beyond the GN2 project, but should evolve into a federated organization, where each domain implements its own PERT function.

Table of Contents

0	Executive Summary	iv
1	Introduction	1
2	Origins of the PERT	2
2.1	Background	2
2.2	The Trial PERT	2
2.3	The GN2 PERT	3
3	PERT Organization	4
3.1	PERT Roles and Users	4
3.1.1	PERT Manager (PM)	4
3.1.2	Duty Case Manager (DCM)	4
3.1.3	Special Case Manager (SCM)	5
3.1.4	Subject Matter Expert (SME)	5
3.1.5	PERT Users	5
3.2	Staffing Levels	6
4	PERT Support Systems	7
4.1	PERT Ticket System	7
4.1.1	Accessing PTS	7
4.1.2	PTS User Privileges	8
4.1.3	Ticket Management	8
4.2	PERT Knowledge Base	8
4.3	PERT Mail Lists	9
5	PERT Policy	10
5.1	Services	10
5.2	Operations	11
5.2.1	Reactive operations	11
5.2.2	Proactive Operations	12
6	PERT Experiences and Lessons Learned	13

6.1	PERT Workload	13
6.2	Types of PERT Cases	13
6.3	Investigations	14
6.3.1	Complexity of cases	14
6.3.2	Progress of Cases	14
6.3.3	Resolution of Cases	15
7	Conclusion and Next Steps	16
8	References	17
9	Acronyms	18
Appendix A	GN2 PERT Cases	19
Appendix B	Logistical Session Layer (LSL)	32

Table of Figures

Figure 9.2:	LSL v direct transfers, University of Delaware to University of S. Carolina Beaufort	34
Figure 9.3:	LSL locations on Torun-JIVE path	36

0 Executive Summary

The Performance Enhancement and Response Team (PERT) is a virtual organization made of networking experts who help end-users diagnose and solve network performance issues. The term was first coined by the Internet2 project in the US, but it remained just a concept until it was taken up by the European NREs. Following a successful trial in 2003, a funded, full-time PERT was included in the plans for the GN2 project, as part of its “End to End QoS” Service Activity (SA), SA3.

Nine NREs plus DANTE contribute to the GN2 PERT. Between them they have put in place the PERT procedures, set up the required systems (a ticket system and an on-line Knowledge Base), produced all the necessary documentation and staffed the PERT with Duty Case Managers (DCMs).

The PERT’s Duty Case Managers (DCMs), who rotate on a weekly basis, are the first point of contact for the users of the PERT (DCMs should respond to all legitimate new requests within two working hours). Initially, to avoid the risk of the PERT being overwhelmed, only a limited number of organizations (the European NREs and a few other large, international projects) were permitted to contact the PERT directly. Experience suggests that the PERT could handle more requests than are currently being made, so the new version of the PERT Ticket System (PTS) allows anybody to register themselves and request assistance. It will then be up to the DCM to determine who is an Eligible user and who is Ineligible, based on what their parent organization and/or project is. An Ineligible user might still get help from the PERT, but Eligible users would always have priority.

Since the DCMs change on a weekly basis it can be hard to maintain momentum in a complex case, as by the time the DCM has familiarised themselves with the case much of the week may have passed. Therefore the concept of Special Case Managers (SCMs) was introduced. An SCM volunteers to adopt a given case and then manages it through to resolution. Because a SCM cannot be expected to work full time the DCM should also track the progress of SCM cases and offer to help wherever progress is not being made.

In addition to the co-funded DCMs the PERT also benefits from the help of unfunded, volunteer Subject Matter Experts (SMEs), who give their time as and when they are able to. SMEs come from a wide variety of organizations, including universities, research institutes and equipment manufacturers, and they have access to the PTS and PERT mail list. It had been hoped that the number of SMEs would grow steadily and so create a large community, such that there be a high probability that at any given time there was an SME both able and willing to help in a complex case. In fact, the SME group has remained small and whilst their contribution has been significant, it has all come from just a few individuals.

Project:	GN2
Deliverable Number:	DS3.6.2
Date of Issue:	14/09/06
EC Contract No.:	511082
Document Code:	GN2-06-096v5

The PERT is now looking to the future to consider what form (if any) it should take once GN2 has finished. The consensus is that the PERT should continue, but should devolve to a federated structure, with each network provider putting in place its own PERT function, and helping to troubleshoot problematic en-to-end paths which transit their domain. A new SA3 Work Item (WI) is planned for GN2 Y3 which will carry out preparatory work for this decentralization, including offering advice on how to set up a PERT.

1 Introduction

Until recently the bottleneck in a networked system would always be the data rate¹ of one of the Wide Area Network (WAN) connecting circuits. In comparison to the WAN circuit, the data rate in the rest of the system (Local Area Network (LAN), end systems' Network Interface Cards (NICs), end systems' internal buses, etc) would be 10s or 100s of times greater. This had two significant effects. First, end users assumed (normally correctly) that system performance was always limited by the WAN, and second, any sub-optimal performance of other system components was masked by the WAN's performance.

As WAN data rates catch up (and in some cases overtake) the data rates of LANs, the bottleneck of a networked system is no longer always the WAN. This is particularly true in the world of research and education networking, where international WAN links are often SDH STM-16 (2.5Gbps) or greater, compared to typical LAN data rates of Gigabit or Fast Ethernet (1Gbps or 100Mbps respectively). However, even though the network bottleneck may have increased, many end users have found that their applications are not seeing a matching increase in throughput. This is typically because the users' end-systems are not optimised for high speed data transfer, and the root cause of the problem can lie in one or more of a wide variety of areas (operating system, TCP/IP stack, Network Interface Card (NIC), LAN equipment, WAN equipment).

¹ Frequently, but incorrectly, referred to as "speed"

Project:	GN2
Deliverable Number:	DS3.6.2
Date of Issue:	14/09/06
EC Contract No.:	511082
Document Code:	GN2-06-096v5

2 Origins of the PERT

2.1 Background

As early as 1996, Internet2 in the US recognised that achieving higher throughputs over a shared network would be a future challenge, but at that stage it was expected that over-provisioning and (when deployed) Quality of Service (QoS) mechanisms would ensure that all users would receive the network performance they required. However, by the year 2000 it was realised that over-provisioning and QoS would not by themselves solve all throughput problems, and Internet2 put in place their End-to-End Performance Initiative (E2Epi). One aspect that E2Epi looked at initially was how to provide direct support to end-users experiencing network performance problems. At the [Internet2/APAN/TransPAC/NLANR](#) Technical meeting held in Hawaii in January 2001, the term PERT was coined to describe a team of networking experts who would, at the request of end users, investigate possible network performance issues and offer advice on how to improve under-performing networked applications. The name PERT was chosen because of its similarity to the term CERT, as it was hoped and expected that in time the PERT function would be to network performance what CERT is to network security.

2.2 The Trial PERT

For Internet2 the PERT remained just a concept, as they instead focussed their attention on software applications and tools for measuring network performance. However, at the [TF-NGN-Jul-02] meeting of the European Task Force Next Generation Networks (TF-NGN) in July 2002, the subject of the PERT was raised and a group of European research networks from Italy, Switzerland and the UK discussed beginning a trial PERT.

The trial began in late 2003, when the first case was presented to the trial PERT. At this point the PERT was simply a mailing list of networking experts who would discuss performance issues as and when they were raised. A problem with this set up was that a cases were often discussed on the mailing-list for long periods without any feedback begin given to the end-user, so that after a period of time the user would assume their case had been forgotten and would disengage from the PERT. To address this, in March 2004 an issue tracking application, called RoundUp, was deployed by SWITCH.

Project:	GN2
Deliverable Number:	DS3.6.2
Date of Issue:	14/09/06
EC Contract No.:	511082
Document Code:	GN2-06-096v5

2.3 The GN2 PERT

Although only a small number of cases were investigated during the trial PERT, there was sufficient interest from the European NRENs that a production-quality PERT was included in the plans for GN2.

The GN2 PERT (so written to distinguish it from the general concept of the PERT function), is a sub-activity of the GN2 Service Activity 3 (SA3). Other than the PERT, GN2 SA3 also supports multi-domain Premium IP (PIP) and the operation of the GN2 network performance monitors. SA3 was originally titled “End to End Quality of Service” but because its responsibilities have recently expanded to include support to GN2’s [perfSONAR] applications, it has been re-titled “Support for Multi-domain Services”. Colloquially within the community, SA3 is known by the acronym PACE (Performance and Allocated Capacity for End-users).

The GN2 PERT work began with the design and set-up of the organization. This was Work Item (WI) 2 of SA3 and was lead by DANTE. Setting up of the PERT included a PERT pilot period, during which time the duty PERT staff worked on a part-time basis (3 hours a day). The PERT pilot ran from October 2004 through to March 2005, at which point the PERT started as a production service.

There are 9 GN2 member organizations that participate in the PERT: CARNet, CESNET, DANTE, FCCN, GARR, HUNGARNET, PSNC, RENATER and SWITCH. Between them they are providing over 75 Man Months of effort to run the day to day operations of the PERT (which is WI 6 of SA3) over the 4 year project .

3 PERT Organization

The PERT is a virtual organization – it has no geographical location and (to date) it has no full time employees (even though the PERT itself operates full time). Although virtual organizations are not new, they present a specific set of challenges and these are discussed in the ‘Lessons Learned’ section later.

3.1 PERT Roles and Users

There are four main roles in the GN2 PERT and each member of the PERT fulfils one or occasionally more roles.

3.1.1 PERT Manager (PM)

There are currently 2 PERT Managers (one from DANTE, the other from SWITCH) who between them are responsible for the smooth running of the PERT. It is purely an administrative role and there is no reason why in the future there could not or should not be more PERT Managers, though to avoid confusion the maximum number in the PERT management team should probably be limited to 4. The PMs are directly accountable to the GN2 Technical Committee, and are ultimately accountable to the GN2 partners.

3.1.2 Duty Case Manager (DCM)

Duty Case Managers (DCMs) are the first point of contact for a user wanting to open a new case. DCMs assess new requests for help, open a case if it is warranted, and initialize the investigation. DCMs are responsible for the day to day running of the PERT so they also oversee and manage the progress of all unresolved cases (but see Special Case Manager, below). Duty Case Managers are provided by the 9 GN2 member organizations participating in SA3 Work Item 6 and they work in accordance with a weekly changing roster.

3.1.3 Special Case Manager (SCM)

Special Case Managers (SCMs) are volunteers who adopt a specific PERT case and manage the rest of its investigation through to resolution. SCMs normally come from the DCM group, but this does not have to be the case. An SCM will normally be someone with an affinity to the case in question – they might have been the DCM when the case was first opened, or the customer might be from their country. SCMs and DCMs are jointly referred to as Case Managers (CMs).

3.1.4 Subject Matter Expert (SME)

A Subject Matter Expert (SME) is someone with extensive knowledge in one or more areas of network engineering, or other areas related to network engineering. Unlike the DCMs they are not co-funded by the GN2 project, nor would they be even if they were to volunteer as an SCM (which is possible but not expected of them). SMEs will either be invited to join the PERT or may offer their services directly. In the latter case the PERT management team will assess that person's suitability before accepting the offer. To date the PERT has had mixed success with SMEs. There have been few SME volunteers, and even fewer who have been proactive in PERT investigations. There are however three non-GN2 SMEs who deserve particular recognition and thanks for the significant contributions they have made to the PERT, namely Larry Dunn (Cisco, USA), Even Baruch (Hamilton Institute, Republic of Ireland) and Richard Hughes-Jones (University of Manchester, UK).

3.1.5 PERT Users

Until recently people and organizations who call upon the PERT's services have been termed 'customers'. 'Customer' was chosen in order to distinguish between those people or organizations who requested the PERT's assistance, and those end-users who the PERT later dealt with as part of their investigation. Customers were further divided into 'Primary customers' (which were basically those organizations which directly connected to the GÉANT2 network, such as the European NRENs), and 'End customers', who were the users of the Primary customers. This distinction was useful since to date only the Primary Customers have been able to directly open a PERT case – End Customers have had to rely on Primary Customers to forward their requests to the PERT.

However, it has since been noted the term 'customer' has commercial connotations, which could be misleading in the context of the PERT. As such, in future those requesting the PERT's assistance will no longer be referred to as customers but will instead be termed 'users', and people who operate and/or administer the end-systems involved in a PERT case will be termed 'end-users'.

3.2 Staffing Levels

The PERT operates during core European working hours, which is to say weekdays 0900-1700 CET, with the exception of Europe-wide public holidays (such as Christmas).

The success of the PERT is primarily dependent on users being given effective and prompt help when they request it. To ensure a prompt and appropriate response to all new requests for help there must always be at least one Duty Case Manager (DCM) ready to receive new cases during the PERT's hours of operation. When the PERT was first set up it was thought that the number of cases submitted to the PERT would grow steadily and so the plan was to increase the number of on duty DCMs from 1 Full Time Equivalent (FTE) to approximately 3 FTE by the end of the project. In fact, the rate of new PERT cases being opened is slower than originally forecast and so far there has been no justification in increasing the DCM staffing level above 1 FTE. Since this may change in the future it was decided not to change the participants' co-funding.

4 PERT Support Systems

Two dedicated systems have been put in place to support the PERT. The PERT Ticket System (PTS) has been wholly developed by PSNC and was designed to replace RoundUp, the trial PERT's issue tracker. The PERT Knowledge Base (PERT-KB) is a Wiki website containing advice and guidance on a wide range of subjects relating to network performance. The PTS and the Knowledge Base can be reached directly (see below) or via the dedicated PERT website [PERT website]. The PERT web-site acts as a container for links to the various PERT resources (systems, PERT pages of the GN2 website, etc). The PERT web-site also includes a web forum which, if and when demand for the PERT grows, can be used by non-academic users (who are not automatically eligible for the PERT's assistance) to discuss performance issues amongst themselves, and receive assistance from PERT staff when they are not otherwise busy with prioritised PERT cases. Since the forum is not yet active, to avoid potential confusion this forum is currently hidden from public view.

4.1 PERT Ticket System

The PERT Ticket System (PTS) was developed by PSNC under SA3 WI 4. PSNC also maintain and administer the system. PTS fulfils all the basic requirements of a generic issue tracker (in as much that cases can be added, updated and closed) plus there are other features specific to the PERT's needs (e.g. DCM roster, diary function). A brief overview of PTS is given below. Full information on PTS v2 (which is due to enter service later this year) is given in [DS3.5.2]

4.1.1 Accessing PTS

The PTS is a web-based system located in the GN2 Poznan PoP² but publicly reachable via [PERT website], or directly at <http://www.pert.geant2.net/pert>. In PTS v1 only pre-registered users could access PTS, but later versions of PTS allow anyone to register themselves, and then submit a request for assistance. To ensure requests are genuine anyone registering must provide a legitimate e-mail address, to which their initial

² As of 28 Aug 06 the PTS application is temporarily being hosted on a PSNC platform, following corruption of the GN2 server. Once the GN2 server has been repaired and re-installed, PTS will be migrated back.

password will be sent. Regular users of the PTS (see below) will use an X.509 certificate to simplify and speed up their access to PTS.

4.1.2 PTS User Privileges

The PTS allows for different user groups, with different privilege levels. The Case Managers group have full access to all tickets, whilst the PERT Managers group also have access to the PTS settings – together these two groups form the super-group Managers. The SME group can view and edit all tickets, and can open but not close a case. ‘Primary Customers’ (which in future will be called ‘Users’) have access to all tickets which have been created by someone associated with their organization. Like Managers and SMEs, Primary Customers will be issued with an X.509 certificate, which they can install in their browser, to make it easier to access the PTS. Conversely, the ‘End Customer’ group use their e-mail address and password in order to access the PTS, and they can only see and edit tickets to which they are associated, which can be either because they created the ticket themselves or because a Manager or SME has explicitly added them to that ticket’s end-user list.

4.1.3 Ticket Management

Once a ticket has been created in PTS, and acknowledged by the DCM, it will be assigned a state of “Waiting for PERT action”. It will stay in this state until such time as the PERT staff request more information or action from the end-user or another Third Party. At this point the DCM will add an appropriate note to the ticket and change its state to “Waiting for User [or Third Party] action”. The ticket will then alternate between these three states, with PERT staff and authorised users adding and editing notes, until the case has been resolved. At this point, with the end-users agreement, the Case Manager (be it SCM or DCM) will close the case.

4.2 PERT Knowledge Base

As important as the PTS is the PERT Knowledge Base (PERT-KB). The PERT-KB [KB] was implemented by SWITCH, who are the leaders of SA3 WI-3, and they are responsible for both the system host (a SWITCH server) and the application itself (which is based on the open source software TWiki). Though SWITCH are not directly responsible for the PERT-KB content, they are responsible for the registration of new users, and only registered users can add or edit PERT-KB content.

The primary purpose of the PERT-KB is to help end-users to help themselves, which is to say to give them the knowledge to diagnose and fix their own network performance problems. TWiki includes a comprehensive search function that enables a user to find any and all PERT-KB pages which mention a given string. For example, if a user wants to get information on measuring bandwidth, they might search on “bandwidth” or “bandwidth measurement”. Either way, within the first 5 returned links there is a topic called ‘Bandwidth Measurement Tools’, which itself lists 8 of the most common bandwidth measurement tools, beginning with iperf, which is probably the most common and popular tool of its kind.

Project:	GN2
Deliverable Number:	DS3.6.2
Date of Issue:	14/09/06
EC Contract No.:	511082
Document Code:	GN2-06-096v5

The secondary purpose of the PERT-KB is to help a non-expert to learn about network performance, both from the knowledge base articles and the links that lead to other, carefully selected external web pages. As such, the PERT-KB's first page is a contents page, which groups articles by subject and begins with an introduction to networking performance (its importance, metrics for network performance etc). Furthermore, the content of the PERT-KB has been used to produce hard copy User guides and Best Practice guides, [DS3.3.2] and [DS3.3.3] . These are publicly available on the GN2 website, and have been well received in the NREN community.

4.3 PERT Mail Lists

The PERT mail lists are not a dedicated system such as PTS or the PERT-KB, but are nevertheless very important as they tend to be where most ideas are exchanged. The mail lists are part of DANTE's e-mail system, but are administered by the PERT Managers (specific firewall filters were put in place in DANTE to allow this). At the start of the PERT there were 3 mail-lists:

- pert-managers@geant2.net – A list for just the PERT Managers
- pert-report@geant2.net – A list for primary customers to report problems to the PERT duty case-manager
- pert-discuss@geant2.net – A list for the PERT (Case Managers and SMEs) to discuss open cases

Currently there are still the same 3 lists, though the [pert-managers](mailto:pert-managers@geant2.net) list has been expanded to include the Case Managers as well. However, [pert-report](mailto:pert-report@geant2.net) will soon be discontinued, as users will be able to submit cases directly to the PTS.

5 PERT Policy

PERT Policy is documented in GN2-05-018. The current policy, [GN2-05-018v6], and the previous versions, [GN2-05-018v5], [GN2-05-018v4] and [GN2-05-018v3], are all available from the SA3 section of the GN2 intranet at [GN2-SA3 intranet]. Versions 1 and 2 of the policy were drafts and therefore not formally published.

5.1 Services

The PERT provides three services. The Primary Service, so called because the primary purpose of the PERT is the investigation and resolution of network performance issues directly or indirectly affecting the PERT's primary users. Case Mangers will spend the majority of their time providing this service and SMEs are expected to prioritise the primary cases over other PERT related work. Note that the PERT's primary users are predominantly the European NREN NOCs, who will generally be creating a case on behalf of one of their own end-users. However, end-users may now contact the PERT directly, in which case the DCM will need to ascertain whether or not the end-user qualifies for PERT assistance. If an end-user is a European research or academic network user then they do qualify for PERT assistance and they are termed 'Eligible' end-users. Together, Eligible end-users and primary users are termed 'Eligible users'. Any one submitting a request who is not an Eligible user is termed an Ineligible user.

The secondary purpose of the PERT is to add to and maintain the content of the PERT Knowledge Base (PERT KB). In some ways this service could be considered a sub-set of the Primary service, since part of the procedure for closing a case is to add a suitable entry to the PERT knowledgebase, if appropriate.

A third service originally envisaged for the PERT was the provision and moderation of a public forum for the discussion of all matters relating to network performance. Experience to date suggests that the PERT can manage all cases submitted in the PTS without having to offload low priority cases to such a public forum, however should circumstances change then starting up such a system remains an option. At first the moderators of the forum would be PERT DCMs and SMEs, but then regular contributors to the forum would be invited to become so called PERT Associate Members (AMs), and they could take on the bulk of the forum's operation.

Project:	GN2
Deliverable Number:	DS3.6.2
Date of Issue:	14/09/06
EC Contract No.:	511082
Document Code:	GN2-06-096v5

5.2 Operations

The PERT is staffed between 0900 and 1700 CET on working days (which is to say not weekends or Europe-wide Bank Holidays). There is no official out of hours support, but the PERT systems (PERT Ticket System and PERT Knowledge Base) will normally be available '24x7'.

The PERT will operate both reactively and proactively.

5.2.1 Reactive operations

In its reactive mode, the PERT will respond to network performance issues brought to their attention by Eligible users.

5.2.1.1 New Cases

A Duty Case Manager should read, assess and acknowledge all new requests from Eligible users within 2 hours of the request being made, except if the request is made outside of working hours, in which case the request should be acknowledged by 11.00 CET of the next working day. It is the responsibility of the principal DCM to ensure these deadlines are met. All new requests should be given a priority of 'Urgent', 'Normal' or 'Low', based on the Duty Case Manager's judgment of the issue's urgency and importance. In choosing a priority level the Duty Case Manager should take into account the impact of the problem and the urgency for a solution. The PTS will automatically create a login for the end customer (if they don't already have one), so they can directly access their case, but if appropriate the Duty Case Manager will need to create a login for any other end-users involved.

Depending on the circumstances (and in particular the work load of the PERT), a Duty Case Manager may still open a case for a request from an Ineligible end-user, but normally this should be marked as a Low priority. If the DCM rejects a request for help from an Ineligible end-user then if possible and if appropriate the DCM should sent a short e-mail reply explaining the reason for rejecting, and directing the user to the PERT-KB, where they might find useful information.

5.2.1.2 Ongoing Cases

The Duty Case Manager should proactively investigate and/or monitor all open, unadopted³, unresolved issues, and keep their status up to date. After first being 'Submitted' and then 'Acknowledged', an issue's status is determined by who is next expected to provide information or results – thus the case can be either 'Waiting for PERT action', 'Waiting for User action' or 'Waiting for Third Party action'. The status will change to 'Resolved' once the CM believes the issue has been resolved and 'Closed' once the user is satisfied. An issue can also be 'Cancelled' if an issue is opened in error or not concluded satisfactorily.

³ An adopted case is one which has been taken over by an SCM. Conversely an unadopted case is one which has not (yet) been adopted by an SCM.

The Duty CM should monitor all unresolved cases (even SMC adopted cases) to see whether or not progress is being made. If an SCM is not actively progressing their issue(s) then the CM should offer to do so themselves (if they can) or attempt to stimulate further discussion amongst the PERT staff (perhaps by offering to find out more information).

5.2.1.3 Resolving Cases

Only Case Managers and PERT Managers may mark a case as resolved, closed or cancelled. A case should normally only be closed or cancelled with the agreement of the end-user.

Once a case has been resolved the Case Manager should assess whether the case, or any part of the case or investigation, is worth recording in the PERT KB. Any appropriate, approved person may add the entry to the PERT KB but it is the responsibility of the Case Manager to ensure it happens.

5.2.2 Proactive Operations

When there are no open cases that can be usefully progressed the Duty Case Managers should work proactively.

5.2.2.1 PERT KB Review

The PERT KB will benefit from continual review, so as to:

- 1) Correct any missed errors
- 2) Update old articles with new relevant information
- 3) If appropriate, retire or downgrade articles which are no longer relevant

5.2.2.2 Performance Search

When not otherwise busy, Case Managers should use available tools and logs to search for and investigate any network anomalies or trends they might discover.

6 PERT Experiences and Lessons Learned

6.1 PERT Workload

At the beginning of the GN2 project the management of GN2 SA3, which is responsible for the PERT, were concerned that the PERT might initially be overloaded with requests for help from end-users. This was of particular concern since for the main part the PERT Case Managers were not experienced in network performance and therefore needed a gentle introduction in order to build up their knowledge and confidence.

In order to avoid the embarrassment of a case overload, it was determined that only NRENs and other significant partners of the GN2 project should be able to raise a PERT case, meaning that all end-users would have first to contact their local NREN, who would act as a filter. The 'significant partners' were termed 'Primary Customers', and only they had access to the PTS, and the PERT's main e-mail address.

This strategy worked, in that the PERT was not overloaded. In fact, the strategy worked a little too well, since at an average of one case per month the PERT is a little under utilized, and would like more cases (and consequent exposure). As such it was collectively decided by the PERT that access to the PTS should be opened up, and that the Duty Case Managers should use their discretion as to what cases should or should not be investigated, based on the organization making the request and the nature of the request.

6.2 Types of PERT Cases

Not surprisingly the majority of requests submitted to the PERT relate to a user experiencing less than expected throughput. Many of these throughput problems are related to high-speed trans-Atlantic TCP connections, where the user knows that the path bottleneck is large (say 1Gbps), but they are still only able to achieve, say, 100Mbps. The PERT's experience to date is that these throughput problems are caused by one or more of the following issues:

- Sub-optimal (untuned) TCP parameters – the default TCP settings for most current end-systems do not support long-distance, high data rate connections and therefore need to be tuned. This is a simple task and details are given in the PERT Knowledge Base

- Low rate packet loss – For long distance, high data rate TCP connections, even a packet loss of 0.1% will have a noticeable impact on the achievable throughput. Recognising the seriousness of packet loss, the PERT has been working with researchers from the University of Delaware in trying to evaluate their newly developed system for reducing the effects of packet loss, the Logistical Session Layer (LSL) device. LSL is described in full in Appendix B of this document.
- Congested bottlenecks – Sometimes the end-user is not aware of the actual path bottleneck; they assume the bottleneck is the data rate of their own uplink to the backbone network (say, 100Mbps), but in fact there is a lower capacity link in the path (perhaps a 34Mbps E3 circuit) which itself is congested (and therefore causing packet loss).
- Routing changes – If a routing change has increased a path's RTT then this will often be reflected in decreased throughput. If there is severe route flapping, with potential packet loss or serious re-ordering of packets, then this could cause even greater problems. Fortunately these problems tend to be short-lived (as they are normally related to circuit failures, which themselves are normally quickly fixed).

Another frequent cause of network performance issues is Ethernet duplex mismatching, whereby an end-system and the LAN switch to which it is connected do not correctly auto-negotiate Ethernet full-duplex operation. This is known to be a common issue in LANs so once end-users can directly access the PTS the PERT can expect to see a significant proportion of cases involving duplex mismatch.

6.3 Investigations

6.3.1 Complexity of cases

Because of the procedures that have been in place to date (that is, the PERT can only be contacted by NRENS or international European research projects), the cases which have been submitted to the PERT tend to be difficult to diagnose, since any easy issues would have been solved by campus or NREN network engineers before they reached the PERT. This is something that should be borne in mind when assessing the percentage success rate of the PERT (which is good, but admittedly not perfect),

6.3.2 Progress of Cases

PERT cases are often long lived, sometimes because they are complex and need careful study by the PERT, but often because the end-user is slow to respond to the PERT. Again, this is understandable behaviour on the part of the end-user, as the degraded service the PERT are investigating for them is often something they can live with and so may not be a high priority for them to deal with. Nevertheless, the PERT now wish to distinguish between cases which are waiting for PERT action and those which are waiting for a response from an end-user, or some other third party. Therefore, in the new version of PTS additional ticket states have been created, so that, once an investigation has started, a ticket is marked as either 'Waiting for PERT action', or 'Waiting for Customer [User] Action' or 'Waiting for Third Party Action'.

Project:	GN2
Deliverable Number:	DS3.6.2
Date of Issue:	14/09/06
EC Contract No.:	511082
Document Code:	GN2-06-096v5

6.3.3 Resolution of Cases

When a Case Manager closes a case they assign a resolution rating. Currently this is a simple percentage value (0, 20,40, 60 or 100%), which have the following respective meanings:

- 0% - PERT unable to explain cause of problem
- 20% - Possible explanation, no solution
- 40% - Explanation, no solution
- 60% - Explanation, work-round suggested
- 80% - Explanation, work-round implemented
- 100% - Problem fully solved

However, it has been found that this simple, single value rating does not accurately reflect the outcome of many PERT cases. For example, it is quite common for a problem to be transient, such that by the time the PERT begin their investigation the problem has disappeared – the end-user is happy that normal service has been restored and for them the problem is fully solved (100% resolution). However, the PERT has little or no idea what would have caused that problem (0% or 20% resolution). Indeed, even if the PERT would like to investigate the root cause of a problem, once the problem has disappeared (by itself or via a work around) the end-user is unlikely to want to spend time and effort helping the PERT learn more about a problem which no longer affects them. This is an entirely understandable attitude on the part of the user, and so in future versions of PTS the resolution rating will be made up of 2 parts, namely a 'correction rating' (to what extent has the problem been corrected) and a 'comprehension rating' (to what extent were the PERT able to diagnose the problem).

7 Conclusion and Next Steps

Overall the first two years of the PERT should be considered a success. Whilst there are been slightly fewer cases submitted than might have been liked, those users who have sought the PERT's assistance have generally be pleased with the help they have received. The lessons learned to date are being used to improve both the PERT procedures and the PERT systems - in particular, the latest guides on network performance (derived from the PERT Knowledge Base) have been praised for their high standard and general usefulness.

The PERT will continue a similar service through Y3 and Y4 of GN2, and it is hoped and expected that it will carry on after the project too. Over the last few months the PERT participants have been considering what form the post-GN2 PERT should take. In keeping with the initial spirit of the PERT, it is planned to change the PERT from being a centrally-managed organization into a federated organization, where each network has its own PERT function, and each PERT will be ready and able to assist in troubleshooting problems which occur on paths which pass through their domain. At this stage it is thought there will still be a requirement for a few centralized services (a Knowledge Base plus some kind of case tracking system), but generally each network will be responsible for managing and progressing the cases of their own users. To support this transition, SA3 is starting a new WI in Y3, WI 12, which prepare for the decentralization of the PERT. Its tasks include running a workshop to help NRENs put in place their own PERT teams, and producing a Deliverable that describes how a decentralized PERT would operate.

8 References

[TF-NGN-Jul-02]	http://www.terena.nl/activities/tf-ngn/tf-ngn8/minutes.pdf
[perfSONAR]	http://wiki.perfsonar.net/jra1-wiki/index.php/Main_Page
[DS3.5.2]	“GN2-06-094 - DS3.5.2 PERT Troubleshooting Procedures version 2” B Belter et al, Aug 06
[PERT website]	http://www.pert.geant2.net
[KB]	http://pace.geant2.net/cgi-bin/twiki/view/PERTKB/WebHome
[DS3.3.2]	“GN2-06-094 - DS3.5.2 PERT Troubleshooting Procedures version 2” B Belter et al, Aug 06
[DS3.3.3]	“GN2-06-094 - DS3.5.2 PERT Troubleshooting Procedures version 2” B Belter et al, Aug 06
[GN2-05-018v6]	“GN2-05-018v6 – PERT Operating Policy” T Rodwell, Aug 06
[GN2-05-018v5]	“GN2-05-018v5 – PERT Operating Policy” T Rodwell, Jul 06
[GN2-05-018v4]	“GN2-05-018v4 – PERT Operating Policy” T Rodwell, Oct 05
[GN2-05-018v3]	“GN2-05-018v3 – PERT Operating Policy” T Rodwell, Mar 05
[GN2-SA3 Intranet]	http://intranet.geant2.net/server/show/nav.704

9 Acronyms

[AM]	Associate Member (of the PERT)
[APAN]	Asia Pacific Advanced Network
[CM]	Case Manager
[DCM]	Duty Case Manager
[EC]	European Commission
[E2Epi]	End-to-End Performance Initiative
[FTE]	Full Time Equivalent
[GN2]	GÉANT2 Project
[NIC]	Network Interface Card
[NLANR]	National Laboratory for Applied Research
[NREN]	National Research and Education Network
[LAN]	Local Area Network
[LSL]	Logistical Session Layer
[PACE]	Performance and Allocated Capacity for End-users
[PERT]	Performance Enhancement and Response Team
[PERT-KB]	PERT Knowledge Base
[PIP]	Premium IP
[PM]	Duty Case Manager
[PTS]	PERT Ticket System
[QoS]	Quality of Service
[SA3]	Service Activity 3
[SCM]	Special Case Manager
[SDH]	Synchronous Digital Hierarchy
[SME]	Subject Matter Expert
[STM]	Synchronous Transport Module
[TF-NGN]	Task Force Next Generation Networks
[WAN]	Wide Area Network
[WI]	Work Item

Appendix A GN2 PERT Cases

There have been 17 GN2 PERT cases submitted to the PTS over the last 16 months, all but 3 of which have been closed. Each case is described below. As might be expected, the cases varied significantly in their complexity and the extent of the associated PERT investigation. Each section below begins with a summary of the case, in terms of when it was opened, when closed, the assigned resolution rating, and the author of the article (who was often, but not always, also involved in the investigation itself).

A.1 Low data throughput for e-VLBI project

Case Number	Case Opened Date	Case Closed Date	Rating (%)	Author
2 ⁴	6 May 2005	21 Jul 2006	100	Bartosz Belter (PSNC)

A.1.1 Background

Very Long Baseline Interferometry (VLBI) is a technique in radio-astronomy for correlating data from geographically diverse radio-telescopes in order to emulate a single very large telescope. Currently this is achieved in Europe by sending data to a local hard disk array then mailing the hard disk array to the central Joint Institute for VLBI in Europe (JIVE) in Dwingeloo, NL. The European VLBI Network (EVN) hope to improve this process by sending the data direct to JIVE over data networks (NRENs and GÉANT2). For this to be successful the sustained data rate from each site back to JIVE should be at least 128Mbps, and ideally 512Mbps. However, early tests were not been able to *consistently* sustain even the low rate of 128Mbps. Tests conducted include live data transfers and 'iperf' tests, using both TCP and UDP.

The EVN case has been complicated by the fact that they rely on very specialist data processors for collecting or transmitting (depending in whether the transfer will be via hard disk pack or the network) the astronomical

⁴ Due to an error when the first case was opened, the 17 cases are actually numbered 2 through to 18.

data. These devices, Linux-based PCs called Mk5A systems, are located at each radio-telescope site and in JIVE (in JIVE there are multiple units – one is needed for each remote site involved in a given observation).

The Mk5A system was designed to be compatible with the existing data correlators, and as such the data mimics the format of the original VLBI magnetic tapes. Specifically, each VLBI data frame consists of a 20 byte header and 2480 bytes of data. There is one such data frame per track, and many tracks are recorded simultaneously. The actual make up of the application protocol data units (PDUs), is hidden from the operators, and is done by proprietary hardware and software. VLBI data needs to be transferred at a specific rate, which may be 16Mbps or any harmonic of 2 greater than it, up to 1024Mbps (so 16, 32, 64, 128, 256, 512 or 1024 Mbps). Although TCP is the normal method used for transporting VLBI data, it is possible to use UDP, EVN experience had been that there was no great improvement in using it (from the end-user's point of view).

Because in bench tests Mk5A systems (connected back to back) have had no problems with high speed data transfers, the e-VLBI community suspected the root cause of the problems as being network related.

A.1.2 Investigation

In order to investigate network performance separately from any other issue, bwctl (a wrapper program for the bandwidth measurement tool iperf) was installed on various machines across the e-VLBI network, including the Mk5As themselves. Bwctl was also available on two Linux workstations, in the GÉANT UK and IT PoPs. This meant that pure network tests could be run between EVN sites, between GÉANT sites, or between any combination of the two (it should be noted however that the UK-IT GÉANT route is not part of any actual e-VLBI path).

Bwctl showed that the poorest TCP performance was experienced on the path between the Torun radio-telescope (in Poland) and JIVE (with traffic tests only reaching 200-300 Mbps), so this was where effort was concentrated.

The PERT's investigation (lead by PSNC) found the bottleneck to be caused by core network devices installed in three locations in PIONIER network. These switches (10 GE Black Diamonds BD6808) offer two queuing regimes for the 10GE card: packet-based and flow-based. The former schedules each incoming packet to a different queue, introducing significant reordering. The latter preserves packet order, but limits the single flow capacity to the queue size (and this policy is implemented in PIONIER backbone). This is very unfortunate, because in ideal conditions (empty network) the size of a single flow cannot exceed 1Gbit/s (7Gbit/s remains unused and unavailable for that flow). The situation is even worse in the presence of Internet traffic (multiple different flows), where each queue already has some background traffic scheduled. As an example, if the single link has 4Gbit/s of traffic load, it means that average queue load is 500Mbit/s. Each new flow will encounter congestion conditions when its size reaches 500Mbit/s, even if there is still 4 Gbit/s of free bandwidth on the switch. The issue was fully described in the paper presented at TERENA 2005 Networking Conference: "Shall we worry about Packet Reordering?" available from the TERENA website at: http://www.terena.nl/events/tnc2005/programme/presentations/show.php?pres_id=74

At this stage it was clear that the most promising way forward is to make more use of advanced TCP implementations, and/or utilities such as Tsunami, which adds a reliable transfer layer on top of standard UDP. However it has been discovered that the Mk5As cannot use the newer Linux 2.6 kernel, because of problems with the required Jungo drivers. The designers of the Mk5A (based at the MIT Haystack observatory, MA, USA) were notified and are trying to resolve this problem.

In the meantime to work round this problem the PERT recommended a trial deployment of an American-developed system called the Logistical Session Layer, which sub-divides long TCP connections into several concatenated short connections, which has been shown to improve TCP performance in packet loss conditions. Full details of the LSL and its use in the EVN are given in Appendix B of this document.

A.1.3 Outcome

Although the Black Diamond switches are still in place, in May PSNC commissioned a new circuit (wavelength) between Poznan and Gdansk, and thereby decreased the load on the path Torun to Poznan. As a result, there was a significant improvement for the Torun to JIVE traffic, such that now a consistent 700Mbps plus can be achieved with iperf.

A.2 DEISA TCP throughput suffers from cross traffic

Case Number	Case Opened Date	Case Closed Date	Rating (%)	Author
3	30 May 2005	01 Mar 2006	80	Toby Rodwell (DANTE)

DEISA is a consortium of leading national supercomputing centres that currently deploys and operates a persistent, production quality, distributed supercomputing environment with continental scope (<http://www.deidsa.org>). To ensure the ongoing performance of their network connectivity DEISA sites perform single-stream TCP measurements (using iperf) between one another. One set of such tests run between Juelich (DE), IDRIS (FR) and CINECA (IT). While these tests normally run at 900 Mbps throughput, it was seen that this could slow down to ~400 Mbps when an unrelated project performs high-rate (2-3 Gbps) UDP tests between Karlsruhe (DE) and CERN. This was particularly surprising because the cross-traffic was 'Less than Best Effort' traffic, and was therefore not expected to compete for bandwidth with the normal, 'best effort' DEISA traffic. By the time the PERT investigation had begun the cross traffic itself had stopped, and was not expected to resume anytime soon. This, coupled with a change in GÉANT routing (as the GÉANT network evolved in to the GÉANT2 network) lead to the case being closed.

A.3 Performance Problems between FNAL and DESY

Case Number	Case Opened Date	Case Closed Date	Rating (%)	Author
4	19 Jul 2005	-	-	

A.3.1 Background

The German Electron Synchrotron (DESY) is one of the leading accelerator centres in the world. DESY has locations in Hamburg and Zeuthen (Brandenburg). The Fermi National Accelerator Laboratory (Fermilab) is an accelerator centre in Illinois, USA.

DESY contacted the PERT because although they were connected to the German NREN G-WiN with 1 GE, and Fermilab were connected to ESNET with an OC-12 (622Mbps) link, data transfer between the two was limited to approximately 150Mbps.

DESY have reported that they have no problem to exchange data with CERN (CH) at almost line rate (close to 1 Gbps) which indicates that there is no bottleneck on the route DESY - G-WiN - GÉANT - CERN.

A.3.2 Investigation

Initial progress was slow. Several iperf tests were run using 3rd party devices i.e. a device different from the A and B end, but these were inconclusive and it was deemed necessary to run all subsequent tests from the LANs in question (FNAL and DESY). Even then, regular access to a suitable machine was not always possible (this is quite often the case when investigating).

In order to get some continuity in the PERT investigation, Chris Welti from SWITCH volunteered to become the 'Special Case Manager' (SCM) for this issue, which meant he and not the weekly-changing Duty Case Manager would concentrate on the case. This was good in one sense, but it also added a further delay to the investigation since Chris was of course not always available to work on the case.

The first breakthrough came at the end of November. The SCM co-ordinated an end-to-end iperf test, with people from all the involved networks on a conference call, and each checking how many packets made it across their network (packet counters were set up on the boundary interfaces, so packets-in could be compared with packets-out). At this point it became apparent some packets were being lost on the GÉANT FR-UK link. A careful look at this circuit showed that there were consistent framing errors on the FR receive side. COLT (FR-UK provider) were slow to fix the problem so as a temporary work round an MPLS LSP was used to engineer FNAL-DESY traffic away from the FR-UK link. Eventually (26 Dec) the faulty hardware was traced and replaced, and the LSP bypass removed. End to end performance was still less than expected (13MB/s compared with 70MB/s to CERN), but on 27 Jan after getting root access to 2 DESY machines the PERT SCM was able to find out a problem with the line card of the switch the 2 workstations were connected to.

When running 500Mbps UDP streams between the 2 hosts, which were in the same subnet, packet loss of up to 1% was seen (in comparison, the UK-FR packet loss was 0.2%). The DESY staff then connected the two hosts to a newer high performance line card which was also less congested and it seemed that this helped to eliminate the issue. Afterwards the SCM able to reach a steady 700Mbps from fntst-1.fnal.gov to both end hosts for a period of 5 minutes with a iperf TCP test with 10 parallel streams and 2MB window size at the sender and the receiver (the use of parallel streams helps to reduce the effect of cross traffic - 10 small streams get better service than one large stream if there is other traffic on a link).

A.3.3 Outcome

The UK-FR framing errors had not been detected by the GN2 NOC because the router did not record them as 'input errors'. Juniper (the router manufacturer) were asked about this and they said that there was an internal debate as to what exactly ought to be classified as an input error. Significantly, the current MIBs did not allow the counting of framing errors, and as result of the PERT's comments they raised the priority of the Problem Report (PR) that should fix this omission. In the meantime GÉANT NOC will start monitoring low threshold framing errors which should identify future problems of this sort.

Currently the DESY network connection is fully committed to running tests with CERN which means that it has not been possible to carry out a final verification test. As such the case is suspended.

A.4 Low throughput MU - LSU

Case Number	Case Opened Date	Case Closed Date	Rating (%)	Author
5	29 Jul 2005	27 Sep 2005	40	Toby Rodwell (DANTE)

The PERT was contacted by CESNET who reported low IP throughput (about 32Mbps at UDP) between workstation at Masaryk University (MU) and Louisiana state university(LSU). Both workstations have 1GE card, but the bottleneck was the OC3 (155Mbps) link between LSU and Abilene.

In August the PERT was informed by LSU of the existence of an old, software-based router suspected of causing the problems. Before the PERT could progress the matter hurricane Katrina struck Louisiana and, understandably, LSU turned their attention to other matters. Out of respect for their situation the PERT did not press LSU for final verification, but rather closed the case with an appropriate resolution rating.

A.5 Slow upload from DANTE offices to PERT TWiki at SWITCH

Case Number	Case Opened Date	Case Closed Date	Rating (%)	Author
6	8 Sep 2005	25 Sep 2005	80	Toby Rodwell (DANTE)

DANTE staff noted that uploading documents to a SWITCH server was taking an unacceptably long time, for example a 43kB document took over 15 minutes to transfer. Closer inspection determined that the problem was local to the DANTE source host and that, for reason which are not understood, the work round was to disable (on the source) TCP window scaling. Although this limits the host's TCP window to less than 64kB, this is not a serious concern for a desktop computer user.

A.6 Poor ftp performance to ftp.ncbi.nlm.nih.gov

Case Number	Case Opened Date	Case Closed Date	Rating (%)	Author
7	14 Nov 2005	23 Nov 2005	100	Toby Rodwell (DANTE)

IUCC users reported they were only able to get about 3Mbps FTP throughput to NIH (US National Institutes of Health) and it seemed the NIH ftp site repeatedly changed address between 130.14.29.30 (via Abilene) to 165.112.7.10 (via the commercial Internet). However, after a week of running tests between various European FTP servers IUCC were satisfied that, first, there was nothing wrong with their FTP servers and host, and second, the problems with NIH were of a temporary nature.

A.7 DEISA-Teragrid performance

Case Number	Case Opened Date	Case Closed Date	Rating (%)	Author
8	14 Nov 2005	5 Jan 2006	20	Toby Rodwell (DANTE)

DEISA contacted the PERT via DANTE to request help with the poor TCP performance they were experiencing in data transfers between their site in Garching and the San Diego Supercomputer Centre (SDSC), which were being run as part of the SC2005 Teragrid tests.

Between Garching (connected with a GigE to DFN's network) and the San Diego supercomputing centre the performance measured by IPerf showed that UDP could easily achieve 800Mbps (with only 0.006% packet loss) but TCP only reached around 30Mbps.

The network path between Garching and SDSC, which had a Round Trip Time of approximately 180ms, was as follows:

- Garching computer centre was connected with a Gigabit Ethernet to DFN's network. This interface was dedicated to this purpose.
- DFN routed the IPv4 traffic to the GÉANT router, de1.de.geant.net
- LSPs in GÉANT sent the traffic via the Amsterdam-NY link
- ny1.ny.geant.net was connected to MANLAN (Cisco 6513) with a 10GE. VLAN 520 was assigned for this demo.
- Internet2 (I2) carried this VLAN between NY and Chicago by using a CCC tunnel.
- I2 was connected via 10GE to a Teragrid router in Chicago.
- Teragrid routed the traffic to the San Diego supercomputing centre.

By increasing "tcp_sendspace" and "tcp_recvspace" DEISA were able to get some performance improvement, but the demo finished before fully satisfactory results could be achieved.

A.8 DEISA decrease in throughput between IDRIS and other sites

Case Number	Case Opened Date	Case Closed Date	Rating (%)	Author
9	21 Nov 2005	10 mar 2006	100	

Beginning in October, DEISA observed a decrease in the achievable throughput (as measured by iperf) between IDRIS and their other sites; pre-October 800Mbps was the norm, after October this fell to 600-640Mbps.

The PERT's investigation determined this decrease in throughput was entirely due to the routing changes in the international backbone network that occurred as a result of the migration from the GÉANT to GÉANT2 networks. Once the new GÉANT2 CH-FR link became operational (a few weeks later than planned), the measured throughputs returned to their normal, high levels.

A.9 Opera Oberta: preliminary measurements

Case Number	Case Opened Date	Case Closed Date	Rating (%)	Author
-------------	------------------	------------------	------------	--------

Project:	GN2
Deliverable Number:	DS3.6.2
Date of Issue:	14/09/06
EC Contract No.:	511082
Document Code:	GN2-06-096v5

10	12 Dec 2005	13 Dec 2005	80	Simon Leinen (SWITCH)
----	-------------	-------------	----	--------------------------

A.9.1 Background

Ecole Polytechnique Fédérale de Lausanne (EPFL) wanted to receive high-quality multicast transmissions from the Liceu opera in Barcelona, as part of the "Òpera oberta" project. Past tests had been unsatisfactory, with packet loss and reordering. EFPL contacted the PERT to get help and advice on what they should do to improve their reception.

A.9.2 Investigation

The PERT duly oversaw some preliminary tests. The first transmission, to IP address 227.142.142.1, experiences some performance problems (visible artefacts), apparently due to overload of the router at the sending site in Barcelona. When the transmission was started again using a different address, 227.142.142.100, these problems disappeared. It was not clear why the choice of IP group address had an impact on performance. Subsequent testing passed off satisfactorily.

A.9.3 Outcome

On successful completion of the tests the PERT case was closed, and the following notes made. First, 227.*.* addresses should not be used on the Internet, because this range of addresses has not been allocated by IANA. Some networks/sites may actually filter multicast traffic to such addresses. An interesting alternative would be "GLOP" address space. RedIRIS can assign addresses to the Òpera oberta project from RedIRIS' GLOP range, 233.2.254.0/24 - derived from RedIRIS' AS number 766 ($766 = 2 \times 256 + 254$). Second, it would be good to have a monitoring infrastructure in place that would allow to quickly detect degradation of performance for multicast transmissions such as Òpera Oberta's. Transmission quality could be measured at the endpoints, using packet trace analysis or using mechanisms integrated in the decoding programs. Also, traffic could be counted in the backbone using ACL (access control list) entries, Netflow, or other accounting mechanisms. Finally, it should be possible to get some indication of transmission quality using `mtrace`.

A.10 Large packets lost between Onsala and Stockholm

Case Number	Case Opened Date	Case Closed Date	Rating (%)	Author
11	23 Jan 2006	14 Feb 2006	60	Ana Pinto (FCCN)

A.10.1 Background

Onsala observatory contacted the PERT about a problem with traceroute - it did not work as expected between an end-system in Onsala (a Mk5A system) and a SUNET test server in Stockholm. It seemed that traceroute packets between the Onsala Mk5A and the SUNET test system were being dropped

A.10.2 Investigation

Discussions between the PERT and SUNET determined there were no MTU problems (which had been the primes suspect) on the path between the test server and Onsala Mk5A - the SUNET network core had an MTU of 8192 bytes and the end-systems Gigabit Ethernet cards used an MTU 4470.

It was determined that traceroute only did not work when run after iperf test, and this was true when running iperf and traceroute between the Onsala Mk5A and any other end system, not just the SUNET test system. To 'cure' traceroute the Mk5A had to be re-booted. It was also determined that the problem was particular to traceroute and did not affect, for example, tracepath.

A.10.3 Outcome

Given that this problem only affected this Mk5A, and because there was a perfectly good work-around (which was, to use tracepath instead of traceroute), the end-user (Onsala Observatory) agreed that no further investigation was necessary and the case could be closed.

A.11 Peering problem with AS3300?

Case Number	Case Opened Date	Case Closed Date	Rating (%)	Author
12	16 Feb 2006	16 Feb 2006	100	Toby Rodwell (DANTE)

IUCC reported large packet loss on the path between IL and AS3300 (Eqip, France), which is via GN2. Traceroute results were misleading because of the use of MPLS by Eqip. Before the PERT had a chance to locate the anomaly the packet loss stopped by itself and the end-user agreed to closing the case.

A.12 Low performance between Budapest and Trieste

Case Number	Case Opened Date	Case Closed Date	Rating (%)	Author
13	7 Mar 2006	23 Jun 2006	60	Toby Rodwell (DANTE)

HUNGARNET reported poor performance between Budapest and the International Centre for Theoretical Physics (ICTP) in Trieste. Using SCP (secure copy) only 1.5Mbps could be achieved, which was less than expected or hoped for. However, the PERT discovered that the ICTP access link to GARR was heavily congested, sometimes even saturated, and this was determined to be the root cause of the poor throughput.

A.13 Below expected throughput, Budapest to New York

Case Number	Case Opened Date	Case Closed Date	Rating (%)	Author
14	12 Apr 2006	-	-	Toby Rodwell (DANTE)

A.13.1 Background

DANTE contacted the PERT to report they were experiencing below expected TCP throughput between the GN2 Measurement Point (MP) in Budapest and the MP in New York. Because the MPs are co-located with the GN2 routers in the GN2 PoPs, then on the non-lossy GN2 backbone TCP rates of 900Mbps should be consistently achievable, even over long distances. However, the TCP iperf tests run consistently achieved only around 600Mbps.

A.13.2 Investigation

An in-depth study of the iperf TCP connection between HU and NY, achieved used a script that automatically collected netstat statistics, determined that once the TCP congestion window had grown to approximately 6000 segments there was some kind of congestion event which seemed to set an upper limit on the size the window could grow to (and thus the achievable throughput). Checks showed the congestion event was not packet loss, nor was it Ethernet flow control. There was clearly some packet re-ordering (which is anyway to be expected when sending high flows through a Juniper M160) because SACK (Selective ACKnowledgement) packets were being generated. However, the SACKs themselves did not cause the problem since disabling SACK did not improve the throughput. It was thought that the inability of TCP to exceed the rate when it first inferred congestion might be due to a BIC bug (BIC was the TCP congestion avoidance algorithm in use). However,

changing for BIC to the older Reno algorithm did improve the situation. The next step is to try other high-speed variants of TCP (such as H-TCP), but this will require an operating system upgrade to Linux 2.6.16

A.14 ITER VPN connection

Case Number	Case Opened Date	Case Closed Date	Rating (%)	Author
15	24 Apr 2006	14 Jul 2006	60	Toby Rodwell (DANTE)

The ITER project reported performance problems between their sites in Garching and Cadarache. Their connection ran over a 100Mbps Internet VPN tunnel Checkpoint Firewall1 firewalls at each site and was enhanced with Packeteers SkyX system. However, they reported they could only achieve 10-20Mbps throughput (peaking at approximately 35Mbps).

The PERT discovered that contrary to ITER's understanding of the path between the two sites, there was actually an E3 (34Mbps) bottleneck between Grenoble and Cadarache. This link was often congested, which explained why ITER often only achieved 10Mbps throughput.

The case was then closed, as the only feasible solution was to increase the capacity of the Grenoble-Cadarache circuit.

A.15 Routing to USA going via Japan

Case Number	Case Opened Date	Case Closed Date	Rating (%)	Author
16	16 May 2006	31 May 2006	80	Toby Rodwell (DANTE)

IUCC reported a sub-optimal path to a network in the US, that was being routed via GÉANT2 then TEIN2. It was found that this prefix was being advertised by Mich[igan]Net to Abilene and Starlight, but whilst Abilene discarded the route, Starlight advertised it on to APAN, and on to TEIN2 and GÉANT2. APAN and TEIN2 have now stopped advertising this route and a more direct path is taken to the network.

A.16 Slow file transfer UK to various European DEISA sites

Case Number	Case Opened Date	Case Closed Date	Rating (%)	Author
17	30 Jun 2006	13 Jul 2006	100	Toby Rodwell (DANTE)

The European Centre for Medium-range Weather Forecasting (ECMWF), which is based in Reading, UK is part of the DEISA project. They contacted the PERT when they noticed slow file transfers to DEISA sites over GÉANT network (typically only 40Mbps) which were very poor compared to an FTP test to a JANET (UK) server (which achieved 300Mbps). The poorest results were between the UK and a site in Barcelona (ES).

The poor TCP transfer rates (10-40Mbps) seen between ECMWF (UK) and the sites in DE, IT, FR and FI, were due to sub-optimal TCP tuning of the ECMWF host. After the PERT-recommended TCP settings were applied transfer rates increased by up to ten times (250-450Mbps). The path to ES was also affected by packet loss on the access circuit between RedIRIS and GÉANT2. Having been identified by the PERT, the degraded circuit was repaired and the packet loss stopped.

A.17 Slow file transfers between TIGO (CL) and JIVE (NL)

Case Number	Case Opened Date	Case Closed Date	Rating (%)	Author
18	16 Aug 2005	-	-	

A.17.1 Background

JIVE is currently involved in a program to track the SMART-1 spacecraft in its orbit around the moon. This involves telescopes in South America (TIGO in Chile and Fortaleza in Brazil). Several test observations have already been done, and data has been transferred electronically, for express analysis. The results were a little disappointing. So far there have been only a few, short transfers of TIGO data. The first test was done with FTP, and was subject to a locally imposed 5Mbps rate-limitation – 1.6Mbps was achieved. The second test used the Mark5 file transfer utilities and a tuned TCP stack, and the 5 Mbps restriction (imposed by the TIGO provider) was relaxed. Even though the path bottleneck was (reportedly) 100Mbps, only 7Mbps was achieved. A week later there was a third, FTP test, and this achieved approximately 1.4Mbps. In separate tests JIVE deduced that there was little congestion on the 5Mbps TIGO link, since they were able to send UDP test packets at a rate of up to 4.6Mbps before packet loss was seen.

A.17.2 Investigation

For convenience the PERT began by running iperf tests in the reverse direction of normal data flow (Europe to Chile rather than Chile to Europe). It was seen that TCP parameters were negotiated well, including window scaling and large windows. It was also noted that Selective Acknowledgements (SACK) had been deactivated on the TIGO machine, and it was recommended this be reactivated. This seemed to have a positive effect, as subsequent tests achieved about twice the throughput, (3Mbps compared with 1.5Mbps). It has not yet been possible to see if the SACK will have a similar benefit for an unlimited data transfer (that is to sat, as and when the 5Mbps limitation is lifted).

Studies indicate that the 5Mbps limiter is a simple policer, with no Random Early Detection (RED) or traffic shaping. These methods might improve throughput (if the rate limiter needs to be kept in place).

Appendix B **Logistical Session Layer (LSL)**

By Professor Martin Swany, University of Delaware

B.1 **LSL Overview**

The Logistical Session Layer (LSL) is a layer of network middleware that enables improved network throughput via short-term buffering and cooperative forwarding in the network. The basic model is that end-to-end communication between hosts is no longer bound directly to the Transport layer, but rather to a Session layer that is semantically similar to that defined in the OSI model. Recall that a transport-layer connection may consist of multiple network-layer hops. Analogously, a session-layer connection may consist of multiple transport-layer connections. In fact, the ISO standard specifically allows the binding of a single session "connection" to multiple transport connections. The session layer as designed by the ISO was never widely deployed.

The architecture of the Logistical Session Layer is relatively straightforward. Each session begins with a header that includes a 128-bit session identifier. The header also includes a source and destination IP address (version 4 currently); a 16-bit port number; and a 4-bit Version and Type fields to allow for future modification of the header format. Finally, there is a header length field, as the size of the header will vary when it contains options.

The connection from source to sink can transit one or more session routing processes (depots) and make use of multiple TCP "sublinks". This manner of using multiple TCP connections can be thought of as "serial" rather than "parallel" connections. For purposes of this discussion, we assume that all connections occur synchronously, i.e. the sender and receiver exist at the same time. We note that an asynchronous session is possible with the receiver discovering the session identifier and reading the data from the last depot. The architecture of the system and performance early performance experiments have been presented in IEEE workshops in Cambridge, MA⁵, and Pittsburgh, PA⁶ ..

⁵ M. Swany and R. Wolski, Data Logistics in Network Computing: The Logistical Session Layer, IEEE Network Computing and Applications, Cambridge, MA, October 2001

⁶ M. Swany, Improving Throughput for Grid Applications with Network Logistics, Proceedings of IEEE/ACM Conference on High-Performance Computing and Networking (SC2004), Pittsburgh, PA, November

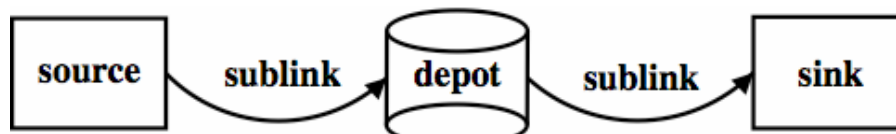


Figure 1

One significant benefit of this approach is that it has predictable impact on the network. The stability and fairness are known as the system relies on TCP connections between depots. The impact on the network is not in question and the system is safe for incremental deployment atop the existing network infrastructure. These deployments can be made in places where network performance is critical, but insufficient at present.

TCP is known to have performance issues on networks with high bandwidth-delay product. Modern high-performance backbones such as GÉANT2 are extremely well engineered and have very low loss rates. Many shared access links do, however, have some amount of loss. When that loss is coupled with a long latency, the performance of TCP suffers. LSL can help mitigate this issue by dividing areas of the network into different sublinks.

Early results have demonstrated that end-to-end throughput can be improved by applying the LSL technique. Figure 9.2 shows the results from a test running from coast to coast in the US. The distance is approximately 2700 miles and the network path exhibits about 85ms round trip time. The machines at either end were running Linux 2.4 and were configured with 8MB TCP kernel buffers. The hosts were connected to 100Mbit Ethernet. The LSL path in this case used depot nodes at the edge of the Internet2 backbone so essentially the edges (where some congestive loss is likely) were isolated from the long-haul backbone (where provisioning is sufficient to insure that loss is unlikely.) As Figure 9.2 illustrates, LSL can dramatically improve the observed throughput across a range of data transfer sizes.

LSL vs. Direct Transfers from U. Del to UCSB
 LSL Nodes in Washington D.C. (WASH) and Los Angeles (LOSA)

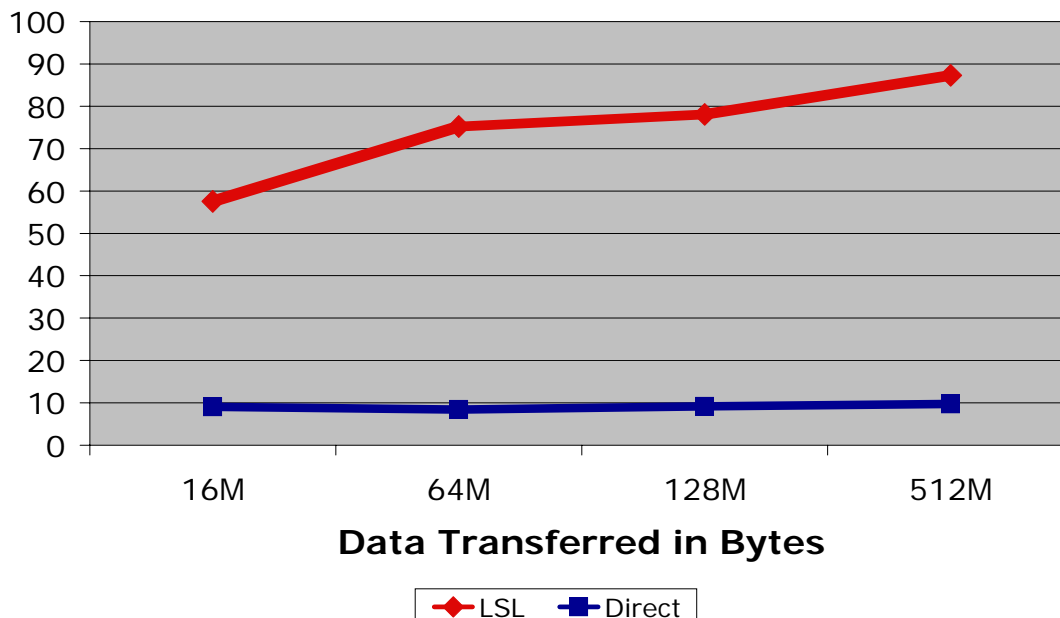


Figure 9.2: LSL v direct transfers, University of Delaware to University of S. Carolina Beaufort

B.2 LSL and EVN

The European VLBI Network (EVN) is an interferometric array of radio telescopes spread throughout Europe. Network connectivity is a critical for the EVN's effort in real-time VLBI analysis (termed e-VLBI), as the data must be sent from the various telescopes to the central facility at the Joint Institute for VLBI in Europe (JIVE) in Dwingeloo, The Netherlands. Among the issues facing the EVN was the fact that the eVLBI data capture and processing machines, known as Mk5As, are very specialized and somewhat underpowered for high-performance networking. In particular, for a long time the site in Torun, Poland did not enjoy sufficient bandwidth to the JIVE facility, and so the JIVE-Torun path seemed like an excellent candidate upon which to evaluate the LSL approach.

While the immediate goal of this effort was to improve the performance from Torun to JIVE, the PERT sought a general solution that could be deployed for other critical applications running on GN2. Thus, one goal of the planned evaluation was to minimize the effort necessary to change between LSL-enabled and standard operation. That is, it needed to be possible for the LSL to be transparently and easily enabled and disabled. The longer-term goal for investigating the integration of LSL into the GN2 environment was to explore the possibility to deploy LSL systems for applications that require additional performance.

Project:	GN2
Deliverable Number:	DS3.6.2
Date of Issue:	14/09/06
EC Contract No.:	511082
Document Code:	GN2-06-096v5

To achieve our goals for deployability, several different models for enabling LSL operation were investigated. The native API for LSL is based on the BSD sockets interface and the most basic way to use LSL is replace the socket API calls with corresponding ones prefixed by "lsl_". To make things easier for applications running on Unix and Linux, a mechanism was developed to "preload" the LSL library and have all the socket calls in an application replaced by their LSL counterparts. This can take place without recompilation of the application and can be as simple as a shell script that is run before the application is started.

The application-level library "preloading" is simple enough, but for this e-VLBI application, even more transparency was desired. To this end, the concept of an LSL machine as an "appliance" in the network was devised. In this model, the LSL machine would have two interfaces and would be placed between the router and the application machine. This mode of operation is referred to as the "transparent" mode and it takes advantage of the Linux IPTables packet-filtering infrastructure. IPTables allows certain packets to be matched and processed. This is the technology that allows Linux machines to act as transparent NAT boxes, for example. The 'transparent mode' approach is very similar, except that rather than simply translating addresses, the LSL box will cause an outgoing TCP connection to be processed as an LSL session. In this way, the LSL machine can be simply placed between host and router and is easy to remove if necessary.

B.3 Early Evaluation of LSL in GN2

Evaluating the effectiveness of the LSL approach for eVLBI was somewhat challenging. First, powerful PC-class systems had to be purchased. While LSL has been previously proven for speeds up to 100Mbit, it was not certain how close the LSL could come to fully utilizing a 1 Gigabit link with PC hardware. To give maximum chance of success, JIVE were advised to purchase relatively high-end systems (based on the AMD Opteron processor). These systems were chosen since their memory bandwidth is high and the memory performance is key to the current LSL implementation. Although Intel Ethernet cards had been specified, the vendor actually shipped motherboards with an integrated Ethernet solution from nVidia. Support for this chipset was relatively new in the Linux kernel and early tests found some problems with their use. Specifically, although the chip used supports 2 Gigabit Ethernet ports, the LSL model places makes big demands on its inbound and outbound interfaces and using both ports of the nVidia chip caused the driver to become unstable. Later versions of the kernel have at least partially addressed this problem and recent tests have shown the machines able to support nearly 700Mbps.

Supporting the transparent mode of operation for LSL also proved to be a challenge. The first issue that became apparent was that complete spoofing of addresses is quite difficult. Essentially, the initial plan called for an LSL node to route from a "fake" address X on one interface, to the "real" address X on another interface. Given a single routing table in the kernel, this proved to impossible without kernel-level modifications to Linux (although alternative solutions are still under investigation.) The fallback plan was to have the LSL node on the same subnet as the application machine and simply change the "default route" on the host machine from e.g. X.Y.Z.1 to X.Y.Z.2. This has the added benefit of avoiding having to re-cable in order to enable LSL, plus this change is at the operating system configuration level and requires no modification to the application whatsoever. This mode is further judged to be acceptable since the Torun site is accustomed to changing routing tables for production and testing runs.

Project:	GN2
Deliverable Number:	DS3.6.2
Date of Issue:	14/09/06
EC Contract No.:	511082
Document Code:	GN2-06-096v5

This approach proves completely workable, but pointed toward another issue. The model now has an LSL machine with two interfaces in the same subnet. Figure 9.3 depicts the current configuration. The issue is that both interfaces of “Torun LSL1” are in the same IP subnet and the basic routing mechanism on Linux provides no way to specify that a connection coming in one interface should go out the other. After identifying the problem, it has been resolved at the LSL layer.

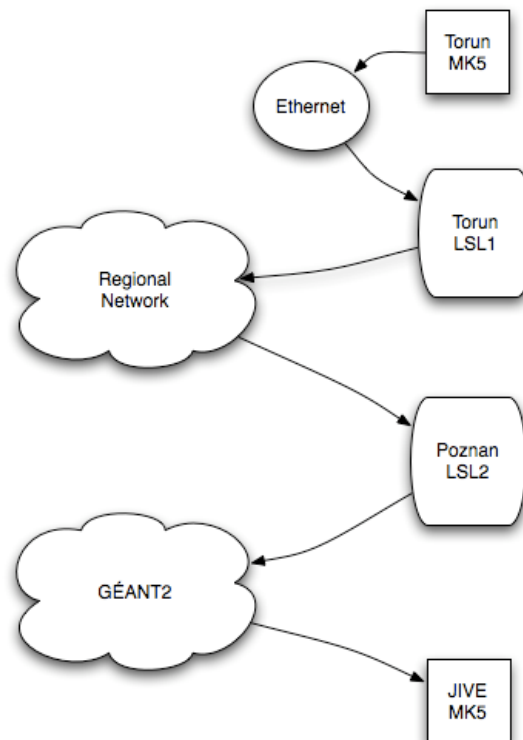


Figure 9.3: LSL locations on Torun-JIVE path

Figure 9.3 shows the current status of the testbed. Due to various timing issues with eVLBI experiments and tests, to date it has not been possible to perform a complete test. The machine room in Torun is currently offline for physical renovation but it is hoped that experiments will be able to restart again soon.

Project:	GN2
Deliverable Number:	DS3.6.2
Date of Issue:	14/09/06
EC Contract No.:	511082
Document Code:	GN2-06-096v5