

05.05.06

Deliverable DS3.7.3: Report on Performance Monitoring Tools



Deliverable DS3.7.3

Contractual Date:	28/02/06
Actual Date:	05/05/06
Contract Number:	511082
Instrument type:	Integrated Infrastructure Initiative (I3)
Activity:	SA3
Work Item:	WI7
Nature of Deliverable:	R (Report)
Dissemination Level	PU (Public)
Lead Partner	DFN
Document Code	GN2-06-016v5

Authors: Roland Karch (DFN-Verein), Sven Ubik (CESNET), Toby Rodwell (DANTE), Vladimir Smotlacha (CESNET)

Abstract

In this report it is described how an array of performance monitoring devices has been deployed and set to work in the GÉANT2 network. The report begins with an overview of the different types of network performance metrics of interest to the GÉANT2 project and then looks at what tools were initially available to meet our needs. The evaluation processes used to select systems for deployment are then explained, and some of the issues discovered during testing are examined. Finally, the improvements and enhancements planned for future releases of the chosen applications are presented.

Table of Contents

0	Executive Summary	v
1	Introduction	1
2	Monitoring Overview	2
2.1	Active IPPM	2
2.1.1	One-way Delay	2
2.1.2	IP Packet Delay Variation	3
2.1.3	One-way Packet Loss	3
2.2	Passive Monitoring	3
2.3	Available Bandwidth	4
2.4	Timing Requirements	4
3	Evaluation and Deployment	6
3.1	Passive Monitoring	6
3.1.1	Overview	6
3.1.2	Software Architectures	7
3.1.3	Hardware Monitoring Adapters	8
3.1.4	Passive Tool Selection	9
3.2	Available Bandwidth	12
3.2.1	Overview	12
3.2.2	Tools Selection	13
3.3	IPPM	13

3.3.1	Tool Selection	14
3.4	Hardware Evaluation	16
3.5	Software Validation and Integration	19
3.5.1	Optimizing measurements	20
3.6	Configuration	23
3.7	Roll-out Plan	23
3.8	Roll-out Status	24
4	Recommendations for Enhancements and Extensions	26
4.1	Available Bandwidth	26
4.1.1	iperf	26
4.1.2	BWCTL	26
4.2	HADES	27
4.2.1	Time To Live	27
4.2.2	Operator's Display	28
4.2.3	Information Presentation	28
4.2.4	Alerts	28
4.2.5	System Monitoring	29
5	Conclusion and Next Steps	30
6	References	31
7	Acronyms	32
Appendix A	Network Performance Metrics	34

Table of Figures

Table 3.1: Comparison of CoralReef and SCAMPI	10
Table 3.2: Comparison of packet capture hardware	12
Figure 3.1: GEANT Dell MP OWD Measurement pattern	15
Table 3.3: Test results of candidate IPPM/AB measurement points	18
Figure 3.2: Back to back measurements using GPS based time synchronization and two identical measurement boxes to reveal accuracy and precision of OWD measurements	19
Figure 3.3: Measurement of OWD. The first packet of a packet group always shows an increased delay due to wakeup	20
Figure 3.4: Influence of the rate of CPU interrupts on achievable data rate	21
Figure 3.5: Correct achievable bandwidth measurements	22

0 Executive Summary

Traditional network monitoring systems typically measure average circuit load and status. Whilst these are important metrics for detecting equipment failure, and helping with capacity planning, they are not the best indicators of what users can expect to achieve in terms of high speed, long distance flows. This is because such flows are very susceptible to any kind of packet loss (even fractions of a percent), and any kind of congestion (even if it lasts only a few seconds).

Given GÉANT2's status as a high capacity, international backbone network, performance monitoring has always been important to its success, as without such systems it would not be possible to properly monitor the service given to users. The project's approach has been to evaluate a range of performance monitoring tools, and select the most suitable for further development and deployment in the network. A variety of systems for One Way Delay (OWD) measurement, Available Bandwidth (AB) measurement, and passive packet capture were assessed, and 2 applications were chosen for deployment - DFN's HADES OWD measurement system, and Internet2's BWCTL AB measurement system. A third system, CESNET's SCAMPI packet capture and analysis application, was selected for further study. There was then an additional evaluation process to select the hardware that would be used for the HADES/ BWCTL Measurement Points (MPs), which resulted in the German bee system being chosen.

Based on experience to date several improvements are planned in the design and use of HADES and BWCTL, including a better display for the operator (for HADES) and an expanded to schedule for BWCTL measurements so that AB measurements are taken during the working day.

Project:	GN2
Deliverable Number:	DS3.7.3
Date of Issue:	05/05/06
EC Contract No.:	511082
Document Code:	GN2-06-016v5

1 Introduction

In TCP/IP based networks, such as GÉANT2, to achieve and sustain high data rates over long distances requires that there be zero (or very low) packet loss, and no additional packet delay as the result of prolonged queuing in the network. Where before only specialist network researchers tried to obtain such high speed, long distance flows, it is now becoming much more common for academics from other disciplines to rely on such connections for their everyday work. Traditional network monitoring systems (which check for circuit status and average load) are not well suited to predicting high speed performance, and so a new breed of tools has been developed, using both active and passive techniques, in order to monitor and log the parameters important to high speed networking. In light of GÉANT2's commitment to international ventures with demanding network requirements, from the outset the project has a significant interest in both the research and deployment of Performance Monitoring Systems (PMS). These PMS are an integral part of the project's QoS initiatives, as they provide the means to verify the service being offered. With this in mind, the activity JRA1 has been dedicated to the research and development of performance monitoring tools, visualization techniques and a common framework, whilst SA3 Work Item 7 ("Evaluate and conduct initial deployment of the Performance Monitoring System") has concentrated on deploying systems for collecting performance data (called measurement points (MPs)) in the GÉANT2 network, with a particular view to validating Premium IP performance. Though they operate independently, the two activities agreed to select a common set of monitoring systems in order to allow a smooth transition of the systems, frameworks and tools developed within JRA1 to SA3.

The first section of this document gives an overview of the technologies that JRA1 and SA3 were interested in. Chapter 2 then describes the hardware evaluation process, by looking at the range of products considered for use by JRA1 and SA3, the decisions taken, the tool deployment issues, and the most relevant technical details of the measurement point and archive installations. Chapter 3 lays out the requirements and plans for the future development of the deployed tools, and finally Chapter 4 concludes with a summary of the activity's next steps.

Project:	GN2
Deliverable Number:	DS3.7.3
Date of Issue:	05/05/06
EC Contract No.:	511082
Document Code:	GN2-06-016v5

2 Monitoring Overview

This section describes the network performance monitoring methodologies and metrics which were of interest to SA3 and JRA1. The tools SA3 and JRA1 sought needed between them to provide the measurements described below.

2.1 Active IPPM

In the Measurement Methodology section of the IETF's IPPM Framework [RFC2330] the first example given for collecting IP performance data is that of "...Direct measurement of a performance metric using injected test traffic". Active IPPM refers to low bandwidth active probing, and it is able to give accurate results on the following metrics without adversely affecting production traffic on the measured path.

2.1.1 One-way Delay

One Way Delay (OWD) is defined in [RFC2679]. It is the time taken for a packet to travel from a point A in the network (typically designated by its IP address) to a different point B reachable through this network. OWD is sometimes confused with the Round Trip Time (RTT) which is the metric provided by, amongst other applications, the well-known measurement tool 'ping'. RTT measures 'two way delay', which is to say the delay of a packet being sent from point A to B, plus the delay of a different, reply packet making its way back from B to A. OWD is only concerned with the first leg of this journey and the advantage of this is that, the two paths (A->B and B->A) can be considered and measured separately. In addition to this very basic view, there are a lot of details visible only by measuring the OWD. While it is obvious that asymmetric bandwidth links (like ADSL lines) or routes that differ between A->B and B->A most certainly will show interesting OWD results, even some not quite obvious cases will profit. NTP as an application for example will only produce equal times between two networked hosts under the assumption that $OWD(A \rightarrow B) = OWD(B \rightarrow A)$. Bulk data transfers will be more vulnerable to differences in OWD on the path from source to destination.

Project:	GN2
Deliverable Number:	DS3.7.3
Date of Issue:	05/05/06
EC Contract No.:	511082
Document Code:	GN2-06-016v5

However, unlike RTT, in order to reliably measure OWD, both measurement points (hosts A and B) must be precisely synchronized. Any difference between the system clocks of hosts A and B will lead to an error in the measurement. Because of this, OWD measurement points benefit significantly from an external hardware time source, such as a GPS, CDMA or PZF receiver. All of these provide a time signal whose inaccuracy is less than that introduced by the host hardware and processing.

2.1.2 IP Packet Delay Variation

IP Packet Delay Variation (IPDV) is also known as One-way Delay Variation (OWDV) and, more colloquially, 'jitter'. IPDV is fully described in [RFC3393] but put simply it is the derivative of OWD. Because IPDV is concerned with differences in timing rather than absolute values, the timing requirements are somewhat relaxed. Instead of precise host system synchronization, the hosts need only to have an accurate clock frequency, so the test packets are sent at a precise and constant rate.

2.1.3 One-way Packet Loss

One Way Packet Loss is specified in [RFC2680]. Although packet loss is a relatively straight forward concept to understand, RFC2680 makes clear that packets which are not lost but have an unexpectedly very large OWD compared to that path's mean OWD must also be considered as lost. Other cases of received packets that should be considered lost are:

- Corrupted packets
- Fragmented packets which cannot be reassembled.

2.2 Passive Monitoring

Passive monitoring is a general term that describes the process of capturing and analysing passing traffic. This might be done on the same platform as the traffic's source and/or destination, or it might be done by a third (intermediate) device on the path between source and destination. An intermediate device might be placed in-line with the measured path or (more likely) it will not be physically in the path but rather take a copy of the measured path. The latter method can be achieved by using switch port mirroring (where all the frames sent out switch port 1 are also sent out of switch port 2, to which an intermediate passive monitoring device is attached) or by an optical splitter (which as its name suggests simply splits a fibre-optic in two).

Whilst even a modest PC is able to collect tcpdump data when the data rate is around 1 Mbps, when monitoring high speed links (1Gbps and above) then capturing and storing packets becomes a complex and expensive task. Most GEANT2 backbone links are 10Gbps and the specialist hardware required to capture packets at this rate costs in the order of €20,000. Of course, storing the captured data is also an issue. If you

Project:	GN2
Deliverable Number:	DS3.7.3
Date of Issue:	05/05/06
EC Contract No.:	511082
Document Code:	GN2-06-016v5

were collecting data at the rate of 10Gbps of data it would take just 7 minutes to fill up a 500GB hard disk. Partly for this reason, and partly for privacy reasons, it is usual to store only packet headers. In order to further reduce the volume of data to store on a hard disk and to reduce the load of host computer CPU, it is desirable to perform some monitoring functions, such as computing statistics, directly on the monitoring adapter and retrieve only results of these functions instead of packet headers. Classification of packets based on header filtering and packet header anonymization can also be done in hardware of the monitoring adapter.

A different approach to reduce the difficulty of high speed passive monitoring is to sample the data flow rather than collect every packet. This technique is already used by GÉANT2 routers which create NetFlow data from incoming traffic sampled at the rate of 1 in 1000.

2.3 Available Bandwidth

Many people believe that the term Available Bandwidth (AB) is today being used improperly. They would argue that in its purist sense AB is quite simply that portion of a link's capacity which is not in use but that term is now used to refer to the end-to-end data rate that a given system can achieve, and that this is more properly called Achievable Throughput. Be that as it may, AB is now commonly accepted to be synonymous with Achievable Throughput, and it is used as such in GÉANT2.

AB can be measured directly or indirectly. A direct measure of AB can be considered a destructive test, in that it (the AB application) measures available bandwidth by consuming it, and thus making that bandwidth temporarily unavailable! In contrast an indirect AB measurement tool infers AB from the analysis of other metrics, based on the fact that throughput is a function of path capacity, delay and loss. Despite the benign nature of indirect AB measurement tools compared to direct, they are not in common use, mainly because there are no widely available, effective, easy-to-use products. We can also compute AB as a complement of used bandwidth (measured by passive monitoring) to the fixed installed (or administratively constrained) bandwidth, but in this case we also need to consider and quantify burstiness of used bandwidth.

2.4 Timing Requirements

All monitoring sites need to know exact time in order to:

- timestamp and schedule events (e.g. start of measurement session). A time accuracy in the order of millisecond is good enough for this purpose.
- give time sensitive measurement (e.g. measurement of OWD) or assign accurate timestamps to captured packets. In this case the absolute time accuracy is essential and should be not worse than 10 microseconds, in order to make measurements with an acceptable error.

This significant difference in acceptable time accuracy for these two requirements indicates that different methods of clock synchronization are appropriate:

- standard NTP synchronization with remote time servers is suitable when only millisecond accuracy is required,
- an external clock (GPS, GSM or DCF receiver) allows computer clock synchronization with an accuracy of 10 microseconds or better,
- a local primary time server provides accuracy of about 50 microseconds, however a directly connected external clock should be used whenever possible.

When NTP time synchronization is used for time sensitive measurement, the network path under measurement must not to overlap with the path between the time server and the point of measurement.

Hardware monitoring adapters such as DAG and SCAMPI are not dependent on the system clock of host computer. They use high quality clocks with external pulse per second (PPS) signal synchronization so that the absolute time accuracy is about 1 microsecond. The clock resolution is better than 50 ns which allows the assignment of unique timestamps to each captured packet even on a loaded 10 Gbps line.

3 Evaluation and Deployment

This chapter describes the system evaluation and selection process SA3 conducted jointly with JRA1. There is an emphasis on the Measurement Points (MPs) deployed to date, which measure OWD and AB. Although SA3s budget allows for the procurement of a number of Passive Monitoring devices, the large cost of such equipment (particularly the 10Gbps equipment) has dissuaded SA3 from deploying them until JRA1 has completed a full and detailed assessment of their use for performance monitoring. A list of existing and planned measurement points is given as well as an overview of the security policies, access privileges and technical details.

3.1 Passive Monitoring

3.1.1 Overview

There are two comprehensive software architectures for the processing of passively captured data:

- CoralReef developed by CAIDA
- SCAMPI developed by the SCAMPI project

For capturing high data rate traffic flows specialist hardware is required. Such hardware monitoring adapters are currently available from the following sources:

- DAG cards produced by Endace
- Combo cards developed by SCAMPI project
- Napatech cards
- Force10 cards

3.1.2 Software Architectures

3.1.2.1 CoralReef

CoralReef was first released in March 1999, with development continuing steadily until 2003. It is a set of drivers, libraries and tools to process data captured by FORE ATM cards, or by DAG cards (up to OC-12 speed), or obtained from standard libpcap library (that is, using a standard NIC).

Drivers are available for Linux (up to 2.2 kernel) and FreeBSD.

Libcoral library provides functions to request header filtering (implemented in software by libpcap library) and IPv4 address anonymization.

The following applications are available:

- To store captured packet trace in tcpdump format or ERF format used by DAG cards
- To compute interface statistics (count of IPv4, IPv6 and non-IP packets)
- To compute volume of traffic flowing between networks, Autonomous Systems (ASes) (using routing tables) and countries, using a related tool that reads the whois database (note, development of this tool has stopped)
- Port summary matrices for TCP and UDP, packet and byte counts by IP length and protocol
- DNS message statistics
- Flow-based statistics

3.1.2.2 SCAMPI

SCAMPI (Scalable Monitoring Platform for the Internet) is a framework developed by the SCAMPI project that allows the easy creation of passive monitoring applications using MAPI (Monitoring API). A user defines one or more “flows”, where a flow is a sequence of all packets arriving on a specified network interface. Then a sequence of monitoring functions can be applied to each flow. These functions perform header filtering, statistics computing, sampling, payload searching and Netflow generation. Applications can read either the packets themselves (after header filtering, sampling and payload searching) or just the statistics from the monitoring adapter.

SCAMPI includes (in sequence from the monitored network line up to the application) a hardware monitoring adapter, drivers, libraries, MAPI and sample applications. Applications are independent of the used hardware monitoring adapter, which can be a COMBO card with SCAMPI firmware (developed within SCAMPI project), a DAG card, or even a regular NIC. MAPI automatically takes advantage of whatever hardware acceleration is available in the used hardware monitoring adapter and implements the remaining monitoring functions in software.

3.1.3 Hardware Monitoring Adapters

3.1.3.1 DAG cards

DAG cards (<http://www.endace.com/networkMCards.htm>) are currently available for Asynchronous Transfer Mode (ATM), Ethernet up to 10 Gbps, and Packet Over SONET (POS) up to 10 Gbps. They can capture packets at a line rate for all packet sizes, but do not process them in any way. A coprocessor is available for Gigabit Ethernet card and for POS cards up to 2.5 Gbps, which can do certain processing (particularly header filtering). DAG cards are a production quality commercial product and they are currently the most readily available and widely deployed of all known monitoring cards.

3.1.3.2 COMBO cards with SCAMPI firmware

Combo cards (<http://www.ces.net/doc/2004/research/proghw.html>) are available for Gigabit Ethernet (there is a twisted-pair version and an SFP version, which uses interchangeable transceivers for either twisted pair or optics) and for 10 Gigabit Ethernet (which uses XFP interchangeable transceivers for different wavelengths). POS is currently not supported but is under development. All cards process packets by themselves (they do not require a coprocessor). All cards support header filtering, sampling and packet-length statistics. 10 Gbps cards additionally support payload searching and packet-time statistics.

3.1.3.3 Other cards

Napatech (www.napatech.com) also produces 1 Gbps and 10 Gbps monitoring cards. The company did not respond to questions about these cards' features or capabilities, nor their price. From unofficial sources it seems that Napatech cards are comparable with DAG cards in their technical features and price, but harder to obtain and so are much less deployed.

Force10 is also starting to produce 1 Gbps and 10 Gbps monitoring cards of potential interest to GN2, but they are not yet commercially available. However they may still be considered if GN2 purchases more passive monitoring hardware.

3.1.3.4 Libpcap library with nCap

While it is not possible to capture and process all packets at 1 Gbps speed with regular NIC and standard operating system drivers and networking layer, there is a tool called nCap (developed by Luca Deri, of the University of Pisa), which significantly enhances packet capture with a regular NIC. With nCap, it is possible to capture full 1 Gbps of packets with zero packet loss at most packet sizes (there is small packet loss with very short packets) with still a lot of CPU time remaining for packet processing.

3.1.3.5 Hardware accelerated Netflow probe

The Netflow probe is based on COMBO6 programmable hardware card. Up to 512,000 simultaneous active flow records can be kept in the flow cache. Selected packet header fields are used to calculate 64-bit hash key, which identifies the flow.

Currently, a Gigabit Ethernet probe exists and 10 GE and PoS/STM-16 versions are in development.

3.1.4 Passive Tool Selection

The SCAMPI architecture has been chosen by the LOBSTER project [LOBSTER], which will extend it with user data anonymisation, support for distributed monitoring from multiple monitoring sensors by DIMAPI (Distributed MAPI), then and deploy it across Europe (primarily for security monitoring purposes).

A comparison of CoralReef and SCAMPI is summarized in Table 3.1:

Feature	CoralReef	SCAMPI
Supported monitoring adapters	FORE ATM cards, DAG cards, NIC	DAG cards, COMBO cards, NIC
API for applications	Low-level packet-based: CoralReef C API, CoralReef Perl API	High-level flow-based: MAPI or libpcap over MAPI
Transparency of user monitoring adapter	No, does not use any monitoring functionality of monitoring adapters	Yes, automatically uses header filtering, sampling, payload searching and statistics if COMBO card is used. Automatic support of header filtering in DAG cards is under development. Software implementation is transparently taken if NIC is used (and currently also DAG card).

Feature	CoralReef	SCAMPI
Applications available	Packet and flow statistics (indicating data volumes for networks, protocols, port numbers, Ass and countries)	Packet length and time statistics for up to 256 classes of packets classified based header filtering, NetFLOW generation tool, intrusion and DoS detection and hardware packet header anonymisation. In development are applications for quantification of bandwidth used by different user-defined flows, protocols and applications in short timescale (peak detection), packet loss monitoring, extensible IPFIX record generation, TCP analysis
User interface	Command line tools and tools with their own graphical GUI.	Command line tools, IDS application uses SNORT interface, flow generation tool and all new applications will use [STAGER].
Further development and support	Version 3.7 released in April 2005, but development seems to have stopped after 2003.	SCAMPI project completed in March 2005. Software is being further developed in a follow-up LOBSTER project, which is scheduled to run until December 2006. Current versions of COMBO cards can be ordered from CAMEA company, new versions will be developed by a start-up company which is yet to be created.
More information	[CORAL]	[SCAMPI], [LOBSTER]

Table 3.1: Comparison of CoralReef and SCAMPI

Alternative hardware monitoring adapters are compared in Table 3.2:

Project:	GN2
Deliverable Number:	DS3.7.3
Date of Issue:	05/05/06
EC Contract No.:	511082
Document Code:	GN2-06-016v5

Feature	DAG cards	COMBO cards with SCAMPI firmware	NIC with nCap
Supported speeds	ATM, E, FE, GE, OC-48, 10GE/OC-192	GE, (10GE, OC-48 and OC-192 are in development)	Works with Intel GE cards, support for Intel 10GE cards is in development
Supported port numbers	2 port on GE cards, 1 port on OC-48 and 10GE/OC192	4 ports on GE cards, but only 1 port works for monitoring (two cards are therefore needed for bidirectional monitoring), 1 port on 10GE card	Can support more cheap Intel GE cards at the same time
1 Gbps full packet capture	Yes, tested zero packet loss at all packet sizes	There is packet loss at high packet rates, see [D3.4v4] on [SCAMPI], will have to be improved in a commercial version	There is packet loss at high packet rates and small packet sizes, we prepare a comparison technical report, generally packet capture performance is worse than DAG card and better than COMBO card (as of June 2005)
10 Gbps full packet capture	Not tested by us, as it uses PCI-X 64-bit/133 MHz, there is a limit of about 8 Gbps	Not tested by us, as it uses PCI-X 64-bit/64 MHz, there is a limit of about 4 Gbps.	
Header filtering	In hardware (readily available in coprocessor for 1Gbps DAG cards, simpler filtering should be available soon for existing 10 Gbps DAG cards)	In hardware	In software, can keep up to 1 Gbps
Sampling	No	In hardware	No
Payload searching	Described in data sheet, but it seems that the required firmware version is not yet available	In hardware	No

Feature	DAG cards	COMBO cards with SCAMPI firmware	NIC with nCap
Statistics	Interface statistics	Per-traffic-class length statistics (and also time statistics on 10GE card)	No
More information	[ENDACE]	[LIBROUTER]	[LUCA NTOP]

Table 3.2: Comparison of packet capture hardware

The SCAMPI application was chosen as the preferred Passive Monitoring system because compared to CoralReef it provides a more powerful API, transparent support of different hardware monitoring adapters, and is more actively developed.

With regards to hardware, if and when COMBO cards have developed to the extent they remove packet loss and are more readily available, they may be considered again as the preferred hardware, because they seem to provide more monitoring functionality. Until that time, DAG cards are recommended for use with SCAMPI.

Due to the relatively prohibitive cost of the Passive Monitoring hardware, especially for 10Gbps, SA3 has deferred the procurement of any equipment until JRA1 have completed their in-depth testing of SCAMPI/DAG system, and made plans for the inclusion of passive monitoring in their perfSONAR framework.

3.2 Available Bandwidth

3.2.1 Overview

Almost all Available Bandwidth test tools in common use today are based on iperf (developed by the US's National Laboratory for Network Research (NLANR)) [IPERF]. The iperf tool is backwards compatible with the traditional "tcp" tool, which was used on systems for many years. Other AB tools considered by SA3 researchers were:

- Bing (<http://fgouget.free.fr/bing/index-en.shtml>): This is a bandwidth test program that calculates available bandwidth by comparing the RTT of different sizes of ICMP echo request packets. It was not a serious contender for the SA3/JRA1 AB tool because it does not give results for TCP or even UDP, which are the main two protocols in use on IP networks.
- Tptest (<http://tptest.sourceforge.net/>): Tptest is an interesting program developed for broadband customers in Sweden who want to test their network connectivity to different places. While this is a

simple-to-use product for the end customer, it was not considered flexible enough for GÉANT2's more sophisticated requirements.

- `ttcp`, `nttcp` (<http://sd.wareonearth.com/~phil/net/ttcp/>): `ttcp` was developed to help DARPA evaluate different TCP stacks running in machines in Berkeley and BBN (the operators of ARPANET). `Nttcp`, developed at Silicon Graphics Inc (SGI) is an improved version of `ttcp` with more tuneable options, including the important ability to set transmit and receive window sizes. `Ttcp` and `nttcp` have effectively been superseded by `iperf`.
- `iperf` (<http://dast.nlanr.net/Projects/Iperf/>): `iperf` has evolved from `ttcp` and `nttcp`. It is the most widely used and reliable AB measurement tool, and has an active user support forum. Of particular interest to GÉANT2 participants is the fact that it allows different Quality of Service (QoS) settings to be used. This is particularly important for emulating Premium IP traffic (as opposed to normal Best Effort traffic).
- Bandwidth Test Controller (BWCTL) [BWCTL] is a wrapper for `iperf` which removes the need for the tester to login to the traffic source device. BWCTL has other benefits, such as the use of user groups to limit the size and length of tests given users are allowed to run.

In addition to the established tools described above, Sunet have written a web-based front-end for some common test tools to make it possible for end-users to test their machines against central test systems in a simple way.

3.2.2 Tools Selection

BWCTL (and its underlying `iperf` application) was chosen to be the AB tool for use by SA3 and JRA1. BWCTL is well suited for periodic tests of networks, and also gives the opportunity to run tests that require more privileges than average users have. BWCTL is also the tool of choice for other major research networks, such as Abilene, and therefore it is the preferred tool for periodic and specialized tests.

The SUNET web-based throughput test tools are very simple for non-specialist users to use, and gives such users the ability to run tests against their own equipment without the need for additional assistance. As and when the GÉANT2 MPs are opened to wider, more public access then these tools are expected to be very useful. In the meantime they will be installed for further evaluation.

3.3 IPPM

IP Performance Monitoring (IPPM) refers to the range of metrics that can be derived from the transmission and receipt of time-stamped test packets. These metrics are OWD, IPDV and packet loss.

Project:	GN2
Deliverable Number:	DS3.7.3
Date of Issue:	05/05/06
EC Contract No.:	511082
Document Code:	GN2-06-016v5

3.3.1 Tool Selection

In the run up to the GN2 project, during its predecessor GÉANT (GN1), there were 3 candidate systems for IPPM:

- RIPE-TTM [RIPE-TTM]
- OWAMP [OWAMP]
- DFN-IPPM⁴.

An important requirement for GN2's IPPM system is that it could be modified, both to add functionality and to integrate it with the monitoring framework JRA1 were due to develop. For this reason, RIPE-TTM was quickly ruled out of contention, since its license prevents the distribution of modified versions of the code without the consent of RIPE, and it was deemed an unnecessary risk to create a dependency on RIPE in this way.

Of the remaining two systems, Internet2's OWAMP was the more mature and widely-deployed system, but DFN-IPPM was nevertheless in service on DFN's G-WiN network. To further determine whether DFN-IPPM was capable of providing the network performance measurement data needed for GÉANT2, DFN and DANTE installed DFN-IPPM systems in 5 sites spread across Europe during GÉANT (GN1) Year 4, namely 2004.. These devices were located in Rome, Italy (GARR), Poznan, Poland (PSNC), Tel Aviv, Israel (IUCC), Paris, France (RENATER) and Frankfurt, Germany (DFN) and had the following hardware specifications:

Vendor: Dell

Make: Poweredge 1750

Height Units: 1U

CPU: P4

Bus: PCI 64 bit

OS: Linux Red Hat Fedora Core release 1 (Yarrow), Kernel: Linux, release 2.4.21-NANO

As soon as they were introduced the Dell systems began showing frequent, significant outliers⁵ in the range of hundreds of milliseconds for OWD (compared to the average dimension of the measured OWD which was in the range of tens of milliseconds). The outliers, which made up approximately 0.05% of all data points, could not be sensibly explained by network performance problems, and it was eventually determined that they must in fact be an artefact of the IPPM systems, probably introduced by the Network Interface Cards (NICs). In addition to this, the servers showed some heavy clock instability due to thermal issues which were related to the way the machines controlled their own cooling. The effects of these two problems are seen in Figure 3.2,

⁴ DFN-IPPM has since been re-named HADES – HADES Active Delay Measurement System

⁵ An outlier is a measured value which significantly deviates from the mean.

which is a typical distribution of measurement points (in this case for the path between Athens and Paris). The figure shows that whilst the majority of measurements taken were at and around 40ms, over the course of one day there were about 100 points (or about 0.5%) which were significantly more than this average value.

Such results made it clear that if HADES was to be deployed in GÉANT2 then different hardware would be required.

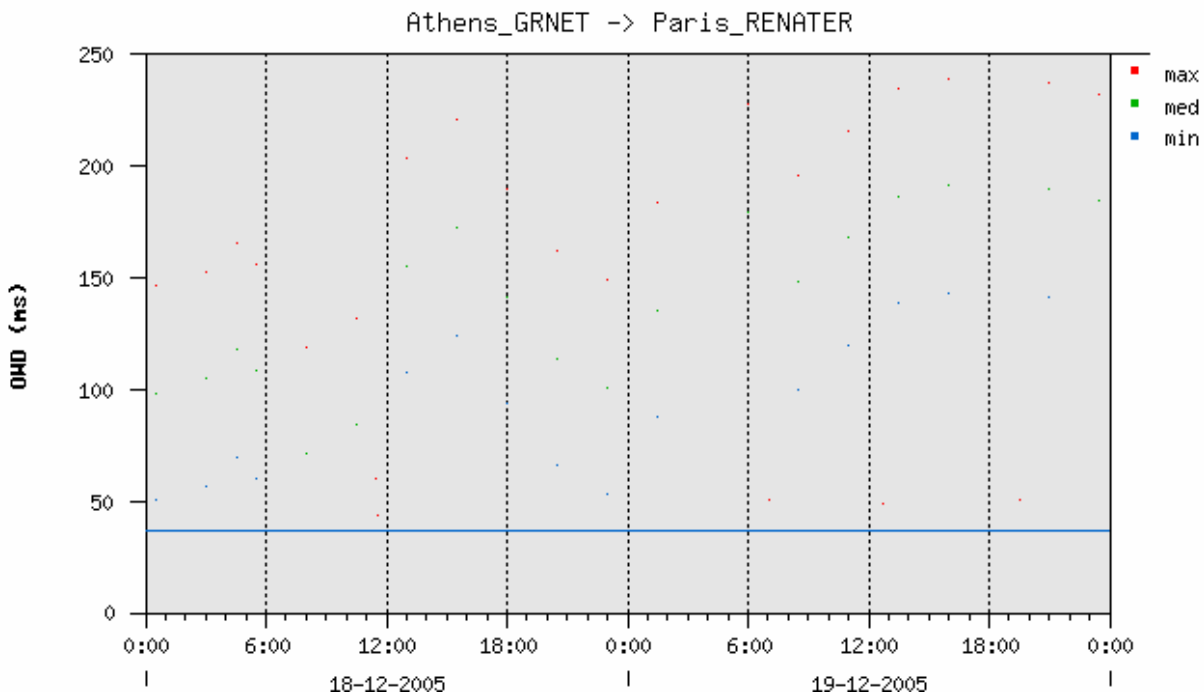


Figure 3.1: GÉANT Dell MP OWD Measurement pattern

As a result of this work, it was clear that both DFN-IPPM and OWAMP would be suitable for GN2's needs (OWAMP was not subjected to the same level of scrutiny as DFN-IPPM since it is a well-established and commonly deployed system). The systems were similar in many respects (for example, both supported both IPv4 and IPv6, both supported QoS) and the most notable difference between the two was that OWAMP sacrificed precision for flexibility (it can run satisfactorily on a wide range on platforms, but the 95th percentile is ignored), whilst DFN-IPPM was potentially much more accurate, but required careful configuration. From an operational, practical point of view the OWAMP approach was slightly more attractive but not so much to outweigh the fact that the GÉANT2 IPPM development team had a large amount of experience and knowledge of DFN-IPPM and would find it quicker to modify and enhance DFN-IPPM.

Project:	GN2
Deliverable Number:	DS3.7.3
Date of Issue:	05/05/06
EC Contract No.:	511082
Document Code:	GN2-06-016v5

3.4 Hardware Evaluation

As a result of the experience of using DFN-IPPM in GN1 Y4, it was recognised that in order to get the most accurate and consistent results from DFN-IPPM then its hardware platform would need to be chosen carefully so as to avoid problem such as the thermal instability seen in the pilot phase. During the hardware evaluation process, four different models of PC were tested.

- Fujitsu-Siemens (FSC) Computers GmbH, D-90451 Nürnberg (FSC)
- Hewlett Packard (HP), Bechtle IT Systemhaus Nürnberg, D-90579 LangenzennDell (provided by Dante)
- bee Baastrup EDV-Entwicklung GmbH, D-44135 Dortmund / Germany

The main goal of the tests was to verify that the hardware will work properly together with the external time sources (GPS or PZF clocks, PPS signals etc) and that the NIC will work properly for the requirements active measurement has.

The specific attributes checked were:

- Remote access to BIOS
- Remote hardware reset/power
- Remote access to OS in text mode
- Remote access to OS in graphics mode
- Remote installation of OS
- Installation of Meinberg GPS card
- NTP operation with Meinberg GPS card
- NTP PPS with built-in serial port
- NTP PPS with additional serial port
- OS compatibility
- Network interface performance with IPPM

Other factors taken in to account included:

Project:	GN2
Deliverable Number:	DS3.7.3
Date of Issue:	05/05/06
EC Contract No.:	511082
Document Code:	GN2-06-016v5

- Price
- Size (Form Factor)
- Maintenance cost and format
- Support performance
- Previous experience with the make/vendor

At the beginning of the project, it had been thought that AB and IPPM measurement would be run on separate devices. During the hardware evaluation process it was decided to combine the IPPM and AB measurement systems on the same platform, and therefore candidate PCs had to meet both sets of requirements.

Test:	Result (FSC):	Comment:	Result (bee):	Comment:	Result (HP):	Comment:	Result (Dell):	Comment:
Remote access to BIOS:	Good		Good		Not tested		Good	Tested via serial port only since we didn't have the Dell RSC card
Remote hardware reset/power:	Good		Good		Not tested		Not tested	
Remote access to OS in text mode:	Good		Good		Not tested		Good	
Remote access to OS in graphics mode:	Good		Good		Not tested		Not tested	
Remote installation of OS:	Good		Good		Not tested		Not tested	
Installation of Meinberg GPS card:	Good		Good		Not tested	No card available during test period	Good	
NTP operation with Meinberg GPS card:	Good		Good		Not tested		Satisfactory	
NTP PPS with built-in serial port:	Bad		Good		Bad	As this test failed, the other tests were not needed anymore	Satisfactory	
NTP PPS with additional serial port:	Bad	PC didn't boot with the PCI card available at the time There was not enough time to	Good		Not tested		Satisfactory	

Project:	GN2
Deliverable Number:	DS3.7.3
Date of Issue:	05/05/06
EC Contract No.:	511082
Document Code:	GN2-06-016v5

		test a different card						
OS compatibility	Satisfactory		Satisfactory		Satisfactory		Satisfactory	
Network interface performance with IPPM:	Good		Good		Not tested		Good	
Result Summary	+		++		--		-	

Table 3.3: Test results of candidate IPPM/AB measurement points

As can be seen from Table 3.2, the device that performed best overall was the bee device, and so bee GmbH were invited to supply the hardware. The following machine was chosen to be the combined IPPM/AB measurement device:

bee Linux-Server in a 19"-server-chassis (3HE) with a 460watt power supply

It was configured as follows:

- 19", 3RU chassis
- one Intel® Pentium4 3,0 GHz CPU, FSB800 »Boxed«
- 64bit/66 MHz PCI-X and 32bit/33 MHz PCI slots
- two 10/100/1000Mbps onboard network cards
- 512MB DDR-RAM main memory, PC400, ECC (2 x 256MB)
- one SATA 80GB hard disk drive, mounted in the SATA hot-swap-frames (8) of the chassis
- one additional SATA 80GB hard disk
- one ATAPI DVD-drive
- one 3,5" floppy drive
- redundant 2x 460watt power supply
- [Optional] eRIC express (short eRIC X), remote management solution

Project:	GN2
Deliverable Number:	DS3.7.3
Date of Issue:	05/05/06
EC Contract No.:	511082
Document Code:	GN2-06-016v5

3.5 Software Validation and Integration

The G-WiN Lab Erlangen contains a testbed for hardware and software.

The measurement setup consists of two IPPM/AB measurement boxes connected back-to-back with a cross-over cable between their two network interface cards (NIC). Time-synchronization is done via Network Time Protocol (NTP) with two different hardware clock systems based on high precision time signals from the Global Positioning System (GPS) and the long wave time signal DCF77 provided by the Physikalisch Technische Bundesanstalt Braunschweig (PTB), Germany's national institute for metrology [PTB].

Performance measurements without additional background traffic between these two boxes were carried out to collect statistical data. These data were the base for understanding the hardware specific behaviour of the measurements. The results can be influenced by the NIC, the CPU, the operating system, but most of all by the clock synchronisation. In addition, environmental factors such as temperature and temperature changes must be taken into consideration. To discover possible influences of measurements running concurrently, available bandwidth tests were performed concurrently with the OWD measurements. Finally, the connection was loaded with generated network traffic.

Figure 3.2 shows the result of OWD measurements with the test bed. It can be seen, that over the plotted period of time (2 days), no significant drift of the measured values occurred.

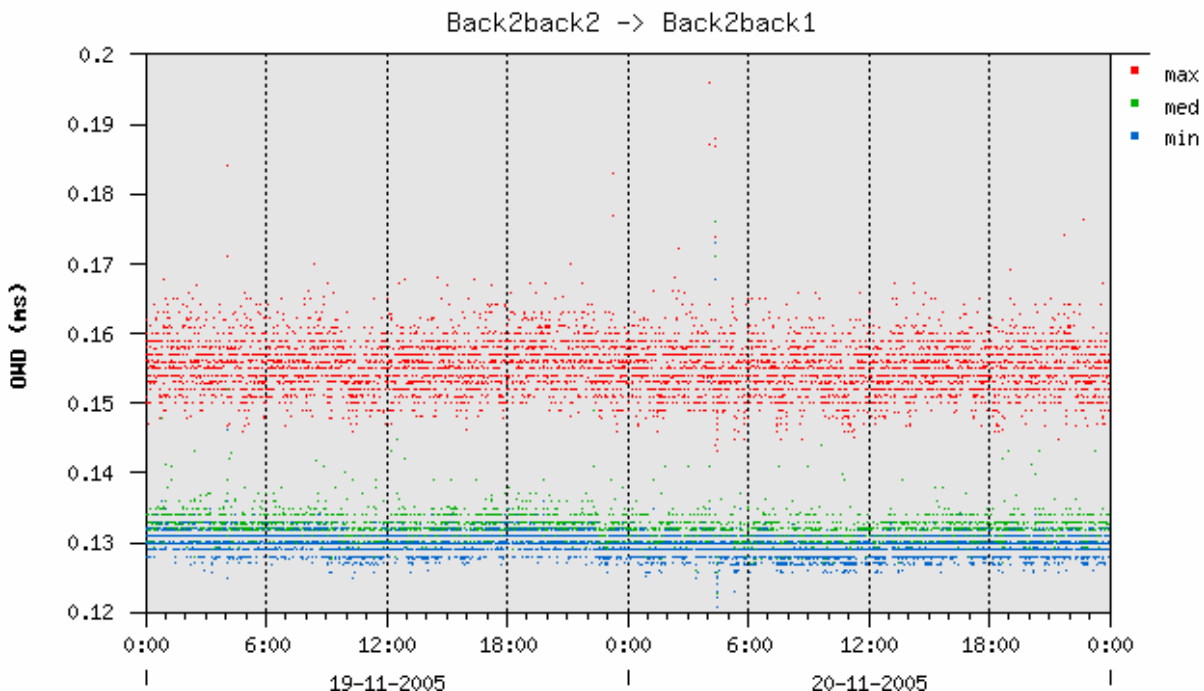


Figure 3.2: Back to back measurements using GPS based time synchronization and two identical measurement boxes to reveal accuracy and precision of OWD measurements

Project:	GN2
Deliverable Number:	DS3.7.3
Date of Issue:	05/05/06
EC Contract No.:	511082
Document Code:	GN2-06-016v5

The absolute value of the measurement data is in the range of 200 microseconds max. Nevertheless, the system can differentiate between Minimum, Maximum and Median values and the values are stable within 30 microseconds.

In Figure 3.3, OWD values depending on the packet number (always five packets are sent for one measurement to determine Maximum, Minimum and Median) are plotted. The first packet of one group always shows markedly higher OWD. This could be caused by a wake-up process triggered by the first packet. The exact reasons were still under investigation at the time of writing.

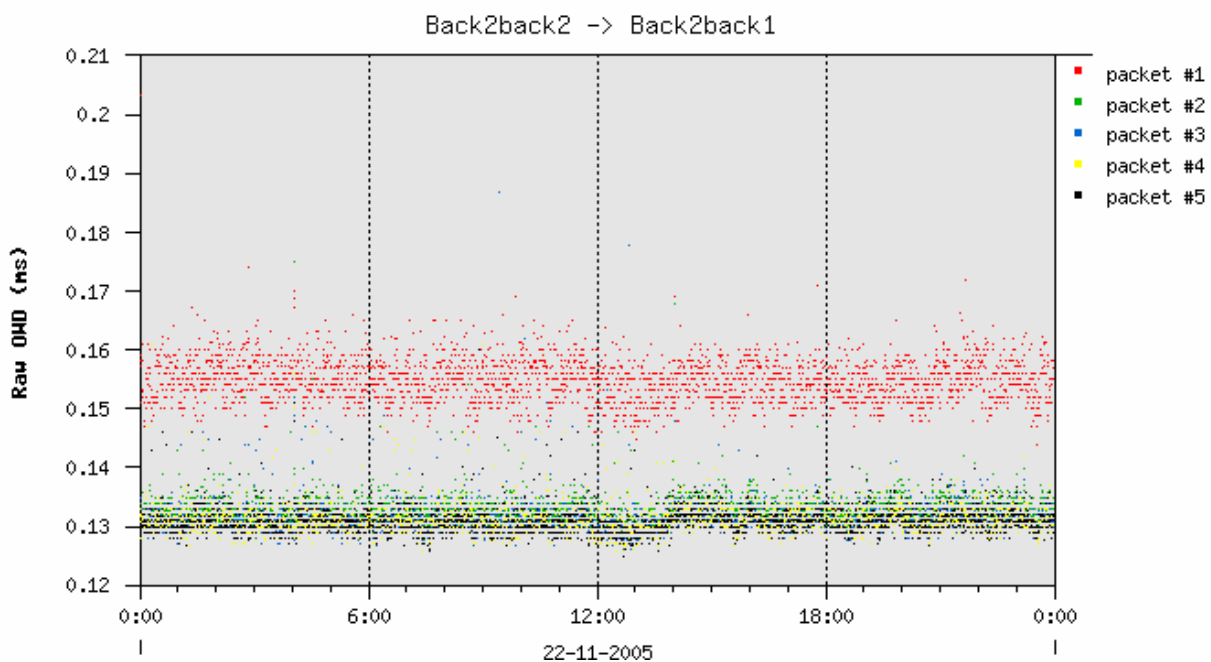


Figure 3.3: Measurement of OWD. The first packet of a packet group always shows an increased delay due to wakeup

3.5.1 Optimizing measurements

Setting hardware parameters on the NIC can improve the accuracy of OWD measurements. Measurements in the German NREN G-WiN showed that the interrupt handling on the NIC should be set to generate immediately an interrupt to accelerate the data processing.

Conversely, for iperf achievable bandwidth measurements, setting immediate interrupt generation results in degraded performance. Figure 3.4 demonstrates this effect, showing iperf results for specific window sizes and protocol versions (IPv4, IPv6). During the course of the test the parameters RxIntDelay, TxIntDelay and

Project:	GN2
Deliverable Number:	DS3.7.3
Date of Issue:	05/05/06
EC Contract No.:	511082
Document Code:	GN2-06-016v5

InterruptThrottleRate⁶ were varied (these parameters were different for different hardware, so they are not stated explicitly). Most noticeably at x = 15:00, the Rx_intdelay is increased to a higher value, with positive results. For IPv4 measurement, the effect can be seen by an increase of the achievable bandwidth from approximate 870 Mbps to 950 Mbps. Even more obvious is the influence on IPv6 measurements. As can be seen, the increase of the data is from near 0 Mbps to values in the range of 1 Gbps. (The value of 0Mbps was very unexpected and it is not understood why the throughput should be quite so poor, even allowing for a sub-optimal configuration. This may be the subject of a future study.)

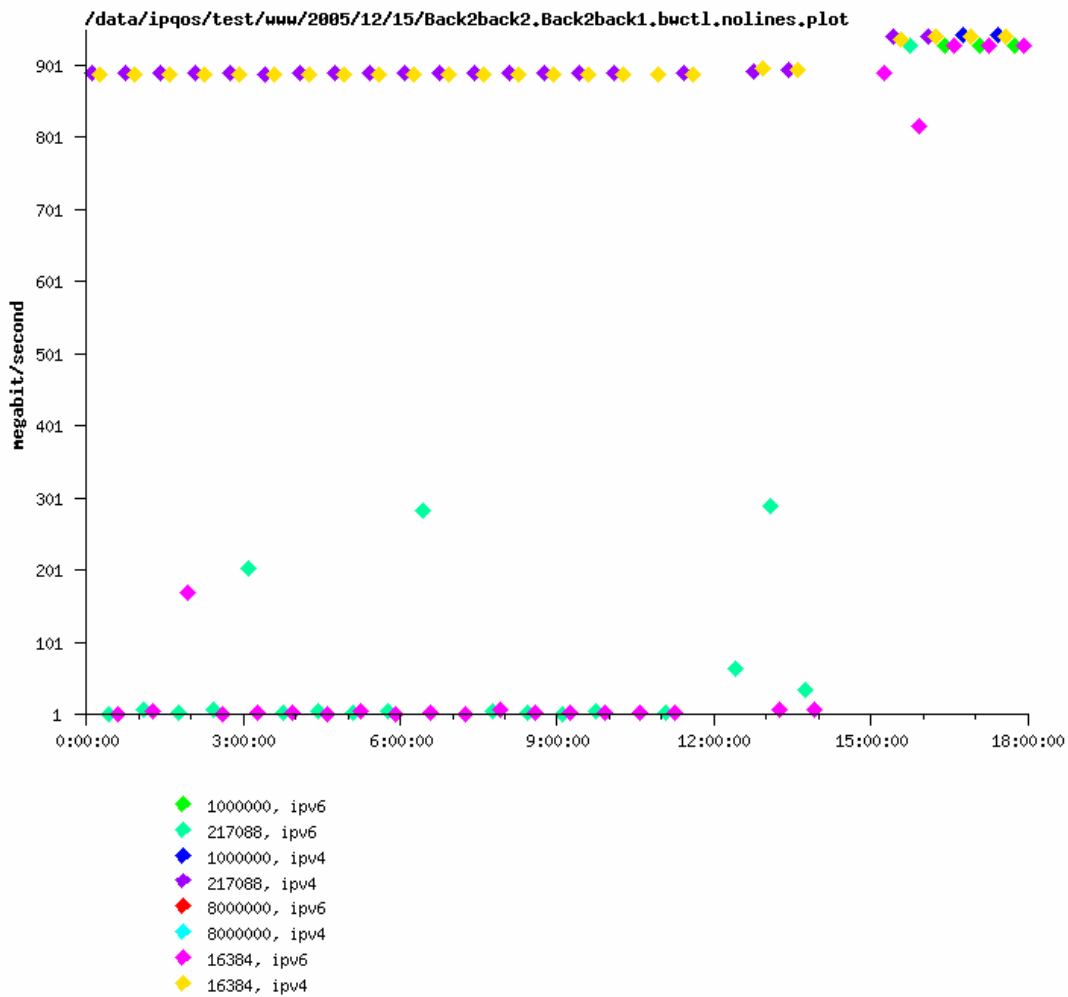


Figure 3.4: Influence of the rate of CPU interrupts on achievable data rate

⁶ RxIntDelay and TxIntDelay are, respectively, Receive and Transmit Interrupt Delay, and are the delay between a packet being received (transmitted) by the NIC and the NIC controller raising a CPU interrupt. InterruptThrottleRate limits the number of times per second the NIC controller can generate interrupts, and therefore limits the amount of time the CPU spends handling NIC interrupts.

This leads to the conclusion, that AB and IPPM measurements should not be performed on the same interface. Properly adjusted, undisturbed achievable bandwidth measurements using BWCTL will show results as shown in Figure 3.5. The values are in the expected range for a 1 Gbps interface. Slightly lower values can be achieved using IPv6.

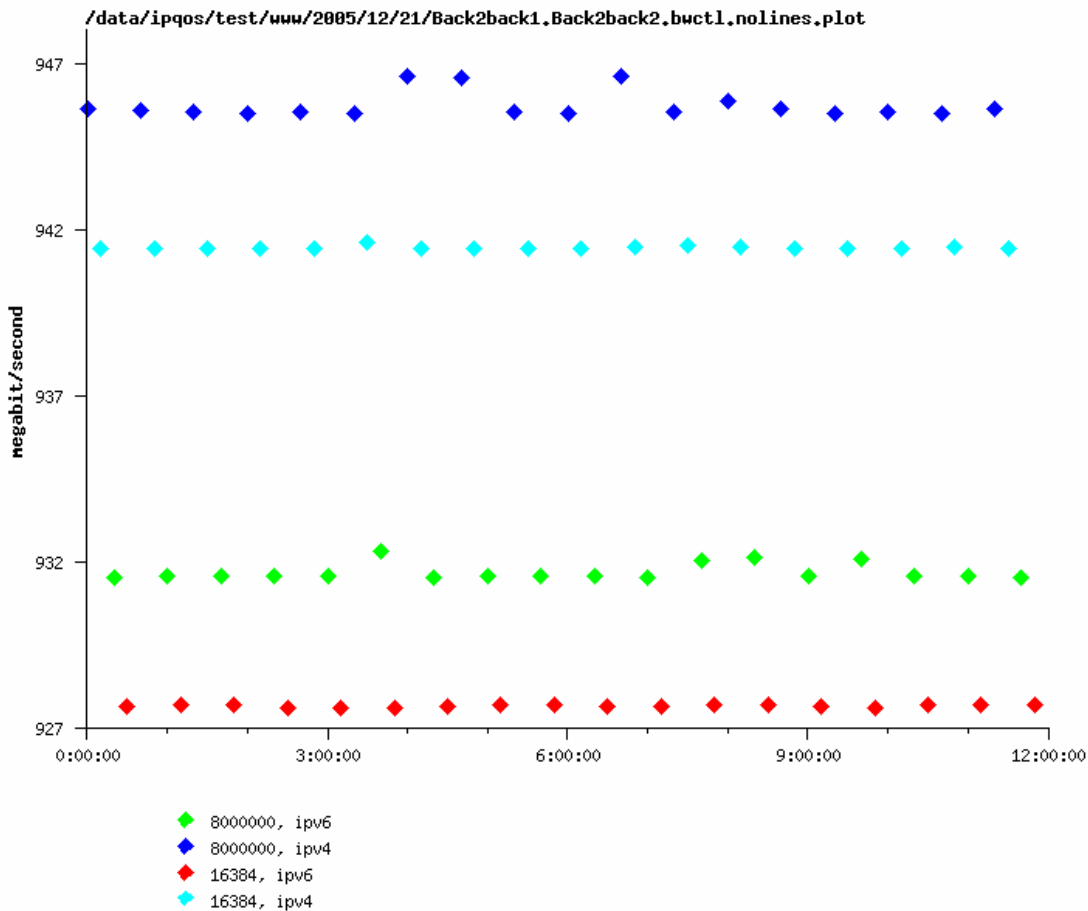


Figure 3.5: Correct achievable bandwidth measurements

One other issue is the selection of the interface used for the measurements. The two interfaces of the selected main board are connected to different bus systems on the main board. Measurements showed, that the (slower) PCI bus lead to lower bandwidth values, whereas the faster Communication Streaming Architecture (CSA) bus has a higher performance (as explained at [INTEL CSA]). So for AB measurements, the CSA bus should be used. The CSA attached interface is eth0 for the GN2 MPs.

Project:	GN2
Deliverable Number:	DS3.7.3
Date of Issue:	05/05/06
EC Contract No.:	511082
Document Code:	GN2-06-016v5

3.6 Configuration

In the GÉANT2 PoPs router filters (access control lists) prevent unauthorised access to the MP. In non-GÉANT sites iptables is used to control access. The basic restrictions of the boxes are everything is denied except:

- Connecting via SSH from specified hosts.
- IPPM measurements running on ports > 50000 for all networks allowing only UDP.
- Open IP addresses for NTP, when raw time synchronization is needed.
- HTTPS for remote management (for specified hosts).

User accounts are generated as and when required. Currently there is a PERT Case Manager's account on all GÉANT2 devices.

3.7 Roll-out Plan

SA3 had initially planned to deploy 10 IPPM MPs and 4 AB MPs. By combining IPPM and AB SA3 could procure up to 14 MPs, which represented a significant improvement over the original plans, particularly because AB measurements are currently the most useful tool for investigating network performance issues. Because the support offered by bee GmbH is "Return To Base" one of the 14 devices was nominated to be a spare, which can be deployed if an operational device fails and needs to be sent to bee for repair.

Since SA3 is concerned with End-to-end Quality of Service, and the use of Premium IP to support this, the sites chosen for deploying the SA3 procured devices were based primarily on their involvement with Premium IP. So for example, each of the NRENs which are taking part in the multi-domain Premium IP trial (GARR, GRNET and UKERNA) were allocated their own MP, to deploy on their own networks, and an MP was planned for the GÉANT2 PoP in each of the trial Premium IP countries. The deployment plan for SA3 MPs was

- Thessalonica GRNET
- London UKERNA
- Bologna GARR
- Athens GÉANT2
- London GÉANT2
- Milan GÉANT2

- Amsterdam GÉANT2
- Geneva GÉANT2
- Budapest GÉANT2
- Poznan GÉANT2
- Frankfurt GÉANT2
- Paris GÉANT2
- New York City GÉANT2

Boxes delivered to GÉANT2 PoP location will not be equipped with a remote management solution, as there is a local KVM-over-IP (KVMoIP) solution available.

All MPs are assembled and configured at DFN in Erlangen – this includes the installation of the GPS PCI hardware clock, and the DFN-IPPM and iperf/BWCTL software. The operating system currently used is Linux version 2.6.13.2, Fedora Core release 3 (Heidelberg).

For those sites which will use GPS synchronisation (the normal, preferred method), a Meinberg GPSANT antenna needs to be installed on the outside of the building. This uses RG58 coax cable to connect to the MP and the installation of both antenna and cable can lead to long delays. Typically, the cost of this work (not including the cost of the antenna, which is purchased directly from Meinberg and then sent to site) is in the order of a few thousand Euros. In commercial locations, there is normally also an ongoing monthly charge for the use of the roof space where the antenna is located. This is typically around 100 Euros per month.

3.8 Roll-out Status

The boxed MPs and GPS antennas are collected from Erlangen and shipped to their destinations. To date, of the 13 planned MPs, 7 are operational, 4 are awaiting installation and 2 are delayed due to on-site issues.

The operational MPs are:

- Milan GÉANT2
- Thessalonica GRNET
- New York City GÉANT2
- Budapest GÉANT2

Project:	GN2
Deliverable Number:	DS3.7.3
Date of Issue:	05/05/06
EC Contract No.:	511082
Document Code:	GN2-06-016v5

- Frankfurt GÉANT2
- Amsterdam GÉANT2
- Paris GÉANT2

The following MPs have been delivered and are awaiting installation:

- Geneva GÉANT2
- Bologna GARR
- Poznan GÉANT2
- London UKERNA

The GÉANT2 Athens MP has not been shipped yet at the time of writing because the POP is being re-located approximately in mid July 2006, and the cost of installing and then moving the MP is not considered justified. In addition, the current PoP has very limited space available. In the GÉANT2 London PoP there were some initial problems with the installation of a GPS antenna. It was eventually agreed that Telecity would supply an L1 GPS signal (a raw GPS signal) on a co-axial cable, which can be converted to the format required by the Meinberg PCI GPS receiver. The London MP should thus be operational approximately in mid-June 2006.

4 Recommendations for Enhancements and Extensions

Based on experiences to date, DFN and other participants of SA3 Work Item 7 have made the following recommendations. Those recommendations which are achievable in the short term and at no extra cost to the project have been accepted by the Activity Leader – longer term plans will be included in the draft GN2 Year 3 Technical Annex.

4.1 Available Bandwidth

4.1.1 iperf

The current publicly available version of iperf (version 2.0.2) does not include support for the Web100 kernel enhancements which the GÉANT2 MPs have. Staff at Internet2 have before now written patches for older versions of iperf to take advantage of Web100 features, and they have made this code available so that GN2 developers can update it for iperf v2.0.2. Although the amount of effort required is expected to be relatively small (in the region of 2 to 4 weeks), it is not thought there will be resource available within the next 4 months and so this work will be proposed for GN2 Y3.

4.1.2 BWCTL

BWCTL (and the underlying iperf) is currently the most useful tool for the investigation of performance issues, and as such its proper deployment and operation is worth close attention. In addition to “on demand tests” (normally run by PERT engineers when they are investigating a problem), some form of regular, scheduled testing should be implemented in order to base-line network performance, and show a history of achievable throughput for a given path. Currently, BWCTL tests are run on GÉANT2 overnight, during the quiet period

Project:	GN2
Deliverable Number:	DS3.7.3
Date of Issue:	05/05/06
EC Contract No.:	511082
Document Code:	GN2-06-016v5

between midnight and 6.00AM, but to get a more accurate idea of achievable bandwidths some scheduled tests should be run during the core working day. Such tests should be carefully planned, in order to minimise the risk that they might adversely impact production traffic. For GÉANT2 this should be possible, as long as the following points are observed:

- The iperf flows are limited to a maximum of 1Gbps (which is anyway the current maximum possible, because the MPs have 1GE interfaces, not 10GE interfaces)
- The test paths are chosen such that, under normal routing conditions, they only traverse 10Gbps links.
- The tests are co-ordinated such that no tests share the same link concurrently.

The impact of scheduled iperf tests could be further mitigated by sending all AB test traffic as Less than Best Effort.

The early results of the nightly BWCTL tests show a surprisingly wide variation in results. Some of these differences are undoubtedly caused by un-tuned TCP/IP configurations, and these will soon be corrected by DFN. However, there appear to be some under-performing paths across GÉANT2 whose end-points are properly tuned MPs. More investigation is required here but one cause of reduced throughput between MPs within GÉANT2 could be the topology of the LAN through which the MP connects to the router. Although the standard LAN switch (a Hewlett Packard ProCurve 3400 series) was specifically tested for 1Gbps transfers before procurement (and is even capable of 10Gbps with the appropriate expansion module fitted) there is only a single, Gigabit Ethernet connection between the LAN switch and the router, which means that effectively the MP has to share its 1Gbps link with other devices. Although most GÉANT2 PoPs have very little traffic on the LAN, even a small amount of traffic contesting the Gigabit Ethernet uplink could lead to packet loss and thus significantly degrade long-distance TCP data transfers. In order to determine whether this is indeed a contributing cause to below-expected performance it is recommended that anyone deploying MPs connect such devices to their routers using dedicated connections.

4.2 HADES

Whilst HADES has been proved to measure OWD accurately to within fractions of a millisecond, it has not yet, within GÉANT2, established itself as a major tool for performance monitoring. In order to do so several shortcomings must be addressed, and a number of enhancements made.

4.2.1 Time To Live

HADES uses Linux's TCP/IP stack for the sending and receiving of the IPPM active probes. Whilst this is convenient and simple, one effect is that it is not possible to see the probe packet's Time To Live (TTL) value. Since the TTL value is directly related to the number of hops the packet has taken along the path, if one packet

(or more likely, one set of five packets) has a different TTL from the those at tests run 30 seconds before and after the packet(s) in question, this would be confirmation of a short-lived change in routing, which would explain any difference in OWD. Checking the TTL of every packet would be an improvement on HADES current method of detecting changes in route, which relies on running a traceroute run every 5 minutes. Of course, checking the TTL will identify if there has been a routing change but will not identify **what** the new route may be, so the TTL check should be used in addition to, rather than instead of, the traceroute.

4.2.2 Operator's Display

Currently, to access HADES information it is necessary to drill down through selecting a date, a measurement point, one or more measurement paths and one or more metrics (OWD, OWDV, packet loss). This effectively means that HADES is only of use if the operator is already aware of a problem on a given link and is investigating it. Whilst this is a useful feature in itself, it would be even better if there were some form of graphical display for an operator to get an overview of network health. Such a HADES graphical display is already being used by DFN for their G-Win network, which has coloured links showing OWD between adjacent network nodes. An equivalent map for GÉANT2 would need to allow for the fact that the network is not fully meshed – however, this should not be an insurmountable problem.

4.2.3 Information Presentation

The information presented at the time of writing is simply a graph of OWD measurements (minimum, median and maximum of a train of 5 probes) against time of day. DFN are working on enhancements so that more than one day's worth of results can be shown, and the size of data points can be changed to make outliers more obvious. Equally important will be the calculation and presentation of statistical analysis, so that an operator can immediately see, say, whether the 99th percentile of Premium IP traffic is the same or better than the 99th percentile of normal Best Effort traffic. This particular comparison could be used to advise an end customer whether or not they would benefit from using the Premium IP service.

4.2.4 Alerts

Once statistical analysis has been put in place, it will be possible to identify thresholds, the exceeding of which would indicate a potential problem. Once such a threshold has been exceeded, a mail alert should be sent out to the network operator, and/or a suitable change made to the HADES operator's main display.

4.2.5 System Monitoring

When a measurement path is not available, or is recording anomalous results, this needs to be clearly displayed on the administrator's web interface; currently this interface simply shows whether or not a node is reachable. For added robustness, the HADES network administrator should be sent an e-mail when an abnormal condition exists (loss of MP, loss of measurement path, loss of time sync unreliable data values etc).

5 Conclusion and Next Steps

After some delays and complications a network of performance monitoring Measurement Points (MPs) has now been established in the GÉANT2 network, and beyond. In order to measure OWD (one of the fundamental metrics of network performance) accurately the MPs need a very precise time reference. This is provided in a variety of ways, the most common of which is GPS. Installing GPS in some sites can be difficult and prohibitively expensive, so in some cases other methods were required. Out of 22 GN2 router-equipped PoPs 7 are already equipped with MPs and 4 more are planned. For complete visibility of network conditions in GÉANT2 it is recommended that the 11 remaining GN2 PoPs are also fitted with MPs.

Whilst the accuracy of the GPS synchronised HADES is beyond doubt, in order to make HADES an effective tool for network operators, improvements must be made to the analysis and display of collected data. To a lesser extent, this is also true of the BWCTL/iperf AB measurement application which is co-deployed with HADES on the MPs – scheduled tests are currently run overnight on the MPs, but they could also be run during the working day, if appropriate care is taken to mitigate the impact bulk tests might have on production traffic. Initial tests have shown a wide variation in results, some of which are certainly due to un-tuned TPC/IP configurations and others may be due to congestion in the LAN. To remove the latter effect would require installing an extra Gigabit Ethernet port on each GÉANT2 router.

Two passive monitoring stations have been deployed in JRA1, monitoring the connections of CESNET and ARNES NRENs to the GÉANT2 network. The performance of these two passive monitoring stations will be evaluated, in terms of the monitoring cards, software framework and developed applications. If good experience is acquired then it is suggested that GN2 deploys more passive monitoring stations to monitor connections of other NRENs to GÉANT2.

6 References

[RFC2330]	http://www.ietf.org/rfc/rfc2330.txt
[RFC2679]	http://www.ietf.org/rfc/rfc2679.txt
[RFC3393]	http://www.ietf.org/rfc/rfc3393.txt
[RFC2680]	http://www.ietf.org/rfc/rfc2680.txt
[LUCA NTOP]	http://luca.ntop.org
[DS3.4v4]	Sven Ubik, "DS3.4 Description of Experimental Results [SCAMPI], v4", April 2005
[LOBSTER]	http://www.ist-lobster.org
[SCAMPI]	http://www.ist-scampi.org
[STAGER]	http://stager.uninett.no
[CORAL]	http://www.caida.org/tools/measurement/coralreef
[ENDACE]	http://www.endace.com
[LIBROUTER]	http://www.liberouter.org
[RIPE-TTM]	http://www.ripe.net/ttm/about.html
[OWMAP]	http://e2epi.internet2.edu/owamp/
[INTEL CSA]	http://www.intel.com/design/network/events/idf/csa.htm
[IPERF]	http://dast.nlanr.net/Projects/lperf
[BWCTL]	http://e2epi.internet2.edu/bwctl
[PTB]	http://www.ptb.de/en/zieleaufgaben/dieptb.html

7 Acronyms

[AB]	Available Bandwidth
[API]	Application Programming Interface
[AS]	Autonomous System
[ATM]	Asynchronous Transfer Mode
[BWCTL]	Bandwidth Test Controller
[CDMA]	Code Division Multiple Access
[CPU]	Central Processing Unit
[CSA]	Communication Streaming Architecture
[DAG]	Data Acquisition and Generation
[DCF77]	Deutschland C-Band Frankfurt [source] 77.5 kHz
[DDR-SDRAM]	Double Data Rate Synchronous Dynamic Random Access Memory
[DNS]	Domain Name System
[DoS]	Denial of Service [attack]
[DVD]	Digital Versatile Disk
[FSC]	Fujitsu Siemens Corporation
[GB]	Gigabyte
[Gbps]	Gigabits per second
[GE]	Gigabit Ethernet
[GN2]	GEANT2 Project
[GPS]	Global Positioning System
[GSM]	Global System for Mobile communications
[HADES]	Hades active delay evaluation system
[HP]	Hewlett Packard
[IP]	Internet Protocol
[IPDV]	IP Packet Delay Variation
[IPPM]	IP Performance Metrics
[JRA]	Joint Research Activity
[KVMoIP]	'Keyboard, Video, Mouse' over IP
[Mbps]	Megabits per second
[NIC]	Network Interface Card
[NLNR]	National Laboratory for Network Research
[NREN]	National Research and Education Network

Project:	GN2
Deliverable Number:	DS3.7.3
Date of Issue:	05/05/06
EC Contract No.:	511082
Document Code:	GN2-06-016v5

[NTP]	Network Time Protocol
[OC]	Optical Carrier
[OS]	Operating System
[OWD]	One Way Delay
[OWDV]	One Way Delay Variation
[PCI]	Peripheral Component Interconnect
[PIP]	Premium IP
[PM]	Passive Monitoring
[PMS]	Performance Monitoring System
[PoP]	Point of Presence
[POS]	Packet over SONET
[PPS]	Pulse per second
[PTB]	Physikalisch Technische Bundesanstalt Braunschweig
[RTT]	Round Trip Time
[RU]	Rack Unit
[SA]	Service Activity
[SATA]	Serial Advanced Technology Attachment
[SCAMPI]	Scalable Monitoring Platform for the Internet
[SFP]	Small Form Factor Pluggable [interface]
[SSH]	Secure Shell
[TCP]	Transmission Control Protocol
[UDP]	User Datagram Protocol

Appendix A Network Performance Metrics

A.1 One Way Delay Metrics

The definition of the metrics (from JRA1, DJ1.2.3) determined by the DFN-IPPM tool are:

Packet Loss

The packet loss metric is measured on the network layer. Consequently, for each measurement one has to define explicitly the network layer parameters applied during the measurements, e.g. source/destination node IP addresses, packet size, packet Type of Service value, time-to-live value, etc. Also, as discussed in the definition, a packet is considered as lost only if it is not delivered to the destination node within a specific time period. Consequently, a timeout value has to be chosen prior performing measurements. If not defined, the timeout interval is considered an extremely large time value (e.g. 255 s). In active measurements, artificial traffic is inserted in the network by a sender and collected by a receiver at the end of the path. By measuring the packet loss of the artificial traffic, estimations of the packet loss of real traffic can be made.

One Way Delay

One-Way Delay is the time (in milliseconds) between the occurrence of the first bit of a packet on the first observation point, e.g. transmitting monitor interface, and the occurrence of the last bit of a packet on the second observation point (see RFC 2679). In active measurements, monitoring packets are time stamped as they “enter” the network at the source node. The receiver node compares the time-stamp time information carried in each monitoring packets with the time that each packet was received and calculates the delay.

One Way Delay Variation

Project:	GN2
Deliverable Number:	DS3.7.3
Date of Issue:	05/05/06
EC Contract No.:	511082
Document Code:	GN2-06-016v5

Given a stream of at least two packets crossing observation point A and observation point B, we define One Way Delay Variation OWDV (also known as IPDV, Inter-Packet Delay Variation) as the difference in the OWD of a selected pair of packets in the streaming. OWDV measurements may be performed by using a pre-defined train of packets between two observation nodes (active monitoring). The receiving node knows exactly the traffic profile characteristic of the monitoring traffic generated by the source node, i.e. the packet rate, the packet size distribution, the intervals between the packets, etc. The receiver node timestamps the incoming packets and using the information regarding the packet train produced by the source, it can accurately estimate the jitter exhibited among each pair of packets.

For measuring IPPM related data, the boxes are equipped with a GPS (Global Positioning System) based hardware clock (Meinberg GPS169PCI) for time synchronization. The time synchronization is the most sensitive task for measuring OWD and related metrics. All systems in a measurement network must have a stable time base to ensure accurate measurement. The main point here is to avoid time drifts of the local clock.