Lighting the Blue Touchpaper for UK e-Science - Closing Conference of ESLEA Project

March 26-28, 2007 - Edinburgh

Abstract

ESLEA, an EPSRC-funded project, aims to demonstrate the potential benefits of circuit-switched optical networks (lightpaths) to the UK e-Science community. This is being achieved by running a number of “proof of benefit” pilot applications over UKLight, the UK’s first national optical research network.

UKLight provides a new way for researchers to obtain dedicated "lightpaths" between remote sites and to deploy and test novel networking methods and technologies. It facilitates collaboration on global projects by providing a point of access to the fast growing international optical R&D infrastructure.

A diverse range of data-intensive fields of academic endeavor are participating in the ESLEA project; all these groups require the integration of high-bandwidth switched lightpath circuits into their experimental and analysis infrastructure for international transport of high-volume applications data. In addition, network protocol research and development of circuit reservation mechanisms has been carried out to help the pilot applications to exploit the UKLight infrastructure effectively.

Further information about ESLEA can be viewed at http://www.eslea.uklight.ac.uk. ESLEA activities are now coming to an end and work will finish from February to July 2007, depending upon the terms of funding of each pilot application. The first quarter of 2007 is considered the optimum time to hold a closing conference for the project.

The objectives of the conference are to:

(a) Provide a forum for the dissemination of research findings and learning experiences from the ESLEA project.

(b) Enable colleagues from the UK and international e-Science communities to present, discuss and learn about the latest developments in networking technology.

(c) Raise awareness about the deployment of the UKLight infrastructure and its relationship to SuperJANET 5.

(d) Identify potential uses of UKLight by existing or future research projects.

The deliverables of the conference will be electronic and printed conference proceedings.

Editorial Board

Sessions

Session 1: Protocols

Session 2: Protocols and Performance Testing

Session 3: RealityGrid

Session 4: Resource Scheduling and Hosting

Session 5: HEP

Session 6: e-VLBI

Session 7: Arts & Humanities

agenda
PoS(ESLEA)031 Vincent Gaffney

Session 8: Closing Session

Exploitation of switched lightpaths for e-Health: constraints and challenges
PoS(ESLEA)032 Lee Momtahan and Andrew Simpson

Posters

Monitoring the UKLight network
PoS(ESLEA)037 Barnaby Garrett

VLBI_UDP
PoS(ESLEA)038 Simon Casey

Authors

Contributions

PROCEEDINGS OF SCIENCE

Network Developments and Monitoring in Internet2

Eric Boyd[1]
Internet2
1000 Oakbrook Drive, Suite 300, Ann Arbor, Michigan, USA
E-mail: eboyd@internet2.edu

Susan Evett
Internet2
1000 Oakbrook Drive, Suite 300, Ann Arbor, Michigan, USA
E-mail: sevett@internet2.edu

Given that performance is excellent across backbone networks, and that performance is a problem end-to-end, it is clear that problems are concentrated towards the edge and in network transitions. To achieve good end-to-end performance, we need to diagnose (understand the limits of performance) and address (work with members and application communities to address those performance issues). We envision readily available performance information that is easy to find, ubiquitous, reliable, valuable, actionable (analysis suggests course of action), and automated (applications act on data received).

The Internet2 End-to-End Performance Initative (E2Epi) currently focuses on the development and widespread deployment of perfSONAR [1][2], an international consortium developing a performance middleware architecture and a set of protocol standards for inter-operability between measurement and monitoring systems. perfSONAR is a set of open source web services that can be added, piecemeal, and extended to create a performance monitoring framework. It is designed to be standards-based, modular, decentralized, and open source. This makes it applicable to multiple generations of network monitoring systems and encourages outside development while still allowing it to be customized for individual science applications. perfSONAR is a joint effort of ESnet, GÉANT2 JRA1, Internet2, and RNP.

The Internet2 Network is a hybrid optical and IP network, that offers dynamic and static wavelength services. The Internet2 Network Observatory supports three types of services: measurement, co-location, and experimental servers to support specific projects. The Observatory collects data and makes it publicly available.

Lighting the Blue Touchpaper for UK e-Science - Closing Conference of ESLEA Project
The George Hotel, Edinburgh, UK
26-28 March, 2007

[1] Speaker

PoS(ESLEA)001

1. Internet2 Network

The Internet2 Network is a hybrid optical and IP network, that offers dynamic and static wavelength services. The fiber and equipment on which the Internet2 Network runs is dedicated to Internet2's use; Level 3 provides the network and service level maintenance. The platform supports production services (IP, static waves, and dynamic waves) as well as experimental projects, such as HOPI (Hybrid Optical/Packet Infrastructure).

The deployment of the new Internet2 Network will be complete by July 15; 2007. In March, Internet2 began exploring a merger with National Lambda Rail (NLR) with a goal of consolidating national higher education and research (R&E) networking organizations. A technical team began considering what the merged technical infrastructure would look like. A decision on this potential collaboration is expected by June 2007.

2. Performance Middleware

Science is a global community; networks link scientists and collaborative research occurs across network boundaries. For the scientist, the value of the network is the *achieved* network performance. Scientists should not have to focus on the network; good end-to-end performance should be a "given." For example, the Large Hadron Collider is an international physics facility located at CERN in Switzerland. This effort involves major U.S. involvement, with 2 major U.S. data repositories (PetaBytes/year), 17 U.S. institutions providing data analysis and storage, and 68 universities and National Laboratories with scientists looking at the data. There are dedicated transatlantic networks connecting the U.S. to CERN; advanced network services are required over existing campus, connector/regional, and national networks.

2.1 Good End-to-End Performance

To achieve good end-to-end performance, we need to diagnose (understand the limits of performance) and address (work with members and application communities to address those performance issues) [3]. Internet2 consists of campuses, corporations, regional networks, and the Internet2 backbone network. Our members care about connecting with other members, government labs and networks, and international partners. The Internet2 community cares about making *all* of this work. Given that performance is excellent across backbone networks, and that performance is a problem end-to-end, it is clear that problems are concentrated towards the edge and in network transitions.

We envision readily available performance information that is easy to find, ubiquitous, reliable, valuable, actionable (analysis suggests course of action), and automated (applications act on data received). We assume a security system to authenticate and authorize users of the system as well as deter attacks. The goal of all this testing is to eliminate the mystery by increased network awareness. This allows us to set user expectations accurately, reduce diagnostic costs, and notice performance problems early, and address them efficiently. It also allows network engineers to see and act outside their "turf" as well as transforming application design to incorporate network intuition into application behavior.

The strategy Internet2 is employing is building and empowering the community. This requires developing analysis and visualization tools [4], and encouraging performance data generation and sharing. For this to happen, we need clean APIs and protocols between each layer and widespread deployment of both the measurement infrastructure and a set of common performance measurement tools. Internet2 provides diagnostic information for its "U.S. backbone" portion of problem, and we are working closely with GÉANT2 to ensure that we are providing the type of data they are likely to require, as well as encouraging them to provide data needed by U.S. diagnosticians. Internet2 has created a few diagnostic tools (BWCTL, OWAMP, and Thrulay) and makes network data public.

In the coming year, the focus will be on making performance data more widely available via perfSONAR. Currently, Internet2 is contributing to the 'base' perfSONAR development effort in partnership with ESnet, European NRENs, and RNP (Brazil). Staff contributes to the development of standards for performance information sharing via the Open Grid Forum Network Measurement Working Group and are working to integrate the Internet2 diagnostic tools as examples of perfSONAR services to encourage development of tools in the community.

A network engineer or application easily can discover additional monitoring resources, authenticate locally, be authorized to use remote network resources to a limited extent, acquire performance monitoring data from remote sites via standard protocol, innovate where needed, and customize the analysis and visualization process.

For the radio astronomy community, the use of integrated network monitoring when conducting tests to determine the feasibility of e-VLBI (electronic Very Long Baseline Interferometry) helped enable identification of a hardware problem in time to correct it. Ongoing, automated monitoring allowed a view of network throughput variation over time and is now used to schedule short-term data transfers. The visualization of the data highlights route changes, network outages, and any throughput issues at end points. This integrated monitoring provides an overall view of network behavior at a glance for astronomers in the U.S., Japan, and Sweden; the goal is to extend this network of testing servers to all participating radio telescopes.

2.2 Internet2 End-to-End Performance Initiative (E2Epi)

E2Epi, which includes Internet2 staff, members, and federal and international partners building performance monitoring tools and frameworks, is supported via network revenues, partnerships, and grants from NSF and NLM. Work focuses on the development and widespread deployment of perfSONAR, an international consortium developing a performance middleware architecture and a set of protocol standards for inter-operability between measurement and monitoring systems. Overall, perfSONAR is a set of open source web services that can be mixed-and-matched and extended to create a performance monitoring framework. Design goals include being standards-based, modular, decentralized, open source, and extensible. This makes it applicable to multiple generations of network monitoring systems and encourages it to grow 'beyond our control,' while still allowing it to be customized for individual science applications. perfSONAR is a joint effort of ESnet, GÉANT2 JRA1, Internet2, and RNP.

For more information, either contact Eric Boyd <eboyd@internet2.edu> or see http://e2epi.internet2.edu/, http://www.perfsonar.net/, and http://nwmg.internet2.edu.

3. Network Measurement

The original Abilene racks included measurement devices: a single PC coordinated early OWAMP and Surveyor measurements. The primary motivation was an understanding of how (and how well) the network operates. This was, largely, a NOC function but access to the measurements was available to other network operators to better understand the network. Over time, it became apparent that the datasets were valuable as a network research tool.

3.1 The Abilene Observatory

There are two components to the Observatory: 1) co-location (network research groups are able to co-locate equipment in the Abilene router nodes) and 2) measurement (data is collected by the NOC, the Ohio ITEC, and Internet2, and made available to the research community). During the Abilene upgrade in 2002; the network was expanded to two racks, one of which was dedicated to measurement. This provided the potential for the research community to co-locate equipment, which was immediately taken advantage of by several projects, including PlanetLab. For more information, see: http://abilene.internet2.edu/observatory/research-projects.html

3.2 The Internet2 Network Observatory

The Internet2 Network provides greater IP services than were available with Abilene and it requires a new type of Observatory to measure and monitor network activity. After seeking input from the community, Internet2 is supporting three types of services: 1) measurement, 2) co-location, and 3) experimental servers to support specific projects. The Internet2 Network Observatory will support both optical and router nodes. At this time, the Observatory collects data on one-way latency, jitter and loss; regularly-scheduled TCP and UDP throughput tests; SNMP; flow, with anonymized addresses; routing updates; router configuration; and dynamic updates (syslog, alarm generation, and polling via router proxy).

The Observatory uses Dell 1950 and Dell 2950 servers; they have Dual Core 3.0GHz Xeon processors, 2GB memory, Dual RAID 146GB disks, integrated 1GE copper interfaces, 10GE interfaces, and Hewlett-Packard 10GE switches. There are nine servers at router sites, with three servers at optical-only sites. Data is collected locally and stored in distributed databases; there are databases for usage, NetFlow, routing, latency, throughput, router, and syslog data. Some uses of existing datasets and tools include quality control, network diagnostics, network characterization, and network research. Modifications to the Observatory may be made in response to feedback from researchers.

3.2.1 Quality Control (QC)

Latency and throughput tests are required for any quality control effort. For Internet2 and its users, QC has been vital for some user communities (such as radio astronomers; see http://e2epi.internet2.edu/case-studies/VLBI/), network peerings, and IP backbone integrity. QC measurements for the latter have been vital to ensure problems are resolved in a timely manner. There are a full mesh of IPv4 and IPv6 tests on machines with 1GE interfaces (9000 MTU); the expected result is greater than 950Mbps TCP flows; if any path falls below 900Mbps for two successive testing intervals, this generates an alarm at the NOC.

QC is also important for peerings; Internet2 and ESnet watch the latency across peering points; Internet2 and DREN also will conduct some throughput and latency testing. During the setup, engineers found interesting routing and MTU size issues that they are investigating.

3.2.2 Network Diagnosis and Characterization

This includes testing to end hosts and more generic testing. For the former, Internet2 uses NDT and NPAD servers; users can run a quick check from a host with a browser to eliminate (or confirm) last mile problems, such as buffer sizing, duplex mismatch, etc. NPAD finds switch limitations if the server is close enough. For the latter, tests generally look for configuration and loss. Flow data is collected with a flow-tools package; all data that is not used for security alerts and analysis [REN-ISAC] is anonymized by truncating the addresses. Reports from the anonymized data (and some engineering reports) are publicly available.

3.2.3 Research Projects

The four topics of primary interest to network researchers are: major consumption, flows, routes, and configuration. Thanks to NSF, a Network Research Facilities Grant made it possible for Internet2 to provide access to this information to network researchers for 1.5 years. The Internet2 Network Observatory (replacing the Abilene Observatory) contains inside racks set for initial deployment, including new research projects (NetFPGA, Phoebus). The software and links are easily changed and Internet2 easily could add or change hardware depending on costs. The new Observatory will provide researcher tools and new datasets. For more information, see http://www.internet2.edu/observatory/.

References

[1] Boote, J. W., Boyd, E. L., Durand, J., Hanemann, A., Kudarimoti, L., Lapacz, R., Simar, N., Trocha, S., *Towards Multi–Domain Monitoring for the European Research Networks*, in *Selected Papers from the TERENA Networking Conference*, TERENA, ISBN 90-77559-04-3, 2005; also published in *PSNC's Computational Methods in Science and Technology* series (Volume 11(2)).

[2] Hanemann, A., Boote, J. W., Boyd, E. L., Durand, J., Kudarimoti, L., Lapacz, R., Swany, D. M., Zurawski, J., Trocha, S., *PerfSONAR: A Service Oriented Architecture for Multi–Domain Network Monitoring*, in *Proceedings of the Third International Conference on Service Oriented Computing*, Springer Verlag, LNCS 3826, pp. 241–254, ACM Sigsoft and Sigweb, Amsterdam, The Netherlands, December, 2005.

[3] Hanemann, A., Liakopoulos, A., Molina, M., Swany, D. M., *A Study on Network Performance Metrics and their Composition*, in proceedings of *TERENA Networking Conference 2006*; also in special edition of *Campus-Wide Information Systems* (Volume 23 – 4 – 2006 – ISSN 1065-0741), Emerald Publishing Group Ltd.

[4] Hanemann, A., Jeliazkov, V., Kvittem, O., Marta, L., Metzger, J., Velimirovic, I., *Complementary Visualization of perfSONAR Network Performance Measurements*, in proceedings of *International Conference on Internet Surveillance and Protection (ICISP)*, IARIA/IEEE, Cap Esterel, France, IARIA/IEEE, August, 2006.

PROCEEDINGS OF SCIENCE

TCPDelay: Constant bit-rate data transfer over TCP

Stephen Kershaw*
The University of Manchester
E-mail: Stephen.Kershaw@manchester.ac.uk

Richard Hughes-Jones
The University of Manchester
E-mail: R.Hughes-Jones@manchester.ac.uk

Transmission Control Protocol (TCP) is a reliable transport protocol which guarantees that data sent will be perfectly replicated at the receiver. In order to deliver on this guarantee, TCP transmits data according to well defined rules which create reliable transfers and attempt to ensure that our traffic can fairly co-exist with other data flows. However, this can result in delayed data transmission and highly variable throughput. We investigate the use of TCP for the transfer of real-time constant bit-rate data and report on the effects of the protocol on the flow of data. We discuss the requirements for TCP to be successfully applied in this situation and the implications for applications such as e-VLBI, where we are more concerned with timely arrival of data than guaranteed delivery.

Experiments show that for a lossy TCP connection using standard bandwidth-delay sized buffers the packet arrival times for a constant bit-rate flow diverge away from real-time arrival. Increasing the buffer sizes, by orders of magnitude in some situations, allows timely arrival of data with only temporary, though possibly lengthy, delays.

Lighting the Blue Touchpaper for UK e-Science - Closing Conference of ESLEA Project
March 26-28, 2007
Edinburgh

*Speaker.

PoS(ESLEA)002

1. Introduction

Transmission Control Protocol (TCP) is the most widely used transport protocol on the Internet, due in the most part to its reliable transfer of data. However, it is not ideal to use for constant bit-rate applications because TCP throughput can vary wildly in a lossy environment. Many applications using constant bit-rate data transfer desire timely arrival of data but the rate fluctuations of TCP mean that timely arrival of data is not guaranteed. We examine the effect of packet loss on packet arrival times and investigate whether packet loss and the consequent effect on throughput delays the data irrecoverably. The performance of TCP from the perspective of data arrival time will determine the suitability for real-time applications, such as e-VLBI.

Electronic Very Long Baseline Interferometry (e-VLBI) is a technique used for high-resolution observations in radio astronomy which involves the transmission of constant bit-rate data streams which are generated in real-time. Timely arrival of data is a fundamental requirement of e-VLBI and data are often transmitted using TCP, hence tests were conducted using constant bit-rate flows at rates of up to 512 Mbit/s to be representative of e-VLBI observations.

2. Transmission Control Protocol

2.1 Properties of TCP

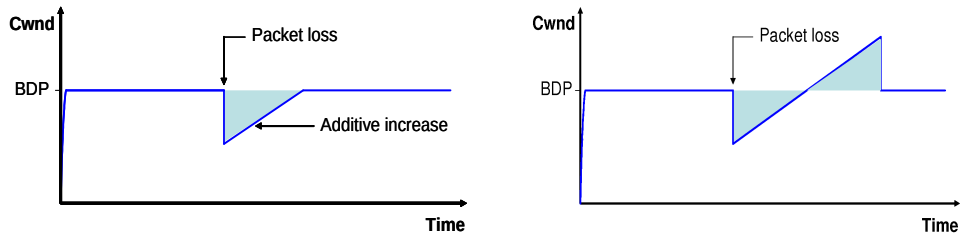
TCP is connection-oriented and reliable, ensuring that data sent will be perfectly replicated at the receiver, uncorrupted and in the byte-order sent. From the perspective of the application TCP ensures that the byte stream sent is the same as the byte stream received. Data corruption is detected by checksums and the receipt of all data (reliability) is ensured by using automatic repeat-request (ARQ), whereby the receiving system sends messages (ACKs) back to the sending system to acknowledge the arrival of data and hence indicate the missing data to be retransmitted. TCP assumes lost data packets are due to network congestion and attempts to mitigate congestion by varying the transmit rate - a process known as *congestion avoidance*, of great importance and described in more detail later.

2.2 High performance TCP

To make effective use of TCP, especially with high-capacity networks, it is often necessary to tune certain parameters. The end-hosts maintain windows over the data and to use the full capacity of a link the windows must be sized to the *bandwidth-delay product* (BDP) to allow sufficient "in-flight" unacknowledged segments [1]. In this investigation, a desired constant bit rate *CBR* was considered, where bandwidth delay product, *BDP*, is expressed as:

$$BDP = CBR \cdot RTT \tag{2.1}$$

where RTT = round-trip time. With windows sized to the BDP, steady line-rate throughput is achievable if we have no packet losses and so this practice is a generally recommended step to tune a TCP connection. In the event of packet loss the size of one window, the congestion window, on the sending host is adjusted to limit the maximum instantaneous throughput.



(a) BDP buffer size (b) Expanded buffer size

Figure 1: Theoretical action of TCP congestion avoidance with CBR data transfer. Comparing (a) and (b) we see the effect of dramatically increased buffer size. The shaded area of delayed data in (a) is compensated for in (b), transmitted faster than the constant bit-rate, having been buffered on the sending host

2.3 Reaction to loss

When TCP detects a lost packet it is assumed that the loss was due to network congestion and TCP enters a congestion avoidance phase, altering the achievable transmit rate dramatically by adjusting the congestion window. This feature of TCP was implemented to prevent congestion collapse of the Internet where competing flows reduce the useful throughput to zero. It is the congestion avoidance behaviour of TCP that creates problems for constant bit-rate flows.

The standard NewReno response to congestion is a decrease of the congestion window by a factor of 2, followed by an additive increase of 1 packet per round-trip time. This gives the throughput a characteristic sawtooth shape when a packet loss is detected - a sudden dramatic reduction of the congestion window, followed by a gradual linear increase. Considering this sawtooth congestion avoidance response, as shown in Figure 1, the amount of data that is delayed can be calculated.

2.4 Delayed data

When a packet is lost, the equations of TCP congestion avoidance determine both the decrease of the congestion window and the rate of increase. If we consider a pre-loss thoughput of *CBR*, it can be calculated that the time taken to regain *CBR* throughput is given by:

$$t_{recovery} = \frac{CBR \cdot RTT^2}{2MSS} \tag{2.2}$$

where *MSS* is the maximum segment size, the maximum amount of data that TCP can encapsulate in one packet.

The shaded triangular area in Figure 1(a), whose presence is due to a packet loss, has area proportional to the amount of data that has been delayed. The area is proportional to the recovery time and can be represented simply as:

$$\frac{CBR^2 \cdot RTT^2}{8MSS} \tag{2.3}$$

For applications like e-VLBI, where data are transferred over large distances at high rates it is essential to note from Equation 2.3 that the delayed data scales with the square of the throughput and the square of the round-trip time.

3. Constant bit-rate data over TCP

It is often said in the literature that TCP is largely unsuitable for real-time applications and constant bit-rate flows because of the variable rate of TCP over a lossy connection due to congestion avoidance [2],[3],[4].

If CBR data is streamed over TCP, as with some multimedia applications or e-VLBI, the reduced throughput due to packet loss leads to a data arrival rate on the receiver of less than the original constant bit-rate. If the processing or playback at the application level is not to stall then sufficient data must be stored in a playout buffer to compensate for the lower data-rate at the transport level, allowing CBR data arrival at the application level. This is common practice for streaming multimedia applications, requiring an initial buffering period and hence a delay between the start of the transfer and the start of the playback.

The situation of bulk data transfer is quite well researched and understood,[5],[6], in contrast to the equivalent situation but where the CBR data is generated and transferred in real-time. When a CBR data stream is generated in real-time and cannot be stalled then we must transfer the data at a steady CBR else we have to either discard or buffer the data at the sending end.

3.1 Regaining timely arrival

If we temporarily store the delayed data on the sending host and can subsequently transfer it faster than the constant bit-rate then we should be able to regain timely arrival of data at the receiving host. We require the data to be buffered and the maximum window size must permit transfers at a rate higher than the CBR. In the investigation that follows, both functions are performed using the socket buffers in Linux.

Data from a Linux application, destined for a network, is buffered in a socket buffer, the size of which we can specify through kernel and application parameters. The socket buffer in Linux serves two purposes: to retain data for windowing and also as an application buffer, designed to isolate the network from effects of the host system, such as scheduling latency of the Linux kernel. Therefore, on the sending host, the socket buffers which are an integral part of TCP/IP in the Linux kernel can be used to buffer the data that is delayed in the event of a loss.

4. Experimental configuration

4.1 Network setup

The network links used to test constant bit-rate performance over TCP were dedicated fibre optic lightpaths with connections to UKLight in the UK peering with NetherLight in the Netherlands. The links were tested to have a very low bit-error rate, allowing loss-free data transfers with stable delay and jitter characteristics, making for an ideal protocol testing configuration. The lightpaths were used to connect to hosts in Manchester, Jodrell Bank Observatory and JIVE (Joint Institute for VLBI in Europe, Dwingeloo, Netherlands) with dedicated point-to-point 1 Gbit/s connections. From Manchester the round-trip times (RTT) were 1 ms and 15 ms for Jodrell and JIVE respectively, with a connection from Manchester to Jodrell, looped back in the Netherlands giving a RTT of 27 ms.

The computers used as end-hosts were server-quality SuperMicro machines, with all configurations tested to give 1 Gbit/s throughput using UDP/IP or TCP/IP over Gigabit Ethernet interfaces. The systems used Intel Xeon CPUs and were running Red Hat or Fedora distributions of Linux. Test were performed using kernel versions 2.4.20 and 2.6.19 with negligible difference in TCP performance seen between kernel versions. All systems were equipped with onboard Intel e1000 Gigabit Ethernet ports.

4.2 Diagnostic software

TCPdelay [7] is an application written by Richard Hughes-Jones, used to conduct tests using memory-to-memory TCP streams, sending data to a socket at regular intervals so as to attain a specified average data rate, emulating a CBR data stream. TCPdelay measures timings of packet sending and arrival at the application level, allowing measurement of whether the data stream is arriving at the receiver in a timely manner.

In order to gain more insight into the behaviour of TCP the *web100* kernel patch was used. Web100 [8] is a kernel patch which provides extended TCP instrumentation, allowing access to number of useful TCP related kernel variables, such as the instantaneous value of the congestion window.

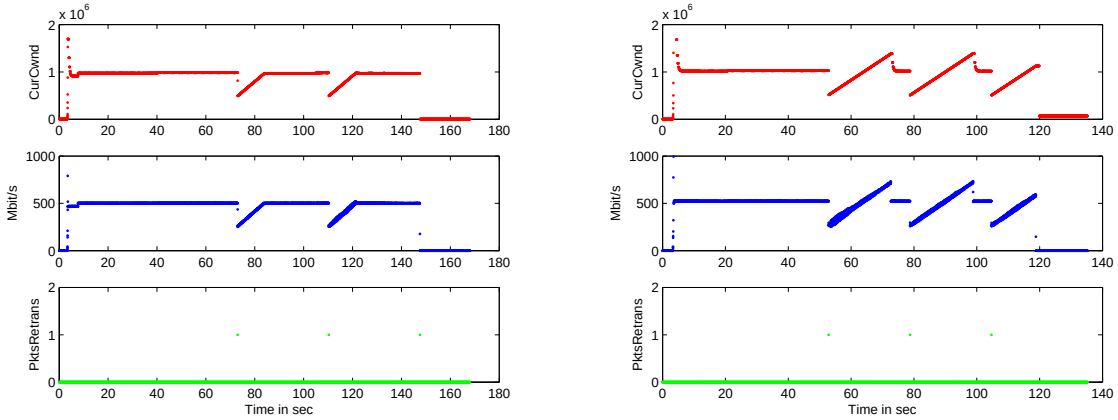
Packet loss on the test networks is rare, so we simulated packet loss in the receiving hosts using a Linux kernel patch to discard packets at a configurable, regular rate.

5. Results

Using the often recommended socket bandwidth-delay product buffer size, the behaviour of a 512 Mbit/s CBR TCP stream over a lossy 15 ms connection was studied with 0.9 Mbyte (BDP) socket buffers. The observed behaviour is shown in Figure 2(a). In the event of a lost packet (deliberately dropped on the receiving host) we see the reliable TCP protocol retransmitting a packet (lowest plot) and we see the expected congestion window evolution, as detailed earlier and illustrated in Figure 1(a). The rapid decrease and additive increase of the congestion window is apparent, with recovery of the constant bit-rate transfer taking around 10 seconds. We see an amount of delayed data of around 160 Mbyte, in agreement with Equation 2.3 when delayed acknowledgements are accounted for.

Data is further delayed with every subsequent packet lost, the cumulative effect of multiple losses shown in Figure 3, which demonstrates the effect of loss rate on message arrival time. The lowest curve in Figure 3 shows the observed timely arrival of data, with higher curves showing lossy transfers diverging rapidly away from this ideal. As one may expect, with the throughput dipping below the desired rate many times and never exceeding it, the amount of delayed data increases and the data arrives later as the duration of the transfer increases.

Figure 1(b) shows the same network configuration of the test in Figure 1(a) but with increased socket buffers of 160 Mbytes, which is the calculated amount of delayed data. As explained previously, the intention was that the delayed data was stored in the TCP socket buffer, to eventually be transmitted at a rate in excess of the constant bit-rate. We see in Figure 1(b) that we initially have the same post-loss behaviour as (a) but the buffered data means that we can transmit faster as the transfer from the buffer memory is not limited to the constant bit-rate. Once the buffered data has



(a) 0.9 Mbyte (BDP) buffer

(b) 160 Mbyte (180x BDP) buffer

Figure 2: Plots of TCP parameters, logged using web100. Kernel patch used to drop packets.
Top: TCP Congestion window (bytes)
Middle: Achieved throughput (Mbit/s)
Bottom: Number of packets retransmitted

been exhausted, we transmit new data at the CBR once more, as seen in the figure. For the duration of the sawtooth the receiver experiences delayed data arrival, but subsequent data arrives in a timely manner once more, until the next loss. In this situation, with a constant bit-rate of 512 Mbit/s and a 15 ms RTT, we can use a 160 Mbyte buffer on the sending side to allow timely delivery to be resumed at the receiver.

Instead of never resuming timely arrival and the departure for timely arrival becoming increasingly worse with time, which is the situation with conventionally sized buffers, we can use larger buffers to instead suffer only a temporary period of delayed data arrival. One must consider however the logistics of providing such large buffers and be able to cope with the temporary period of delay.

6. Conclusions

In a lossy environment, using TCP/IP to transfer CBR data with normal TCP buffer settings (BDP) leads to delayed data arrival. The delay is to the entire stream of data as all data arriving after the first loss will be delayed, with subsequent losses compounding the problem. For an application such as e-VLBI this not acceptable and can lead to a loss of correlation and lower quality results as multiple streams become unsynchronised.

In theory, to regain timely delivery of data following a loss, it is necessary to store the delayed data and subsequently transmit it at a rate exceeding the constant bit-rate to achieve an average CBR throughput. This can be demonstrated in practice in a relatively simple manner by using the socket buffers for this temporary data storage. The practicalities of providing buffers depend strongly on the parameters of the application and network. For a 512 Mbps flow over a connection with a round trip time of 15 ms we are required to buffer 160 Mbytes of data. The scaling apparent

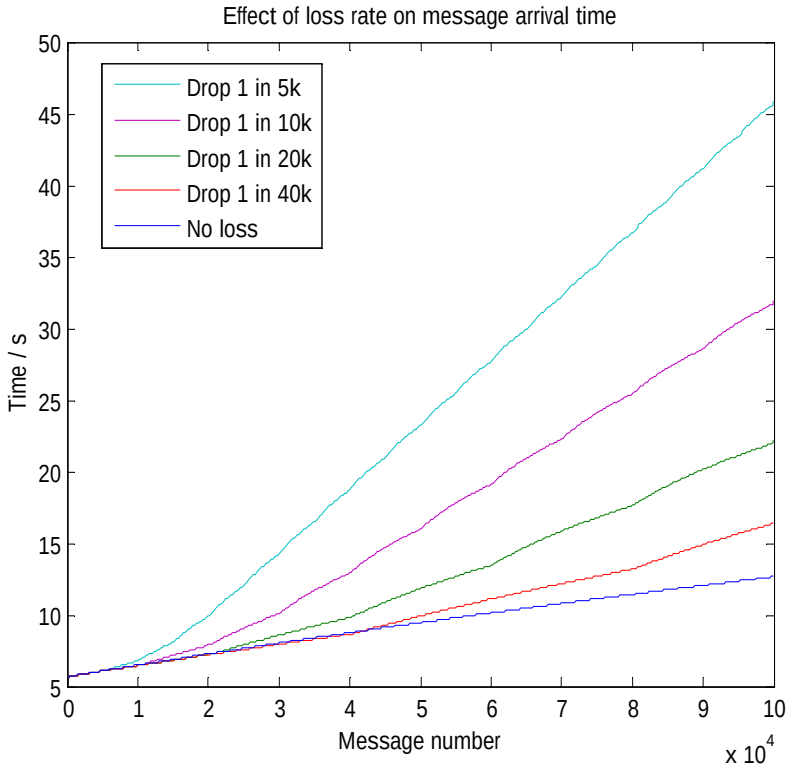


Figure 3: The effect of packet loss on message arrival time. Manchester to Jodrell Bank, looped through Amsterdam, 27ms RTT. TCP buffer size 1.8 Mbytes (BDP)

in Equation 2.3 is an important consideration, with transatlantic distances requiring the buffering of upwards of 5 Gbytes of data and temporary departure from timely delivery of tens of minutes. This will often prove impractical, with alternative protocols or TCP variants being options to consider.

References

[1] W. R. Stevens, *TCP/IP illustrated: the protocols*, Addison-Wesley Publishing Company, Reading, Mass., 1994

[2] K. Li et. al, *The Minimal Buffering Requirements of Congestion Controlled Interactive Multimedia Applications*, *Lecture Notes in Computer Science*, 2158, 2001

[3] C. Krasic, K. Li, J.Walpole, *The Case for Streaming Multimedia with TCP*, *Lecture Notes in Computer Science*, 2158, 2001

[4] B. Wang et. al., *Multimedia streaming via TCP: An analytic performance study*, *Performance Evaluation Review*, 32, 2004

[5] M. Mathis, J. Semke, J. Mahdavi, and T. Ott, *The Macroscopic Behavior of the TCP Congestion Avoidance Algorithm*, in *Computer Communications Review*, 27(3), July 1997

[6] J. Padhye et. al., *Modeling TCP Throughput: A Simple Model and its Empirical Validation*, in proceedings of *ACM SIGCOMM*, September 1998.

[7] R. Hughes-Jones, *TCPdelay Home Page*. Available at :
http://www.hep.man.ac.uk/u/rich/Tools_Software/tcpdelay.html

[8] Various authors, *The Web100 Project*. http://www.web100.org/

PROCEEDINGS OF SCIENCE

Implementing DCCP: Differences from TCP and UDP

Andrea Bittau*
University College London
E-mail: a.bittau@cs.ucl.ac.uk

Mark Handley
University College London
E-mail: m.handley@cs.ucl.ac.uk

We describe our experiences in contributing to the implementation of a new protocol, DCCP, in the Linux kernel. Being the first implementation in a main-stream operating system, we are the first ones to explore the implications and the unexpected issues that could arise from developing this protocol. We will focus on how the DCCP implementation differs from that of TCP and the performance issues that we have encountered.

Lighting the Blue Touchpaper for UK e-Science - Closing Conference of ESLEA Project
March 26-28, 2007
Edinburgh

*Speaker.

PoS(ESLEA)003

Component	lines
Ack vectors & feature negotiation.	1,162
Rest of DCCP core.	2,876
Total DCCP core.	4,038
CCID2.	583
CCID3.	1,839
Minimum DCCP (core & CCID2).	4,621
TCP implementation.	8,042
UDP implementation.	1,160

Table 1: Source lines of code for protocol implementations in Linux 2.6.19.

1. Introduction

The *Datagram Congestion Control Protocol* (DCCP) is a transport protocol that does not provide delivery guarantees and has built-in congestion control [3]. One may think of it as UDP with congestion control, or TCP without reliability. This makes DCCP ideal for multimedia applications that prefer, upon packet loss, sending new data instead of old (and now useless) retransmissions. The congestion control algorithm in DCCP is not fixed and applications may choose which one to use by selecting the appropriate *Congestion Control Identifier* (CCID). Currently, there are two CCIDs defined: CCID2 which is TCP-like and CCID3 which is TFRC [2].

In this paper we focus on the implementation issues of DCCP rather than on its design. Our work has been carried out as protocol research in the context of e-VLBI, an application where multiple data streams from different telescopes are correlated to produce an image. The requirements for e-VLBI are transmitting large amounts of data at a (very high) constant bit-rate, and packet loss can be tolerated. TCP is inadequate for e-VLBI due to its bad performance when dealing with large windows. UDP is partially suitable for e-VLBI since it may not be used on shared links (*e.g.* Internet, or shared academic networks) due to its lack of congestion control. DCCP is the best fit—it can transmit data at a constant bit-rate (CCID3) and in the case of congestion, it will back off. In the following sections we will discuss the major implementation differences in Linux between DCCP, TCP and UDP, followed by some performance considerations.

2. Differences from TCP and UDP

Table 1 summarizes the lines of code of different protocols in Linux 2.6.19. DCCP is core-complete, but still missing some optional parts. The code is $\approx 57\%$ of TCP's code size (UDP is much simpler). Part of the reason is that DCCP needs a state machine and mechanisms equivalent to those of TCP in order to be robust against attacks, *e.g.* it needs to detect whether a reset packet is valid (in sequence) before terminating a connection. UDP does not have this complexity and it is generally left to application protocols (if necessary). Complexity in DCCP is added by *ack vectors* and *feature negotiation*. Because DCCP's delivery is unreliable, the protocol may not make use of cumulative acknowledgments as TCP does. Instead, a map representing which packets have been received and which not, much like TCP's SACK, needs to be transmitted and processed (ack

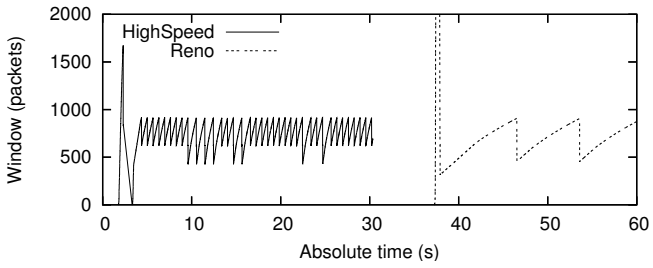


Figure 1: DCCP CCID2 running with TCP's HighSpeed algorithm.

vectors). Thus, detecting loss in DCCP is more complex than in standard TCP. Feature negotiation is a mechanism for negotiating options and may be thought of like TCP options, although the mechanism is much more versatile in DCCP. Together ack vectors and feature negotiation comprise $\approx 29\%$ of the core DCCP code. Because of these extra mechanisms, we believe that the complete DCCP implementation will approach the complexity as TCP's—it is not a trivial protocol like UDP. Currently missing in the implementation are some protocol options (*e.g.* slow receiver) and the handling of special cases such as detecting and dealing correctly with unidirectional data flows.

The CCIDs in DCCP are quite large because they share no code (unlike in TCP's case). This is so because the algorithms are fundamentally different—CCID2 is window based and CCID3 is rate based. We developed an experimental patch which allows TCP congestion control modules to be used by CCID2. Figure 1 shows DCCP's congestion window when using the HighSpeed TCP algorithm [1]. The result is as expected, a higher frequency of losses and a more aggressive window increase when HighSpeed is compared with Reno. It was an interesting result that the same congestion control code worked correctly in protocols which have totally different semantics—reliable *vs.* unreliable. Although DCCP's CCID2 and TCP have very different mechanisms for detecting congestion, the actions taken are very similar. The congestion control modules between TCP and DCCP turned out to be compatible because they only need to be notified about loss—they do not need to detect loss themselves.

3. Performance

We were able to transfer at 1Gb/s, as reported by *iperf*, by using DCCP with CCID2 in a lab experiment. We connected, back-to-back, two Intel Xeon 3GHz boxes with e1000 1Gbit PCIe network cards. On the transmitter, we emulated a 200ms delay using *netem* in order to give us a large bandwidth-delay product (window). This stressed the implementation since the amount of required state, *e.g.* ack vector size, grows proportionally to the window size. We still need to further optimize the code since the CPU utilization is $\approx 90\%$ when transmitting at gigabit rates. In TCP's case, the CPU utilization is lower and this is mainly due to the fact that TCP does not have to process a large ack vector upon receiving every packet.

After profiling the kernel, we discovered that the CPU was spending most of its time calculating checksums ($\approx 25\%$). Checksum offloading to the network card will definitely reduce CPU utilization and we are planning to support it in the future. Figure 2 shows how fast two different boxes can calculate checksums using the Linux kernel code. As the packet size grows, the number

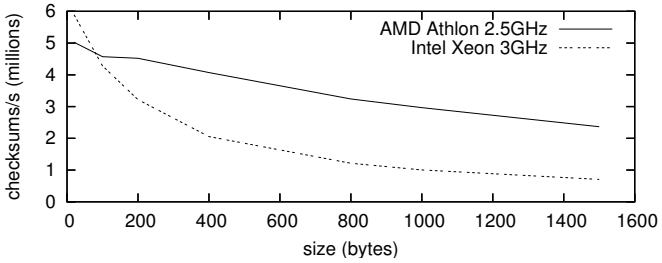


Figure 2: Checksum calculation speed of an (older) Intel and AMD box.

of checksums that can be calculated (thus packets sent) decreases significantly. One difference between DCCP and TCP is that the DCCP protocol allows a sender to specify, via the *checksum coverage* field, which bytes are to be included in the checksum calculation. For example, it is possible to checksum only the header and not the payload but with the drawback of sacrificing some protection. This is tolerable by some applications, such as e-VLBI, and most likely is not an issue in practice if the MAC layer has a checksum too. Thus, by using these simpler to calculate checksums, it is possible to decrease the load on a system. This cannot be done with TCP

In DCCP, packet framing is done by the application so the kernel does not need to worry about segmentation. This leads to a problem in DCCP which is absent in TCP. When sending large chunks of data with TCP, it is possible to invoke a single *send* system call that will cause multiple packets to be sent out. With DCCP, a single *send* call will send out only a single packet. Thus, to transfer large amounts of data, many *send* calls need to be invoked and the context switch overhead is no longer negligible. It is our intent to research APIs which suit DCCP better and are optimized for high-speed networks. For example, we are thinking about a *sendv* system call which will enqueue multiple packets with a single call.

4. Conclusion

The DCCP implementation approaches the complexity of TCP's because of the rich set of features supported by the core protocol. We intend to unify the congestion control algorithms used by TCP and CCID2 in order to share the (at times complex) congestion control code.

In the current implementation, the largest performance hit is checksum calculation. This can be mitigated by offloading checksum calculations to the network card, or in some cases by sending out checksums based only on the packet header and not the entire payload. We also believe that existing APIs need to be extended in order to achieve even greater performance with DCCP, for example, by adding APIs to enqueue multiple packets with a single system call.

References

[1] S. Floyd. HighSpeed TCP for large congestion windows, 2002.

[2] M. Handley, S. Floyd, J. Padhye, and J. Widmer. TCP Friendly Rate Control (TFRC): Protocol Specification, January 2003.

[3] E. Kohler, M. Handley, and S. Floyd. Designing DCCP: Congestion Control Without Reliability. In *SIGCOMM '06*, September 2006.

PROCEEDINGS OF SCIENCE

Testing of DCCP at the Application Level

Richard Hughes-Jones*
The University of Manchester
E-mail: R.Hughes-Jones@manchester.ac.uk

Stephen Kershaw
The University of Manchester
E-mail: Stephen.Kershaw@manchester.ac.uk

Datagram Congestion Control Protocol (DCCP) is a recently developed transport protocol whose development and implementation in Linux is being aided by the work of Mark Handley and Andrea Bittau in the ESLEA project. The protocol is attractive to many applications where data is transferred with tight constraints on the timing of data delivery, such as internet telephony and e-Science applications such as e-VLBI.
Porting test programs to DCCP has allowed the investigation of the DCCP implementation in recent releases of the Linux kernel and reporting of performance test results. A suggested approach for the use of DCCP for e-VLBI is discussed, with a proposal for a new CCID.

Lighting the Blue Touchpaper for UK e-Science - Closing Conference of ESLEA Project
March 26-28, 2007
Edinburgh

*Speaker.

PoS(ESLEA)005

1. Introduction

Datagram Congestion Control Protocol (DCCP) is a recently developed transport protocol, similar in parts to both TCP and UDP with the intention that certain applications and the transport of certain types of data may benefit. Where congestion control is required but reliability is not, DCCP provides a transport level option attractive to many applications such as VoIP and e-VLBI. The congestion control algorithm, CCID, used by DCCP is selectable, allowing DCCP to be tuned more closely to the requirements of a particular application. CCID2 is *TCP-like Congestion Control*, closely emulating Reno TCP while CCID3 is *TCP-friendly rate control*, minimising rate fluctuations whilst maintaining long-term TCP friendly behaviour.

DCCP has been in the Linux kernel since 2.6.14, with recent kernel releases such as 2.6.19 and 2.6.20 having an implementation, incorporating the code developed by ESLEA, that is often considered as fairly stable and high-performance. We report on the porting of a network testing application to DCCP, experiences with creating a stable DCCP testbed and results from initial performance tests.

2. Porting of test software

In order to test the performance of DCCP, software tools were required hence *DCCPmon* is a port of *UDPmon* by the original author, Richard Hughes-Jones [1]. Guidance was given by Andrea Bittau to help with the port to DCCP and the resulting application is being used and proving to work well. However, the process was not entirely trouble-free - some problems were encountered that were perhaps indicative of an implementation that is in development, rather than complete and polished. DCCP related #defines were not to be found in the userland include files, an issue mitigated by creating specific include files. Some system calls were noted to be missing and the API was in a state of flux with functions changing between kernel releases 2.6.19 and 2.6.20. For this reason, and due to limited testing, *DCCPmon* is currently still considered by the author as experimental.

During the development of *DCCPmon* and for corroboration of results, a patched version of *iperf* [2] was used. In addition to the information from the main test application it is desirable to gather data from as many other sources as possible. One useful window into the kernel networking stack is though the kernel SNMP statistics, however there are currently (as of kernel 2.6.21) no SNMP counters for DCCP variables. These statistics would also have been invaluable when problems became apparent with certain kernel versions and it would certainly be a worthy addition to the implementation at the earliest opportunity.

3. End-host setup

The computers used as end-hosts were server-quality SuperMicro machines, with all configurations tested to give 1 Gbit/s throughput using UDP/IP or TCP/IP over Gigabit Ethernet interfaces. The systems used Intel Xeon CPUs and were running Scientific Linux or Fedora distributions of Linux. We had systems using two Dual Core Intel Xeon Woodcrest 5130 CPUs clocked at 2 GHz, dual-booting 32-bit and 64-bit distributions of Fedora Core 5. We also had systems with two Intel

Xeon 2.4 GHz Hyper-Threaded CPUs using a 32-bit distribution of Scientific Linux 4.1. All systems were equipped with and DCCP tested with on-board Intel e1000 Gigabit Ethernet ports. Tests with UDP and TCP gave stable line-rate performance over all tested networks, including 1 Gbit/s over a transatlantic lightpath.

4. Experiences with the Linux DCCP implementations

While developing the *DCCPmon* program and preparing for performance tests, several different Linux kernels have been used, often displaying undesirable effects. With such a new implementation of a new protocol it has often been unclear whether we are seeing problems with the DCCP implementation or something specific to our systems, however we report on our findings and some of the steps taken to achieve a stable DCCP test bed.

4.1 Kernel version 2.6.19-rc1

This kernel version is a release candidate for stable kernel version 2.6.19, which was tested before the stable kernel version was released. Using both *DCCPmon* and *iperf* it was found that we were not getting a working DCCP connection - *tcpdump* showed that the connection was successfully made, with packets exchanged both ways but no ACKs were sent in response to data packets received. In the absence of feedback the sender-side DCCP transmit timer progressively fell back until a threshold upon which DCCP terminated the connection.

We conducted many diagnostic tests to establish the cause of the problem. Advanced features of the network interface card were disabled and DCCP data was sent though a tunneled connection to prevent possible discrimination of the new protocol. Eventually, inserting debugging code into the kernel showed that data were incorrectly being discarded due to header checksum errors, a problem that was later fixed in the network development tree and merged into the stable 2.6.19 kernel release.

4.2 Kernel versions 2.6.19 and 2.6.20

As previously noted, the API calls changed slightly, necessitating further development of the test software code, after which, with the checksum problems resolved it was hoped that interesting tests could be run.

The initial results were promising, with CCID2 showing short-term line-rate throughput - a useful data rate of around 940 Mbit/s after header overheads. CCID3 had an average rate of around 300 Kbit/s but unfortunately DCCP proved to be unstable using either CCID on our 64-bit systems. Transfers would often only last for a few seconds before the receiving system hung with a kernel panic. Some tests would continue for longer, a few minutes with the same throughput performance, but all would trigger a kernel panic within four minutes and repeating tests with larger packet sizes would lead to a quicker crash. The crash dumps associated with the panic generally indicated that the crashes were occurring most regularly in the region of the packet reception code of the network interface card (NIC), where memory is allocated to store incoming packets.

Repeating the tests using a 32-bit distribution and kernel on the same computers yielded the same behaviour, however the older systems running Scientific Linux on Hyper-Threaded Xeon processors proved to be more stable, with extended runs possible, with the majority of transfers

persisting until deliberately terminated after many tens of minutes. The system logs, however, showed that everything was not perfect, with many zero order page allocation failures logged, in a similar context to the panics - close to the receive interrupt of the NIC.

5. Towards a stable test bed

Analysis of crash dumps and kernel messages, showed that most error messages were generated when memory was being allocated in NIC RX IRQ handler. To attempt to fix the problem the operation of the NIC driver was analysed together with aspects of the kernel memory management code.

In general, when a request is made for memory allocation, the request will either be serviced immediately (if memory is available) or it will be blocked while sufficient memory is reclaimed. However, when memory allocation is requested in an interrupt context, for example memory allocation to store received packets, blocking is forbidden. In order that the memory allocation has a higher chance of succeeding, the kernel reserves some memory specifically for this situation where the allocation is classed as *atomic*. The amount of memory reserved for atomic allocations is determined by the value of the *min_free_kbytes* sysctl variable.

Increasing the *min_free_kbytes* parameter in the receiving host from the default value of 5741 to 65535 proved to prevent all the previously seen error messages, though it is not entirely clear to us why the memory allocation problems originally occur. It is possible that the default value of *min_free_kbytes* is not sufficient relative to the time between scheduled runs of the memory management daemon (e.g. kswapd), which are scheduled to keep that minimum amount of memory free. A larger value of *min_free_kbytes* may mean that the reserved memory is never filled before the memory management routines can be run. As we do not encounter similar problems with UDP and TCP, it is possible that the higher CPU utilisation of DCCP could cause such a situation by using more CPU time. It is strange that on one system the allocation failures prompted errors messages while on another the result was a fatal system crash.

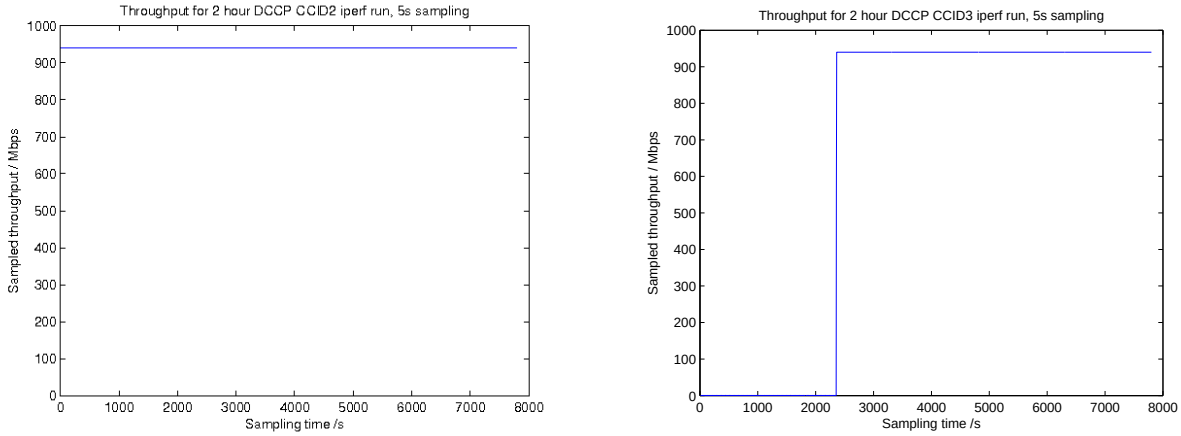
The problem is not entirely mitigated though as even with the increased value of *min_free_kbytes* crashes persist if the packet size is increased sufficiently. More investigation is needed to gain a full understanding of this unwanted feature of our DCCP test.

6. Results of recent tests

With an increased value of *min_free_kbytes* on the receiving hosts, the systems proved to be stable with 1500 Byte packets, with no tests generating error messages of any kind. Every test, with flow durations of up to 2 hours, remained stable and was terminated gracefully at the predetermined time, using all kernel versions of 2.6.19 or later. Having a stable test bed has allowed preliminary tests of DCCP throughput performance, as outlined below.

6.1 Back-to-back tests

With any two systems CCID2 can attain the maximum possible data rate of 940 Mbit/s, which is line-rate over Gigabit Ethernet and stable for the duration of the test. This result is illustrated in Figure 1(a), with Figure 1(b) showing the different behaviour of CCID3. With CCID3 there is an



(a) CCID2 (b) CCID3

Figure 1: CCID comparison

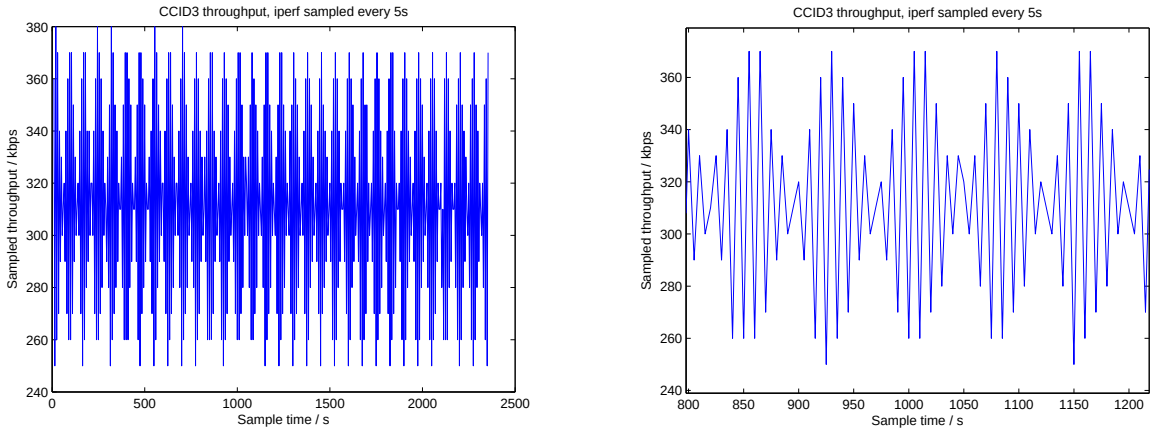


Figure 2: Expanded view of initial CCID3 throughput variation

initial period with an average rate of 300 Kbit/s, with the regular rate variation detailed in Figure 2. After a number of packets (around 65,500) the rate jumped to line-rate and remained steady, as seen in Figure 1(b). This is strange behaviour, with the number of packets being indicative with a 16-bit overflow perhaps, but there has been a lot of patches produced for CCID3 recently which have not yet made it into the stable Linux tree. Using a development tree and patches from numerous authors changes the CCID3 behaviour completely. The most appropriate comment to make is that CCID3 is developing and the performace of current stable kernels is not indicative of what is beiachievedved by developers.

6.2 Tests over extended networks

Over a transatlantic connection, with end-hosts in Manchester and Chicago, using UDP and TCP we can achieve line-rate throughput. Although the back-to-back performance of DCCP be-

tween identical systems gave line-rate, over the 94 ms transatlantic lightpath only a steady 130 Mbit/s was attained.

The performance of DCCP seemed to be CPU limited at the sender, with one CPU showing an average of 98% load, compared to the load at line-rate back-to-back of 82%. Increased CPU load with increasing round-trip time can sometimes be observed with TCP flows but it is not immediately obvious that this should be the case with DCCP and it is curious that the effect seems so dramatic. The performance of DCCP needs to be investigated further over different distances and with different systems. CPU load profiles can hopefully yield further useful information about the performance of DCCP.

7. Developing a new CCID

VLBI has a clear requirement to move constant bit-rate data and can tolerate high-levels of packet loss, making UDP seem like the ideal transport protocol. Other applications have similar requirements, with streaming media and VoIP being examples of applications where constant bit-rate can be advantageous and packet loss is often tolerable. However, there is concern from network providers that UDP traffic could overwhelm other traffic and overload the network. Concerns and opinions have been voiced and mitigating options have been discussed at recent meetings such as the EXPReS & EVN-NREN meeting in Zaandan, NL and PFLDnet 2007 / IRTF workshop in Marina Del Rey, US, with input from Kees Neggers, SURFnet; Glen Turner, AARNET; Aaron Falk, IRTF Chair. One option that the authors support is to use DCCP in combination with a new CCID, initially given the name *SafeUDP*. The proposed CCID aims to address the concerns expresses about using plain UDP by implementing something "UDP like" but with network protection.

SafeUDP would use the DCCP ACK mechanism to detect congestion, following which the congestion would be evaluated: to ensure that congestion is not in the end-host and to determine whether the congestion is transient. This evaluation step is useful to remove the assumption that all losses are congestion events, which is a conservative assumption but in some circumstances often unnecessarily detrimental to performance. The application would be notified of the congestion through modified API calls, with *sendto* and *recv_from*, etc. having new return codes. The application can then take action, with the CCID dropping input from the application and informing the application that it has done so if no action is taken. This idea is being worked on with the long-term aim of a draft RFC.

8. Conclusions

The Linux implementation of DCCP is almost certainly the most mature implementation available. Once we had established a stable test bed we investigated the performance of DCCP using CCID2 and CCID3 with tests conducted primarily using the test program *DCCPmon*, a port of existing application *UDPmon*. Apart from minor troubles due to omissions or changes to the API, the port was relatively straight-forward.

We have seen that the back-to-back performance of DCCP using CCID2 is good, achieving line-rate for extended (multiple hour) back-to-back, memory-to-memory transfers. The throughput

of CCID3 was generally lower though there is much current development with performance changing with every patch. Given the amount of patches being created by developers it is uncertain at what speed the CCID3 implementation in the stable kernel will develop.

Tests of CCID2 over extended networks have been quite limited to date, with early results showing that DCCP uses much more CPU time and achieves a lower rate over a transatlantic lightpath. A rate of 130 Mbit/s to compare with 940 Mbit/s back-to-back has been seen, with further work needed to fully assess DCCP performance over long-distances.

achieveive a stable test setup has not been trivial and there are some issues still to be resolved. We hope that our investigations of the issues with DCCP on our systems can help improve the implementation and make DCCP work "out-of-the" box on more systems. Working round the issues we encountered revealed a protocol implementation that we look forward to investigating more fully in the near future. Many applications can benefit from DCCP and we hope to extend the utility by considering the concerns of and working with network managers to build a new CCID.

References

[1] R. Hughes-Jones, *DCCPmon Home Page.* Available at : http://www.hep.man.ac.uk/u/rich/Tools_Software/dccpmon.html

[2] National Laboratory for Applied Network Research, *NLANR/DAST : Iperf.* Available at : http://dast.nlanr.net/Projects/Iperf/

PROCEEDINGS OF SCIENCE

Using UDT for High Energy Physics Data Transport

Barney Garrett[1]
The University of Edinburgh
James Clerk Maxwell Building, Mayfield Road, Edinburgh. EH54 6TD. UK
E-mail: barney.garrett@ed.ac.uk

Brian Davies
Lancaster University
Department of Physics, Lancaster. LA1 4YB. UK
E-mail: b.g.davies@lancaster.ac.uk

eScience applications, in particular High Energy Physics, often involve large amounts of data and/or computing and often require secure resource sharing across organizational boundaries, and are thus not easily handled by today's networking infrastructures. By utilising the switched lightpath connections provided by the UKLight network it has been possible to research the use of alternate protocols for data transport. While the HEP projects make use of a number of middleware solutions for data storage and transport, they all rely on GridFTP for WAN transport. The GridFTP protocol runs over TCP as the layer 3 protocol by default, however with the latest released of the Globus toolkit it is possible to utilise alternate protocols at the layer 3 level. One of the alternatives is a reliable version of UDP called UDT. This report presents the results of the tests measuring the performance of single-threaded file transfers using GridFTP running over both TCP and the UDT protocol.

Lighting the Blue Touchpaper for UK e-Science - Closing Conference of ESLEA Project
The George Hotel, Edinburgh, UK
26-28 March, 2007

[1] Speaker

PoS(ESLEA)006

1. Introduction

TCP uses what it calls the congestion window to determine how many packets can be sent at one time. The maximum congestion window is related to the amount of buffer space that the kernel allocates for each socket. If the buffers are too small for the network connection the TCP congestion window will never fully open up resulting in never reaching the maximum potential of the network connection. In Long Fat Networks (LFN), such as UKLight the default kernel settings within Linux are inadequate. The Linux kernel can be tuned[1] for LFN's with the Bandwidth Delay Product (BDP) being used to give an appropriate buffer size. This is calculated using:

```
BDP (bytes) = Bandwidth (bytes) * RTT (seconds)
```

The socket receive buffer space is shared between the application and kernel. TCP maintains part of the buffer as the TCP window, this is the size of the receive window advertised to the other end. The rest of the space is used as the "application" buffer, used to isolate the network from scheduling and application latencies. By default this overhead is a quarter of the buffer space that the kernel is configured to use.

The txqueuelen is another buffer in the kernel stack that can affect performance; especially of TCP transfers. When the system is sending out too much data from the IP layer to the Ethernet device driver layer, this buffer, txqueuelen, may overflow. In TCP, this has the effect of a congestion event causing the congestion window to halve.

As can be seen from this description achieving the maximum throughput on a LFN requires an amount of work by the administrator of each end host involved, and if multiple links with different characteristics are involved the settings used may be suboptimal. This work can be avoided by creating multiple simultaneous connections however this can lead to other issues including file fragmentation if using multiple streams for a single bulk data move.

The aim of these experiments is to provide an alternative bulk transport mechanism that can be dropped into a running environment and provide high speed transport without requiring significant tuning to achieve maximum performance.

1.1 UDT

UDT[2] is an application level data transport protocol which uses UDP to transfer bulk data and it has its own reliability control and congestion control mechanism. It is not only for private or QoS-enabled links, but also for shared networks since it is TCP friendly.

1.2 GridFTP and Globus XIO

The Globus toolkit[3] provides the data management component GridFTP and a common runtime component XIO. GridFTP is a high-performance, secure, reliable data transfer protocol base on FTP that is optimized for high-bandwidth wide-area networks. Globus XIO is an extensible input/output library written in C for the Globus Toolkit. It provides a single API that supports multiple protocols, with these protocol implementations encapsulated as drivers. XIO

drivers can be written as atomic units and stacked on top of one another. The latest Globus implementation of the GridFTP server implements XIO which allowed the replacement of TCP with UDT as the layer 3 protocol for the purposes of these tests.

2. System Component Testing

The hardware configuration for these tests is two Supermicro X6DHE-G2's with dual Xeon 3.2Ghz dual core CPUs, 2GB ECC DDR RAM, LSI MegaRAID SATA 300-8x RAID controller. The disks are six Western Digital Raptor 74GB SATA disks connected to the RAID controller and a seventh Western Digital 80GB SATA disk as the system Disk. The RAID controller is seated in PCI-X slot 1 so that it is on a separate PCI bus interface to the Gigabit LAN connection and thus not competing for bandwidth on the bus.

2.1 Disk Subsystem

The disk subsystems were tested using IOZone[4] which is a file system benchmark tool. It generates and measures a variety of file operations including write, rewrite, read and reread.

For the results to have the maximum relevance to the production systems used by the GridPP sites we configured the disks for RAID 5 and used an ext3 file system. IOZone tests were performed using the command:

```
iozone –a –g 16G –i 0 –i 1
```

These showed that once the file size exceeded the cache size write speeds in the order of 125MB/s [Figure 1] and read speeds of approximately 155MB/s [Figure 2] are possible, which is just fast enough so as not to be the bottleneck on a 1Gb/s network connection.

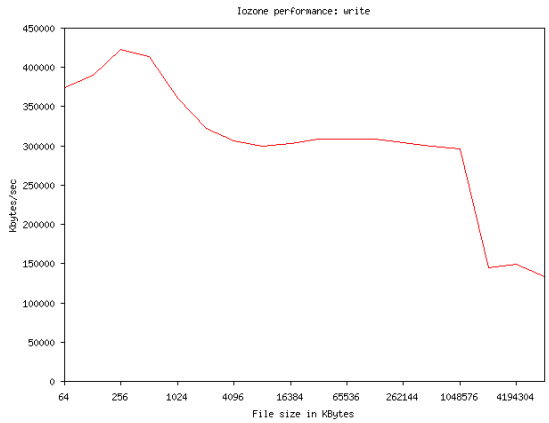


Figure 1

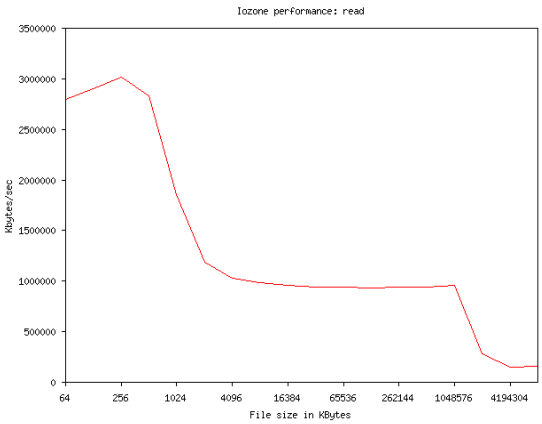


Figure 2

2.2 Networking

The raw network link was tested using Iperf [5] for the TCP and UDP protocols and XIOperf [6] for the UDT protocol. XIOperf is a tool similar to Iperf, and measures the performance characteristics of a transfer and reports them to the user. It is written on top of Globus XIO so it has all of the dynamically loadable transport driver functionality which

allowed the testing of the UDT driver that will be used in GridFTP in later testing. These tests were carried out to profile the UKLight connection and set the baseline for later comparisons.

For the TCP tests tuning was carried out to determine what was required to achieve maximum throughput for a single transfer using only a single stream. Using the calculation for BDP:

```
BDP (bytes) = Bandwidth (bytes) * RTT (seconds)
BDP = 134217728 * 0.01
BDP = 1342177.28
```

Multiple test runs were made using multiples of the BDP to determine the effect on the throughput. TCP transfers were tested both with the buffer sizes as calculated above and then also taking into account the overhead that is reserved for the application.

Figure 3 shows the results of these tests. It can be seen that without any tuning of the kernel UDP, which features no congestion control and is not a reliable protocol, is capable of 957Mb/s which is about 97% of the available bandwidth, UDT is slightly slower at 905Mb/s which is about 92% of its available bandwidth, and finally TCP only manages to achieve 62Mb/s or about 6%. It is only after the buffer sizes have been increased to over 1.5 times the BDP corrected for overheads that TCP reaches its maximum performance of 941Mb/s, almost matching the performance of UDP.

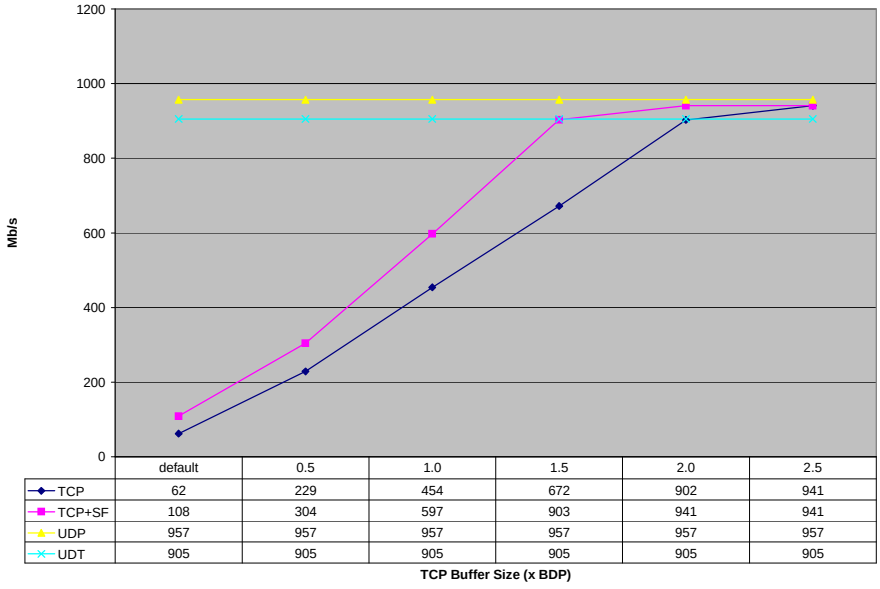
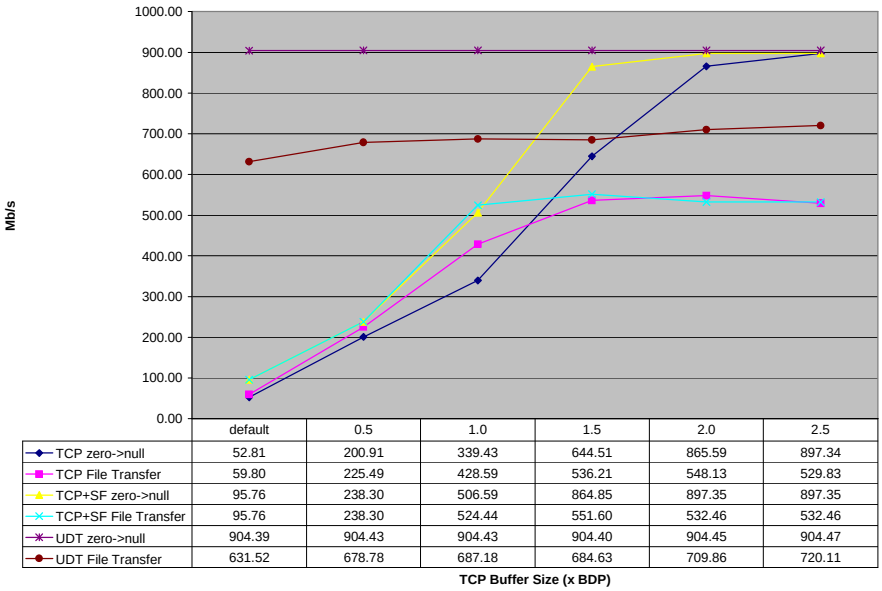


Figure 3

3. Putting it all together

Once the baseline performance of the individual components had been determined tests began on actual file transfers using GridFTP with both UDT and TCP and the layer 3 transport protocol. The first test runs were made using /dev/zero and /dev/null to determine what effect using GridFTP would have on memory to memory transfers, similar to those done using Iperf and Ieper, and they showed that there was a slight drop in throughput when using GridFTP.

Finally transfers were done from file system to file system, again using increasing kernel buffer sizes to maximize the throughput for TCP. Figure 4 show the complete results.



	default	0.5	1.0	1.5	2.0	2.5
TCP zero->null	52.81	200.91	339.43	644.51	865.59	897.34
TCP File Transfer	59.80	225.49	428.59	536.21	548.13	529.83
TCP+SF zero->null	95.76	238.30	506.59	864.85	897.35	897.35
TCP+SF File Transfer	95.76	238.30	524.44	551.60	532.46	532.46
UDT zero->null	904.39	904.43	904.43	904.40	904.45	904.47
UDT File Transfer	631.52	678.78	687.18	684.63	709.86	720.11

Figure 4

By looking at these figures it can be seen that when the network and the file system are being loaded at the same time there is a bottleneck causing a significant slowdown in the throughput of the transfer. Further testing showed that while receiving data from the network and writing to disk was possible at line rate reading data from the disk while sending it to the network is only possible at about 50% - 70% of the available bandwidth. Why this is the case is unknown at present.

References

[1] TCP performance tuning - how to tune Linux http://www.acc.umu.se/~maswan/linux-netperf.txt

[2] UDT http://udt.sourceforge.net/

[3] Globus Toolkit http://www.globus.org/

[4] IOZone http://www.iozone.org/

[5] Iperf http://dast.nlanr.net/Projects/Iperf/

[6] XIOperf http://globus.org/alliance/publications/papers/xioperf.pdf

PROCEEDINGS OF SCIENCE

Trans-Atlantic UDP and TCP network tests

Anthony Rushton*, **Paul Burgess, Richard Hughes-Jones, Stephen Kershaw, Ralph Spencer and Matthew Strong**
The University of Manchester
Jodrell Bank Observatory
Macclesfield
Cheshire SK11 9DL
UK
E-mail: Anthony.Rushton@postgrad.manchester.ac.uk

A VLBI trans-Atlantic connection would greatly extend the resolution of capabilities of eVLBI. So far igrid 2005 and SC 2005 saw the first UKlight connection to the US via Chicago. We report on UDP and TCP network tests performed between Jodrell Bank Observatory, UK, and Haystack Observatory, USA, utilising the UKLight dedicated lightpath, provided by the ESLEA project, to the Starlight connecting node in Chicago. We show near linerate instantaneous UDP throughput over this lightpath, and IPerf TCP bandwidths in excess of 900 Mbps.

Lighting the Blue Touchpaper for UK e-Science - Closing Conference of ESLEA Project
March 26-28 2007
The George Hotel, Edinburgh, UK

*Speaker.

PoS(ESLEA)007

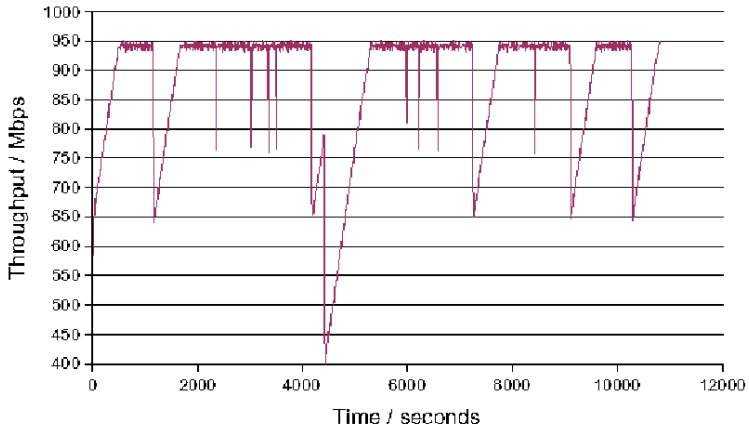


Figure 1: A TCP network bandwidth test between Manchester Computing and a server located in Chicago. The network application tool, Iperf was used to measure the maximum throughput for a period of over 3 hours.

1. Testing the trans-Atlantic link

Very long baseline interferometry (VLBI) generates large rates of data from many telescopes simultaneously observing a source in the sky. This can require the information to traverse intercontinental distances from each telescope to a correlator in order to systhesise high resolution images. With the dramatic development of the Internet and high bandwidth networks, it is becoming possible to transmit the data over large area networks. This allows the correlation of radio astronomical data to be done in real-time, whereas this process would take weeks using conventional disk based recording.

Jodrell Bank Observatory has two 1 Gbps dark fibres from the MERLIN telescope array to the University of Manchester's campus. This local network connects to UKLight at Manchester computing [1] via a Cisco 7600 switch. UKLight is a network of dedicated optical light paths, provided by UKERNA [2]. A guaranteed 1 Gbps bandwidth connects between UKLight and StarLight [3] to a server located in Chicago.

1.1 TCP bandwidth tests with Iperf

TCP throughput rates were measured with the software package, Iperf [4]. Fig. 1 shows the results of a test lasting 3.3 hours, with multiple packet losses observed throughout the test. Despite the network utilising a dedicated lightpath, packet losses are observed and as a result TCP goes into a congestion avoidance phase and reduces the transmitted bandwidth.

1.2 UDP network tests

In order to better understand the network behavior, the UDP protocol was used. Unlike TCP, UDP is not consider a 'fair' protocol and therefore is not widely used on production networks. If a packet is lost in the network, the transmission rate is not reduced, nor is the packet resent. This makes UDP an excellent diagnostic tool for troubleshooting dedicated network paths.

The network analysis software package, UDPmon [5], was used to investigate the link between Manchester and Chicago by transmitting packets at a constant rate, and reporting packet loss; see

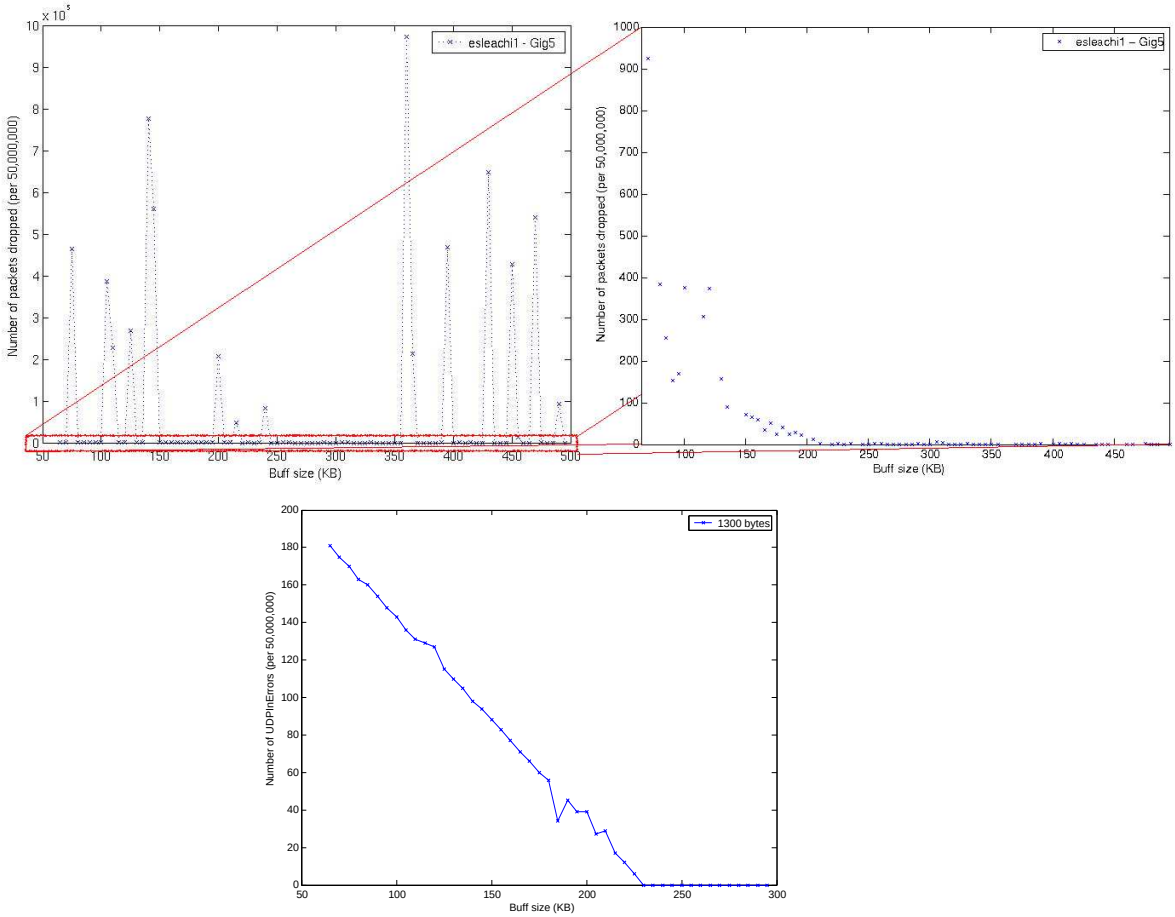


Figure 2: (top left) a) The number of packets lost in a UDPmon test at 940 Mbps from Chicago to Manchester Computing as a function of the receive buffer size. Large intermittent packet losses of >1 % were observed. (top right) b) The same UDPmon data with the packet loss axis limited to 1000 counts. At low receive buffer size constant packet loss is seen. (bottom) c) The number of packets lost between two computers directly connected between network interface cards (i.e. no network). At low receive buffer sizes there is a linear relationship between packets lost per test and buffer size.

results in Fig. 2. In order to emulate the transmission of eVLBI science data, long periods of time were measured, transferring 50 million packets every test. The receive buffer of the application was varied to observe the effects of packet loss.

The application reported large intermittent packet losses, as seen in Fig. 2a. Increasing the receive buffer size at the application layer did not stop this large intermittent packet loss of > 1 %. However varying the buffer does have a small constant effect on the packet loss. Fig. 2b shows during the large transmission of packets, if the receive buffer size in the application is too small then packets will also be lost.

We therefore tested the UDP performance by linking servers directly in a back-to-back test (i.e. without the network). The application transmitted UDP at line rate (940 Mbps). The buffer size of the receive host was varied from 65 KB to 300 KB. Packet loss was observed when the receive buffer was less than $\sim$ 230 KB.

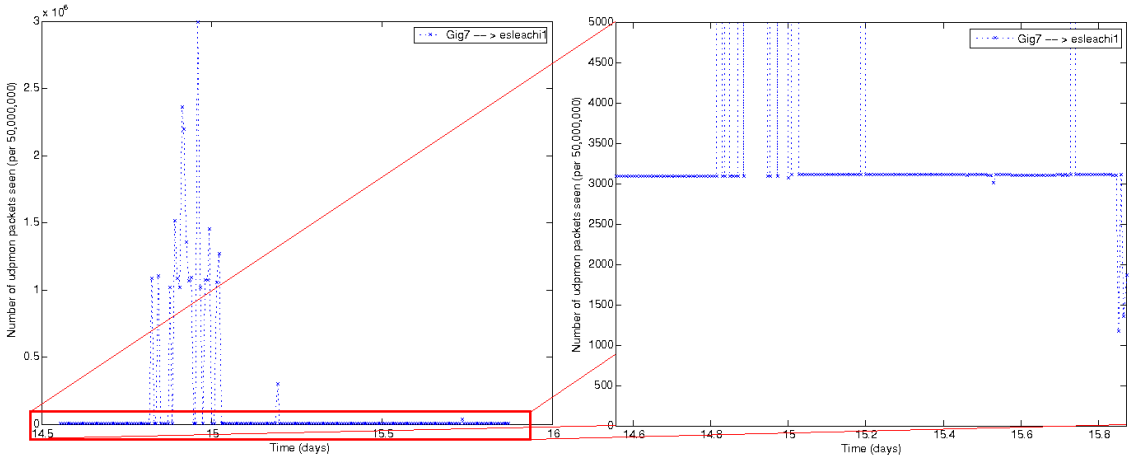


Figure 3: (left) a) The number of packets lost in a UDPmon test at 940 Mbps from Manchester Computing to Chicago as a function of time. Once again large intermittent packet losses of >1 % were observed. (right) b) The same UDPmon data with the packet loss axis limited to 5000 counts. A constant loss of at least 3000 packets per test is observed.

1.3 Constant packet loss when running at line rate

The network tests with UDPmon was repeated in the opposite direction (from Manchester to Chicago) in Fig. 3a. The receive buffer was set to 300 KB and the test was performed for a longer period of time ($\sim$ 30 hours). Once again the application was losing a large fraction of packets (>1%). However this time, as seen Fig. 3b. a constant loss of at least 3000 packets per 50,000,000 sent (0.001%) occurred in every test.

2. Network isolation

In order to have characterised this network link, it was important to examine where the data packets were dropped at the lowest point of the OSI model, i.e. layer 2. To do this we examined the SNMP (simple network management protocol) counters of the network inferace cards and the Cisco 7600 switch connecting to UKLight. The results showed the constant packet loss observed in section 1.3 were within the Cisco 7600. All 50,000,000 packets were received by the switch throughout each test. However if the transmission rate was reduced to 800 Mbps, the switch could transmit all 50,000,000 packets without loss.

We tested the switch's inferace from Jodrell Bank to the Cisco 7600 using a different connection to the computer in the University of Manchester campus. Both the switch's SNMP counts and UDPmon's reports showed the switch transmitted every packet at line rate without packet loss in this configuration.

2.1 UKERNA's maintenance

We concluded through process of elimination that the large intermittent packet loss of > 1% was therefore within UKLight. After private communication with UKERNA it is believed the network issues were due to a broken fiber on the Manchester Computing to StarLight connection. After multiple maintenance tickets were issued by UKLight, we repeated the UDPmon tests.

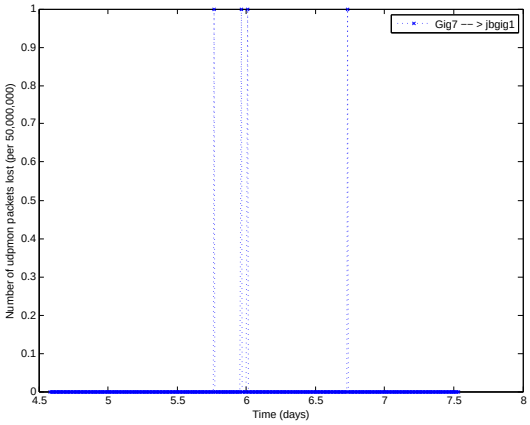


Figure 4: UDPmon tests from Jodrell Bank to Chicago sending UDP packets at 940 Mbps. Over 200 TB of data was transfered over a period of over 60 hours. Only 4 packets were lost throughout the experiment.

The configuration of the local network from Jodrell Bank into UKLight did not give the constant packet loss seen in section 1.3 even at line rate. The results in Fig. 4 show that when running at line rate (940 Mbps), very few packets were lost over (the observed) period of 2.5 days. Over 20 Billion packets were transmitted ($\sim$ 200 TB) with the loss of only four packets in the network.

3. Conclusion

This work demonstrates some of the challenges encounted when using high bandwidth networks. Software application tools have simulated the date rates required by eVLBI science by continually sending large numbers of packets for many hours. This has shown the need for the receive buffers of the applications to be capable enough to collect data at these rates for long periods of time.

Issues have arisen with the Cisco 7600 switch showing, that under certain circumstances the instrument does not perform to the manufacturers specifications. This highlights the requirement to identify and test equipment to maximum abilities. Problems with the link were isolated by eliminating the client, end-host servers and local network by inspecting level 2 SNMP packet counts. This led us to confidently conclude that large packet losses were within UKERNA's UKLight network. After maintenance on UKLight, our tests were repeated for a large time period ($\sim$ 3 days). This successfully showed it was possible to transmit packets between Manchester Computing and Chicago at 940 Mbps without losing a significant number of packets.

References

[1] Manchester Computing, http://www.mc.manchester.ac.uk

[2] United Kingdom Education and Research Networking Association, http://www.ukerna.ac.uk

[3] StarLight, http://www.startap.net/starlight

[4] Iperf home page, http://dast.nlanr.net/projects/Iperf

[5] UDPmon home page, http://www.hep.man.ac.uk/u/rich/net

PROCEEDINGS OF SCIENCE

Performance testing of SRM and FTS between Lightpath Connected Storage Elements

Brian GE Davies[1]
Lancaster University
Physics Department, Lancaster University, Lancaster, Lancashire, UK
E-mail: b.g.davies@lancaster.ac.uk

Roger WL Jones
Lancaster University
Physics Department, Lancaster University, Lancaster, Lancashire, UK
E-mail: Roger.Jones@cern.ch

We describe the configuration, testing and optimisation of file transfers with LCG middleware between the SRM storage systems at two LCG sites using a UKLight connection. We will also discuss recommendations for continued work.

Lighting the Blue Touchpaper for UK e-Science - Closing Conference of ESLEA Project
The George Hotel, Edinburgh, UK
26-28 March, 200

[1] Speaker

PoS(ESLEA)008

1. Introduction

ATLAS[1] is one of 4 large High Energy Physics (HEP) experiments physically based at CERN, Switzerland which will produce tens of PetaBytes of data each year. Since it is impractical to host and process this and associated Monte-Carlo data at a single site, high bandwidth data transfers between the hundreds of LHC Computing Grid[2] (LCG) sites around the world are needed. Within ESLEA[3], the ATLAS exploitation group, with help from the UK GRIDPP[4] community and in coordination with LCG service challenges aimed to establish the ability of the LCG middleware and hardware implementations to transfer large data volumes using the UKLight dedicated lightpath network.

ESLEA used in part UKLight to connect the LCG Tier1 centre at the Rutherford Appleton Laboratory (RAL) to Lancaster, which is a part of a distributed Tier2 (NORTHGRID) within the UK. This tier to tier connection crosses Regional Networks (RNs). The challenge of crossing these boundaries and the need for access to both production and research networks requires good communication between end-site system administrators, the regional network operators and the national network organisations.

2. Configuration

In order to optimise and analyse the use of the available bandwidth, an effective network, hardware and software configuration is needed. Configuration design should minimise obvious bottlenecks in performance. Since neither ESLEA nor the LCG are sole users of the RAL services, ESLEA worked within the RAL production framework to reduce interventions and carried out optimisations at Lancaster, where it has greater control and flexibility of hardware, software and network solutions.

2.1 Network configuration

Many factors affect the useful bandwidth on a production network. These include variable third party usage and bandwidth limitation, increased packet loss and jitter caused by multi-hop and variable routing and variable congestion of links. Dataset size and parallelisation of data streams were investigated. A private point to point link permitted complete control of the number of data flows and connections were allowed and so increased the ability to monitor rates between end-sites.

Dual homing of both hardware and software was initially considered. As this solution was not tested at the time, a network solution was used. By organising static routing and local network configuration, it was possible to allow both traffic flows across the dedicated lightpath for data transport whilst also allowing communication between internal and external services over the production network. One consequence of static routing is the need for all end-hosts' routing tables to be correctly configured and confirmed to ensure appropriate routing. The network configuration allowed an increased available nominal bandwidth from an intentional 100Mbps bottleneck (to avoid non-LCG site network congestion) over the production network

to 1Gbps over lightpath. We were also able to bypass a 400Mbps firewall. The lightpath also reduced router hops to two from twelve which leads to less potential for lost/reordered packets and associated network performance effects. With a round trip time (RTT) of 6.5ms and nominal bandwidth of 1Gbps, accepted standard TCP tuning techniques such as TCP window and memory buffer size optimisation of the native version of the linux-2.4 kernel to improve line usage were applied[5].

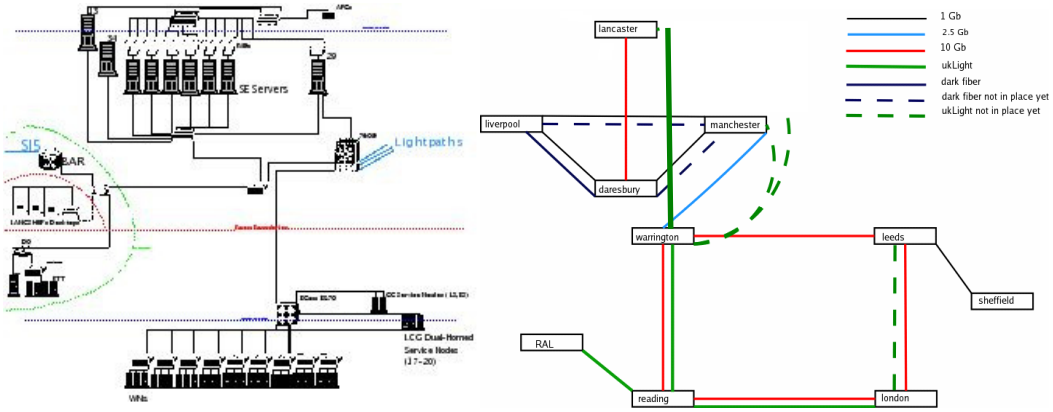


Figure 1- The Local Lancaster Network topology and the network topology connecting NORTHGRID Tier2 sites to RAL Tier1

2.2 Hardware/Software configuration

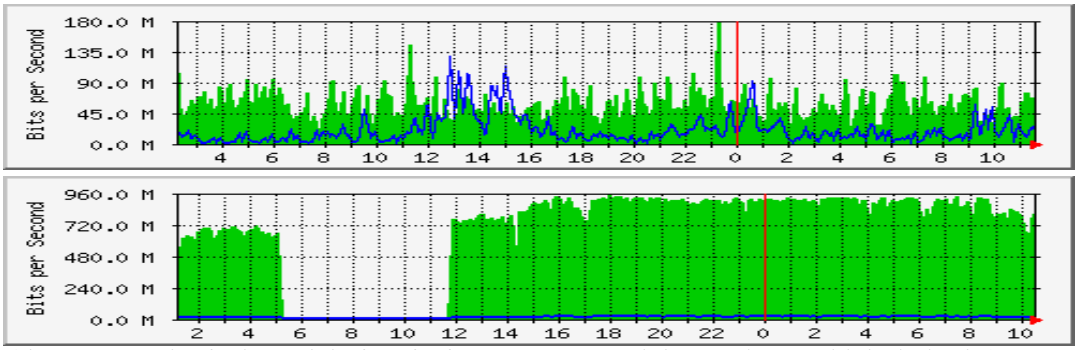
LCG file transfers use a storage resource manager (SRM) interface as a front end to an extended disk system. At Lancaster, we deployed a dCache[6] storage element (SE) installation onto a system consisting of a head node and six I/O servers, each with two 6TB RAID5 arrays. This allowed us to test both single and parallel concurrent transfers. CASTOR[7] (an alternative SE system) and dCache were both deployed and tested as the RAL end-point system. These systems had various numbers of file servers assigned to each endpoint throughout the testing procedure. In addition to an SE, several other LCG services and clients were installed to progress towards full distributed data management. Of particular importance were the File Transfer Service (FTS) and user interface (UI). The FTS, in conjunction with a UI allows file transfers from both disk-SRM and SRM-SRM. The SRM copies themselves are controlled by the dCache srmcp command. Transfers were initiated and controlled by two methods. The first method used a BASH command line script to initiate copies and deletions of files using loops and system sleeps. FTS managed transfers were controlled using the filetransfer.py script supplied by GRIDPP storage group. This incorporates another level of complexity of the software stack, as it requires extra communication with external servers leading to additional overhead. The FTS uses “channel management” to control gsi secure file transfers. This adds the ability to manipulate the number of concurrent streams and files transferred between two LCG sites, whilst channel status control enables complete transfer initialisation, cessation and retries. However, the FTS also increases the communication overhead of the transfer compared to a single SRM initiated command which in turn has its own overhead. The overhead from BASH onto srmcp is smaller than FTS and the filetransfer.py script. However long term functionality that FTS provides is needed for long term experimental use and so cannot be ignored.

2.3 Monitoring

Monitoring of rates, file storage and system diagnostics were achieved with various tools. MRTG and similar RRD tools were useful for both instantaneous rates and recording historical data of network traffic flows. Files copied using BASH scripts were checked with the commands ls and du. Python controlled FTS transfers were also capable of giving timing and failure rates. Both storage completions and rates werer cross-checked with Ganglia monitoring of SRM servers and SNMP walk information of routers.

3 RAL dCache to Lancaster transfers

For a single 1GB file transfer using srmcp an instantaneous rate of 330Mbps was achieved. However when incorporating the srmcp communication overhead, this rate dropped to an aggregate sustained rate of 195Mbps. Parallelisation of files transfers using a BASH script allowed a sustained rate of near line speed of over 900 Mbps with a peak rate of 946Mbps. This was accomplished with 20 concurrent file transfers from RAL to Lancaster. This rate also produced a back traffic rate of 18Mbps (2% of forward flow) which is presumed to be a summation of ACK packets inter-gridFTP door communication. Staggering of transfer initialisation also improved data rates by avoiding the concurrent dead time caused by concurrent initialisation/cessation of individual transfers. Further evidence of the effect of transfer dead time comes from studying the effect of file size on aggregate rates. The rate for a single file test between two particular servers increased from 150 to 180 Mbps with an increase from 1 to 10 GB file size. Sustained rates of 800 Mbps for 24 hours (Figure 2) and over 500 Mbps aggregated for a one week period were obtained. This corresponded to 8TBytes and 36TBytes of data transferred. Figure 2 also shows typical current production network rate usage.



(Figure 2- MRTG plot of 32 hour periods of transfers from RAL to Lancaster during normal usage and during load testing).

The drop from 800 to 500 Mbps between 24 hour and the week test was caused by authentication errors due to the user's grid certificate proxy expiring. Fail-over to the production network, caused by lightpath downtime, was successful in that no manual intervention was needed. The 100Mbps bottleneck imposed on the system led to full congestion of the production link with only a few concurrent transfers. This led to communication and timeout errors between FTS, UI and SRM services leading to dramatic drop in successful transfers. FTS controlled transfers for a single-stream, single-file transfer gave a rate similar to that of manual srmcp transfers (150Mbps). However, competition with production traffic using the FTS

channel led to an uncertain and unstable number of concurrent test files being transferred with FTS. Additional FTS server load from other experiments and end-sites caused lower transfer rates than from BASH script controlled transfers.

The modes and argument values of the dCache srmcp command and its effects needs to be studied. Variations in the direction of transfer and end-host initiation may explain directional rate variance. This observed change in rate may be an effect of the passive/active nature of the GridFTP or an effect of multiple/single stream transfers. It may also be a result of different SRM setups (such as pool balancing and file location.); or could also be an effect of disk I/O limitations of particular file-systems involved in the transfers.

3.1 RAL Castor to Lancaster Tests

Tests of the Castor system at RAL to Lancaster were successful but no extensive data loading rates are currently available. This confirms that the lightpath network topologies can allow multiple SEs to function at a single site. Initial rates obtained gave 600Mbps into Lancaster and 400Mbps out of Lancaster for single direction transfers. Rates of 200Mbps (in) and 300Mbps (out) for bi-directional tests with similar parallelisation were achieved. Initial tests of failure rates give a figure of 51 failed transfers from 851 1GB files transferred in a 12 hour period.

4 Future Plans

Following completion of the ESLEA project, we plan to continue testing of the CASTOR system at RAL. We intend to continue file transfers to the Netherlands over UKLight, connecting to a dCache system hosted by the LCG site at SARA. We hope to implement the UDT transport protocol into the LCG's Disk Pool Manager. The effects on data transport rates of additional LCG and ATLAS services, such as file catalogues and the ATLAS Distributed Data Manager software system (DDM) needs to studied. This work will be within the UK GRIDPP community and within LCG Service Challenge 4. An analysis of the effect of optimising the operating system (with particular focus on kernel version and automated TCP/IP tuning) might be studied in conjunction with the planned upgrade of the LINUX kernel version to 2.6.

References

[1] ATLAS homepage: http://atlas.web.cern.ch/Atlas/index.html

[2] LCG homepage: http://lcg.web.cern.ch

[3] ESLEA homepage: http://www.eslea.uklight.ac.uk

[4] GRIDPP homepage: http://www.gridpp.rl.ac.uk

[5] TCP-tuning: http://dsd.lbl.gov/TCP_Tuning

[6] dCache homepage: http://www.dcache.org

[7] CASTOR homepage: http://www.castor.org

PROCEEDINGS OF SCIENCE

Working with 10 Gigabit Ethernet

Richard Hughes-Jones [1],
The School of Physics and Astronomy,
The University of Manchester,
Manchester,
M13 9PL
UK
E-mail: R.Hughes-Jones@manchester.ac.uk

Stephen Kershaw
The School of Physics and Astronomy,
The University of Manchester,
Manchester,
M13 9PL
UK
E-mail: stephen.kershaw@manchester.ac.uk

Network technology is always moving forward and with the recent availability of 10 Gigabit Ethernet (10GE) hardware we have a standard technology that can compete in terms of speed with core or backbone Internet connections. This technology can help deliver high-speed data to the end-user but will systems that are currently used with Gigabit Ethernet deliver with 10GE?

We investigate the performance of 10 Gigabit Ethernet network interface cards in modern server quality PC systems. We report on the latency, jitter and achievable throughput and comment on the performance of transport protocols at this higher speed.

Lighting the Blue Touchpaper for UK e-Science - Closing Conference of ESLEA Project
The George Hotel, Edinburgh, UK
26-28 March, 2007

[1] Speaker

PoS(ESLEA)009

1. Introduction

Current developments in Radio Astronomy being researched by the EXPReS [1] project will require multi-gigabit flows across Europe using the National Research Networks interconnected by GÉANT. This paper reports on detailed measurements to determine of the performance and behaviour of 10 Gigabit Ethernet NICs when used in server quality PCs.

2. Methodology

The methodology follows that described in [2]; measurements were made using two PCs with the NICs directly connected together with suitable fibre or copper CX4 cables. UDP/IP frames were chosen for the tests as they are processed in a similar manner to TCP/IP frames, but are not subject to the flow control and congestion avoidance algorithms defined in the TCP protocol and thus do not distort the base-level performance. The packet lengths given are those of the user payload[1].

2.1 Latency

To measure the round trip latency, UDPmon [3] was used on one system to send a UDP packet requesting that a response of the required length be sent back by the remote end. Each test involved measuring many (~1M) request-response singletons. The individual request-response times were measured by using the CPU cycle counter on the Pentium [3] and the minimum, average and maximum times were computed. For all the latency measurements the interrupt coalescence of the network interface cards (NICs) was turned off. The measurements thus provide a clear indication of the behaviour of the host, NIC and the network.

Transfer Element	**Inverse data transfer rate µs/byte**	**Expected slope µs/byte**
Memory access	0.00004	
8 lanePCI-Express	0.000054	
10 Gigabit Ethernet	0.0008	
Memory, PCI-Express & 10 Gigabit		0.00268

Figure 1. Table of the slopes expected for PCI-Express and 10 Gigabit Ethernet transfers.

The latency was plotted as a function of the frame size of the response. The slope of this graph is given by the sum of the inverse data transfer rates for each step of the end-to-end path [2]. Figure 1 shows a table giving the slopes expected for PCI-Express and 10 Gigabit Ethernet transfers. The intercept gives the sum of the propagation delays in the hardware components and the end system processing times. Histograms were also made of the singleton request-response measurements. These histograms show any variations in the round-trip latencies, some of which may be caused by other activity in the PCs.

[1] Allowing for 20 bytes of IP and 8 bytes of UDP headers, the maximum user payload for an Ethernet interface with a 1500 byte Maximum Transfer Unit (MTU) would be 1472 bytes.

2.2 UDP Throughput

The UDPmon tool was used to transmit streams of UDP packets at regular, carefully controlled intervals and the throughput and packet dynamics were measured at the receiver. On an unloaded network, UDPmon will estimate the capacity of the link with the smallest bandwidth on the path between the two end systems. On a loaded network, the tool gives an estimate of the available bandwidth. These bandwidths are indicated by the flat portions of the curves.

In these tests, a series of user payloads from 1000 to 8972 bytes were selected and for each packet size, the frame transmit spacing was varied. For each point, the following information was recorded:

- the throughput;
- the time to send and the time to receive the frames;
- the number of packets received, the number lost, and the number out of order;
- the distribution of the lost packets;
- the received inter-packet spacing;
- the CPU load and the number of interrupts for both transmitting and receiving systems.

The "wire"[2] throughput rates include an extra 66 bytes of overhead and were plotted as a function of the frame transmit spacing. On the right hand side of the plots, the curves show a 1/t behaviour, where the delay between sending successive packets is the most important factor. When the frame transmit spacing is such that the data rate would be greater than the available bandwidth, one would expect the curves to be flat (often observed to be the case).

2.3 TCP Throughput

The Web100 [5] patch to the Linux 2.6.20 kernel was used to instrument the TCP stack allowing investigation of the behaviour of the TCP protocol when operating on a 10 Gigabit link. Plots of the throughput, TCP Congestion window (Cwnd), the number of duplicate acknowledgements (DupACK) and the number of packets re-transmitted were made as a function of time though the flow. A further patch to the kernel allowed incoming TCP packets to be deliberately dropped.

3. Hardware

The Supermicro [6] X7DBE motherboard was used for most of the tests. It was configured with two dual-core 2 GHz Xeon 5130 Woodcrest processors, 4 banks of 530 MHz FD memory and has three 8-lane PCI-Express buses connected via the Intel 5000P MCH north bridge, as shown in the left hand block diagram of Figure 2. Each processor is connected by an independent 1.33GHz front side bus. For one of the tests, a Supermicro X6DHE-G2 motherboard was used at one end of the link. This had two 3.2 GHz Xeon CPUs with a shared

[2] The 66 "wire" overhead bytes include: 12 bytes for inter-packet gap, 8 bytes for the preamble and Start Frame Delimiter, 18 bytes for Ethernet frame header and CRC and 28 bytes of IP and UDP headers

800 MHz front side bus to the Intel 7520 chipset. It has two 8-lane PCI-Express buses, as indicated in the right hand block diagram of Figure 2.

Myricom [7] 10 Gigabit Ethernet NICs were used for all of the tests. These are 8-lane PCI-Express devices and both fibre and copper CX4 versions were tested. Version 1.2.0 of the Myricom myri10ge driver and version 1.4.10 of the firmware was used in all the tests. Also check summing was performed on the NIC and Message Signalled Interrupts were in use for all of the tests.

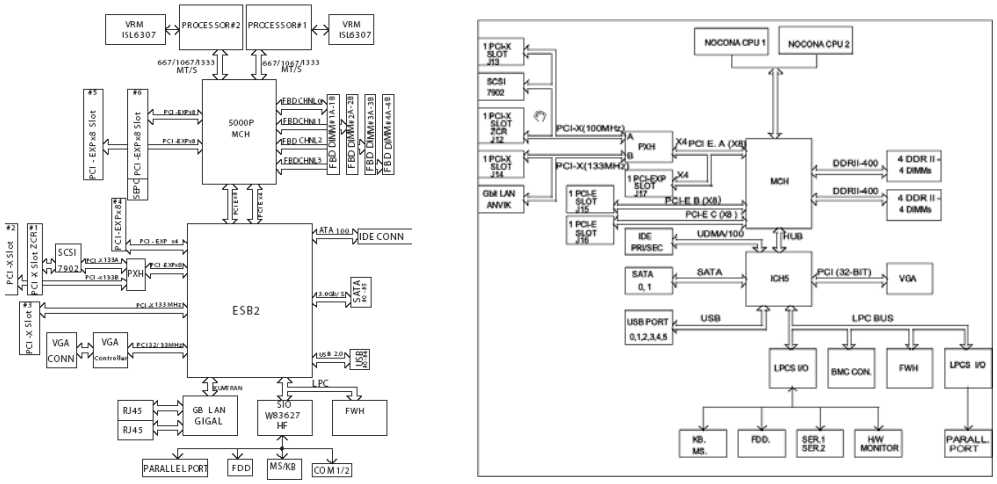


Figure 2. Block diagrams of the Supermicro motherboards used in the tests.
Left: X7DBE dual-core Xeon motherboard Right: X6DHE-G2 motherboard.

4. Measurements made with the Supermicro X7DBE Motherboard

4.1 Latency

Figure 3 shows that variation of round trip latency with the packet size is a smooth linear function as expected, indicating that the driver-NIC buffer management works well. The clear step increase in latency at 9000 bytes is due to the need to send a second partially filled packet. The observed slope of 0.0028 µs/byte is in good agreement with the 0.00268 µs/byte given in Figure 1. The intercept of 21.9 µs is reasonable given the NIC interrupts the CPU for each packet received.

Figure 3 also shows histograms of the round trip times for various packet sizes. There is no variation with packet size, all having a FWHM of ~1 µs and no significant tail.

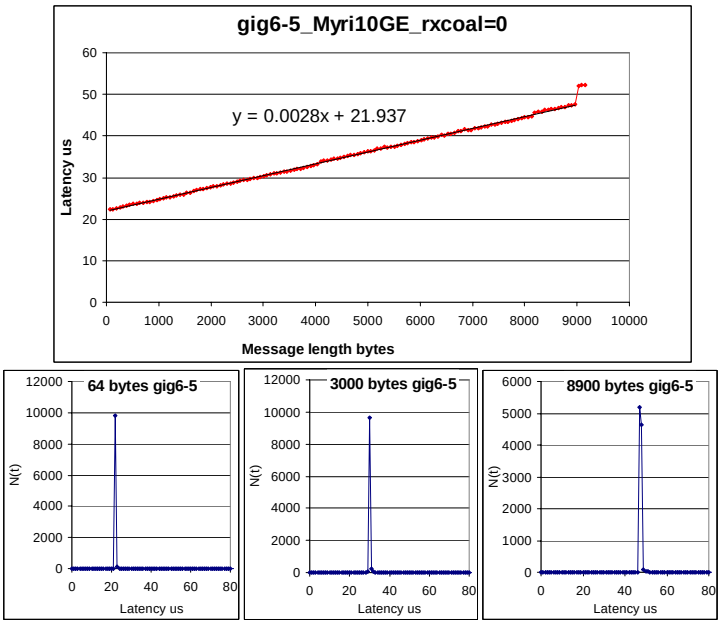


Figure 3. Top: The UDP Request-Response latency as a function of packet size. Bottom: Histograms of the latency for 64, 300 and 8900 byte packet sizes.

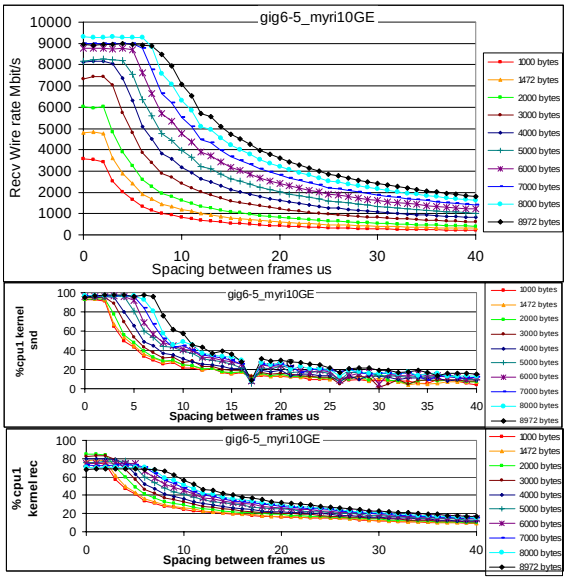


Figure 4. Top: UDP throughput as a function of inter-packet spacing for various packet sizes. Middle: Percentage of time the sending CPU was in kernel mode. Bottom: Percentage of time the receiving CPU was in kernel mode.

4.2 UDP Throughput

For these measurements the interrupt coalescence was set to the default value of 25 µs. Figure 4 shows that the NICs and host systems performed very well at multi-gigabit speeds, giving a maximum throughput of 9.3 Gbit/s for back to back 8000 byte packets. For streams of 10 M packets, about 0.002% packet loss was observed in the receiving host. For packets with

spacing of less than 8 µs, one of the four CPU cores was in kernel mode over 90% of the time, and the other three CPUs were idle. Similarly for the receiving node, where one of the four CPU cores was in kernel mode 70-80% of the time and the other three CPUs were idle. As the packet size was reduced, processing and PCI-Express transfer overheads become more important and this decreases the achievable data transfer rate.

It was noted that the throughput for 8970 byte packets was less than that of 8000 byte packets, so the UDP achievable throughput was measured as a function of the packet size. The results are shown in Figure 5.

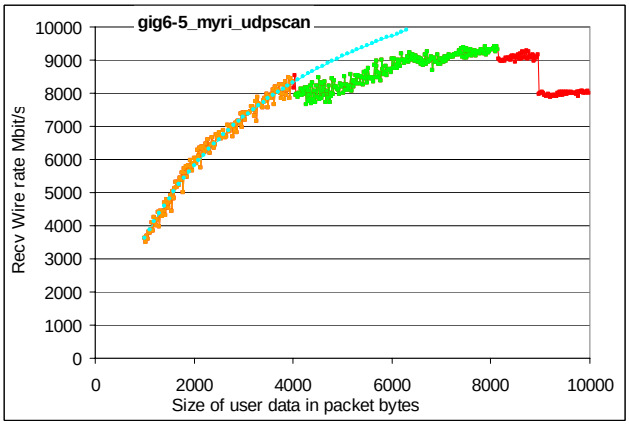


Figure 5. Measurement of UDP throughput as a function of packet size.

5. Measurements made with the SuperMicro X6DHE-G2 Motherboard

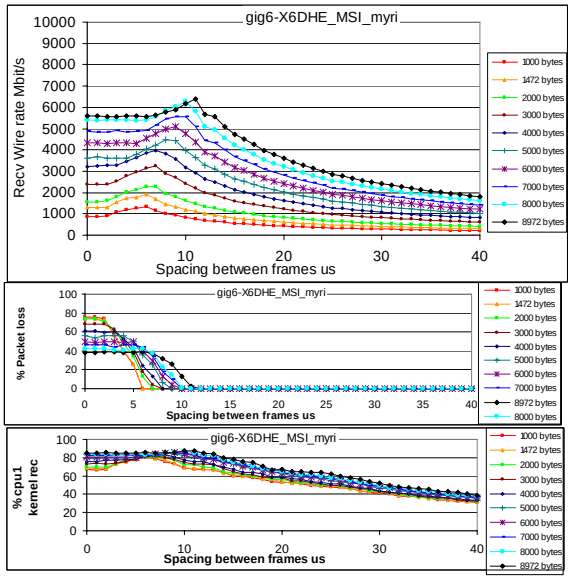


Figure 6. UDP throughput as a function of inter-packet spacing for various packet sizes using the Supermicro XDHE-G2 motherboard.
Middle: Percentage of time the sending CPU was in kernel mode.
Bottom: Percentage of time the receiving CPU was in kernel mode.

Figure 6 shows the achievable UDP throughput and CPU usage when packets are sent from a Myricom NIC in a Supermicro X7DBE motherboard to one in a X6DHE-G2 motherboard. The maximum throughput is only 6.5 Gbit/s with no clear plateau indicating a simple bottleneck. Given the lower CPU usage for the sending CPU than that shown in Figure 4, it is possible that limitations in moving received packets from the NIC to the memory result in queues building up in the receiving NIC, which then sends Ethernet pause packets to the sender. Clearly not all motherboard and chipsets provide the same input-output performance.

6. Protocol Performance

6.1 TCP flows

Plots of the parameters taken from the web100 interface to the TCP stack [2] are shown in Figure 7 for a memory to memory TCP flow generated by iperf and demonstrate the congestion avoidance behaviour in response to lost packets. This TCP flow was set up between two systems connected back-to-back and used a TCP buffer size of 256 kbytes, which is just smaller than the bandwidth-delay product, BDP, of 300 kbytes. The packets were deliberately dropped using a kernel patch in the receiving host. The upper plot in Figure 7 shows Cwnd decreasing by half when a lost packet is detected by the reception of multiple duplicate acknowledgments (DupACKs), shown in the second plot. The third plot shows that one packet is re-transmitted for each packet dropped, while the bottom plot indicates that there is not much reduction in the achievable TCP throughput. This is due to the short round trip time when the systems are connected back-to-back.

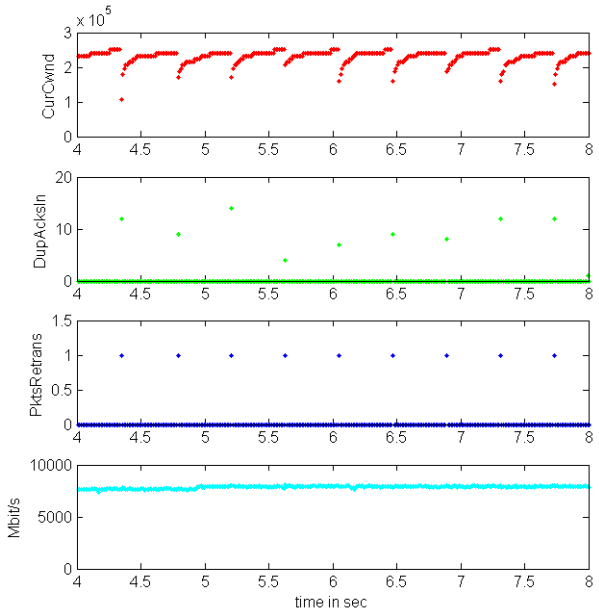


Figure 7. Parameters from the Reno TCP stack recorded by Web100 for an iperf flow. Packets were dropped in the receiving kernel.

6.2 UDP flows with Concurrent Memory Access

As discussed in section 4.2 and shown in Figure 4, a 9.3 Gbit/s UDP flow uses one of the four CPU cores in kernel mode 70-80% of the time, but the other three were unused. Tests were made to determine if the other three CPUs could be used for computation at the same time as sustaining a multi-gigabit flow. Figure 8 shows measurement of the achieved UDP throughput and packet loss under three conditions for a series of trials. On the left are the results for just a UDP flow, in the centre a process that continually accesses memory was run on the second core of the CPU chip processing the networking, and on the right this process was run on the second CPU. In both cases when the load process was run there was a reduction of ~200 Mbit/s in the throughput and about 1.5% packet loss. Figure 9 shows the percentage of time the four CPUs were in different modes when the memory load process was run on CPU3. These results demonstrate that useful work can be done in the end host with minimal effect on the network flow.

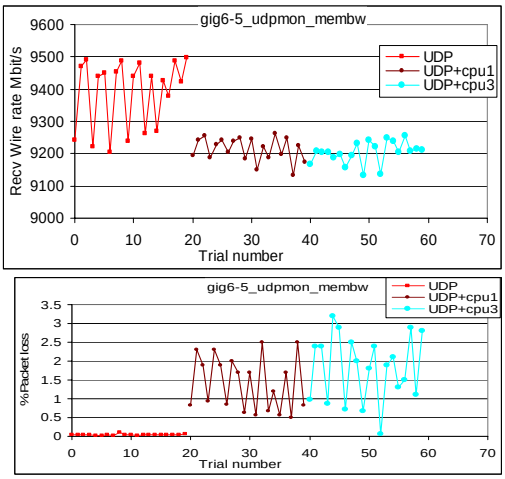


Figure 8. The UDP thoughput and packet loss for a network flow only and when a CPU-Memory process is run on another CPU.

```
Cpu0  :  6.0% us, 74.7% sy,  0.0% ni,   0.3% id,  0.0% wa,  1.3% hi, 17.7% si,  0.0% st
Cpu1  :  0.0% us,  0.0% sy,  0.0% ni, 100.0% id,  0.0% wa,  0.0% hi,  0.0% si,  0.0% st
Cpu2  :  0.0% us,  0.0% sy,  0.0% ni, 100.0% id,  0.0% wa,  0.0% hi,  0.0% si,  0.0% st
Cpu3  : 100.0% us, 0.0% sy,  0.0% ni,   0.0% id,  0.0% wa,  0.0% hi,  0.0% si,  0.0% st
```

Figure 9. The percentage of time the four CPUs were in different modes when the memory load process was run on CPU3.

7. Conclusions

This work has demonstrated that the Myricom 10 Gigabit Ethernet NICs can deliver UDP flows of 9.3 Gbit/s and TCP flows of 7.77 Gbit/s when using PCs using the Supermicro X7DBE. However not all motherboard and chipsets provide the same input-output performance.

Even though one of the CPU cores is occupied in driving the network, these results also show that useful work can be done in the other CPUs. We conclude that these Myricom- X7DBE will be suitable to evaluate the performance of 4 Gigabit UDP flows over lightpaths provisioned over the GÉANT2 network.

References

[1] EXPReS, Express Production Real-time e-VLBI Service Three year project, started March 2006, funded by the European Commission (DG-INFSO), Sixth Framework Programme, Contract #026642 www.express-eu.org .
[2] R. Hughes-Jones, P. Clarke, S. Dallison, "Performance of 1 and 10 Gigabit Ethernet Cards with Server Quality Motherboards," Future Generation Computer Systems Special issue, 2004
[3] UDPmon: a Tool for Investigating Network Performance, http://www.hep.man.ac.uk/~rich/net
[4] R. Hughes-Jones and F. Saka, *Investigation of the Performance of 100Mbit and Gigabit Ethernet Components Using Raw Ethernet Frames*, Technical Report ATL-COM-DAQ-2000-014, Mar 2000. http://www.hep.man.ac.uk/~rich/atlas/atlas_net_note_draft5.pdf
[5] web100 interface to the TCP stack Web100 Project home page, http://www.web100.org/
[6] SuperMicro motherboard reference material, http://www.supermicro.com/products/motherboard/matrix/
[7] Myricom home page http://www.myri.com/

PROCEEDINGS OF SCIENCE

Application-based Network Performance Profiling

Robin Pinning*
University of Manchester
E-mail: pinning@manchester.ac.uk

The large scale modelling of many physical phenomena increasingly requires the model to be of a size that is too large for one HPC resource. MPICH-G2 is a grid-enabled MPI implementation that allows the coupling of multiple machines for the running of a single MPI-based application. This work aims to show how the use of light-switched optical networks, such as UKLight, affect codes of this class by running a series of benchmarks using the Intel MPI benchmarking suite

Lighting the Blue Touchpaper for UK e-Science - Closing Conference of ESLEA Project
March 26-28 2007
The George Hotel, Edinburgh, UK

*Speaker.

PoS(ESLEA)011

1. Introduction

Traditionally networking, particularly for academic scientific researchers, has been over best effort TCP/IP packet switched networks such as SuperJANET. UKLight is an optical network comprising of a 10Gbit/s backbone connecting participating academic institutions in the UK and connecting to global optical networks such as Starlight and NetherLight.

The ESLEA [1] project aims to demonstrate the potential of circuit-switched optical networks by allowing the exploitation of the UKLight infrastructure for a range of scientific application-led projects. One of these sub-projects is RealityGrid.

The RealityGrid [2]infrastructure provides the computational scientist with a framework for computational steering and on-line visualization. The use of these interactive techniques provides some unique demands on the networking infrastructure, requiring both high bandwidth (TeraGyroid experiment [3]) and high QoS (SPICE [4]) depending on the requirement of the scientific application being used.

Recently the project has extended the scientific applications used by researchers to make use of meta-computing via MPICH-G2 middleware [5]. To overcome the inherent bottleneck of including a relatively slow, compared with the internal HPC network, wide-area network the scientific application needs both high bandwidth and high QoS. Along with careful porting of the application optical networks are essential. One such approach that has been successfully demonstrated at SC2005 across a trans-Atlantic link which included UKLight, is that taken by the Vortonics project [6]. For this application, another lattice-Boltzmann code with similarities to the LB3D code used by RealityGrid researchers, the memory requirements are typically larger than a single computational resource can provide. The code is parallelised with MPI with data distributed accross the available processors according to a scheme called geographically distributed domain decomposition or GD^3 [7].

Given the importance of the use of grid-enabled application codes using MPICH-G2 this paper presents performance figures gathered using the Intel MPI benchmarking suite run on the UKLight network. A brief analysis, some conclusions and some suggestions for future work are also presented.

2. Methodology

2.1 Equipment

The experiments involved three linux workstations situated in London and Edinburgh. The machines were connected by a dedicated UKLight link provisioned at 300Mbps. Two of the machines were configured as compute nodes with a third, running a NIST Net-instrumented linux kernel, acting as a traffic shaper connected in-between the two compute nodes. One compute node and the NIST Net box were situated in Edinburgh, connected via gigabit ethernet. The NIST Net box was then connected via the UKLight router and link to the machine at UCL.

2.2 Software

NIST Net [8] version 2.0.12 was run as a kernel module. All kernels were version 2.6.9x and all OS level TCP-stack parameters were left as standard. When using HPC machines there is

usually no way for the user to alter these parameters therefore all tuning was done at the user-level. The network parameters altered in the NIST Net module were chosen so as to emulate conditions seen in packet-switched 'production' networks were α, the packet transmission delay, to emulate different network latencies; β, the packet loss, to emulate loss due to congestion and $\delta(\alpha)$, the jitter, to emulate variability in packet delivery.

The middleware stack on each machine consisted of Globus Toolkit v4.0.3 (pre-WS components) and MPICH-G2 v1.2.6. The Intel MPI Benchmark software version 3.0 was built against MPICH-G2 with custom parameters to allow for large data sizes (to emulate the size of data that application codes such as LB3D pass around). Due to the use of only two processors only the Ping-Pong test was used. The parameter "MPICH_GLOBUS2_TCP_BUFFER_SIZE" was set, based on instructions from Brian Toonen (MPICH-G2 developer), to 524288 in the RSL used to launch each run. This was based on a measured RTT of 5ms.

3. Analysis

The data for the added latency can be seen in table 1 and plotted in figure 1. As the latency, α, is increased the time for the transfer increases. This is unsurprising as the performance of MPI-based codes is very sensitive to latency in the connection between processes. As the packet loss rate, β, is increased the transfer time also increases, again an expected result.

Results of increasing jitter, $\delta(\alpha)$, can be seen in table 2 and are plotted in figure 2. Overall the results present few surprises other than the slower than expected transfers for 16MB transfers when no jitter is present at high latency.

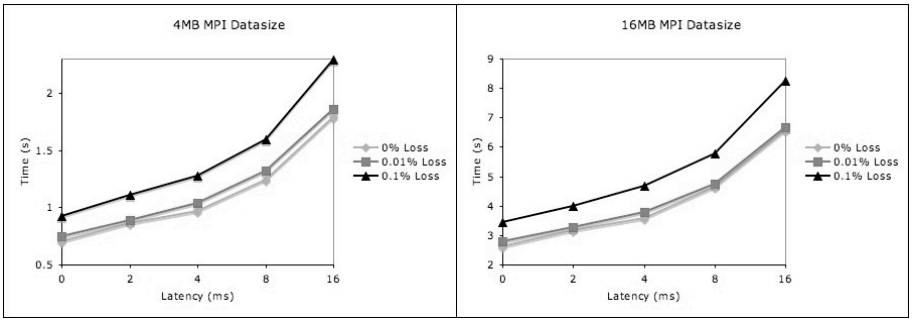


Figure 1: Plots showing transfer time in seconds for a two MPI message sizes for varying values of packet delay, α, and packet loss, β.

4. Summary

The data collected, due to time constraints, is a small snapshot of what could be achieved using these methods. In fact the difficulty faced in setting up, and maintaining, the UKLight link used illustrates how difficult it still is for application scientists to utilise these links. Especially when complex software stacks are sat on top of them, as debugging problems becomes very difficult. It is also clear that if these tests were to be repeated using HPC-class machines that some form of advance reservation and co-allocation [9] of those resources is essential for this kind of work.

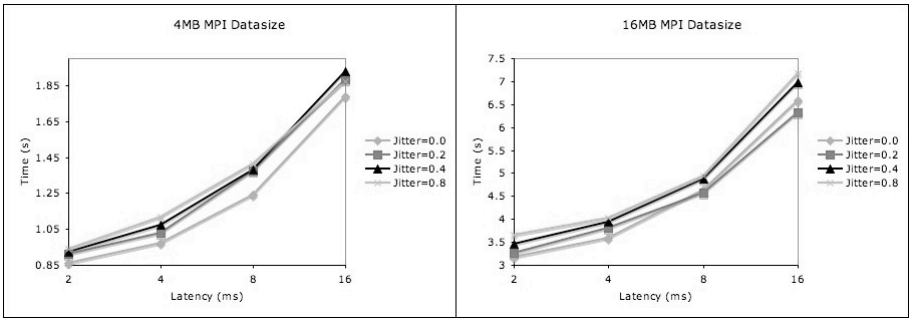


Figure 2: Plot showing transfer time in seconds for two MPI message sizes for varying values of packet delay, α, and jitter, $\delta(\alpha)$.

	Latency α (*ms*)				
Packet Loss β %	0.0	2.0	4.0	8.0	16.0
4*MB MPI Datasize*					
0.0	0.703	0.860	0.971	1.242	1.789
0.01	0.743	0.889	1.043	1.318	1.863
0.1	0.927	1.106	1.279	1.592	2.288
1.0	1.997	2.415	2.799	3.489	4.956
16*MB MPI Datasize*					
0.0	2.618	3.182	3.590	4.630	6.576
0.01	2.777	3.260	3.777	4.760	6.682
0.1	3.449	4.002	4.691	5.774	8.229
1.0	7.760	9.248	10.592	13.694	19.427

Table 1: Table showing transfer time in seconds for respective values of packet delay, α, and packet loss, β.

	Latency α (*ms*)			
Jitter $\delta(\alpha)$	2.0	4.0	8.0	16.0
4*MB MPI Datasize*				
0.2	0.910	1.027	1.372	1.881
0.4	0.922	1.071	1.379	1.928
0.8	0.940	1.117	1.416	1.872
16*MB MPI Datasize*				
0.2	3.266	3.810	4.577	6.329
0.4	3.464	3.940	4.879	6.977
0.8	3.649	4.029	4.947	7.165

Table 2: Table showing transfer time in seconds for respective values of packet delay (α) and jitter $\delta(\alpha)$.

This work points the way to further studies such as tuning of the TCP stack on host machines to get better bandwidth utilisation or investigation of the effect the underlying protocol (TCP) has on performnce compared with newer protocols such as Reliable-Blast UDP (RBUDP).

This research has been funded by ESLEA grant GR/T04465. I would like to acknowledge the invaluable help of Barney Garrett, Clive Davenhall and Nicola Pezzi in configuring the network and hardware infrastructures; and Radhika Saksena for application-specific discussions.

References

[1] ESLEA, *Exploitation of Switched Lightpaths for eScience Applications*, http://www.eslea.uklight.ac.uk

[2] S. M. Pickles, R. Haines, R. L. Pinning and A. R. Porter, *A Practical Toolkit for Computational Steering, Phil Trans R Soc*, **363**, 1833, pp. 1843-1853, 2005, http://dx.doi.org/10.1098/rsta.2005.1611

[3] M. Mc Keown, S. M. Pickles, A. R. Porter, R. L. Pinning, M. Riding and R. Haines, *The Service Architecture of the TeraGyroid Experiment*, *Phil Trans R Soc*, **363**, pp. 1743-1755, 2005.

[4] S. Jha, P. V. Coveney, M. J. Harvey, and R. Pinning, *SPICE: Simulated Pore Interactive Computing Environment, Proceedings of the 2005 ACM/IEEE conference on Supercomputing*, p. 70, 2005, http://dx.doi.org/10.1109/SC.2005.65

[5] N. Karonis, B. Toonen, and I. Foster, *MPICH-G2: A Grid-Enabled Implementation of the Message Passing Interface*, *Journal of Parallel and Distributed Computing (JPDC)* **63**, No. 5, pp. 551-563, 2003.

[6] B. M. Boghosian, P. V. Coveney, S. Dong, L. I. Finn, S. Jha, G. Karniadakis and N. Karonis, *Nektar, SPICE and Vortonics: Using Federated Grids for Large Scale Scientific Applications*, *Proceedings of Challenges of Large Applications in Distributed Environments*, 34-42, 2006. IEEE Catalog Number: 06EX1397 ISBN 1-4244-0420-7 Library of Congress 2006925560.

[7] B. Boghosian, L. I. Finn and P. V. Coveney, *Moving the data to the computation: multi-site distributed parallel computation*, 2006, http://www.realitygrid.org/publications/GD3.pdf

[8] M. Carson, D. Santay, *NIST Net - A Linux-based Network Emulation Tool*, *Computer Communication Review*, **6**, 2003.

[9] J. MacLaren, M. McKeown and S. Pickles, *Co-Allocation, Fault Tolerance and Grid Computing*, *Proceedings of the UK e-Science All Hands Meeting 2006*, 155-162, 2006.

PROCEEDINGS OF SCIENCE

Large-scale lattice-Boltzmann simulations over lambda networks

Radhika S. Saksena*
Centre for Computational Science, Department of Chemistry, University College London
20 Gordon Street, London WC1H 0AJ, United Kingdom
E-mail: r.saksena@ucl.ac.uk

Peter V. Coveney
Centre for Computational Science, Department of Chemistry, University College London
20 Gordon Street, London WC1H 0AJ, United Kingdom
E-mail: p.v.coveney@ucl.ac.uk

Robin L. Pinning
Manchester Computing, University of Manchester
Oxford Road, Manchester M13 9PL, United Kingdom
E-mail: pinning@manchester.ac.uk

Stephen P. Booth
Edinburgh Parallel Computing Centre, University of Edinburgh
Edinburgh EH9 3JZ, Scotland, United Kingdom
E-mail: s.booth@epcc.ed.ac.uk

Amphiphilic molecules are of immense industrial importance, mainly due to their tendency to align at interfaces in a solution of immiscible species, e.g., oil and water, thereby reducing surface tension. Depending on the concentration of amphiphiles in the solution, they may assemble into a variety of morphologies, such as lamellae, micelles, sponge and cubic bicontinuous structures exhibiting non-trivial rheological properties. The main objective of this work is to study the rheological properties of very large, defect-containing gyroidal systems (of up to 1024^3 lattice sites) using the lattice-Boltzmann method. Memory requirements for the simulation of such large lattices exceed that available to us on most supercomputers and so we use MPICH-G2/MPIg to investigate geographically distributed domain decomposition simulations across HPCx in the UK and TeraGrid in the US. Use of MPICH-G2/MPIg requires the port-forwarder to work with the grid middleware on HPCx. Data from the simulations is streamed to a high performance visualisation resource at UCL (London) for rendering and visualisation.

Lighting the Blue Touchpaper for UK e-Science - Closing Conference of ESLEA Project
March 26-28 2007
The George Hotel, Edinburgh, UK

*Speaker.

PoS(ESLEA)012

1. Introduction

Our objective is to simulate the rheological properties of ternary amphiphilic mixtures which undergo self-assembly into cubic and non-cubic periodic structures. Lattice-Boltzmann simulations of cubic and non-cubic self-assembled mesophases exhibit interesting rheological behaviour. Such discoveries can be exploited to design functional materials with specific rheological properties. Self-assembled mesophases are also finding application in the synthesis of mesoporous nanomaterials which have interesting structural and electronic properties. These simulations involve traversing complex parameter spaces in order to identify regions where self-assembled mesophases are formed. Self-assembly phenomena are known to suffer from hysteresis whereby a meta-stable state can persist for long times and one requires long simulations to identify the final self-assembled phase. Additionally, defects in the periodic mesophases are known to significantly influence rheological properties: in order to simulate defect dynamics correctly, one needs to perform simulations of large system sizes which can be deemed to be free of finite-size effects. Thus, in order to simulate physically realistic system behaviour, large-scale and long simulations need to be performed, making this an extremely computationally demanding endeavour.

2. Simulation Code

We use the lattice-Boltzmann code, LB3D, to perform large-scale lattice-Boltzmann simulations. The implementation of the lattice-Boltzmann model for ternary amphiphilic fluids in LB3D has been discussed previously [1]. LB3D correctly simulates the self-assembly dynamics [2] and rheology [3] of cubic and lamellar mesophases in these systems. The code has been under development for over 7 years and is widely deployed on the UK NGS [4], the UK supercomputing resource HPCx [5] and on various supercomputing resources on the US TeraGrid [6]. LB3D is a scalable parallel MPI code written in Fortran 90. It has been parallelised according to the domain decomposition scheme [7]. LB3D is a memory intensive application code requiring approximately 1 kilobyte of memory per lattice-site to store state data. The compute-intensive part of the algorithm consists of the collision and propagation steps. Because of the non-local interaction forces between different species in the amphiphilic mixture, two communication steps per cycle are required to exchange state data for lattice-sites on the sub-domain boundaries between neighbouring processors. LB3D checkpoints system state and visualisation data-sets at regular intervals. For large lattices, the size of checkpoints becomes non-trivial. For a 1024^3 lattice-sites system, LB3D requires 1.07 TB of total memory to run; writing checkpoint files requires O(TB) of disk space. Each visualization step requires emission of a 4.3 GB visualisation dataset which has to be transferred to and rendered by a high performance visualisation resource.

3. Technical Challenges

In order to simulate physically realistic rheological behaviour of multi-domain, defect containing, ternary amphiphilic mixtures we need to perform long-time simulations of large systems containing at least 1024^3 lattice sites. As mentioned in the previous section, the simulation checkpoint and visualisation data-sets can reach up to terabytes and require significant network bandwidth for

transfer to storage and visualisation resources. Such large-scale data transfers have been performed for these lattice-Boltzmann simulations over UKLight [9] and in other RealityGrid projects, e. g. [8]. The large amount of memory required to carry out these simulations is often not available to us on a single supercomputer. Here we discuss a new network-intensive meta-computing approach called geographically distributed domain decomposition or GD^3 [10] that can overcome this memory bottleneck. In this approach, a single MPI simulation is split across processors on geographically distributed supercomputers. Network provisionability, bandwidth, latency and reliability during the simulation run are all critical in the GD^3 approach. The grid-middleware that we use to launch cross-site GD^3 simulations is called MPICH-G2 [11] and its newer pre-release version called MPIg.

Our initial aim is to split a 1024^3 lattice-sites simulation across supercomputers, the obvious candidates for this being the US TeraGrid resources and HPCx in the UK. HPCx is connected via the UKLight optical network [12] and Starlight network to the TeraGrid optical backbone. From an application scientist's perspective there are many technical challenges that need to be overcome in order to efficiently run cross-site simulations. Firstly, the simulation code needs to achieve maximum overlap between computation and communication by taking advantage of MPI's asynchronous communication calls. Unlike the previous MPICH-G2 version, MPIg implements asynchronous communications and is well-suited to take advantage of latency hiding optimisations in the code. Also UDT communication protocol was proposed in future versions of the grid middleware instead of the currently used TCP protocol. This is estimated to improve cross-site performance by a factor of two [10]. Secondly, in order to be included in MPICH-G2/MPIg cross-site framework, the participating machines need to have externally addressable nodes. This poses a problem for relatively less grid-enabled machines like the Cray XT3 machine (Bigben) at the Pittsburgh Supercomputing Center and the new UK HEC resource, HECToR [13]. Thirdly, we face issues in using MPICH-G2 on the UK's HPCx machine due to the port-forwarding mechanism in place on that machine. However, within the Vortonics project at SuperComputing Conference 2005 [14], a trans-atlantic cross-site run, over UKLight on TeraGrid machines and the now decomissioned Newton machine at CSAR on the UK NGS [4], was performed by Boghosian *et al* using their lattice-Boltzmann simulation code which has a similar communication pattern as LB3D [10]. In situations where the memory requirements of the simulation are too large to fit onto a single supercomputer, their results provide support for the viability of GD^3 as compared to alternatives like swapping portions of the simulation to disk or worse, waiting for a bigger machine to become affordable. From a usability point-of-view, cross-site simulations depend critically on the availability of automatic mechanisms for the advanced reservation and co-scheduling of compute resources and networks. To this end, tools have been developed for automated reservation and co-scheduling of grid resources [15] and of dynamically provisioned networks (within the ESLEA project). These tools used in conjunction with the grid application hosting middleware like the Application Hosting Environment [16] can allow scientists to efficiently and frequently schedule and launch cross-site simulations. Although there is a significant effort on middleware development, the support from grid resource managers on this front is less forthcoming. The UK NGS and EU DEISA [17] grid resource providers have not shown any serious indication of providing such a facility. There is a mechanism in place in the US TeraGrid to request for advanced reservation of resources through a web page, however, this requires manual intervention on the application scientist's part and as far

as we understand also on the resource manager's part. The beta version of the NAREGI [18] grid software stack, however, has a super-scheduler component to support co-scheduling and advanced reservation of grid resources. This is an encouraging development and further efforts like this are required to allow scientists to exploit advances in various areas of computing in a coherent fashion and be able to do science that was not possible before.

4. Summary

In this paper, we have described the scientific motivation for our lattice-Boltzmann simulations and the primary simulation issues that need to be overcome. We describe the main features of the LB3D code which determine network requirements. Finally we discuss a new meta-computing approach called geographically distributed domain decomposition (GD^3) for which high bandwidth, low latency, reliable network connections are critical and the technical challenges in deploying these simulations on transatlantic grids connected via UKLight.

This research has been funded by ESLEA project's EPSRC grant GR/T04465/01 and the EPSRC grants GR/R67699, EP/C536452/1, EP/E045111/1, GR/T27488/01 and through the OMII Managed Programme grant GR/290843/01. Access to the US TeraGrid resources was provided under the NRAC and PACS grants MCA04N014 and ASC030006P. We would also like to acknowledge useful discussions with Giovanni Giupponi, Marco Mazzeo and Steven Manos, and Nicola Pezzi's help with network issues.

References

[1] J. Harting, J. Chin, M. Venturoli, and P. V. Coveney. *Phil. Trans. R. Soc. A.*, 1833:1895–1915, 2005.

[2] J. Chin and P. V. Coveney. *Proc. R. Soc. London Series A,*, 462:3575–3600, 2006.

[3] G. Giupponi, J. Harting, and P. V. Coveney. *Europhys. Lett.*, 73:533–539, 2006.

[4] http://www.ngs.ac.uk.

[5] http://www.hpcx.ac.uk.

[6] http://www.teragrid.org.

[7] W. Gropp, E. Lusk, and A. Skjellum. In *Using MPI*, pages 59–97. MIT Press, 1994.

[8] M-A. Thyveetil, S. Manos, J. L. Suter and P. V. Coveney *in Lighting the Blue Touchpaper for UK e-Science - Closing Conference of ESLEA Project, POS (ESLEA), 013*, 2007.

[9] M. Venturoli, M. J. Harvey, G. Giupponi, P. V. Coveney, R. L. Pinning, A. R. Porter, and S. M. Pickles. *Proceedings of the UK e-Science All Hands Meeting 2005*, 2005.

[10] B. Boghosian, L. I. Finn, and P. V. Coveney. *http://www.realitygrid.org/publications/GD3.pdf*.

[11] N. Karonis, B. Toonen, and I. Foster. *J. Parallel and Distributed Computing*, 63(5):551–563, 2003.

[12] http://www.uklight.ac.uk.

[13] http://www.hector.ac.uk.

[14] http://hilbert.math.tufts.edu/ bruceb/VORTONICS/index.html.

[15] J. Maclaren and M. Mc Keown. HARC: A Highly-Available Robust Co-scheduler, 2006. Proceedings of the 5th UK e-Science All Hands Meeting.

[16] P. V. Coveney, R. S. Saksena, S. J. Zasada, M. McKeown, and S. Pickles. *Comp. Phys. Comm.*, 176:406–418, 2007.

[17] http://www.deisa.org.

[18] http://www.naregi.org.

PROCEEDINGS OF SCIENCE

Use of UKLight as a Fast Network for Data Transport from Grid Infrastructures

M.-A. Thyveetil*, **S. Manos, J. L. Suter and P. V. Coveney**
Centre for Computational Science, Department of Chemistry, University College London, 20 Gordon Street, London, United Kingdom, WC1H 0AJ.
E-mail: m.thyveetil@ucl.ac.uk

Large-scale molecular dynamics simulations run on high-end supercomputing facilities can generate large quantities of data. We simulate mineral systems up to 10 million atoms in size in order to extract materials properties which are otherwise difficult to obtain through existing experimental techniques. Simulating clay systems this large can generate large files up to 50GB in size. These simulations were carried out on remote sites across the US TeraGrid and UK's HPCx supercomputer. Using the dedicated network, UKLight, connected to these high-end supercomputing resources has significantly reduced the time taken to transport large quantities of data generated from our simulations. UKLight provides excellent quality of service, with reduced packet loss and latency. This enhanced data transfer method paves the way for faster communication between two coupled applications, such as in the case of real-time visualisation or computational steering.

Lighting the Blue Touchpaper for UK e-Science - Closing Conference of ESLEA Project
March 26-28 2007
The George Hotel, Edinburgh, UK

*Speaker.

PoS(ESLEA)013

1. Introduction

Computational grids [1, 2, 3] provide an attractive environment in which to undertake the intensive compute tasks required for large-scale molecular dynamics (MD) simulation. Large-scale atomistic simulations, which we define as containing more than 100,000 atoms, provide a bridge between atomistic and mesoscopic scale simulations [4]. Operating over at least tens of thousands of atoms, emergent mesoscopic properties are observed in full atomistic detail. Large-scale simulations of clay nanocomposites reveal long wavelength, low amplitude thermal undulations [5]. Small simulation sizes implicitly inhibit long wavelength clay sheet flexibility due to the periodic boundaries used in condensed matter molecular dynamics, which effectively pin the clay sheet at the edges of the simulation cell. It is often difficult to predict how nanocomposites will behave from theories of conventional composite behavior due to the disparity of dimensions; hence the need for large-scale molecular simulation to sample all possible length scales[6, 7, 8, 9]. We perform many simulations at various system sizes; the largest approaches that of realistic clay platelets and contains upwards of a million atoms. Using computational grid resources allows the turnaround on the large number of simulations required to be on a feasible timescale. We utilise Grid resources on the US TeraGrid and the UK's flagship machine HPCx. Jobs are launched remotely using the Application Hosting Environment [10][11].

With increasingly large simulation sizes now possible, new problems appear as our largest simulations can generate files up to 50GB in size. These files contain important atomistic data which need to be retrieved from the remote Grid resource to a local machine, but this can become time consuming. An answer to the problems of slow and unreliable networks is to use switched-circuit networks. In a switched-circuit network the user has sole use of a dedicated network connection, thus eliminating contention with other traffic and providing excellent, predictable, Quality of Service (QoS). Switched-circuit networks can be implemented in various ways, though there has been much recent work on allocating users or groups sole use of individual wavelengths (lambdas) in multi-wavelength optical fibres[12]. In the UK dedicated connections are available via the UKLight network which uses manually-configured SDH circuits. In this paper we present network tests intended to optimise the use of local machines at UCL connected by UKLight to external Grid resources. We conclude with a summary of our findings and future plans we have with UKLight.

2. Performance Testing of UKLight

Currently, a lambda network operates in the UK called UKLight[1]. It provides a fast connection between UCL and various supercomputing resources such as EPCC's HPCx[2] and the TeraGrid[3]. We carried out a series of tests on the performance of the link between UCL and HPCx, as well as UCL and the TeraGrid. Preliminary results showed that the link was not as fast as expected. Two methods could have been used in order to improve the bandwidth of the link. The first was to use GridFTP in order to use multiple streams over one link. The problem with this method is that it can be difficult to implement on a network such as UKLight. This led us to use common software

[1] http://www.uklight.ac.uk
[2] http://www.hpcx.ac.uk
[3] http://www.teragrid.org

such as SSH, along with network tuning to achieve maximum bandwidth over a single stream. Specifically, we needed to tune the TCP window size in order to achieve maximum bandwidth. This section describes the methodology we used to test the system and the results we obtained once network parameters were tuned.

2.1 Methodology

Currently high-performance dedicated network connections are provided in the UK by the UKLight[4] network. The first connection studied was between a Linux box connected to the same UKLight switch as UCL's SGI Prism and a box connected to the UKLight switch of HPCx. The second connection was between UCL and the TeraGrid's IA-64 Linux cluster NCSA. NCSA's network parameters were already tuned, therefore a separate machine was not needed for testing.

Iperf is a network tool which we used to test UDP and TCP bandwidth between networked computers[5]. Iperf UDP tests provided the maximum bandwidth of the connection before packet loss is seen. This test was carried out in both the production network and UKLight. In all UDP tests the Grid resource acted as the client while the UCL Linux machine was the server. The result of the UDP tests can be used to calculate the bandwidth delay product (BDP) of the link, which is found by: bandwidth $\times$ round trip time. The round trip time is the time elapsed for a message to travel to a remote place and back again. The BDP helps determine the maximum window size for TCP communication.

In order to get the best performance from a network, the TCP window size defined on the kernel and application side, needs to be tuned. The Linux kernel parameters which we needed to adjust were as follows:

$/proc/sys/net/core/wmem\_max$
$/proc/sys/net/core/rmem\_max$
$/proc/sys/net/ipv4/tcp\_rmem$
$/proc/sys/net/ipv4/tcp\_wmem$

In addition, we used the application called High Performance Enabled SSH/SCP (HPN-SSH)[6]; this is a patch for recent OpenSSH releases which allows adjustment of the TCP window size within the application. With these changes in place, we then tested the performance of SCP on all connections, including the production network, UKLight with untuned network parameters and also with tuned network parameters.

2.2 Results

The UDP tests showed that the connection between UCL and other grid resources could be as large as 40MB/s, as summarised in Table 1. The TCP tests showed that a maximum of 34.4MB/s could be achieved, as shown in Table 2. Using these parameters, the Linux kernel parameters were tuned and HPN-SSH was used to compare the data transfer rates for the production network, as well

[4] http://www.uklight.ac.uk
[5] http://dast.nlanr.net/Projects/Iperf
[6] http://www.psc.edu/networking/projects/hpn-ssh

as the untuned and tuned UKLight network. As summarised in Table 3, a massive improvement can be seen using tuned network parameters. The academic Super Janet production network operates at 4.5MB/s for connections within the UK and 600KB/s from UCL to TeraGrid's NCSA, whilst with tuned network parameters, the bandwidth of the connection goes up to 16MB/s for the NCSA UKLight link and 28MB/s for the HPCx link.

Grid resource	Maximum Bandwidth (MB/s)	Round trip time	Bandwidth delay product
NCSA	40 MB/s	92ms	3.2MB
HPCx Linux	32MB/s	8ms	300KB

Table 1: Results for UDP tests of UKLight connection between UCL and Grid resources. The Grid resource acted as the Iperf client while the UCL Linux machine was the server.

Grid resource	Maximum Bandwidth (MB/s)	Maximum Window Size (MB)
NCSA	34.3 MB/s	4MB
HPCx Linux	34.4MB/s	3MB

Table 2: Results for TCP tests of UKLight connection between UCL and Grid resources. The Grid resource acted as the Iperf client while the UCL Linux machine was the server.

	Maximum bandwidth (MB/s)		
Grid resource	Janet Network	UKLight (untuned)	UKLight (tuned)
NCSA	0.6 MB/s	0.7 MB/s	16 MB/s
HPCx Linux	4.5 MB/s	8 MB/s	28 MB/s

Table 3: Comparison of networks using the maximum bandwidth obtained from SCP. A great improvement is seen using UKLight over the academic Super Janet network.

3. Conclusion

We conclude that UKLight makes a significant impact when transporting large files from Grid resources. In order to achieve the maximum bandwidth of this link over a single stream network parameters must be tuned. The results also highlight the fact that UKLight provides an efficient way to transport data for more complicated uses such as real-time visualisation and computational steering[13]. Visualisation of molecular simulations as presented in this study is very important but with system sizes greater than 1 million atoms, most visualisation is currently carried out only after the simulation has completed. Ideally we would like to be able to carry out real-time visualisation in order to see the evolution of the system while it is running. In addition to this, computational steering provides a way to interact with and monitor a simulation.

Real-time visualisation and computational steering are processes which cannot be carried out on batch systems, used by most high performance computing facilities. This means that advanced reservation and co-scheduling of the resources are needed, so that the user is able to determine when their simulation has started. In the future we hope to be able to use UKLight in order to carry out real-time visualisation and steering of our clay nanocomposite simulations across UKLight with the aid of co-scheduling.

References

[1] P. V. Coveney, editor, "Scientific Grid Computing", *Phil. Trans. R Soc. A* **7** (2005) 24–32.

[2] I. Foster, C. Kesselman and S. Tuecke, "The anatomy of the grid: Enabling scalable virtual organizations", *Intl J. Supercomp. Appl.* **15** (2001) 3–23.

[3] B. Boghosian and P. V. Coveney, "Scientific applications of grid computing", *Comp. Sci. and Eng.* **7** (2005) 10–13.

[4] P. Boulet, P. V. Coveney and S. Stackhouse, "Simulation of hydrated Li^+-, Na^+- and K^+-montmorillonite/polymer nanocomposites using large-scale molecular dynamics", *Chem. Phys. Lett.* **389** (2004) 261–267.

[5] J. L. Suter, P. V. Coveney, H. C. Greenwell, and M.-A. Thyveetil, "Large-Scale Molecular Dynamics Study of Montmorillonite Clay: Emergence of Undulatory Fluctuations and Determination of Material Properties", *J. Phys. Chem. C*, *in press* (2007).

[6] H. C. Greenwell, W. Jones, P. V. Coveney, and S. Stackhouse, "On the application of computer simulation techniques to anionic and cationic clays: A materials chemistry perspective", *J. Mater. Chem.* **16** (2006) 708–723.

[7] H. C. Greenwell, A. A. Bowden, B. Q. Chen, P. Boulet,, J. P. G. Evans, P. V. Coveney, and A. Whiting, "Intercalation and in situ polymerization of poly(alkylene oxide) derivatives within M+-montmorillonite (M = Li, Na, K)", *J. Mater. Chem.* **16** (2006) 1082–1094.

[8] E. S. Boek, P. V. Coveney, and N. T. Skipper, "Monte Carlo molecular modeling studies of hydrated Li-, Na-, and K-smectites: Understanding the role of potassium as a clay swelling inhibitor", *J. Am. Chem. Soc.* **117** (1995) 12608–12617.

[9] H. C. Greenwell, M. J. Harvey, P. Boulet, A. A. Bowden, P. V. Coveney, and A. Whiting, "Interlayer structure and bonding in nonswelling primary amine intercalated clays", *Macromolecules* **38** (2005) 6189–6200.

[10] P. V. Coveney, R. Saksena, S. J. Zasada, M. McKeown and S. Pickles, "The Application Hosting Environment: Lightweight Middleware for Grid-Based Computational Science", *Comp. Phys. Comm.* **176** (2007) 406–418.

[11] S. Zasada, "The Application Hosting Environment: Lightweight Middleware for Grid Based Computational Science", *Lighting the Blue Touchpaper for UK e-Science - Closing Conference of ESLEA Project*, (2007).

[12] A. Hirano, L. Renambot, B. Jeong, J. Leigh, A. Verlo, V. Vishwanath, R. Singh, J. Aguilera, A. Johnson, and T. A. DeFanti, "The first functional demonstration of optical virtual concatenation as a technique for achieving terabit networking", *Future Gener. Comput. Syst.*, 22 (2006) 876–883.

[13] M. Harvey, S., Jha, M.-A. Thyveetil and P. V. Coveney, "Using Lambda Networks to Enhance Performance of Interactive Large Simulations", *Second IEEE International Conference on e-Science and Grid Computing (e-Science'06)*, (2006) 40.

PROCEEDINGS OF SCIENCE

Using lambda networks to enhance performance of interactive large simulations *

Matthew J Harvey

Imperial College London

Shantenu Jha†

Louisiana State University and University College London

Mary-Ann Thyveetil

University College London

Peter Coveney

University College London

The ability to use a visualisation tool to steer large simulations provides innovative and novel usage scenarios, for example, the ability to use new algorithms for the computation of free energy profiles along a nanopore [1]. However, we find that the performance of interactive simulations is sensitive to the quality of service of the network with latency and packet loss in particular having a detrimental effect. The use of dedicated networks (provisioned in this case as a circuit-switched, point-to-point optical lightpath or *lambda*) can lead to significant (50% or more) performance enhancement. When running on say 128 or 256 processors of a high-end supercomputer this saving has a significant value. We discuss the results of experiments performed to understand the impact of network characteristics on the performance of a large parallel classical molecular dynamics simulation when coupled interactively to a remote visualisation tool.

Lighting the Blue Touchpaper for UK e-Science - Closing Conference of ESLEA Project
March 26-28, 2007
Edinburgh

*This talk is a condensed version of work originally presented in Ref. [9].

†Speaker.

PoS(ESLEA)014

1. Introduction

Lambda networking involves using different wavelengths (lambdas) of light in fibres for separate connections. Lambda networks provide high-levels of Quality of Service (QoS) by giving applications and user communities dedicated lambdas on a shared fibre infrastructure. The implementation requires Dense Wavelength Division Multiplexing (DWDM) to accommodate many wavelengths on a fibre, optical switches, and other optical networking equipment. Grid computing applications have so far mostly made use of best-effort, shared TCP/IP networks, *i.e.* the network has simply been the glue that holds the middleware-enabled computational resources together. In contrast, by using lambdas the networks themselves are schedulable "first class" grid resources. These deterministic lambda networks, carrying one or more lambdas of data, form on-demand, end-to-end dedicated networks, often called lightpaths; lightpaths form the basis of the next generation of network-centric applications.

Lightpaths have the ability to meet the needs of very demanding e-science applications as a consequence of their ability to provide several features that are not possible using regular, production best-effort networks. These include providing higher bandwidth connections (*e.g.* [2]), user-defined networks [3], implementation of novel protocols [4] and provide essentially contention-free and high quality-of-service links.

Most applications however, have tended to use lambdas for their high bandwidth alone. For example, an important class of applications driving the development and research of lambdas are visualisation of large and complex data sets. Here we report on one of the first uses of lambdas to couple interactive, steered visualisation with "active" simulations. This work was conducted as part of the SPICE (Simulated Pore Interactive Computing Environment) project details of which have been discussed elsewhere [1, 9] . In the next section, we describe the project's motivating scientific problem and the technical solutions adopted.

2. Simulated Pore Interactive Computing Environment

The transport of bio-molecules like DNA, RNA and poly-peptides across protein membrane channels is of primary significance in a variety of areas. Although there has been a flurry of recent activity, both theoretical and experimental [5, 6], aimed at understanding this crucial process, many aspects remain unclear.

Of the possible computational approaches, classical molecular dynamics (MD) simulations of bio-molecular systems have the ability to provide insight into specific aspects of a biological system at a level of detail not possible with other simulation techniques. MD simulations can be used to study details of a phenomenon that are often not accessible experimentally [7] and would certainly not be available from simple theoretical approaches. However, the ability to provide such detailed information comes at a price: MD simulations are extremely computationally intensive – prohibitively so in many cases. As was discussed in Ref. [1], advances in both the algorithmic and the computational approaches are imperative to overcome such barriers.

SPICE, the Simulated Pore Interactive Computing Environment project [1], implements a method, henceforth referred to as SMD-JE, to compute the free energy profile (FEP) along the vertical axis of the protein pore. This method reduces the computational requirement for the problem

of interest by a factor of at least 50-100, at the expense of introducing two new variable parameters, with a corresponding uncertainty in the choice of their values. Fortunately, the computational advantages can be recovered by performing a set of "preprocessing simulations" which, along with a series of interactive simulations, help inform an appropriate choice of the parameters. To benefit from the advantages of the SMD-JE approach and to facilitate its implementation at all levels – interactive simulations of large systems, the pre-processing simulations and finally the production simulation set – we use the infrastructure of a federated trans-Atlantic grid [8].

Interactive simulations involve using the visualiser as a steerer, *e.g.* to apply a force to a subset of atoms, Figure 1(b), and requires bi-directional communication – there is a steady-state flow from the simulation to the visualiser as well as the visualiser to the simulation. As a consequence of requiring geographically distributed resources, high-end interactive simulations are dependent on the performance of the network between the scientist (visualiser) and the simulation. Unreliable communication leads not only to a possible loss of interactivity, but equally seriously, a significant slowdown of the simulation as it waits for data from the visualiser.

On switching traffic flow from the production network to a lambda network, we found an improvement in the performance of around 50%. The simulation performance over the production network varied, *i.e.* was apparently sensitive to prevailing network conditions. Interactive MD simulations thus require high quality-of-service – as defined by low latency, jitter and packet loss – networks to ensure reliable bi-directional communication. This leads to the interesting situation where large-scale interactive computations require both computational and visualisation resources to be co-allocated with networks of sufficient QoS [8].

3. Experiment

In order to quantify the impact of network performance characteristics on the efficiency of an interactive MD simulation a series of measurements were made under controlled conditions. The full details of our methodology are described in [9] and are presented here in outline.

The two compute resources on which the molecular dynamics simulation (using NAMD[10]) and visualiser were run (VMD[11]), were connected via a dedicated circuit on the UKLight[12] optical network, provisioned at 300Mbps. UKLight provides the user with sole-use of dedicated, manually-configured SDH circuits which hence have similar properties to dedicated lambda networks.

To control the characteristics of the network, a third system was introduced between the visualiser and the UKLight link. This system acted as an IP bridge and employed the NISTnet [13] package to modify the traffic flow. The wall-time per simulation timestep (t_s) was measured for interactive simulations over a range of network characteristics controlled by NISTnet. The parameters varied were 1) Packet transmission delay (α), to simulate different latency network paths, 2) Packet Loss (β), to simulate packet loss due to network congestion.

3.1 Quality of Service: Latency

The effective latency of the UKLight link was varied to emulate different paths, service times and congestion – all characteristic of best effort networks between two given end points.

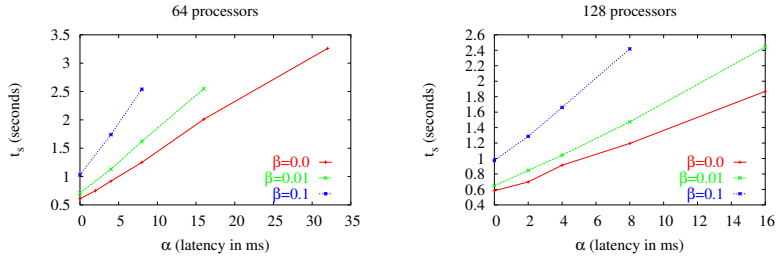


Figure 1: Plots showing the effect of latency on the performance (t_s). An increase in latency leads to a linear increase in the wall-time taken per simulation timestep, independent of the number of processors used. The performance degradation remains linear for different values of β.

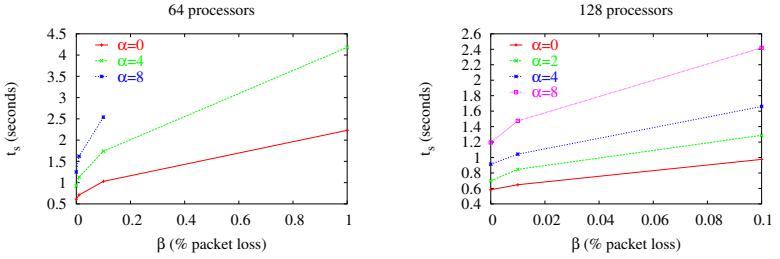


Figure 2: Plots showing the dependence of performance (t_s) on the packet loss (β) for different values of α. The qualitative characteristics remain the same for 64 and 128 processors, *i.e.* at a fixed value of α, the ratio of t_s for a pair of β values is similar for the two processor counts. For example for α = 4, the ratio of t_s at β = 0.1 and 0.01, is 1.54 and 1.59 for 64 and 128 respectively. At the same values of β but for α=8 the values of the ratio are 1.56 and 1.62 respectively.

Not surprisingly the time taken for each timestep increases linearly with greater latency. Our results are plotted in Fig. 1. It is interesting to note, that although the average wall-time taken per simulation time-step is of the order of hundreds of milli-seconds, introducing latencies of a few milli-seconds has a significant effect. This is attributed to the fact that a significant fraction of the simulation time is spent waiting for I/O operations to complete. Thus we conclude that reducing any avoidable latency is a good performance enhancing strategy.

3.2 Packet Loss Effects

Packet loss is interpreted by the TCP protocol as an indication of congestion and causes the window size to be immediately reduced to a minimum size (4096 bytes in this case) and then renegotiated up. The effect of increasing β (as shown in Fig. 2 is to increase the frequency of window size reduction (reducing the average size over time) and consequently reducing effective throughput. We observe that, in general, default settings for TCP window size parameters are unsuited to networks with high bandwidth-delay products.

4. Conclusion

It can be argued that with significant effort, a highly optimised I/O mechanism could be implemented within the NAMD code to withstand performance degradation arising from production networks. Whereas we do not contest that this in principle is possible, doing so would require significant re-factoring of a very complex code which has been developed by the community over many years (we estimate the number of person-years effort to be easily a hundred). Equally important, it is impractical to aim to introduce special-purpose code for every unique usage scenario; thus it is highly desirable to be able to use the same general-purpose code over a wide range of scientific problems and usage scenarios. Our efforts to quantify the advantages of lightpaths need to be understood in the above mentioned context.

Not only can grids use lightpaths to more closely integrate distributed environments, but they *must* use lightpaths to couple distributed environments to overcome some of the bottlenecks of traditional programming methodologies of high performance codes as well as problem solving approaches for challenging scientific problems. SPICE provides an example of a large-scale problem that depends on using algorithms amenable to distributed computing techniques and then implementing them on grids. In order to effectively utilise these algorithms, interactive simulations on large computers are required.

In order to enable meaningful interactive exploration, the responses must be computed in reasonable times. Thus as larger systems are studied – MD simulations of a million atoms are now just appearing in the literature [15] – not only will larger computers be required, but the need for efficient and reliable communication will also grow.

5. Acknowledgements

This work has been supported by EPSRC grant number GR/T04465/01 (ESLEA), by EPSRC Grant EP/D500028 (SPICE) and the EPSRC-funded RealityGrid project (GR/R67699 and EP/C536452/1).

References

[1] S. Jha, P. V. Coveney, M. J. Harvey, and R. Pinning, "SPICE: Simulated Pore Interactive Computing Environment," *Proceedings of the 2005 ACM/IEEE conference on Supercomputing*, p. 70, 2005, dx.doi.org/10.1109/SC.2005.65.

[2] A. Hirano, L. Renambot, B. Jeong, J. Leigh, A. Verlo, V. Vishwanath, R. Singh, J. Aguilera, A. Johnson, and T. A. DeFanti, "The first functional demonstration of optical virtual concatenation as a technique for achieving terabit networking," *Future Generation Computer Systems*, vol. 22, pp. 876–883, 2006.

[3] J. Mambretti, R. Gold, F. Yeh, and J. Chen, "Amroeba: Computational astrophysics modeling enabled by dynamic lambda switching," *Future Generation Computer Systems*, vol. 22, pp. 949–954, 2006.

[4] R. L. Grossman, Y. Gu, D. Hanley, M. Sabala, J. Mambretti, A. Szalay, A. Thakar, K. Kumazoe, O. Yuji, and M. Lee, "Data mining middleware for wide-area high-performance networks," *Future Generation Computer Systems*, vol. 22, pp. 940–948, 2006.

[5] D. K. Lubensky and D. R. Nelson. *Phys. Rev E*, 31917 (65), 1999; Ralf Metzler and Joseph Klafter. *Biophysical Journal*, 2776 (85), 2003; Stefan Howorka and Hagan Bayley, *Biophysical Journal*, 3202 (83), 2002.

[6] A. Meller *et al*, *Phys. Rev. Lett.*, 3435 (86) 2003; A. F. Sauer-Budge *et al. Phys. Rev. Lett.* 90(23), 238101, 2003.

[7] M. Karplus and J. A. McCammon, "Molecular Dynamics Simulations of Biomolecules," *Nature Structural Biology*, vol. 9, no. 9, pp. 646–652, 2002.

[8] B. Boghosian, P. Coveney, S. Dong, L. Finn, S. Jha, G. Karniadakis, and N. Karonis, "Nektar, SPICE and Vortonics – Using Federated Grids for Large Scale Scientific Applications," in *Proceedings of Challenges of Large Applications in Distributed Environments (CLADE) 2006*, vol. IEEE Catalog Number: 06EX13197, Paris, June 2006, pp. 32–42, iSBN 1-4244-0420-7.

[9] M. J. Harvey, S. Jha, M. A. Thyveetil, and P. V. Coveney, "Using lambda networks to enhance performance of interactive large simulations," *2nd IEEE International Conference on e-Science and Grid Computing, 4-6 December 2006, Amsterdam*, 2006.

[10] J. C. Philips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kale, and K. Schilten, "Scalable molecular dynamics with NAMD," *Journal of Computational Chemistry*, vol. 26, pp. 1781–1802, 2005.

[11] W. Humphrey, A. Dalke, and K. Schulten, "VMD – Visual Molecular Dynamics," *Journal of Molecular Graphics*, vol. 14, pp. 33–38, 1996.

[12] JISC, "UKLight Switched Optical Lightpath Network," http://www.uklight.ac.uk

[13] M. Carson and D. Santay, "NISTNet - A Linux-based Network Emulation Tool," *Computer Communication Review*, vol. 6, 2003.

[14] U. o. S. C. Information Sciences Institute, 1981, rFC 793: Transmission Control Protocol `http://rfc.net/rfc793.html`.

[15] K. Y. Sanbonmatsu, S. Joseph, and C.-S. Tung, "Simulating movement of tRNA into the ribosome during decoding," *PNAS*, vol. 102, no. 44, pp. 15 854–15 859, 2005. [Online]. Available:

PROCEEDINGS OF SCIENCE

The ESLEA Circuit Reservation Software

A.C. Davenhall[*,a] **P.E.L. Clarke,**[a] **N. Pezzi**[b] **and L. Liang**[a]
[a]*National e-Science Centre, 15, South College Street, Edinburgh, EH8 9AA, UK*
[b]*Network Group, Information Systems, University College London, 5, Gower Place, London, WC1E 6BS, UK*
E-mail: clive@nesc.ac.uk

We describe the Circuit Reservation Software (CRS) developed by the ESLEA Project. Switched-circuit networks can offer improved performance over conventional packet-switched networks for applications with some demanding types of network requirements. However, if switched-circuit networks are used to connect scarce and expensive resources, such as supercomputers, then advanced reservations or bookings of circuits are necessary if the resources and networks are all to be used efficiently. We describe and discuss the CRS, a demonstrator system for making such advanced bookings.

Lighting the Blue Touchpaper for UK e-Science - Closing Conference of ESLEA Project
March 26-28, 2007
Edinburgh

[*]Speaker.

PoS(ESLEA)015

1. Introduction

We describe the Circuit Reservation Software (CRS) developed by the ESLEA Project for making advanced reservations of dedicated network connections. This software has a distributed and flexible architecture that makes it applicable to a wide variety of types of network. Deployment and use of the software will provide valuable practical experience of reserving and using dedicated network connections.

The ESLEA (Exploitation of Switched Light Paths for e-Science Applications) Project[1][1] is a collaboration comprising groups working in computer science and a variety of scientific disciplines who are all using high-capacity dedicated network connections. Specifically, all the groups are using UKLight,[2] the UK research and development network for switched-circuit networks. The aim of the project is to investigate ways of using these connections and to share the resulting expertise both within the project and more widely. The project runs from February 2005 to July 2007.

The packet-switched Internet has been incredibly successful. However, there are now numerous scientific applications that require high network performance which is difficult to achieve over the conventional, production network. One alternative is to use special-purpose connections dedicated to the flows, and the purpose of ESLEA is to investigate this technique.

Such circuit-switched networks have been around for many years. They have received renewed interest in recent times because of developments in optical network hardware, notably DWDM (Dense Wavelength Division Multiplexing) in optical fibres and all-optical switches. There are, however, many ways to to implement (or simulate) direct connections. For example, emerging protocol standards such as GMPLS (Generalised Multi-Protocol Label Switching; see for example [2]) offer the possibility of extending dedicated flows to campus end-hosts without the provision of special hardware or manual configuration.

Currently circuits are usually configured 'manually' and typically persist for at least a few weeks and usually much longer. Many applications require connections of a shorter duration. It is only practical to provide such ephemeral circuits if they can be configured automatically. Further, it should be possible to 'reserve' or 'book' them in advance. This facility is necessary both to allow work to be planned effectively and to allow the network connection to be provided at the same time as the scarce resources that it will connect, typically super-computers. The CRS has been developed as a prototype to investigate making advance reservations in circuit-switched networks. A previous paper presented our plans early in the Project[3]. The present paper describes and discusses the CRS.

2. The CRS Architecture

The purpose of the CRS is to create a dedicated connection between two end-hosts, one of which is local and the other geographically remote. The CRS adopted the architecture of the EGEE (Enabling Grids for E-Science) BAR (Bandwidth Allocation and Reservation) Project[3] which had

[1] See URL: `http://www.eslea.uklight.ac.uk/`
[2] See URL: `http://www.uklight.ac.uk/`
[3] See URL: `http://egee-jra4.web.cern.ch/EGEE-JRA4/`

developed similar software for making reservations of guaranteed bandwidth between two end-hosts using QoS (Quality of Service) mechanisms operating in conventional networks. This architecture is general, robust, flexible and extensible.

The route of the circuit between the two end-hosts is considered to comprise three segments: the local campus network to which the local end-host is connected; the WAN (Wide Area Network) connecting the local campus network to the campus network where the remote host is located and the remote campus network to which the remote host is connected. The EGEE BAR architecture (see Figure 1) is distributed and has separate components for controlling each of these segments. In outline the architecture is as follows. Requests to create a new reservation are initiated by some client software, which can be either an interactive 'reservations browser' driven by a user or some piece of Grid middleware (usually called 'higher layer middleware' or HLM) acting autonomously. To make a reservation the client invokes the top-level component of the architecture, which is also called a BAR. The BAR component does not itself configure any network hardware. Rather, it controls subsidiary components which configure the hardware. It also maintains a database of existing reservations.

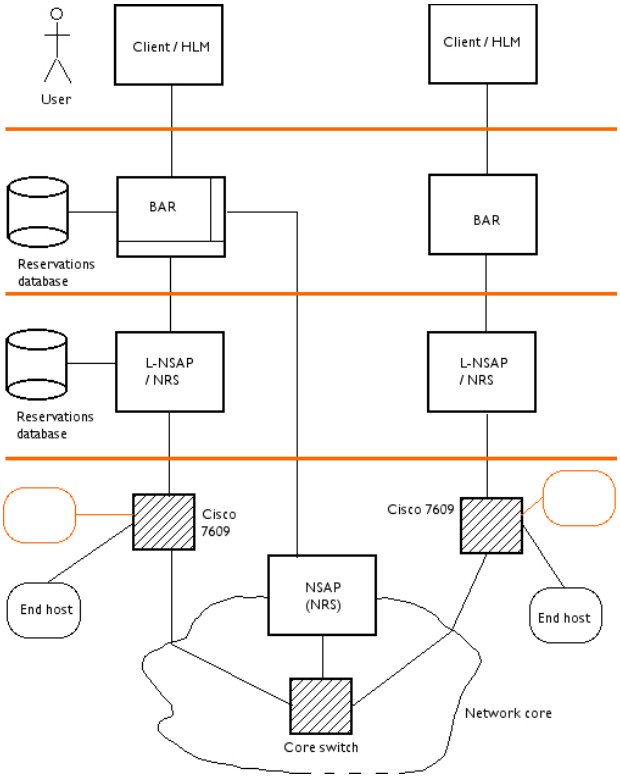


Figure 1: The CRS architecture

When a request to make a new reservation is received the BAR asks the individual subcomponents whether they can handle the request. If all can do so the reservation is accepted and added to the database. If the reservation cannot be accepted the user is informed. The individual components

are the L-NSAP and the NSAP. The L-NSAP (Local Network Services Access Point) configures the local campus network, and the NSAP (Network Services Access Point) configures the WAN.

The BAR directly invokes the L-NSAP that configures the local campus network, but it does not directly invoke the L-NSAP that configures the remote end-site. Rather, it communicates with the BAR responsible for the remote end-site, which in turn invokes its own L-NSAP.

3. The CRS Implementation

Each of the CRS components (see Figure 1) is implemented as a Web service written in Java and running on a Linux PC. The interfaces to the Web services and some of the code were adopted from the EGEE BAR Project. The L-NSAP and NSAP components were also not written *ab ovo*, but rather, are modified versions of the NRS (Network Reservation System)[4][4] developed by Saleem Bhatti *et al.* The CRS only uses items of external software that are usually available free of charge, such as Apache Tomcat,[5] Apache Axis[6] and the databases HSQLDB[7] and PostgreSQL.[8]

Though the CRS architecture is flexible and general, the current *implementation* is quite limited in some respects. Switched-circuit connections to the participants in the ESLEA Project were provided across the UKLight network, which does not support the automatic configuration of circuits. We connected a Cisco 7609 switch/router to the UKLight access points at suitable institutions (these machines were available from the earlier MB-NG Project[9]). The 7609s simulate the LAN connecting the end-host to the WAN. Switching inside UKLight was simulated using an additional 7609 connected so as to be topologically inside UKLight whilst remaining physically outside it. The edge 7609s were controlled by the L-NSAP components and the simulated UKLight switching by the NSAP component.

4. Discussion

The major development of the CRS is complete. It has been tested on a small test network of UKLight circuits. Trial deployments with two collaborating groups in the ESLEA Project are continuing. However, we have already demonstrated the practicality of the basic idea of configuring switched-circuit networks from an advance booking system.

Further work that we anticipate includes collaborating with the UCLP[10] (User Controlled Lightpaths) group to produce an NSAP that interfaces to the UCLP software for creating switched lightpaths across an optical network. We are also collaborating with the developers of the HARC (Highly-Available Robust Co-scheduler)[5, 6][11] co-allocation software to allow our network reservations to be automatically co-scheduled with reservations of other resources.

[4] See URL: `http://www.cs.ucl.ac.uk/staff/S.Bhatti/grs/`
[5] See URL: `http://tomcat.apache.org`
[6] See URL: `http://ws.apache.org/axis/`
[7] See URL: `http://hsqldb.org/`
[8] See URL: `http://www.postgresql.org/`
[9] See URL: `http://www.eslea.uklight.ac.uk/mb-ng`
[10] See URL: `http://www.uclp.ca/`
[11] See URL: `http://www.cct.lsu.edu/~maclaren/HARC/`

High capacity dedicated connections are likely to remain a scarce resource for at least the near future, as are the super-computers and similar devices which typically act as the sources and sinks for the data that flow through them. It seems likely that there will be a continuing need to reserve and schedule these scare and expensive resources in advance. We have demonstrated the basic viability of making advanced reservations over a circuit-switched network.

Acknowledgments

We are grateful to Saleem Bhatti and Richard Smith for discussions about NRS; to Charaka Palansuriya, Kostas Kavoussanakis, Florian Scharinger and Alistair Phipps for discussions about the EGEE BAR effort and to Stefan Zasada, Jon MacLaren and Stephen Pickles for discussions about HARC and co-scheduling.

For assistance in setting up test circuits we are grateful to UKERNA, particularly John Graham and David Tinkler, and to Sam Wilson and his colleagues of EUCS. Stephen Kershaw and Barney Garrett assisted with setting up end-hosts.

ESLEA was funded by the UK Engineering and Physical Science Research Council, with additional contributions from the Medical Research Council and the Particle Physics and Astronomy Research Council.

References

[1] C. Greenwood, V. Bartsch, P. Clarke, P. Coveney, C. Davenhall, B. Davies, M. Dunmore, B. Garrett, M. Handley, M. Harvey, R. Hughes-Jones, R. Jones, M. Lancaster, L. Momtahan, N. Pezzi, S. Pickles, R. Pinning, A. Simpson, R. Spencer, R. Tasker, *Exploitation of Switched Light Paths for e-Science Applications (ESLEA)*, in S.J. Cox, D.W. Walker (eds), *Proceedings of the UK e-Science All Hands Conference 2005*, Engineering and Physical Sciences Research Council (2005), CD-ROM.

[2] A. Farrel, I. Bryskin, *GMPLS: Architecture and Applications*, Morgan Kaufmann, Amsterdam (2006).

[3] A.C. Davenhall, P.E.L. Clarke, N. Pezzi, *The ESLEA Control Plane Software*, in S.J. Cox, D.W. Walker (eds), *Proceedings of the UK e-Science All Hands Conference 2005*, Engineering and Physical Sciences Research Council (2005), CD-ROM.

[4] M. Rio, A. di Donato, F. Saka, N. Pezzi, R. Smith, S. Bhatti, P. Clarke, *Quality of Service Networking for High Performance Grid Applications*, *Journal of Grid Computing*, **1** (2003) 329-343.

[5] J. MacLaren, *Co-allocation of Compute and Network resources using HARC*, in proceedings of *Lighting the Blue Touchpaper for UK e-Science: closing conference of ESLEA Project*, PoS(ESLEA)015 (2007); these proceedings.

[6] J. MacLaren, M. McKeown, S. Pickles, *Co-Allocation, Fault Tolerance and Grid Computing*, in S.J. Cox (ed), *Proceedings of the UK e-Science All Hands Conference 2006*, Engineering and Physical Sciences Research Council (2006), CD-ROM.

PROCEEDINGS OF SCIENCE

Co-allocation of Compute and Network resources using HARC

Jon MacLaren[*†]
E-Science NorthWest (ESNW), University of Manchester, Oxford Road, Manchester M13 9PL
E-mail: jon.maclaren@manchester.ac.uk

HARC—the Highly-Available Resource Co-allocator—is a system for reserving multiple resources in a coordinated fashion. HARC can handle multiple types of resource, and has been used to reserve time on supercomputers distributed across a nationwide testbed in the United States, together with dedicated lightpaths connecting the machines. HARC makes these multiple allocations in a single atomic step; if any resource is not available as required, then nothing is reserved. To achieve this "all or nothing" behavior, HARC treats the allocation process as a Transaction, and uses a phased commit protocol. The Paxos Commit protocol to ensure that there is no single point of failure in the system, which, if correctly deployed, has a very long Mean Time To Failure.

Here we give an overview of HARC, and explain how the current HARC Network Resource Manager (NRM) works, and is able to set-up and tear-down dedicated lightpaths.

Lighting the Blue Touchpaper for UK e-Science - Closing Conference of ESLEA Project
March 26-28 2007
The George Hotel, Edinburgh, UK

[*]Speaker.

[†]Special thanks to Mark Mc Keown who provided the initial design for HARC, while at the University of Manchester.

PoS(ESLEA)016

1. Motivation

The ever-improving availability of high-bandwidth, low-latency optical networks promises to enable the use of distributed scientific applications [7, 3] as a day-to-day activity, rather than simply for demonstration purposes. However, in order to enable this transition, it must also become *simple* for users to reserve *all* the resources the applications require.

The reservation of computational resources can be achieved on many supercomputers using advance reservation, now available in most commercial and research schedulers. However, distributed applications often require guaranteed levels of bandwidth between compute nodes, or between compute nodes and a visualization resource. At the network level there are switches and routers that support the bandwidth allocation over network links, and/or the configuration of dedicated end-to-end lightpaths. These low-level capabilities are sufficient to support the development of prototype middleware solutions that satisfy the requirements of these applications.

However, the development of an booking system for network resources is not a complete solution, as the user is still left with the complexity of coordinating separate booking requests for multiple computational resources with their network booking(s). Even if there is a single system available that can reserve all the required compute resources, such as Moab, or GUR (which can reserve heterogenous compute resources), this does not address the need to coordinate the scheduling of compute and network resources—a co-allocation system that can deal with multiple *types* of resources is required.

2. HARC: The Highly-Available Resource Co-allocator

HARC, the Highly-Available Resource Co-allocator [2, 8], is an open-sourced system that allows users to reserve multiple distributed resources in a single step. These resources can be of different types, e.g. supercomputer time, dedicated network connections, storage, the use of a scientific instrument, etc. Currently, HARC can be used to book High-Performance Computing resources, and lightpaths across certain GMPLS-based networks with simple topologies. The HARC Architecture is shown in Figure 1.

HARC uses a phased commit protocol to allow multiple resources to be booked in an all-or-nothing fashion (i.e. atomically). Paxos Commit [6] is used, rather than the classic 2-Phase Commit (2PC), to avoid creating a single point of failure in the system. Paxos Commit replaces 2PC's single Transaction Manager (TM) with a number of processes, or *Acceptors*, which perform the same function as the TM. The Paxos Consensus algorithm guarantees consistency, so clients can talk to any Acceptor to find the results of their requests. The overall system functions normally provided a majority of Acceptors remain in a working state. This gives a deployed system of five Acceptors a far longer Mean Time to Failure than that of any single Acceptor.

HARC is designed to be extensible, and so new types of Resource Manager can be developed without requiring changes to the Acceptor code. This differentiates HARC from other co-allocation solutions. The assumption is that the underlying resource has a scheduler capable of reserving the resource (or part thereof) for a specific user; the RM should be a small piece of code that interacts with this scheduler on the user's behalf.

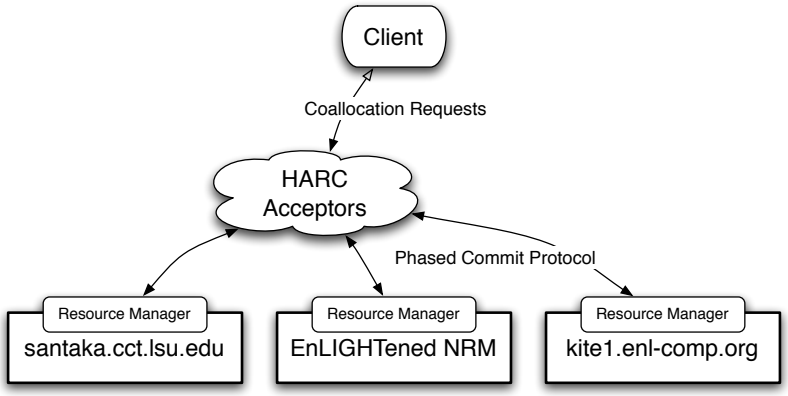


Figure 1: The HARC architecture, showing the relationship between the client, the Acceptors, and the Resource Managers (RMs).

3. Reserving Network Connections using HARC

How best to reserve network connectivity in advance is still a research topic. When the EnLIGHTened Computing project [1] started, there were deployed reservation systems such as the G-lambda project's GNS-WSI2 [5] and EGEE's BAR [9]. However, the project chose to implement a new, simple, timetable-based system, which was embedded in a HARC Resource Manager; this component is referred to as the HARC Network Resource Manager (NRM). There is a single HARC NRM for the entire testbed (centralized).

A schematic for the EnLIGHTened Computing testbed is shown in Figure 2. At the heart of the network are three Calient Diamondwave PXCs (UO1 in Chicago, UO2 in Raleigh, and UO3 in Baton Rouge) and a single Diamondwave PX (UO4 in Caltech). The software on the switches supports GMPLS, and connections across the testbed can be initiated by sending a TL1 command to a switch at either end of the connection. A dedicated lightpath can be set-up between any two entites at the edge of the cloud. These are either routers (UR2, UR3) or compute nodes (RA1, VC1, CH1); the router X1U is a special case, used to connect through to the Japanese JGN II network. All links in the network are 10 Gigabit Ethernet (10 GE), except for the connections to Japan, which are Gigabit Ethernet.

The NRM accepts requests for network connections on a first-come, first-served basis. Requests specify the start and end points of the connection (using the three letter acronyms shown on Figure 2), as well as the required bandwidth, and also the desired setup and teardown times. The example in Figure 3 would be used to request a lightpath between two supercomputers at MCNC in Raleigh and CCT in Baton Rouge.

Typically, GMPLS chooses the best path through the network when a path is set up, dynamically avoiding non-functioning components. However, when scheduling links in advance, it is important for the scheduler to be in control of the routes that each lightpath uses, to ensure that all paths follows the schedule. In the current EnLIGHTened testbed network, for any two endpoints, there is only a single possible path through the network. Even though this is the case, the NRM does specify the full path to the switches during the provisioning process, in an Explicit Route Object (ERO), which is sent as part of the TL1 command.

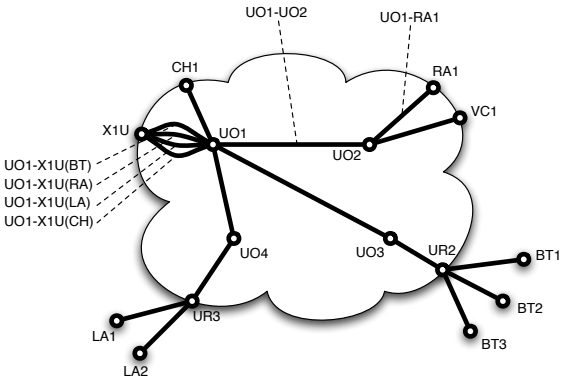


Figure 2: A simplified schematic of the EnLIGHTenend testbed network.

```
<Schedule><TimeSpecification><Exact>
   <StartTime>2007-04-25T21:00:00Z</StartTime>
   <EndTime>2007-04-25T22:00:00Z</EndTime>
</Exact></TimeSpecification></Schedule>
<Work>
   <Path>
      <From>RA1</From><To>BT2</To>
      <BandwidthMbs>10240</BandwidthMbs>
   </Path>
</Work>
```

Figure 3: XML Snippet from a HARC NRM Message.

3.1 The Future of the NRM

The current HARC NRM needs to be split into two components: a pure scheduling component with a service interface, and a much smaller HARC NRM component, which simply becomes an interface between the HARC Acceptors and the network scheduling service.[1] This is consistent with the other HARC Resource Managers that have been developed, as explained in Section 2.

The internals of the current scheduling code are also very simple. Although the network topology has not been hardcoded into the service in any way (all configuration is obtained from a set of files), there is still an assumption that given two endpoints, there is a single path through the network. Soon the EnLIGHTened testbed will be extended with a Calient Diamondwave PXC in Kansas City, creating additional paths between most endpoints; additional code will be required to deal with this correctly.

The NRM also needs to be able to cope with both planned and unplanned downtime of parts of the network, and—where possible—should ensure that users are not permitted to schedule lightpaths for times when the network is not going to be available. This will involve some level of integration between the NRM and relevant monitoring software.

4. Conclusions

There are two deployments of HARC in use today: the EnLIGHTened testbed in the United States; and a second on NorthWest Grid,[2] a regional Grid in England. A trial deployment is planned for TeraGrid,[3] and HARC is being evaluated for deployment on the UK National Grid Service.[4] An alternate Network Resource Manager that interfaces to the ESLEA Circuit Reservation Software [4] is also being considered. This would allow HARC to be used to co-allocate parts of the UK Lite network.

The prototype HARC Network Resource Manager component, described in Section 3, has been used to schedule some of the optical network connections being used to broadcast Thomas

[1] The G-lambda project's GNS-WSI2 [5] interface is currently being evaluated for its suitability for this task.

[2] http://www.nw-grid.ac.uk/

[3] http://www.teragrid.org/

[4] http://www.ngs.ac.uk/

Sterling's HPC Class from Louisiana State University.[5] Previously, HARC was used in the high-profile EnLIGHTened/G-lambda experiments at GLIF 2006 and SC'06, where compute resources across the US and Japan were co-allocated together with end-to-end optical network connections.[6]

Although these early successes are encouraging, if the advance scheduling of lightpaths is to become a production activity, then the network scheduling service(s) need to be properly integrated with the other control/management plane software to ensure that these activities do not interfere with the pre-scheduled lightpaths (and vice-versa).

Acknowledgements

The implementation of HARC took place while the author was employed at the Center of Computation & Technology at Louisiana State University. During this time, the work was supported in part by the National Science Foundation "EnLIGHTened Computing" project [1], NSF Award #0509465.

References

[1] EnLIGHTened Computing: Highly-dynamic Applications Driving Adaptive Grid Resources [Online]. http://www.enlightenedcomputing.org.

[2] HARC: The Highly-Available Resource Co-allocator [Online]. http://www.cct.lsu.edu/∼maclaren/HARC.

[3] R. J. Blake, P. V. Coveney, P. Clarke, and S. M. Pickles. The teragyroid experiment–supercomputing 2003. *Scientific Computing*, 13(1):1–17, 2005.

[4] A. C. Davenhall, P. E. L. Clarke, N. Pezzi, and L. Liang. The ESLEA Circuit Reservation Software. In *Proceedings of "Lighting the Blue Touchpaper for UK e-Science: closing conference of ESLEA Project"*. PoS(ESLEA)015, 2007.

[5] G-lambda Project. Grid Network Service / Web Services Interface, version 2. http://www.g-lambda.net/wordpress/?page_id=19.

[6] J. Gray and L. Lamport. Consensus on transaction commit. *ACM TODS*, 31(1):130–160, March 2006.

[7] A. Hutanu, G. Allen, S. D. Beck, P. Holub, H. Kaiser, A. Kulshrestha, M. Liška, J. MacLaren, L. Matyska, R. Paruchuri, S. Prohaska, E. Seidel, B. Ullmer, and S. Venkataraman. Distributed and collaborative visualization of large data sets using high-speed networks. *Future Generation Computer Systems. The International Journal of Grid Computing: Theory, Methods and Applications*, 22(8):1004–1010, 2006.

[8] J. MacLaren, M. M. Keown, and S. Pickles. Co-Allocation, Fault Tolerance and Grid Computing. In *Proceedings of the UK e-Science All Hands Meeting 2006*, pages 155–162, 2006.

[9] C. Palansuriya, M. Büchli, K. Kavoussanakis, A. Patil, C. Tziouvaras, A. Trew, A. Simpson, and R. Baxter. End-to-End Bandwidth Allocation and Reservation for Grid applications. In *Proceedings of BROADNETS 2006*. http://www.x-cd.com/BroadNets06CD/pdfs/87.pdf, October 2006.

[5] This class is the First Distance Learning Course ever offered in Hi-Def Video. Participating locations include other sites in Louisiana, and Masaryk University the Czech Republic. See http://www.cct.lsu.edu/news/news/201

[6] See http://www.gridtoday.com/grid/884756.html

PROCEEDINGS OF SCIENCE

The Application Hosting Environment: Easing the Scientist's Access to the Grid

P. V. Coveney
University College London
E-mail: p.v.coveney@ucl.ac.ukl

R. S. Saksena
University College London
E-mail: r.saksena@ucl.ac.ukl

S. J. Zasada*
University College London
E-mail: stefan.zasada@ucl.ac.ukl

Current grid computing technologies have often been seen as being too heavyweight and unwieldy from an end user's perspective, requiring complicated installation and configuration steps to be taken that are too time consuming for most end users. This has led many of the people who would benefit most from grid technology, namely computational scientists, to avoid using it. In response to this we have developed the Application Hosting Environment, a lightweight, easily deployable environment designed to allow the scientist to quickly and easily run unmodified applications on distributed grid resources. This is done by building a layer of middleware on top of existing technologies such as Globus, and exposing the functionally as web services using the WSRF::Lite toolkit. The scientist can start and manage an application via these services, with the extra layer of middleware abstracting the details of the particular underlying grid resource in use. The scientist can concentrate on performing their scientific investigations, rather than learning how to manipulate the underlying grid middleware.

Lighting the Blue Touchpaper for UK e-Science - Closing Conference of ESLEA Project
March 26-28 2007
The George Hotel, Edinburgh, UK

*Speaker.

PoS(ESLEA)017

1. Introduction

We define grid computing [1] as distributed computing conducted transparently across multiple administrative domains. Fundamental to the inter-institutional sharing of resources in a grid is the grid middleware, that is the software that allows an institution to share its resources in a seamless and uniform way. While many strides have been made in the field of grid middleware technology [2, 3], the prospect of a heterogeneous, on-demand computational grid as ubiquitous as the electrical power grid is still a long way off. Part of the problem has been the difficulty to the end user of deploying and using many of the current middleware solutions, which has led to reluctance amongst many scientists to actively embrace grid technology [4].

Many of the current grid middleware solutions can be characterised as what we describe as 'heavyweight', that is they display some or all of the following features: the client software is difficult to configure or install; the middleware is dependent on lots of supporting software being installed and requires non-standard ports to be opened within firewalls. To address these deficiencies there is now much attention focused on 'lightweight' middleware solutions, such as [5], which attempt to lower the barrier of entry for users of the grid.

2. The Application Hosting Environment

In response to the issues raised above we have developed the Application Hosting Environment (AHE)[1], a lightweight, WSRF [6] compliant, web services based environment for hosting scientific applications on the grid. The AHE allows scientists to quickly and easily run unmodified, legacy applications on grid resources, managing the transfer of files to and from the grid resource and allowing the user to monitor the status of the application. The philosophy of the AHE is based around the fact that very often a group of researchers will all want to access the same application, but not all of them will possess the skill or inclination to install the application on a remote set of grid resources. In the AHE, an expert user installs the application and configures the AHE server, so that all participating users can share the same application.

The AHE focuses on applications not jobs, with the application instance being the central entity. We define an application as an entity that can be composed of multiple computational jobs, for example a simulation that consists of two coupled models which requires two jobs to instantiate it. An application instance is represented as a stateful WS-Resource[6]. Details of how to launch the application are maintained on a central service, in order to reduce the complexity of the AHE client. The design of the AHE has been greatly influenced by WEDS (WSRF-based Environment for Distributed Simulations)[7], a hosting environment designed for operation primarily within a single administrative domain. The AHE differs in that it is designed to operate across multiple administrative domains seamlessly, but it can also be used to provide a uniform interface to applications deployed on both local machines, and remote grid resources.

The AHE is based on a number of pre-existing grid technologies, principally GridSAM [8] and WSRF::Lite [9]. WSRF::Lite is a Perl implementation of the OASIS WSRF specification. GridSAM provides a web services interface, running in an OMII [10] web services container, for submitting and monitoring computational jobs to a variety of Distributed Resource Managers

[1]The AHE can be downloaded from http://www.realitygrid.org/AHE

(DRM), including Globus [2], Condor [11] and Sun Grid Engine [12], and has a plug-in architecture that allows adapters to be written for different types of DRM. Jobs submitted to GridSAM are described using Job Submission Description Language (JSDL) [13].

The problems associated with 'heavyweight' middleware solutions described above have greatly influenced the design of the AHE. The design assumes that the user's machine does not have to have client software installed to talk directly to the middleware on the target grid resource. Instead the AHE client provides a uniform interface to multiple grid middlewares. The client machine is also assumed to be behind a firewall that uses network address translation [14]. The client therefore has to poll the AHE server to find the status of an application instance. In addition, the client machine needs to be able to upload input files to and download output files from a grid resource, but we assume it does not have GridFTP client software installed. An intermediate file staging area is therefore used to stage files between the client and the target grid resource.

The AHE client maintains no knowledge of the location of the application it wants to run on the target grid resource and should not be affected by changes to a remote grid resource, for example if its underlying middleware changes from Globus version 2 to Globus version 4. The client does not have to be installed on a single machine; the user can move between clients on different machines and access the applications that they have launched. The user can even use a combination of different clients, for example a command line client to launch an application and a GUI client to monitor it. The client must therefore maintain no information about a running application's state.

These constraints have led to the design of a lightweight client for the AHE which is simple to deploy and does not require the user to install any extra libraries or software. It should be noted that this design does not remove the need for middleware solutions such as Globus on the target grid resource; indeed we provide an interface to run applications on several different underlying grid middlewares so it is essential that grid resource providers maintain a supported middleware installation on their machines. What the design does is simplify the experience of the end user.

3. Deploying the AHE

To host an application in the AHE, the expert user must first install and configure it on the target grid resource. The expert user then configures the location and settings of the application on the AHE server and creates a JSDL template document for the application and the resource. To complete the installation the expert user runs a script to repopulate the application registry; the AHE can be updated dynamically and does not require restarting when a new application is added.

The AHE is designed to interact with a variety of different clients. We have developed both GUI and command line clients in Java and is designed to be easy to install. The GUI client uses a wizard to guide a user through launching their application instance. The wizard allows users to specify requirements for the application, such as the number of processors to use, the choice of target grid resource to run their application, staging required input files to the grid resource, specification of any extra arguments for the simulation, and job execution. Once an application instance has been prepared and submitted, the AHE client allows the user to monitor the state of the application by polling its associated WS-Resource. After the application's execution has finished,

the user can stage the application's output files back to their local machine using the client. For further discussion of AHE use cases see [15].

4. Summary

The AHE is designed to facilitate grid based computational science by extension of existing approaches to the use of high-end computing resources. Production codes, few in number but used by significant numbers of people, are installed on a set of grid enabled resources and accessed via a lightweight client which enables simple and also arbitrarily complex application workflows to be developed. We are currently working to integrate the AHE with the HARC [16] co-schedluing system, providing users with a single interface to reserve the resources that they want to use and run their applications.

References

[1] P. V. Coveney, editor. *Scientific Grid Computing*, pages 1701–2095. Phil. Trans. R. Soc. A, 2005.

[2] http://www.globus.org.

[3] http://www.unicore.org.

[4] J. Chin and P.V. Coveney. Towards tractable toolkits for the grid: a plea for lightweight, useable middleware. Technical report, UK e-Science Technical Report UKeS-2004-01, 2004. http://nesc.ac.uk/technical_papers/UKeS-2004-01.pdf.

[5] J. Kewley, R. Allen, R. Crouchley, D. Grose, T. van Ark, M. Hayes, and Morris. L. GROWL: A lightweight grid services toolkit and applications. 4th UK e-Science All Hands Meeting, 2005.

[6] S. Graham, A. Karmarkar, J Mischkinsky, I. Robinson, and I. Sedukin. Web Services Resource Framework. Technical report, OASIS Technical Report, 2006. http://docs.oasis-open.org/wsrf/wsrf-ws_resource-1.2-spec-os.pdf.

[7] P. V. Coveney, J. Vicary, J. Chin, and M. Harvey. Introducing WEDS: a Web services-based environment for distributed simulation. In P. V. Coveney, editor, *Scientific Grid Computing*, volume 363, pages 1807–1816. Phil. Trans. R. Soc. A, 2005.

[8] http://gridsam.sourceforge.net.

[9] http://www.sve.man.ac.uk/research/AtoZ/ILCT.

[10] http://www.omii.ac.uk.

[11] http://www.cs.wisc.edu/condor.

[12] http://gridengine.sunsource.net.

[13] http://forge.gridforum.org/projects/jsdl-wg/document/draft-ggf-jsdl-spec/en/21.

[14] http://www.faqs.org/rfcs/rfc1631.html.

[15] P. V. Coveney, R. S. Saksena, S. J. Zasada, M. McKeown, and S. Pickles. The application hosting environment: Lightweight middleware for grid-based computational science. *Computer Physics Communications*, 176(6):406–418, 2007.

[16] J. Maclaren, M. Mc Keown, and Pickles, S. Co-allocation, Fault Tolerance and Grid Computing. 5th UK e-Science All Hands Meeting, 2006.

PROCEEDINGS OF SCIENCE

Building a distributed software environment for CDF within the ESLEA framework

Valeria Bartsch[*], **Mark Lancaster, Nicola Pezzi**
University College London
E-mail: bartsch@fnal.gov

A fast optical link (UKLight/StarLight) between UCL and the Fermi National Accelerator Laboratory (FNAL) in Chicago was established in 2004. It has been used by the CDF collaboration to send data between the USA and the UK at rates in excess of 500 Mb/sec and forms part of the CDF data-handling infrastructure. The issues involved in setting up the link and the CDF data handling system are described. The distributed grid environment and the problems encountered in establishing a data analysis environment utilising the link are also briefly described.

Lighting the Blue Touchpaper for UK e-Science - Closing Conference of ESLEA Project
March 26-28, 2007
Edinburgh

[*]Speaker.

PoS(ESLEA)020

1. Introduction

ESLEA is an EPSRC funded project which is seeking to establish a proof-of-principle demonstration of the utility of dedicated, guaranteed bandwidth light-paths in applications needing to transfer large volumes of data across Wide Area Networks (WANs). The ESLEA project utilises the UKLight switched circuit optical network to enable guaranteed high bandwidth network links for a range of eScience applications. The CDF experiment [1] is one such application endeavouring to exploit the potential of high speed optical network links.

CDF is a particle physics experiment trying to elucidate the fundamental nature of matter. It is presently taking data from proton anti-proton collisions at the Tevatron collider at FNAL in Chicago. Analysis of 2 Pb of raw data is underway by almost 800 physicists located at 61 institutions in 13 countries across 3 continents. The amount of raw data and the need to produce secondary reduced datasets have required new approaches to be developed in terms of distributed storage and analysis. Grid systems based on DCAF [2] and SAM [3] are being developed with the aim that 50% of CDF's CPU and storage requirements will be provided by institutions remote from FNAL. In order to effectively utilise this distributed computing network it is necessary to have high speed point-to-point connections, particularly to and from FNAL which have a bandwidth significantly higher than commonly available. To this end, as part of the ESLEA project, the use of a dedicated switched light path from FNAL to UCL was investigated. However in order to utilise the data sent to UCL from FNAL it was necessary to deploy the CDF data handling system at UCL and also make the CPU resources available to the rest of the experiment as part of a distributed grid. The issues associated with setting up this infrastructure, which necessarily formed a large part of the project, are also described.

2. CDF/ESLEA Objectives

Figure 1 shows the data flow of the CDF experiment. Raw data from the detector is accumulated at a rate of 2TB/day. This raw data is then *re-processed* which typically involves calibrating the data and applying more sophisticated algorithms to the data such that it can be utilised in physics analyses; this is termed *reconstructed* data. In general a given user then analyses a subsample of this reconstructed data and this analysis is facilitated by comparing the data to simulated or *Monte-Carlo (MC)* generated data. The generation of MC data is CPU intensive and typically performed on grid farms with the output data being stored centrally at FNAL. The re-processing of data for CDF takes place at FNAL. The user-defined datasets need to be made available to users across many institutes and in general they are distributed across different storage locations both within and outside FNAL. The two key objectives of the CDF/ESLEA project were:

- To generate a significant fraction ($\sim$ 10%) of CDF's MC need using grid farms at UCL and use UKLight/StarLight to send the generated MC data to FNAL.
- To make UKLight/StarLight available to users for the rapid transfer of user datasets from FNAL to the UK and then subsequently to German and Italian institutes.

To achieve these objectives, in addition to commissioning the UKLight/StarLight link, it was necessary to modify and port CDF's bespoke data-handling and grid-interface software to UCL so that it could be utilised within the context of the CERN LHC computing grid (LCG) [4] which remains the only supported particle physics grid system at UCL. In the following section, we briefly describe CDF's data-handling and grid interface infrastructure before describing the commissioning and usage we made of the UKLight/StarLight link.

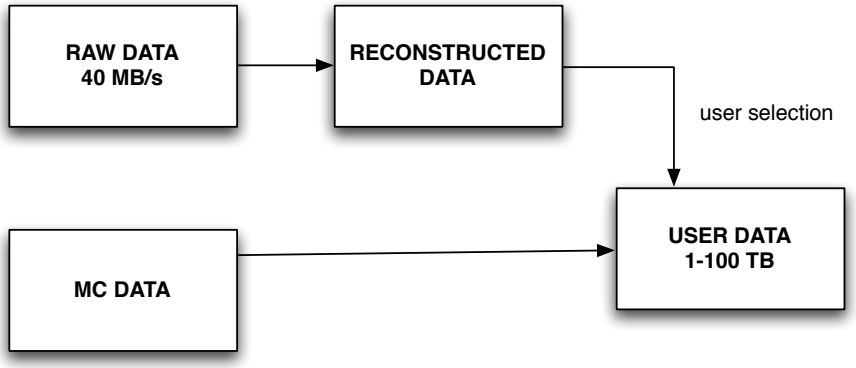


Figure 1: *The data flow and types of data utilised by the CDF experiment.*

3. CDF's Data Handling and Grid Analysis Systems

Transparent access to CDF's data is provided through a custom data-handling and cataloguing system called SAM [3]. SAM is used to store, manage, deliver and track the processing of all data. Each storage location is registered with a central SAM ORACLE database at FNAL and has an associated SAM server/station managing the local data. Users, through CORBA enabled SAM-clients, make requests for data and the data is delivered to a temporary local cache from the appropriate SAM server(s). In this way physicists have transparent access to the CDF data. At present $\sim$ 500 Tb of data is distributed across SAM servers in the US, Europe and Asia. For a given user it is most efficient if the majority of the data of interest to them resides at a local SAM station. The rapid population of a local SAM station with datasets of interest to UK physicists through UKLight/StarLight was one of the key aims of the project. Much of the development work to establish SAM as CDF's default data-handling system was carried out in the UK. Figure 2 shows the amount of data transferred through the SAM system per month for selected SAM stations, including the UCL station that was connected to FNAL via UKLight/StarLight.

Traditionally within CDF the analysis of data was done on dedicated resources of commodity nodes at FNAL managed as Condor [5] pools with a wrapper of CDF specific software around the batch system. This so-called CAF [2] system though has limitations within the context of farms at universities which are not easily customised and which, particularly in Europe, only have an LCG interface to the local batch system. However, the maintenance and monitoring of the bespoke scheduling and job submission software within the CAF system requires a significant manpower

commitment which was not available within this project. Furthermore, the requirements of the CAF system that one node be outside of the firewall were not compatible with local internet security restrictions. It was therefore not possible to mimic the FNAL CAF system at UCL. We therefore devoted considerable time in evaluating and developing a more tractable solution that could make UCL CPU resources available via UKLight/StarLight utilising LCG and the SAM data handling system as opposed to deploying the bespoke CDF CAF system.

We successfully established user authentication with FNAL servers (via kerberos) and LCG servers (via grid certificates and a CDF VO) such that we could utilise the LCG resources at UCL using the gLite [6] interface. However within the time of the project we were unable to make the resources available for specific CDF analysis work owing to an instability in the software (PARROT/SQUID) serving the bespoke CDF analysis software to the batch nodes and the lack of an LCG compliant SRM interface within SAM. These issues are currently being worked on and expected to be resolved before the end of 2007.

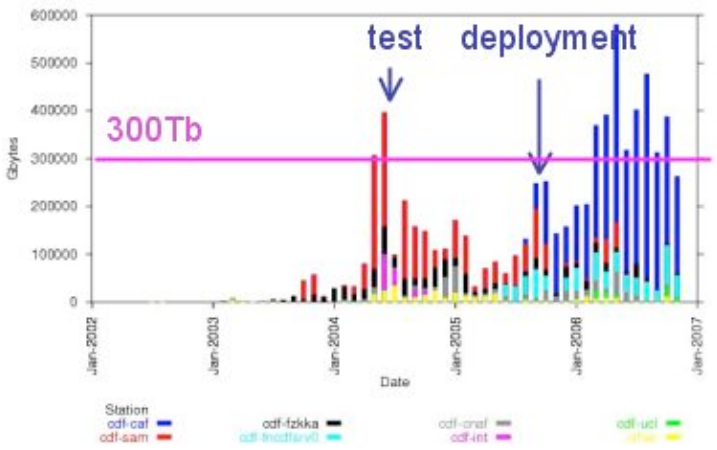


Figure 2: *Data read by the major SAM stations, cdf-caf and cdf-sam are located at FNAL, cdf-cnaf, cdf-ucl and cdf-fzkka are located in Europe.*

4. Data Transfers For CDF using UKLight/StarLight

Over the course of this project there has been a general shift of CPU resources to institutes remote from FNAL and now 50% of CPU resources are remote. The de-centralisation of the CPU has also required a similar migration of the data which in order to not hamper analysis progress has had to utilise the fastest available networks such as UKLight/StarLight. Typical CDF secondary datasets that are used for physics analyses are presently 1-100 Tb in size. Typical transfer rates from FNAL to Europe (UCL) using the standard network are approximately 25 Mb/sec (for multiple streams). A 50 Tb dataset would thus take approximately 6 months to copy from FNAL. This is comparable to the entire time that a CDF physicist would spend analysing the data in order

to produce a publication. CDF produces in excess of 40 publications per annum. The datasets themselves are typically distributed in many files, each approximately 1 GB in size. To be useful to CDF, we required that UKLight/StarLight:

- Deliver a throughput of $>$ 500 Mb/sec such that typical datasets could be made available on a timescale of a week.
- Deliver files without corruption.

Since real-time analysis of a data-stream is not undertaken, a modest retransmission rate of files/packets at the 10% level was acceptable and the order in which files were received was not critical.

A dedicated 1 Gb/s circuit connecting UCL and FNAL utilising UKLight/StarLight infrastructure was setup late in 2004 and ultimately satisfied our requirements with transfer rates above 500 Mb/sec sustained over several days. However this was only achieved after incremental modifications to the hardware and software configurations that we briefly describe below.

In the initial period of the project, the link was unavailable for several months due to technical problems with switches and its prioritised use in demonstrations at conferences. We utilised this downtime to optimise the performance of the dedicated UCL PC connected to the optical network. It was clear from initial tests using "iperf" and "dd" that the PC was not configured in an optimal way for transferring large data volumes across the network and for writing this data to disk. The kernel TCP settings were modified [7], the file-system was re-formatted as an XFS file-system and parameters of the 7.5 Tb SCSI RAID-0 array were modified [8]. After these modifications we were able to make memory to memory transfers between PCs on the UKLight network at 950 Mb/sec and write to disk at 750 Mb/sec.

Ultimately we were interested in the disk to disk transfer rate via gridFTP [9] (as implemented in the SAM software package) from the CDF data pools at FNAL to the disk on the UCL UKLight PC. Initial tests could not deliver a throughput higher than 250 Mb/sec. This low rate was due to three factors:

- A switch within the CDF/FNAL network that was not appropriate for the network.
- Concurrent disk access from other CDF users at FNAL.
- Non-optimal gridFTP settings.

It also became clear that it was extremely important in diagnosing problems of the first type to have a good and regular communication channel with the FNAL experts and people controlling the FNAL hardware. Redundancy in the hardware and control over all steps of the network was vital in achieving a reasonable throughput. This control was ultimately only achieved by establishing a good working relation with the FNAL/StarLight experts over a period of time through regular presentations at their weekly network meetings. Redundancy at the UCL end and the ability to support multiple ftp streams was achieved by adding a second identically configured PC to the UKLight/StarLight network at UCL. Our highest throughput was achieved using 20-25 parallel

gridFTP transfers concurrently to two UCL PCs. Typical transfer rates [10] are shown in Figure 3 which shows peak rates in excess of the required 500 Mb/s. The reductions are due to disk contention from other users accessing the same CDF/FNAL disks that we were copying data from.

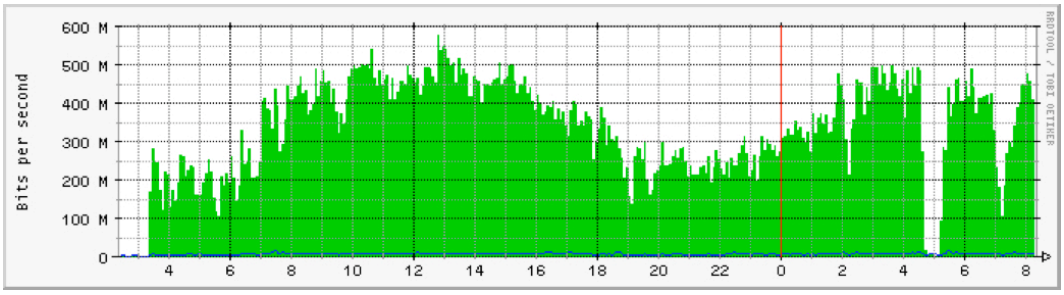


Figure 3: *Typical data transfer rates achieved across UKLight/StarLight over a 24-hour period between the CDF/FNAL storage and the UCL storage.*

5. Summary

The key objective to transfer files from CDF/FNAL to the UK over UKLight/StarLight at a throughput in excess of 500 Mb/s was achieved. It was however not possible to make the UCL CPU resources available to CDF via UKLight/StarLight due to problems encountered in deploying the CDF analysis and data-handling software within the supported LCG environment. As a proof-of-principle the project illustrated that a fast optical network can be deployed within the context of a running high energy physics experiment and deliver the required bandwidth. However the issues associated with trying to deploy highly specialised and custom software from a running experiment within a generic environment such as LCG need further work before the network could be used in a production capacity.

References

[1] The CDF experiment, `http://www-cdf.fnal.gov`

[2] DCAF, Decentralised Analysis Farm for CDF, `http://cdfcaf.fnal.gov`

[3] SAM, Sequential Access to Metadata, see CHEP04 conference contribution, `http://projects.fnal.gov/samgrid/conferences/chep04/chep04.html`

[4] LCG, `http://lcg.web.cern.ch/LCG`

[5] Condor project homepage, `http://www.cs.wisc.edu/condor`

[6] gLite, Ligthweight Middleware for Grid Computing, `http://glite.web.cern.ch/glite`

[7] Modifications were made to 4 files.
/etc/sysctl.conf :

```
kernel.core_uses_pid = 1
vm.max-readahead = 2048
vm.min-readahead = 1024
```

```
kernel.shmmax = 1073741824
```
/proc/sys/net/core/wmem_max :
```
8388608
```
/proc/sys/net/core/rmem_max :
```
8388608
```
/proc/sys/net/ipv4/tcp_rmem:
```
4096 87380 4194304
```

[8] The parameters of the RAID array were modified using "dellmgr" to:
```
Stripe Size = 128kb
Write Policy = WRBACK
Read Policy = READAHEAD
Cache Policy = CACHED_IO
```

[9] gridFTP,
`http://www.globus.org/alliance/publications/papers/GFD-R.0201.pdf.`

[10] Network performance for the UKLight/StarLight was monitored from the URL:
`http://tinyurl.com/2ejd2a`

PROCEEDINGS OF SCIENCE

IS Security in a World of Lightpaths

Robin Tasker
Science and Technology Facilities Council, Daresbury Laboratory
Warrington, Cheshire WA4 4AD, UK
E-mail: r.tasker@dl.ac.uk

IS Security is a cornerstone for the delivery of consistent and reliable services in every aspect of the business of an organisation. The traditional IP network service provided to Institutes is carefully managed and controlled to limit illegal and/or antisocial use to protect the business processes of that Institute. SuperJANET5 has the capability for additional bandwidth circuits - lightpaths - to be provided between specific endpoints across the network to meet specific need. Because these are end-to-end circuits they reach right into the heart of an organisation, typically providing a high bandwidth interconnection, and often at rates that are difficult to police. This paper explores this problem space and provides a strategy to minimise any associated risk through the development of an appropriate Security Policy that can sit alongside an Institute's overall approach in this area.

.

Lighting the Blue Touchpaper for UK e-Science - Closing Conference of ESLEA Project
The George Hotel, Edinburgh, UK
26-28 March, 2007

PoS(ESLEA)022

1. Lightpaths and IS Security - The Problem Stated

Organisations typically invest considerable resources into the provision of Information Systems (IS) Security to protect their core business operation from the many threats posed by their connection to the wider Internet. This provision will be specified in the IS Security Policy of the organisation which should provide clear instruction as to what is, and what is not, permitted with respect to any activity that make use of the IS of the organization [1]. To that end employees will almost certainly have signed their acceptance of these rules of usage as a condition of their employment. In this respect an employee's access to IS resources is conditional upon this agreement.

An IS Security Policy is normally realised through a range of technical solutions which manages the interaction between IS within the organisation and IS located remotely and beyond the jurisdiction of the organisation. The most common of such solutions is the firewall which through its associated rule-set determines what is, and is not, allowed to pass between the trusted organisational network and the untrusted generalised Internet. There are other measures that are commonly used that taken together provide layered "onion skins" of measures to protect the organisation.

The availability of lightpaths – in reality end-to-end circuits provided in addition to, and typically at rates at least equal with, the general commodity provision – challenge the IS Security *status quo*. Such provision will necessarily reach right into the heart of an organisational network with the potential to bypass any or all of the existing security measures in place.

Further, such lightpaths will be provisioned to allow collaborations across communities to develop and thrive. These virtual organisations (VO) and the individuals they represent cannot be bound by the IS Security Policy of any one single organisation. A member of a VO might expect to be bound by the IS Security policy of the VO but of course that person will physically reside within a home organisation where that organisation's IS Security Policy will have primacy [2].

It is clear that this model of operation will increasingly become the norm and there will be an expectation that the potential for conflict outlined here will be routinely handled to everyone's satisfaction.

2. Developing a Lightpath IS Security Policy

The purpose of developing a distinct IS Security Policy for a lightpath is to mitigate those risks associated with the delivery of the service associated with the lightpath. In essence the risks to service are no different from any other production network [3]. There are however two significant additions to this conventional perspective. Firstly the exceptional data rates expected over such a network means that conventional security access devices may not provide sufficient or adequate protection; and secondly, the lightpath network is designed to closely couple administratively distinct institutes to deliver the service that it carries. In setting the scene, the

IS Security Policy should address the purpose of the lightpath network, any associated assumptions that are made, and its intended audience [4].

2.1 Policy Purpose

As with any other IS Security policy it is good practice to state the purpose of the service for which the policy provides protection in terms of what it does and why that makes a difference.

2.2 Policy Assumptions

For lightpath networks the assumptions described in the lightpath IS Secuirty policy are crucial because almost certainly there will be a statement that each site that connects to the lightpath network will take its own view on what is and is not acceptable with respect to Information Security. That is to say the site IS Security policy will take precedence over the lightpath IS Security policy. Furthermore there will almost certainly be the expectation that the lightpath IS Security policy does not supersede or invalidate any local IS policies at any local site, and should the lightpath policy conflict with any local site policy then the local site policy will take precedence for that site. It is reasonable to assume that each site will assess suitability of access to the lightpath based upon the specification and implementation of its own local IS policy. The collorally to this assumption is that sites will only be allowed access to the lightpath network once they have agreed to follow the lightpath IS Security policy.

2.3 Policy Scope

Of crucial importance is a clear detail of the scope of the lightpath IS Security policy with respect to the set of rules which govern the right to transmit or receive data across the lightpath network. Certainly the better the specification of the data flows across the lightpath network, the more precise can be the specification of those rules.

The Scope should also provide a statement that each site ensures that any traffic for which it is responsible is generated in accordance with the lightpath IS Security policy. Further the list of sites - the members - that are authorised to make use of the lightpath network must be clearly specified.

Where a site does not agree to implement the lightpath IS Security policy, all other sites connected to that network and that have agreed to the policy may reject all transmissions from the site and in that manner protect themselves from some undefined or unspecified risk. Where a site attempts to use the lightpath network in a manner beyond its declared and agreed purpose any traffic resulting from such usage may be discarded by any member site without warning or notification.

2.4 Governance - Roles and Responsibilities

It is important that for each member site, a security contact is nominated and advertised and it is that person's responsibility to engage with the local site IS Security officer in all matters relating to the use of the lightpath network at that site. Furthermore it is to be expected that the local site IS Security officer at each site will be satisfied with the mitigation of any

information security risk associated with that site's connection to the specified lightpath network. This mitigation is achieved through the implementation of the lightpath IS Security policy and the nominated representatives will be responsible for all necessary on-site liaisons with the local site to obtain a formal record from the local site's IS Security officer of acceptance and implementation of this policy.

Changes to the lightpath IS Security policy must be discussed and agreed by the nominated representatives with any resulting operational changes taking place only at specified advertised times once agreement has been reached.

2.5 Governance - Legislation and Compliance

It is reasonable to expect that each site will act in accordance with any national or international legislation applicable in that country to the operation of a data network. Further the nominated representatives should be expected to ensure that the member sites are aware of any such matter that bears upon the operation of the network..

The nominated representatives might be expected to work with the local site IS Security officer to demonstrate compliance with the lightpath IS Security policy with the output from such a review being shared with the other nominated representatives to ensure broad dissemination.

2.6 Technical Considerations

An IS Security policy will contain very specific technical detail which provides the guidance on how the policy is to be delivered. The detail is clearly beyond the scope of this paper and will most certainly vary from one policy to another but may include statements with regard to, for example, IP routing, IP protocol usage and access control.

2.7 Procedural Matters – Incident Handling and Reporting

Security incidents are never planned and take no account of convenience. It is therefore vital that an IS Security policy states clearly and concisely the action to be taken should such an incident arise. An IS Security policy for a lightpath network is no different except that it will require the engagement of the nominated representatives together with the IS Security offices for each member site so that the composite risk might properly be assessed and accommodated.

3. Conclusions

At the time of writing and within the UK academic network there are already in excess of a dozen lightpath networks in use, and with the advent of SuperJANET5 this trend will accelerate. In so doing the complexity of the inter-connections between sites will grow as a consequence the associated risks will increase. It is not clear yet whether issues of lightpath network security are taken seriously. Certainly there remains a tension between the site network service providers and users of these lightpath networks who are concerned that precious performance may be impacted by the requirements for security. There is a case to be made on both side of this debate, however it is clear that a site's IS Security policy will be dictated by the business needs of the organisation and its ability to manage risk. To do so account must be

taken of any component that increases the risk to an organisation, and an *ad hoc* lightpath network most certainly increases that risk.

This paper describes a proactive approach which seeks to balance the needs of the service provider with those of a user of a lightpath network. An open and transparent approach serves both sides best and to achieve that end clear lines of communication need to be established and exercised through the lifetime of a lightpath network.

References

[1] UCISA, Information Security Toolkit, Edition 2.0 (2005), ISBN 978-0-9550973-0-4

[2] Kelsey (ed.), Grid Security Policy, https://edms.cern.ch/428008/4

[3] UCISA, Exploiting and protecting the network, Edition 3.0 (2006), ISBN 978-0-9550973-1-2

[4] LHC Optical Private Network Information Security Policy, https://edms.cern.ch/file/708248/LAST_RELEASED

PROCEEDINGS OF SCIENCE

The Contribution of ESLEA to the Development of e-VLBI

Ralph Spencer[1], Paul Burgess, Simon Casey, Richard Hughes-Jones, Stephen Kershaw, Anthony Rushton, Matt Strong

The University of Manchester
Jodrell Bank Observatory
Macclesfield
Cheshire SK11 9DL
UK
E-mail: ralph.spencer@manchester.ac.uk

Arpad Szomoru.
Joint Institute for VLBI in Europe
Postbus 2,
7990 AA Dwingeloo,
The Netherlands
E-mail: aszomoru@jive.nl

Colin Greenwood
National e-Science Centre (NESC)
Edinburgh, EH8 9AA,
UK
E-mail: coling@nesc.ac.uk

e-VLBI - the use of the Internet in real time VLBI high resolution observations in radio astronomy - has become a routinely available technique in this last year. ESLEA has contributed significantly to its development, by improving our understanding of data transmission networks, the limitations of transport protocols and end hosts, and communication of this knowledge to radio astronomers in Europe. A series of tests, organised by JIVE in the Netherlands and ESLEA has gradually led to open call science runs, now considered as a regular part of European VLBI Operations. A major upgrade project - EXPReS - is now underway to equip more European observatories with e-VLBI capability. This paper outlines the work done in ESLEA on e-VLBI and illustrates its success by showing recently obtained astronomical results.

Lighting the Blue Touchpaper for UK e-Science - Closing Conference of ESLEA Project
The George Hotel, Edinburgh, UK
26-28 March, 2007

[1] Speaker

PoS(ESLEA)023

1. Introduction

The radio astronomy technique known as VLBI (Very Long Baseline Interferometry) achieves the highest angular resolution in astronomy by making use of telescopes situated all over the world. The technique traditionally relies on the ability to record the signals from the receivers in the telescopes on magnetic media. The first systems in regular use in the 1970s, e.g. the Mk1, used computer tape. The Mk2 system used helical scan video recorders, starting off with Ampex 2" tape systems [1]. These systems had limited bandwidth, and so the 1980s saw the development of the Mk3 using wideband longitudinal instrumentation recorders able to record at 56 Mbps. This system was later developed into the Mk4 in the 1990s- able to record in principle at rates up to 1024 Mbps, though somewhat unreliable at rates higher than 500 Mbps. More recently the advent of high capacity computer disks (the Mk5 system) now means that reliable recording at 1024 Mbps can be performed. (NB data rates inVLBI systems are in powers of 2 due to technical constraints.)

The recorded data are shipped to a central processor, where the tapes or disks are played back together, and the signals cross-correlated (in case of European VLBI this correlator is at JIVE in Dwingeloo, The Netherlands). The angular structure of radio sources can be obtained by Fourier inversion of the correlated data taken over a number of baselines. The number of Fourier components is effectively increased (and hence improved imaging fidelity) by means of Earth rotation, where the projected baseline length varies with time, tracing out part of an ellipse over ~12 hours [2], and the angular resolution is inversely proportional to the maximum length of the baseline, so trans-world arrays produce images showing the finest detail.

2. e-VLBI

e-VLBI, where the data are transferred to the correlator in real time via the Internet has been developed recently. Initial tests were at relatively low data rates, up to 32 Mbps, using production academic network packet switched connections from telescopes to JIVE, but nevertheless interesting results were obtained [3]. e-VLBI has a number of advantages, in particular the ability to check that everything is working immediately rather than the wait of several weeks or even months for tapes to be shipped and correlated. This can result in a significant increase in reliability for VLBI operations. In addition, the technique lends itself to rapid reaction observations – where subsequent observations on rapidly varying objects can be decided upon within hours rather than weeks after initial measurements are made.

The signal to noise ratio for a wideband continuum source is proportional to $\sqrt{B\tau}$, where B is the bandwidth and τ the total time on source, so bandwidth is as important as time. There is an obvious need for high data rates. Note that typically a data rate of 512 Mbps corresponds to a bandwidth of 128MHz, using 2 bit digitization (possible due to the characteristics of random signals) and a factor of 2 for the Nyquist rate. Very high data rates are in principle possible using fibre-optic technology, well in excess of that accessible in recording techniques, so e-VLBI could eventually give a new level of sensitivity in high resolution astronomy. However the data needs to be continuously streamed, and usage of standard production networks is now

such that congestion is likely to occur at rates of a few hundred Mbps. Protocols such as TCP will drop data rates dramatically even with only one packet lost [4], so though the data loss may be negligible the drop in data rate will seriously degrade the sensitivity of the array, with the effects made worse by the long recovery time of long links running at high rates. The answer is in the use of switched light paths, where bandwidth is more easily guaranteed.

3. ESLEA

The ESLEA project has greatly helped the development of e-VLBI at Jodrell Bank Observatory by providing resources: the provision of two UKLight lightpaths to Amsterdam from Jodrell Bank Observatory (at 1 and 0.63 Gbps), a post-doctoral research assistant (Matt Strong) and funds for computing resource. Our objectives were to demonstrate the advantages of data transport over UKLight compared to that available on production networks, formulate methodologies for optimum use of switched light paths and help bring e-VLBI to the user within the European VLBI context. These aims were achieved by making comparison tests to JIVE on UKLight and production links (see figure 1), developing a UDP based data transfer protocol specifically for e-VLBI [5], optimising TCP parameters for Mk5 data transfer and upgrading Mk5 hardware [6]. In addition tests to USA have been made [7]. Development of e-VLBI has also been helped by investigations of TCP performance [5] and by measurements of correlator performance in the face of high packet loss [8].

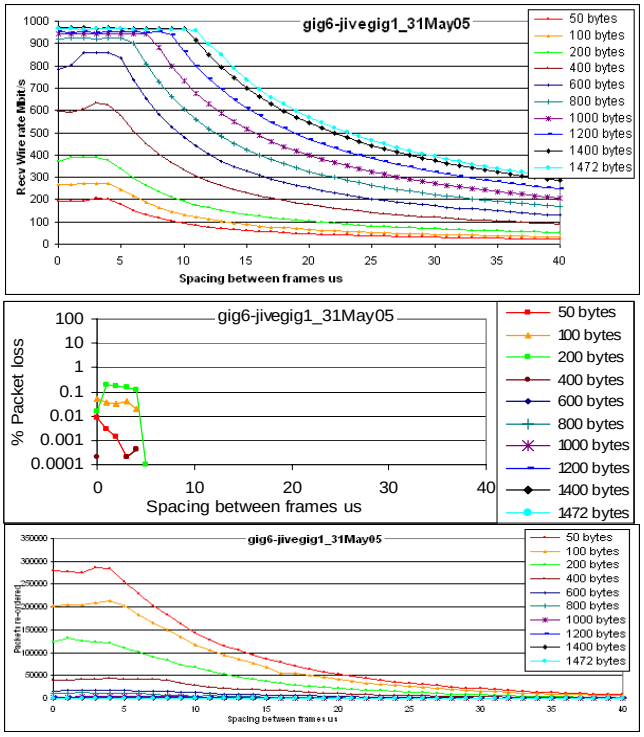


Figure 1. This shows tests on the production academic network between high performance server class machines situated in Manchester and JIVE in the Netherlands. The upper plot shows throughput vs demanded interpacket spacing for a variety of packet sizes. Small packets require more computing and network resource, and this results in a few lost packets (middle plot). The lower plot however shows that significant reordering has taken place. It is significant that tests on UKLight showed no re-ordering, see plot in [6].

Particular milestones in e-VLBI relevant to ESLEA have been:

- The setting up of regular tests with e-VLBI run every ~6 weeks since early 2005 [9],
- A demonstration at the GEANT2 launch in Luxembourg showing 3 way flows up to 800 Mbps across GEANT in June 2005.
- Trans-Atlantic data transfer from Onsala, JBO and Westerbork to the Haystack correlator in Massachussets, USA, during iGRID 2006 and SC2006, producing real time e-VLBI fringes at 512 Mbps. These demonstrations used UKLight connections across the Atlantic and were initiated by JBO staff.
- The first successful open call e-VLBI real time European VLBI session in April 2006.
- The first scientific real time e-VLBI results published in 2007 [10,11] (figure 2), and real time eVLBI science observing sessions becoming a regular part of European VLBI operations in 2007.

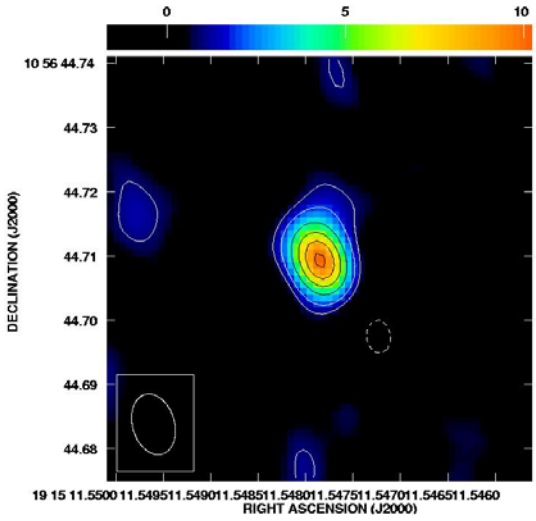


Figure 2. Microquasar GRS1915+105 (11 kpc) on 21 April 2006 at 5 Ghz using 6 EVN telescopes during a weak flare [10]. This object has jets of material which move away from an accretion disk surrounding a central black hole at velocities close to that of light. The jets were quiescent in these observations.

4. Conclusion

In summary, the ESLEA e-VLBI project has been very successful, achieving all its original aims, and even leading to further developments. We are now able to perform routine real time eVLBI measurements at data rates of 256 Mbps, and 512 Mbps tests are underway. An illustration of this was performed recently when the first rapid response experiment was undertaken (investigators A. Rushton and R. Spencer). Here a 6 telescope real time observation was run on 29th Jan 2007, the results were analysed in double quick time, selecting sources for follow up observations on 1st Feb. This kind of operation would be impossible for conventional VLBI. The experiment worked well technically – we successfully observed 16 sources (weak microquasars), but all were <0.5 mJy – too weak to observe in the follow up run – indicating a perverse universe, however the feasibility of the technique was clearly demonstrated. Our work

on e-VLBI is now concentrating on the EXPReS[1] project where with the objective is to create a distributed, large-scale astronomical instrument of continental and inter-continental dimensions.

Acknowledgements

The authors would like to thank the staff at EVN observatories and at JIVE for their patience, forbearance and invaluable assistance in making e-VLBI a reality.

References

[1] Moran, J., *Very Long Baseline Interferometer Systems*, in *Methods of Experimental Physics*, 12, 174, 1976

[2] Thompson, A., Moran, J., and Swenson, G., *Interferometry and Synthesis in Radio Astronomy*, 2nd ed. Wiley, 1991

[3] JIVE, *First Science with e-VLBI* http://www.jive.nl/news/first_science/first_science.html

[4] Kershaw, S. et al., *TCP delay*, this conference

[5] Casey, S., et al., *VLBI_UDP*, this conference

[6] Strong , M., et al., *Investigating the E-VLBI Mk5 End Systems,* this conference

[7] Rushton, A., et al., *Trans-Atlantic UDP and TCP Network Tests*, this conference

[8] Casey, S., et al., *Investigating the Effects of Missing Data on VLBI Correlation*, this conference

[9] Szomoru, A. *E-VLBI Developments at JIVE*, this conference

[10] Rushton, A.; Spencer, R. E.; Strong, M.; Campbell, R. M.; Casey, S.; Fender, R. P.; Garrett, M. A.; Miller-Jones, J. C. A.; Pooley, G. G.; Reynolds, C.; Szomoru, A.; Tudose, V.; Paragi, Z., *First e-VLBI observations of GRS1915+105,* MNRAS 374, L47, 2007

[11] Tudose, V.; Fender, R. P.; Garrett, M. A.; Miller-Jones, J. C. A.; Paragi, Z.; Spencer, R. E.; Pooley, G. G.; van der Klis, M.; Szomoru, A., *First e-VLBI observations of Cygnus X-3,* MNRAS, 375, L11, 2007

[1]EXPReS is an Integrated Infrastructure Initiative (I3), funded under the European Commission's Sixth Framework Programme contract number 026642 EXPReS.

PROCEEDINGS OF SCIENCE

Investigating the e-VLBI Mark 5A end systems in order to optimise data transfer rates as part of the ESLEA Project

Matt Strong[1]
Jodrell Bank Observatory, The University of Manchester, Oxford Road, Manchester, UK
E-mail: matthew.strong@manchester.ac.uk

Richard Hughes-Jones
The University of Machester, Oxford Road, Manchester, UK
E-mail: r.hughes-jones@manchester.ac.uk

Ralph Spencer
Jodrell Bank Observatory, The University of Manchester, Oxford Road, Manchester, UK
E-mail: ralph.spencer@manchester.ac.uk

Simon Casey
Jodrell Bank Observatory, The University of Manchester, Oxford Road, Manchester, UK
E-mail: simon.casey@manchester.ac.uk

Stephen Kershaw
The University of Manchester, Oxford Road, Manchester, UK
E-mail: stephen.kershaw@manchester.ac.uk

Paul Burgess
Jodrell Bank Observatory, The University of Manchester, Oxford Road, Manchester, UK
E-mail: paul.burgess@manchester.ac.uk

Arpad Szomoru
Joint Institute for VLBI in Europe, Dwingeloo, NL
E-mail: szomoru@jive.nl

We report on the development of high bandwidth data transfers for e-VLBI at Jodrell Bank Observatory as part of the ESLEA project. ESLEA is a UK project to exploit the use of switched-lightpath optical networks for various applications, including e-VLBI, HEP, High Performance Computing and e-Health. We show how the CPU power of the Jodrell Bank e-VLBI Mark 5A end systems was limiting the data transfer rate to below 512 Mb/s. Both of the Jodrell Bank Mark 5A end systems have now been upgraded and can now transfer e-VLBI data to JIVE at the required data rate of 512 Mb/s.

Lighting the Blue Touchpaper for UK e-Science - Closing Conference of ESLEA Project
The George Hotel, Edinburgh, UK
26-28 March, 2007

[1] Speaker

PoS(ESLEA)024

1. Introduction to VLBI, and e-VLBI

Very Long Baseline Interferometry (VLBI) is a technique for creating high resolution radio maps using radio telescopes located around the world (and even in space). A defining feature of VLBI is that, historically, due to the large distances involved, the telescopes store data on magnetic tape, or more recently computer disks as they cannot be connected directly. These tapes or disks are then shipped to the correlator, played back, correlated and Fourier transformed in order to create the high resolution images. Recently however, there has been a drive to upgrade the VLBI system to a real time instrument, e-VLBI (e-VLBI in Europe is being developed with funding of the EU project EXPReS). The use of computer networks and the internet are ways of connecting radio telescopes around the world together, and so VLBI astronomy can be performed in real time.

The current VLBI system employs the Mark 5A disk-based recorder [1], which records the astronomical data collected at the telescope to large disk packs with capacities of several hundred gigabytes each. The core of the Mark 5A is a 1.2 GHz standard PC running Linux. The PC contains two interface boards; a StreamStor card[*] for high speed disk reading and writing and an I/O board [2]. As the Mark 5A is simply a custom designed PC which interfaces to the VLBI formatters and disk packs, it is possible to retro-fit a gigabit Ethernet card via the PCI bus. The Mark 5A control software is capable of reading/writing data via the Ethernet card, in a similar way to how it communicates with the disk packs and formatter. It is therefore possible to establish a direct 'link' between the telescope and the correlator and perform real time e-VLBI.

Current Production Goals of e-VLBI

The Mark 5A units are capable of recording data at rates of up to 1 Gb/s, and whilst the Ethernet interface can run at 1 Gb/s, it is not possible to achieve transmission of telescope data at 1 Gb/s. This is due to the fact that the data has to be encapsulated within TCP or UDP packets, which then have to be encapsulated in IP packets and finally in Ethernet packets. Each level of encapsulation adds a little more data that needs to be transmitted, and so for 1 Gb/s of telescope data, there would be 10% more data created by the encapsulation and hence would not be transmitted through a 1 Gb/s Ethernet card. Owing to the nature of the VLBI formatters, and current technical constraints, data rates have to be a power of 2 (32, 64, 128, 256, 512, 1024 Mb/s) and so the next speed down is 512 MBits/s which is technically achievable over a 1 Gb/s link. It is thus the current goal of the e-VLBI community to reliably run e-VLBI experiments at 512 Mb/s, and then to develop the technology and networks in order that this speed can be increased to gigabit levels.

The e-VLBI Network

The Mark 5A units, and other e-VLBI PCs stationed at the telescopes and correlator communicate with each other over the European wide production network for academic research and development, known as GÉANT 2 (with the exception of the Westerbork telescope which has its own fibre connection to the correlator). The Mark 5A end systems are connected to the GÉANT 2 production network through their local and national research and education networks (NRENs). In addition to the GÉANT 2 production network, Jodrell Bank Observatory in the UK also has two dedicated optical links between Jodrell Bank and JIVE. These links are routed via UKLight, and its peering ability with SURFnet and NetherLight. There are dedicated

[*] Made By Conduant

1 Gb/s and 630 Mb/s optical connections between the telescopes at Jodrell Bank and the correlator at JIVE.

2. The 500 Mb/s Bottleneck

Currently, EVN e-VLBI operate a TCP based system, and optimisation of this system is necessary if e-VLBI is to work at the highest data rates (512 Mb/s is the current goal). In e-VLBI system testing, two of the five participating stations (Onsala an Westerbork) have been able to achieve 512 Mb/s using their current Mark 5A machines, but 512 Mb/s transfer between Jodrell Bank and the correlator (located at JIVE in the Netherlands) could not be achieved despite identical hardware and spare capacity on the network links. e-VLBI data transfers of 500 Mb/s could be achieved on both the production link and the 1 Gb/s dedicated UKLight link provided by the ESLEA project, but this point proved to be a bottleneck with the existing hardware. This bottleneck was thought to be caused by the Jodrell Bank Mark 5A system and as such investigation of its performance was necessary. The results of this study are detailed in the following sections.

3. Mark 5A UDP Transfers over UKLight

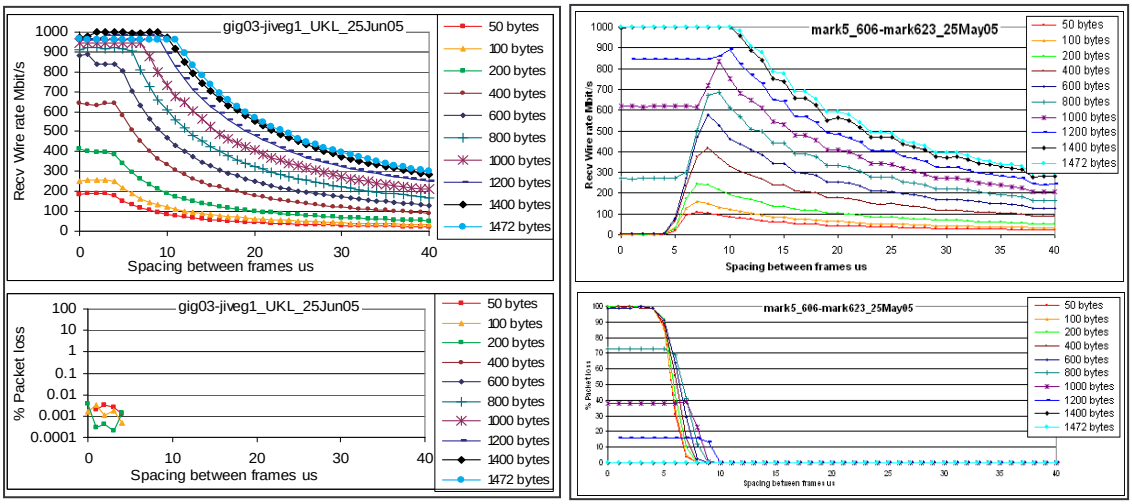


Figure 1 - UDP_mon transmission tests. Left Panel: UDP throughput and packet loss between Jodrell Bank and JIVE using high performance network machines. Right panel: UDP throughput and packet loss between Jodrell Bank and JIVE using the Mark 5A machines.

The performance of the Mark 5A end systems was first investigated by transferring UDP data (using the UDP_mon software package[†]) between Jodrell Bank and JIVE and comparing these results with those from a high performance network machine. Figure 1 shows the results from this study. The left panel shows the results from the high power network machine, and we can see good throughput for all inter-packet spacings, whilst only a very small amount of packet loss is detected at small packet sizes and inter-packet spacings. The right hand panel of Figure 1 shows the corresponding graphs for the Mark 5A machines. It can be seen that the UDP throughput shows its normal signature for larger inter-packet spacings. However, the throughput falls off dramatically for low inter-packet spacings for all packet sizes. Also, from the lower graph it is evident that there is dramatic packet loss corresponding to this loss in throughput. A

[†] Developed by Richard Hughes-Jones

common cause of such a loss of throughput can be the CPU speed of the machines. Below we investigate the CPU performance of the Mark 5A machines to ascertain if it is having an adverse effect on the e-VLBI data transmission.

4. Analysis of the CPU Performance of the Mark 5A End Systems

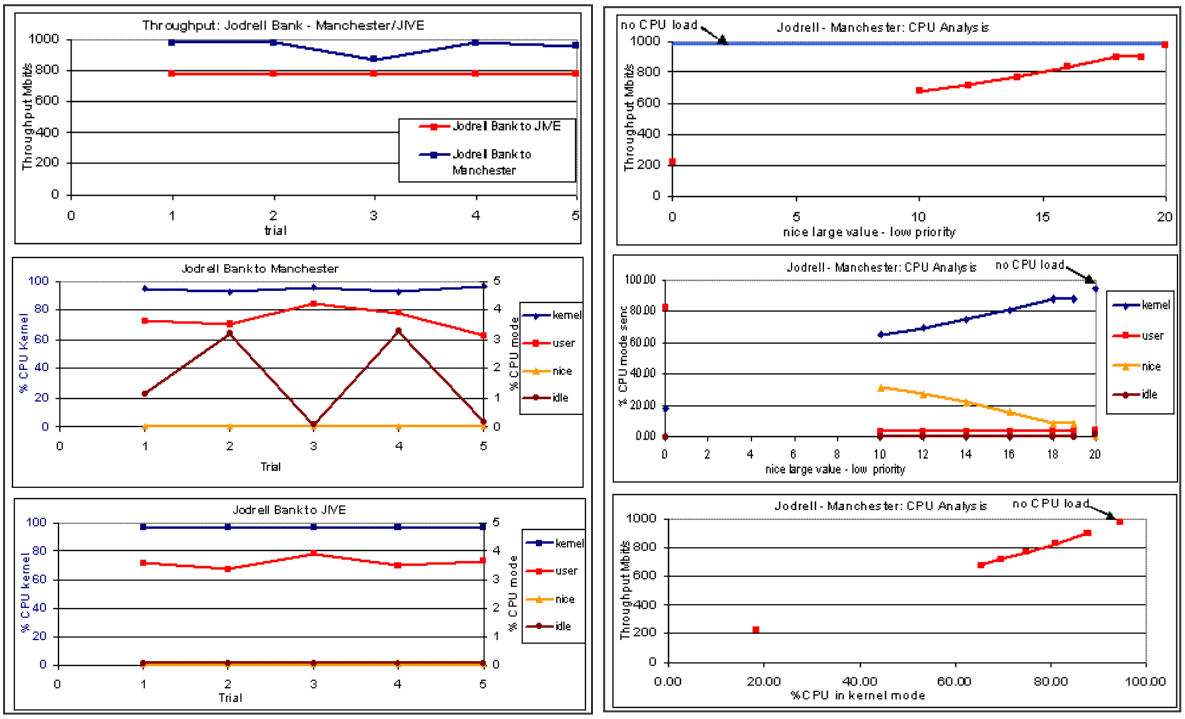


Figure 2 - CPU analysis of the Jodrell Bank Mark 5A. Left panel: Network and CPU performance whilst performing an iPerf TCP transmission between Jodrell Bank and Manchester, and Jodrell Bank and JIVE. Right panel: Network and CPU performance whilst performing an iPerf TCP transmission between Jodrell Bank and Manchester and also running a CPU intensive task.

In order ascertain if the Mark 5A CPU speed was having an adverse effect on e-VLBI transmission, its usage was examined prior to and after each test. In addition to CPU load, network interface, IP, UDP (via UDPmon) and TCP (via iPerf) statistics were also measured just prior to, and just after each test. Thus, this allowed much of the resources to be measured. Tests were performed between Jodrell Bank and Manchester, and then Jodrell Bank and JIVE, both on the dedicated optical network.

For a single, memory to memory TCP stream, the connection between Jodrell Bank and Manchester showed a transmission rate of 950 Mb/s (Figure 2, top left) and a CPU usage of 94.7% kernel, 1.5% idle (Figure 2, middle left), whilst the connection between Jodrell Bank and JIVE showed a transmission rate of 777 Mb/s (Figure 2, top left) and a CPU usage of 96.3% kernel, 0.06% idle (Figure 2, bottom left). Thus, it appears that when transmitting between Jodrell Bank and Manchester, the Jodrell Bank Mark 5A machine just has enough CPU to send the data at line rate. However, when transmitting between Jodrell Bank and JIVE, the Jodrell Bank Mark 5A does not appear to have sufficient CPU power to drive the network at line rate. e-VLBI transfers are obviously not memory to memory transfers, as the data has to be processed

in the Mark 5A machine, resulting in it passing through the PCI bus a number of times. As such, addition CPU is necessary to perform such processing.

Due to the fact that we cannot simulate exact e-VLBI transmission as the receiving machine is not a Mark 5A, we simulated the effect of the Mark 5A processing by adding a CPU intensive task to the sending Mark 5A machine (at Jodrell Bank). Thus, we could then investigate how the throughput and CPU usage varies with respect to this additional task.

From Figure 2 (top right) it can be seen that the throughput of memory to memory TCP flows between Jodrell Bank and Manchester fell from 950 Mb/s with no CPU load process to 900 Mb/s when the CPU load process had the lowest "nice" priority of 19 and to 675 Mb/s when the "nice" priority was 10. The "nice" priority range runs from -20 at its highest to +19 at its lowest, with normal user priority 0. In addition to this, the middle right graph in Figure 2 shows that the available CPU in Kernel mode falls rapidly to approximately 60% when the "nice" priority increases from 19 to 10. The bottom right graph in Figure 2 shows how the throughput of the TCP stream is related to the available CPU power and shows the throughput falling as the available CPU decreases. These graphs clearly show that the addition of a CPU intensive task has an adverse effect on the TCP transmission speed. It is clear that the Jodrell Bank Mark 5A machine does not possess enough CPU power to drive the network at adequate speeds whilst performing other processing. For this reason, the Jodrell Bank Mark 5A was upgraded with an Asus NCCH-DL motherboard, Intel Xeon 2.8 GHz processor and 1 GB of server specification SDRAM. The tests above were then repeated with the new upgraded Mark 5A machine and the results are given in section 5.

5. CPU Performance Tests of the Upgraded Mark 5As

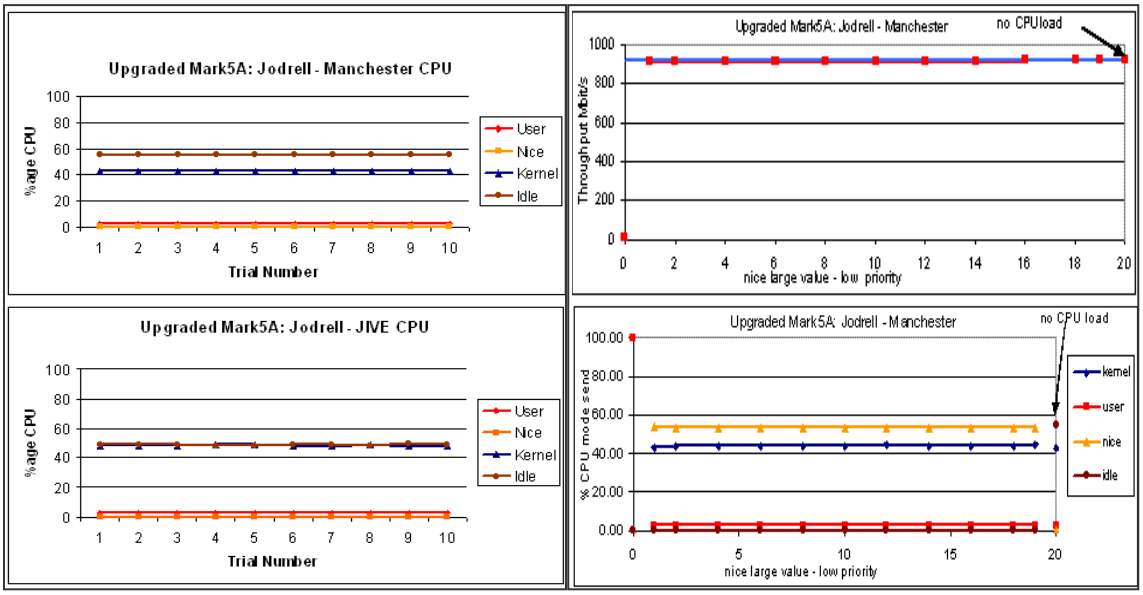


Figure 3 – CPU analysis of the upgraded Jodrell Bank Mark 5A. Left panel: CPU performance whilst performing an iPerf TCP transmission between Jodrell Bank and Manchester, and Jodrell Bank and JIVE. Right panel: Network and CPU performance whilst performing an iPerf TCP transmission between Jodrell Bank and Manchester and also running a CPU intensive task.

Figure 3 shows the CPU performance of the upgraded Mark 5A machine. It is easy to see that the upgraded Jodrell Bank Mark 5A performs extremely well in these tests. Indeed, it maintains a near line rate memory to memory TCP stream whilst the "nice" priority of the CPU intensive is varied over its full range. Indeed, the bottom right graph in Figure 3 shows the upgraded Mark 5A machine to be using less than 50% of its CPU in transferring the data and running the CPU intensive task.

After these tests were performed, standard e-VLBI tests were performed with this upgraded Mark 5A machine, and it achieved e-VLBI data transmission at a rate of 512 Mb/s over the dedicated optical link immediately. It seems obvious that this Mark 5A upgrade has had a positive effect on its data transmission, and now it can achieve 512 Mb/s e-VLBI data transfer as required.

6. Conclusions

From the above tests it is clear that the Jodrell Bank Mark 5A machine did not have sufficient CPU power to transfer e-VLBI data at 512 Mb/s. This result is surprising as both the Onsala and Westerbork Mark 5A machines have been able to transfer eVLBI data at 512 Mb/s. So what is the difference between these machines? Before the Jodrell Bank Mark 5A was upgraded, the specification and components of all the Mark 5A machines were the same. They were Intel P3, 1.2 GHz machines with 256 MBytes of RAM.

It is noted that the difference between the transmission speeds between the Onsala and Westerbork Mark 5As and JIVE, and the initial Jodrell Bank Mark 5A and JIVE was at the 10% level. The specified CPU speeds of the Mark 5A machines are only accurate to ~10% level, and as such, this inaccuracy could be responsible for such a bottleneck. Indeed, if the initial Jodrell Bank Mark 5A's clock speed was 10% lower than its specification, and the Westerbork and Onsala Mark 5A's were 10% higher, this leaves a shortfall on CPU power of ~240 MHz. Such a shortfall in CPU power could have easily resulted in the 512 Mb/s threshold being unattainable from the Jodrell Bank Mark 5A.

Regardless, it is certain the upgrade to the Jodrell Bank Mark 5A has had a dramatic effect on its performance. Indeed, with the upgraded Mark 5A, Jodrell Bank obtained reliable 512 Mb/s data transmission to JIVE immediately. One of the main problems with the Mark 5A machines is that they require CPU power to transfer the telescope data across the PCI bus twice. This results in the Mark 5A machines needing more CPU to drive the network at the necessary rates. Indeed, ultimately the goal of e-VLBI is to transfer data from the telescopes to the correlator at over 1 Gb/s, and as such it is unclear as to whether the Mark 5A machines can manage these rates.

References

[1] Alan Whitney, *Mark 5A disk-based gbps vlbi data system*, viewed 19th July 2006, http://web.haystack.mit.edu/mark5/paper.pdf

[2] Dan L. Smythe. *Mark 5A Memo #007.1. Mark 5A memo series*, viewed 19th July 2006, ftp://web.haystack.edu/pub/mark5/index.html

PROCEEDINGS OF SCIENCE

Investigating the effects of missing data upon VLBI correlation using the VLBI_UDP application

Simon Casey[1]
The University of Manchester, Oxford Road, Manchester, UK
E-mail: scasey@jb.man.ac.uk

Ralph E. Spencer, Matthew Strong, Richard Hughes-Jones, Paul Burgess
The University of Manchester, Oxford Road, Manchester, UK
E-mail: res@jb.man.ac.uk
E-mail: mstrong@jb.man.ac.uk
E-mail: R.Hughes-Jones@manchester.ac.uk
E-mail: pb@jb.man.ac.uk

Arpad Szomoru
Joint Institute for VLBI in Europe (JIVE)
Dwingeloo, NL
E-mail: szomoru@jive.nl

Colin Greenwood
National e-Science Centre (NESC)
Edinburgh, UK
E-mail: coling@nesc.ac.uk

The work presented in this paper describes tests conducted to assess the effect that missing data has upon VLBI correlations. Results obtained show that the correlation is resilient under data losses approaching 20% and that the resulting correlation amplitude decreases as 1.2 times the packet loss rate. The results also indicate that while the Station Units are able to cope with at least 2 consecutive missing headers, they need to see at least 2 correct headers otherwise synchronisation is lost.

Lighting the Blue Touchpaper for UK e-Science - Closing Conference of ESLEA Project
The George Hotel, Edinburgh, UK
26-28 March, 2007

[1] Speaker

PoS(ESLEA)025

1. Introduction

The design of the data transport application VLBI_UDP[1], allows for portions of data to be lost before transmission. This can allow data packets to be dropped selectively in order to investigate the behaviour of the VLBI correlator. The correlator was originally designed to be fed with data from magnetic tapes and so is resilient to a certain amount of data loss[2]. The aim of the work presented here is to determine what effect varying rates of packet-loss has on the correlator output, and also to see if there is a limit at which the output is un-usable. These results may then be programmed back into VLBI_UDP such that it can selectively drop packets if congestion is detected, whilst having a minimal effect on the resulting correlated data.

2. Packet-dropping theory

This section outlines the data format used by the Mark5A data recorders, and explains the decisions behind choosing the tests that were run.

2.1 Mark5A data format

The data fed into the Mark5A data recorder comes from the Mark4 tape formatter. This has a variety of modes[3] which affect exactly how the data stream appears when written to a standard PC file. The data used in these tests were recorded at a rate of 256Mbit/s, 32 tracks with 1-bit sampling. Upon examining the recorded file, it became apparent that the VLBI data tracks were bit-interleaved as in *Figure 1*, such that 2 subsequent bits from any track would be separated by 32 bits in the PC file. Each track is framed in 2500 byte segments, and so 32 tracks give a 'combined frame' size of 80 000 bytes.

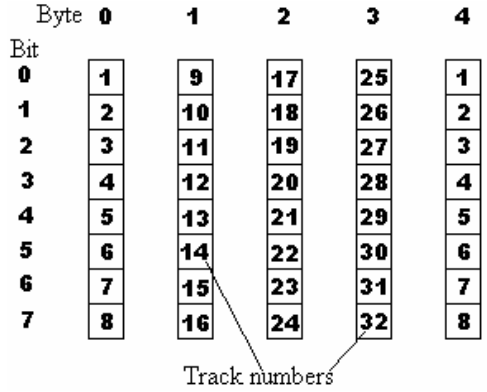


Figure 1: Visual layout of VLBI data tracks within PC file

2.2 Keeping Station Units in synchronisation

The Station Units (SUs) sit between the Mark5As and the correlator. They are a relic from the days of tape and, among other things, serve to servo the incoming data stream to keep all the streams aligned in time and check that the data has valid headers. Since the goal of the Mark5A was to emulate a tape drive, the SUs are still required. To ensure all the data streams are synchronised in time, they read the headers from each frame, perform various checks to ensure

that the headers are valid, and then read the time stamp to ensure synchronisation. Since the data frames are a constant 2500 bytes long, the SUs know exactly where in the stream the next header should be located. If the SU does not find a valid header where it expects to see one, then it is able to flywheel over that frame and check where the subsequent header should be. It is unknown how the SUs behave when there are multiple headers missing, giving rise to an investigation where only the VLBI headers are dropped. The data file used for all tests was approximately 15 minutes long, and different strategies were attempted every 2 minutes. The first 2 minutes were left intact to ensure the system was fully synchronised, then starting at 2 minutes, headers were dropped for 30 seconds followed by 90 seconds of intact data to allow for any resynchronisation if needed. The strategies chosen can be seen in *Figure 2*.

Minutes in file	*Header drop pattern*
2-4	Drop---Keep---Keep
4-6	Drop---Keep---Drop---Keep
6-8	Drop---Drop---Keep---Keep
8-10	Drop---Drop---Keep
10-12	Drop---Drop---Drop---Keep
12-14	Drop---Drop---Drop---Drop---Keep

Figure 2: Strategies for dropping headers

2.3 General packet-loss

These tests are to examine the effect that loss of any part of the data stream has on the resulting correlation. As has been discussed, the SUs are able to cope with 1 missing header, but any more than 1 consecutive missing header and synchronisation may be lost resulting in a temporary loss of correlation.

If, therefore, it is assumed that the SUs will stay in synchronisation when every other header is present, then the following can be used to determine a theoretical maximum packet-loss rate, below which synchronisation will be maintained. The 32 recorded data frames, when interleaved together, create a large 80 000 byte ‘frame’. This can then be broken down into 79 360 bytes of VLBI data sandwiched between a 384 byte header and 256 byte footer. VLBI_UDP allows for 1444 bytes of user data in each UDP packet, hence one 80 000 byte frame will need 55.4 UDP packets. Working on the case where losses are statically distributed, if no more than 1 in every 57 packets (1.75%) is lost, then 2 consecutive headers will never be lost. With this in mind, loss rates of 0.5%, 1%, 1.5%, 1.75%, 2% and 2.25% were chosen for the initial series of tests. Based upon the results of these tests in September 2006, follow up tests were performed in December 2006 at rates of 1%, 2%, 5%, 7.5%, 10%, 15% and 20%.

3. Experimental details

The correlation tests used data provided by JIVE, taken during an NME experiment from 3 antennae: Jodrell Bank, UK; Westerbork, NL; Effelsberg, DE. Only the Jodrell Bank data were modified, giving a total of 2 modified baselines and 1 unmodified.

The Jodrell Bank data were taken from a Mark5A disk pack and written to a standard PC file using the Mark5A software suite. The resulting file was ~28GB and transferred to a PC (Huygens) at JIVE which had a RAID array large enough to accommodate all the subsequent data sets. VLBI_UDP was run in a stand-alone mode, where it simply reads 1444 byte size chunks from one file, filters them through a packet dropping function, and then writes the resulting data out to another file.

For each of the loss rates listed previously in the September tests, two output files were created; the first where single packets were dropped at a constant rate; the second where packets were dropped in bunches. For the December tests, 4 files were created for each loss rate. Two files created as before, with a further 2 created by replacing the missing data with a fill pattern 0x11223344 instead of random data. The fill pattern is processed by the Mark5A's I/O board, and when the pattern is observed, the interface board signifies to the SUs that the data is invalid. The output files were placed back onto Mark5A disk packs, ready for correlation.

4. Results

4.1 September 2006

Unfortunately due to a miscalculation on the position of the headers within the file, the header dropping test was not a success. The varying loss rates however, produced good results. *Figure 3* shows a typical plot produced from one of the runs. It can clearly be seen at 32 minutes (the file starts at 30 minutes) the amplitude differences shoot up as the packets are being dropped. It became apparent that even at rates of 2.25% the SUs had no problem with the packet losses, thus giving rise to the higher loss rates used in the December 2006 tests.

By averaging over all the amplitude differences for each run, it is possible to create a linear plot of packet loss rate against amplitude decrease; *Figure 4* is an example of such a plot.

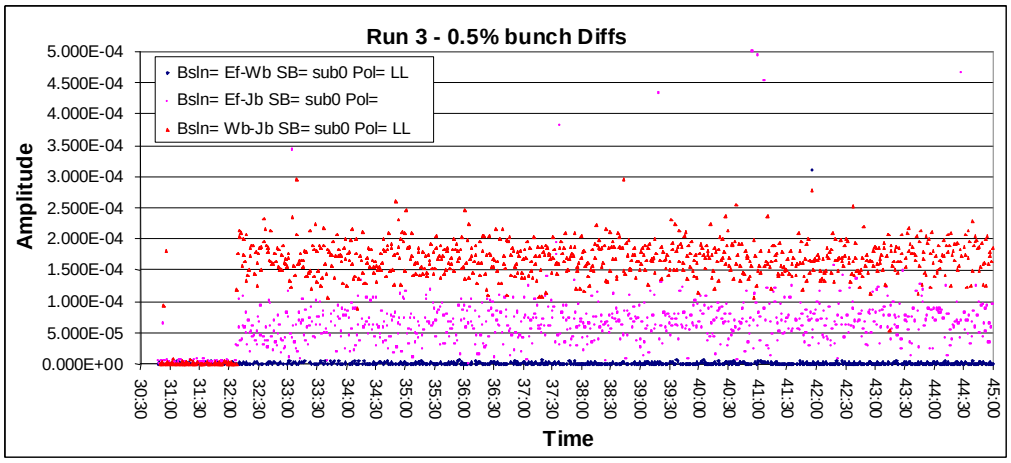


Figure 3: Plot of modified correlation amplitude subtracted from unmodified amplitude

4.2 December 2006

Most of these runs produced good results, but produced spurious amplitudes, thus badly skewing the average plot. Upon closer investigation, it was apparent that the correlation weighting values for the spurious points were close to zero whereas they should be close to 1 for good data. Recalculating the averages using a weighted average based on the correlation weight suppressed the spurious points and brought the average plot back in to line. The set that had this problem was at 20% loss rate, and this would have been caused by the SUs being unable to cope with the amount of lost data and so feeding invalid data to the correlator. It would therefore appear that loss rates of 15% and below can be handled adequately by the SUs and correlator.

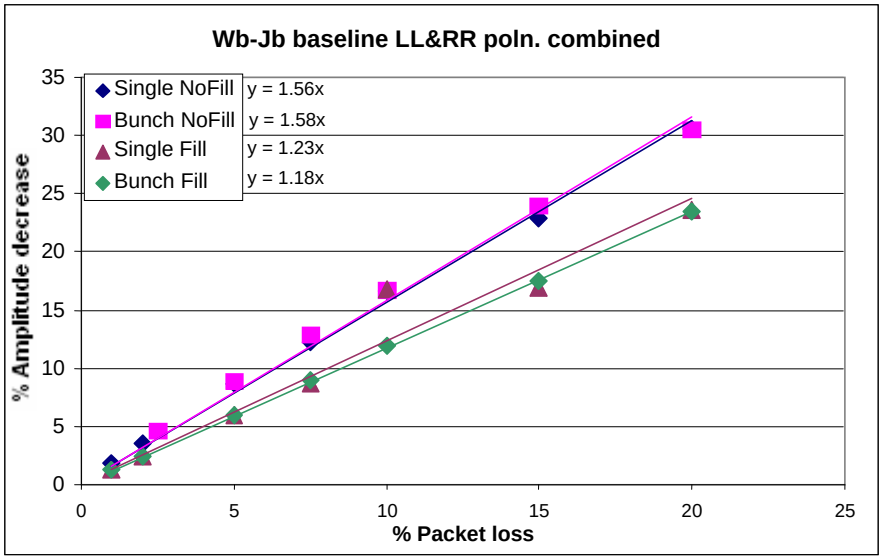
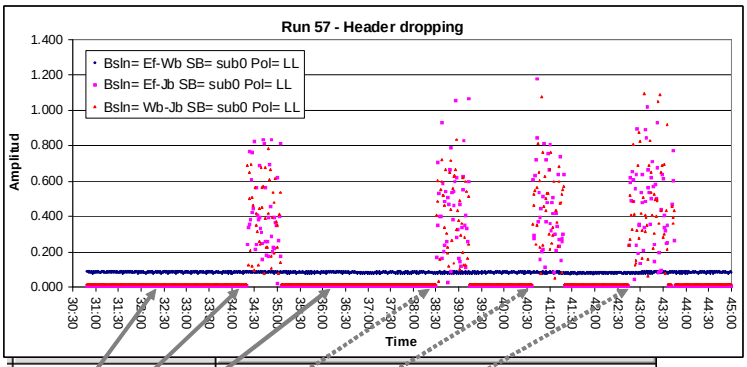


Figure 4: Amplitude decrease with packet loss for all 4 dropping algorithms, with/without fill pattern, single loss/ bunch loss, Left-Left and Right-Right polarisations averaged together

From *figure 4*, it is clear that by replacing missing data with the Fill Pattern instead of random data, the correlator is better able to compensate, the slopes decreasing from 1.5 without the Fill Pattern to 1.2 with. Losing packets singly or in small bunches appears to have a negligible affect on the amplitude.

Finally, *figure 5* shows how the SUs coped with missing headers. It seems from these results that the SUs can cope with more than 1 consecutive missing header (test 3) but must then find 2 or more correct headers since test 3 failed.



Minutes in file	*Header drop pattern*	
2-4	Drop---Keep---Keep	✓
4-6	Drop---Keep---Drop---Keep	✗
6-8	Drop---Drop---Keep---Keep	✓
8-10	Drop---Drop---Keep	✗
10-12	Drop---Drop---Drop---Keep	✗
12-14	Drop---Drop---Drop---Drop---Keep	✗

Figure 5: Results of header dropping tests

5. Conclusion

The work presented above shows that the Station Units and correlator are able to cope with packet loss rates of up to 20% without causing a serious effect on the correlated result. If the packet dropping shown here is implemented during a real VLBI transfer, then it should be

possible to drop perhaps 10% of a 1024Mbit/s VLBI stream and transmit the remainder over a Gigabit Ethernet link.

6. Acknowledgements

I'd like to thank JIVE for giving me access to the correlator and their help in creating the datasets and processing the correlator output, ESLEA for overseeing the UKLight links, and Jodrell Bank Observatory for giving me access to the initial Mark5 recordings.

References

[1] S. Casey et al., *Investigating the effects of missing data upon VLBI correlation using the VLBI_UDP application.* In proceedings of Lighting the Blue Touchpaper for UKe-Science – Closing Conference of ESLEA Project, POS(ESLEA)038, 2007

[2] R.E. Spencer et al., *Packet Loss in High Data Rate Internet Data Transfer for eVLBI* in proceedings of 7th EVN symposium. Toledo 12-15 October 2004

[3] W. Aldrich, *Mark5 memo #013: Mark5A Operating Modes*, *Mark 5 Memo Series*, viewed 26 April 2007, http://www.haystack.edu/tech/vlbi/mark5/memo.html

PROCEEDINGS OF SCIENCE

Lambda Grid developments, History - Present - Future.

Cees de Laat[1]
University of Amsterdam
Kruislaan 403, Netherlands
E-mail: delaat@science.uva.nl

Paola Grosso
University of Amsterdam
Kruislaan 403, Netherlands
E-mail: grosso@science.uva.nl

About 6 years ago the first baby-steps were made on opening up dark fiber and DWDM infrastructure for direct use by ISP's after the transformation of the old style Telecom sector into a market driven business. Since then Lambda workshops, community groups like GLIF and a number of experiments have led to many implementations of hybrid national research and education networks and lightpath-based circuit exchanges as pioneered by SURFnet in GigaPort and NetherLight in collaboration with StarLight in Chicago and Canarie in Canada. This article looks back on those developments, describes some current open issues and research developments and proposes a concept of terabit networking.

Lighting the Blue Touchpaper for UK e-Science - Closing Conference of ESLEA Project
The George Hotel, Edinburgh, UK
26-28 March, 2007

[1] Speaker

PoS(ESLEA)027

1. From Lambda's to Hybrid Networks

In the late 90's new laws slashed the Telco monopolies in many countries. The previously state-owned providers were forced to enter the competitive market. That, combined with the very favourable financial situation for telecom, Internet and ICT led to a huge investment in fiber infrastructure, which became also available to end customers if one had the vision to ask for it. A group of NREN's and scientists gathered in September 2001 at the TERENA offices in Amsterdam to discuss a new approach to networking that would use circuits and colors on fibers. The goal was to start experiments with a first dedicated Lambda between Chicago and Amsterdam, and to use that for tests of new applications and demonstrations at the iGrid2002 [1] event in the Science Park in Amsterdam. The idea was to get a new order of scalability for the Internet by owning the dark fibers and the optical equipment to handle the traffic and, therefore, incur in just small incremental costs for extra capacity later on. The model up to that point was to go to the Telecom provider and ask for a managed service that would typically always be at full cost.

At that time the concept of Lightpaths was introduced [2]. A lightpath is a dedicated connection in an optical network that gives a guaranteed L1 or L2 service to the end user. Nowadays research and education networks around the world offer such connections to e-Science applications that have large bandwidth requirements or are sensitive to network delays. These applications cannot properly function in the traditional shared IP environment where the network behavior is unpredictable due to the large number of concurrent and competing users, and where routing protocols determine the path followed by the traffic. To name a few, in the high-energy physics community, the upcoming CERN experiments in the LHC (http://cern.ch/lhc) use lightpaths for their wide-area data transfers. Also, the visualization applications developed in the OptIPuter [3] transfer images worldwide using SAGE as shared terapixel display workspaces on dedicated lightpaths.

Inversely these huge streams destroy the normal operation of the routed Internet and would require huge investments in routers. However the big streams usually are very long lived flows from same source to same destination and may therefore be mapped to their own circuit and avoid routers. The unit of bandwidth in these applications is Gb/s up to whatever one color can transport, typically 10 Gb/s. Optical photonic switches have huge capacity for much lower prices than routers: a wavelength selective switch that can handle 72 Lambda's in five fibers costs as much as one 10 Gb/s router port on a full fledged backbone router. A network that provides both the routed IP packet-switched services and the lightpath optical circuit-switched services is called a hybrid network [2]. The Dutch SURFnet6 network is one of the first designed with the above ideas in mind. It is basically a nation wide dark fiber network where a mix of DWDM and packet encapsulation equipment takes care of the transport of the traffic on multiple colors in the fibers. Routing is done only in two places in the network. The traffic from all the Universities and connected organizations that needs to be routed is transported as 1 and 10 Gb/s streams over the network to Amsterdam, where the routers are located. On the same basic infrastructure organizations can obtain dedicated lightpaths to form their own optical

private networks or exchange traffic with other organizations for heavily demanding applications. Such lightpaths can be extended via the hybrid exchange NetherLight in Amsterdam to international locations [4]. Current users of this service are notably the Particle Physics Community for distribution of data from CERN and the Astronomers for correlation of radio telescopes. The SURFnet6 hybrid model enables the network to scale to much higher capacities for the same costs compared to a full routed approach.

2. Research and Development in Hybrid networks

The operation of the hybrid networks is two-fold. The routed part is operated in the same way as is current practice in the Internet. The lower layer part for the lightpaths is still mostly done by hand using phone and email as the service layer. A number of pilot implementations exist to aide users and network managers to control the setup of the switches, but they require detailed insight of the local and global setup, of the topology, the interface properties and equipment specifications. They usually also require privileged access to the network devices; any errors could kill someone else's traffic through the same box. Some of the main issues in a multi domain hybrid networking architecture are:

- virtualisation of the network elements by a Network Resource Provisioning System
- authorization (and associated authentication, cost accounting, etc.)
- intra and inter domain topology for path-finding
- addressing schemes of endpoints on the lower layers
- web services versus in band signalling
- connecting together domains with different NRPS's and signalling systems
- applicability of new photonic devices

International projects as Phosphorus, GN2, DRAGON, G-Lambda, Enlightened and organizations as CANARIE and Internet2 try to alleviate several of these issues by developing protocols, interfaces, middleware and tools. In Europe the Phosphorus project is addressing the inter domain lightpath setup signalling where different domains may use different NRPS'es. In the rest of this section we delve into two specific research tracks at the University of Amsterdam. Those are the dynamical DWDM project named StarPlane and the development of a resource description framework for describing the lower layer networks and their interconnections in a format that can be exchanged between domains.

2.1 The StarPlane project

The StarPlane project is designed to add a new level of flexibility to the photonic networking on the dark fiber infrastructure in SURFnet6. The current Lambda based networks usually implement their LightPath services on top of SONET switches that use rather statically configured DWDM infrastructure requiring transponders at switching points. In our opinion a higher level of flexibility in a dark fiber infrastructure of modest size can be achieved operating directly on the photonic level. The challenge in StarPlane is to develop the technology that allows applications running on a grid infrastructure connected to a portion of the SURFnet6 to autonomously switch colors on the photonic layer. In StarPlane applications can optimize the interconnection network topology as function of the computational needs of the application.

The current photonic networks are static for several reasons. Lambda's in fibers can influence each other; e.g. in an amplifier section. Photonic level switching causes the distance, hence dispersion, to change. In the StarPlane project the flexibility is added by introducing NORTEL's wavelength selective switches (WSS) and electronic dispersion compensating transponders (eDCO). The WWS's are functionally a combination of Dense Wavelength Division Multiplexers and micro electro mechanical switches. The switches are capable of selecting any combination of wavelengths out of a number of fibers (in our case 72 colors in 4 fibers) and put them in the outgoing fiber. The actual switching is done by the MEMS part of the WSS and can in principle be very fast. The eDCO works by trying to figure out what the dispersion effect of the fiber is and then modulating the sending laser to generate a kind of inversely deformed signal so that when that signal arrives at the receiver the dispersion is cancelled out. The eDCO can adjust to changing paths. These switches and transponders are now installed in several locations in SURFnet6 and interconnect the DAS-3 [5] computer clusters. The middleware on the clusters is being adapted to enable it to signal to the network its desired network topology.

2.2 Semantic Web Resource Description Framework applied for lightpaths

When applications are in need of a lightpath service the current best practice is to contact the local NREN by phone or email and let them use their knowledge of the worldwide GLIF to construct a connection. In order to get to an automatic setup system it is imperative to somehow distribute the topology of the existing infrastructure. Based on ideas from the semantic web world we constructed an ontology and associated schemas named NDL, the Network Description Language, to exchange, distribute and store this topology information. NDL is an under-development RDF-based ontology that provides a well-defined vocabulary and meaning for the description of optical networks topologies [6].

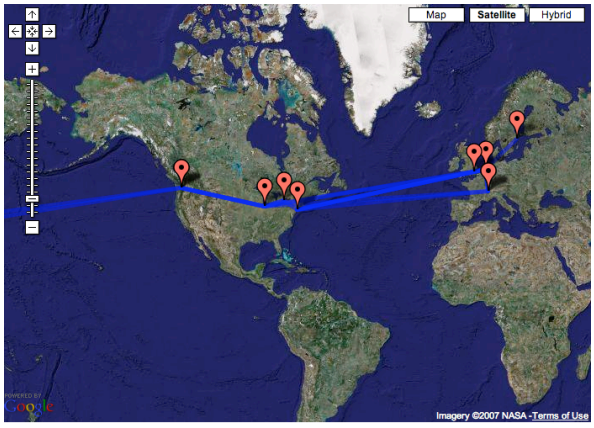


Fig 1 - A Google maps based graph of GLIF lightpaths.

RDF, the Resource Description Framework provides the mechanisms to organize the data. RDF uses a triple-based model to represent resources and the relationship between them. The predicate defines a property of the subject; and the object is the value of such property. RDF identifies subject, object and predicate with URIs - Universal Resource Identifiers. NDL uses RDF/XML to define the ontologies. Currently NDL provides five schemas: topology, layers, capability, domains and physical schemas. Applications for NDL in hybrid networks are: tools

to generate up-to-date network maps, to provide input data to path finding algorithms, and to detect errors and faults in existing lightpaths' setup. We have used the mapping tool to model the GLIF network as can be seen in fig 1.

3. The future.

The developments in Networking technology outpace those in computing and storage. In principle the interconnecting networks do not limit anymore the development of e-Science applications. Events like iGrid2002 and iGrid2005 and projects as GigaPort-NG, CineGrid and OptIPuter demonstrate a completely new wealth of applications if effort is spend to carefully engineer all the systems involved. The e-Science and high quality media applications are now approaching the Terascale. We are currently conceptually thinking about Terabit networking. The Lambda networks have given access to an order of magnitude more bandwidth. The unit of bandwidth has shifted from packets to complete photonic channels of 10 Gb/s. However for application to be able to use the Lambdas effectively the programming models need to be adapted. The Lambdas must be treated in the same way as multiple cores and processors in parallel programs. Middleware as MPI and the use of multithreading turns 1000 computers of 1 GigaFlop each into a TeraFlop machine where one application can get the equivalent of a tera number of multiplications for its own use. We need to think of the equivalent of MPI in Networking such that applications can program an equivalent of a Terabit network for the benefit of one application. Examples of such applications are LOFAR, LOOKING, OPTIPUTER, CINEGRID. The solution is to work at removing or lessening the factors that inhibit the optimal use of the emerging new local and wide are photonic networks and to design systems, protocols, and grid middleware that empower applications to optimally allocate and use the infrastructure. This way applications can become location and distance independent except for the unavoidable limit of the speed of light.

References

[1] Thomas A. DeFanti, Maxine D. Brown, Cees de Laat, "editorial: Grid 2002: The International Virtual Laboratory", iGrid2002 special issue, FGCS, volume 19 issue 6 (2003).

[2] Cees de Laat, Erik Radius, Steven Wallace, "The Rationale of the Current Optical Networking Initiatives", iGrid2002 special issue, FGCS, volume 19 issue 6 (2003).

[3] Larry L. Smarr, Andrew A. Chien, Tom DeFanti, Jason Leigh, Philip M. Papadopoulos, "The OptIPuter," Communications of the ACM, Volume 46, Issue 11, November 2003, pp. 58-67.

[4] Leon Gommans, Freek Dijkstra, Cees de Laat, Arie Taal, Alfred Wan, Bas van Oudenaarde, Tal Lavian, Inder Monga, Franco Travostino, "Applications Drive Secure Lightpath Creation across Heterogeneous Domains", IEEE Communications Magazine, vol. 44, no. 3, March 2006

[5] H.E. Bal et al.: "The distributed ASCI supercomputer project", ACM Special Interest Group, Operating Systems Review, Vol. 34, No. 4, p 76-96, October 2000.

[6] J..J. van der Ham, F. Dijkstra, F. Travostino, Hubertus M.A. Andree and C.T.A.M. de Laat, "Using RDF to Describe Networks", iGrid2005 special issue, FGCS, Vol. 22 issue 8, pp. 862-867 (2006).

PROCEEDINGS OF SCIENCE

MUSIC and AUDIO – Oh how they can stress your network!

Dr R P Fletcher[1]
University of York
Heslington, York, YO10 5DD, UK
E-mail: r.fletcher@york.ac.uk

Nearly ten years ago a paper written by the Audio Engineering Society (AES)[1] made a number of interesting statements:

1. The current Internet is inadequate for transmitting music and professional audio.
2. Performance and collaboration across a distance stress beyond acceptable bounds the quality of service
3. Audio and music provide test cases in which the bounds of the network are quickly reached and through which the defects in a network are readily perceived.

Given these key points, where are we now? Have we started to solve any of the problems from the musician's point of view? What is it that musician would like to do that can cause the network so many problems? To understand this we need to appreciate that a trained musician's ears are extremely sensitive to very subtle shifts in temporal materials and localisation information. A shift of a few milliseconds can cause difficulties. So, can modern networks provide the temporal accuracy demanded at this level?

The sample and bit rates needed to represent music in the digital domain is still contentious, but a general consensus in the professional world is for 96 KHz and IEEE 64-bit floating point. If this was to be run between two points on the network across 24 channels in near real time to allow for collaborative composition/production/performance, with QOS settings to allow as near to zero latency and jitter, it can be seen that the network indeed has to perform very well.

Lighting the Blue Touchpaper for UK e-Science - Closing Conference of ESLEA Project
The George Hotel, Edinburgh, UK
26-28 March, 200

[1] Speaker

PoS(ESLEA)028

1. Introduction

The Audio Engineering Society (AES) published their white paper (AESWP-1001) in 1998, "Networking Audio and Music using Internet2 and Next-generation Internet capabilities"[1]. In this they made some key observations regarding the current state of audio and music over conventional networks:

1. The current Internet is inadequate for transmitting music and professional audio.
2. Performance and collaboration across a distance stress beyond acceptable bounds the quality of service
3. Audio and music provide test cases in which the bounds of the network are quickly reached and through which the defects in a network are readily perceived.

Additionally they observed that the designers of Internet2 foresaw its use for medical researchers, physical scientists and the leading-edge research community. Audio and music were relegated to a "background function". The AES called this "a major, if not disturbing judgement".

2. Discussion

It is important to understand the importance of the role of music in society. All known historical societies have engaged in some form of musical activity. It provides identity, intellectual stimulation and can evoke powerful memories. It is fundamental to many rituals, events and communication. It has also been shown to be a powerful tool for learning and growth. We ignore these facts at our peril.

Trained professional musicians are very sensitive to small shifts of temporal materials in the order of milliseconds. For example, a conductor can isolate minor tuning problems in an orchestral section, and usually can identify the actual musician as well. This is no trick, and is done by localising the position of competing signals reaching his ears using the temporal differences. This fact lends credence to the need for more and better surround sound systems for electronic reproduction or performance. If audio and music is to be transmitted across the new networks, then they need to provide this functionality and level of accuracy.

The Internet2 Quality of Service (QoS) Working Group published a survey of the "Network QoS Needs of Advanced Internet Applications" in 2002[2]. This paper showed the need to facilitate new frontier applications, to explore complex research problems and to enable seamless collaboration and experimentation on a large scale. The notion of a virtual research space and shared virtual reality was also noted. Additionally, real time access was shown to be a major requirement as was new levels of interactivity with multisensory cues.

From the above it does not require a huge leap to see how music and audio now fit into these "advance internet applications". Indeed, they consume most of the areas where higher

levels of QoS are required. Interestingly this paper did find that audio and video were major requirements. They also (for once) differentiated between high quality audio and professional quality audio. The needs of these two categories necessarily overlap, but the additional constraints placed on the network for the latter place huge demands on the networks.

It is important to differentiate between passive and active audio streams. A passive, delivery only system can use compression, jitter buffers and can adapt to bit rate changes. An active stream will most likely have no compression, or at least lossless compression, will work in real time and for some streams they will be two way, and hence round trip times will need to be factored into the equation. There is a plethora of delivery formats, e.g. mp3, aac, real, wav, wma to name a few. In the main we can deliver stereo as a file (listen later) or as a stream (listen now), but what of the other formats, e.g. 5.1, 7.1, or the many ambisonic types?

The "quality" of any audio is always open to debate, and no more so than that delivered over the net as "near CD quality" at 128 kb/s mp3! But on earphones can you really tell the difference and more importantly, do you care? For "better quality", one can always up the encoding bit rate. We will also have surround sound delivery in the near future thanks to the new mpeg surround sound standard published on 12 February 2007[3]. Work is also in progress to encode some ambisonic types into vorbis .ogg streams.

Another debate concerns compression (or not). There are lossless and lossy compression types. Examples of lossless formats are FLAC, Apple lossless, Dolby TrueHD, Monkey's Audio, TTA, wavpak, WMA lossless, and examples of lossy formats are mp3, adpcm, ATRAC, Dolby Digital, Musepack, TwinVQ, Vorbis ogg and WMA. In the lossless world, FLAC and wavpak are very popular and in the lossy world, mp3, ATRAC, vorbis ogg and WMA are popular. Of course, the Dolby standards are ubiquitous as well.

The final debates are about how many bits are required and at what sample rate. Many will argue that 16 bits is enough and can encode all the frequencies we can hear. But, high frequencies (even those we cannot hear) will colour lower frequencies through interference. Whether we can all hear these subtle changes is debatable. However, it is common to use a minimum of 24 bits when recording to allow for headroom and 32 bits is commonly used at the mixing stage. However, the newer range of hardware mixing consoles and software mixing programs have moved to 64-bit pathways. In fact, 64-bit IEEE floating point format is often advocated. There are endless arguments about significant bits in the various representations of floating point numerical data in computers. None more so than the effects of dithering when moving from a full 64-bit mix down to a 16-bit mix ready for a CD!

And, what of sample rates? Again, there those who would contend that 44.1 KHz is acceptable for all, i.e. at 16-bits, this is what is on a CD. However, 48 KHz is used by DAT and audio on DVD-Video (and let us not forget that many consumer grade soundcards run natively at higher rates and down sample from 48 KHz to 44.1 KHz, often with quite poor results!). 96 KHz is used by recording engineers and is the rate on DVD-A, from stereo to 5.1 surround formats. Higher rates can be found on DVD-A, 192Khz for mono and stereo, and interestingly, 88.2Khz and 176.4 KHz is also seen (heard?) on DVD-A.

Musicians have been collaborating over the net for many years exchanging files, and in some cases actively collaborating with "jamming" systems, e.g. RocketNetworks using midi data streams and recently with NINJAM using audio streams. Lossy compression schemes

should not be used due to inherent signal degradation when going through many encode/decode cycles. Lossless compression can be used but the codecs will introduce more latency into the system. Smart systems which detect silence and only send metronome signals for synchronisation are beginning to be used.

From the network point of view the engineers need to look at the worst case data rates, if the network can cope with these, then everyone will be satisfied. Therefore, 192 KHz at 64-bits is the highest data format likely to be used at present – this is 11.7 Mb/s in uncompressed format per channel. Therefore, if we scale up to a basic 24 channel mix, then the data rate works out at 281.25 Mb/s. Thus, for composers to work together at remote sites in near real time we would need the audio to flow between the two sites, most likely at half the data rate (i.e. 96 KHz). The composers would need to trade off compression against extra latency. In the end, the latency due to the laws of physics has to be dealt with in some form, and having a predictable latency with close to zero jitter is the goal.

The new photonic networks open up many possibilities for such collaborative work. 1 Gb/s connections are commonplace. In the UK, the UKLight network (part of SUPERJanet-5 infrastructure) can be used to interconnect sites at rates up to 10 Gb/s. In turn lightpaths can be provisioned to sites in Europe via GEANT2 and to the USA via STARLight and other country-wide infrastructures (e.g. National Lambda Rail) and to Canada via CANARIE.

To date, and unsurprisingly, most of the use of such connectivity has been for "big science" to have access to very large datasets or high performance computing, aka Grid computing. However, there have been a number of successes in the Arts and Humanities fields with musicians taking master classes remotely, using HDTV video streams and full surround sound audio. In a similar vein, remote collaborations have taken place with jazz ensembles and increasingly, cultural exchanges via HDTV and audio across multiple sites across the globe have been happening.

The next "big thing" to encompass all of the above is the push towards digital cinema, and the cinegrid project[4] is just one of these. The video is at least 4096 x 2048, and may use up to 24 channels of surround sound. The data rates for this format are not inconsiderable, as are the constraints put on the QoS of the networks in use. Add to this the requirement for remote collaboration when creating content for this next generation of cinema experience, we have applications in the audio (and video) world which will really stress the best networks.

3. Conclusion

From the above discussion it should be clear that audio data can indeed make considerable demands on the networks on which it will be running. It should be noted that these demands become more critical as the applications in use move more towards the real time environment. The viability of collaborative composition and performance environments and their adoption into mainstream use in the Performing Arts arena will be in part driven by the ability of the network engineers and designers to provide the QoS required by the audio streams. Add the video dimension into this equation, and the requirements become even more demanding. It may be that for many of those engaged in these new exciting areas more dedicated network services

will have to be provided. To date we see this happening in the broadcasting domain where the main players in the field have their own networks for the transmission of TV and radio content. In the academic research domain we are lucky to have access to the new generation of high speed networks, although it is the case that most of the activity on these networks is associated with "big science", e.g. high energy physics, astronomy, medical science etc. Gaining access to these networks is not trivial, and provisioning the appropriate links to a performance space often requires considerable work and expense. It also requires gaining access to a number of professionals not usually engaged in working with "artists", and this in itself can present a considerable "challenge" for all parties concerned.

It is worth noting that the tools required to create content need to be written and we also need a new breed of artists to design the content both in the video (which may be multiscreen) and audio domains (which will be in surround sound and even 3D). Add to this the challenge of remote collaboration/composition/performance, it is clear that these new tools must embrace the network technologies from the ground up. Also, these e-artists need to interwork with the e-scientists to draw upon their existing expertise.

Finally, it is heartening to see these issues being aired, and even more important to be able to work alongside established e-scientists who now appreciate the issues associated with these media on the network. To put it simply, it's just data, lots of it, and it needs to run quickly, very quickly, and we really must not lose any of it.

References

[1] R. Bargar et al. (1998). "Networking Audio and Music Using Internet2 and Next-Generation Internet Capabilities." TC-NAS/98/1. Audio Engineering Society.

[2] Dimitrios Miras. (2002) "Network QoS Needs of Advanced Internet Applications – A Survey.

[3] See www.mpegsurround.com

[4] See www.cinegrid.org

PROCEEDINGS OF SCIENCE

Who "owns" the network?: a case study of new media artists' use of high-bandwidth networks

Frédérik Lesage[1]
Department of Media and Communications
London School of Economics and Political Sciences
Houghton Street
London WC2A 2AE
United Kingdom
E-mail: f.lesage@lse.ac.uk

The objective of this paper is to briefly give an overview of a research project dealing with the social construction of use of information communication technologies among new media artists interested in online collaboration. It will outline the theoretical and methodological tools applied to the case study of the MARCEL Network.

Lighting the Blue Touchpaper for UK e-Science - Closing Conference of ESLEA Project
The George Hotel, Edinburgh, UK
26-28 March, 2007

[1] Speaker

PoS(ESLEA)030

1. Introduction

Near the end of his classic work on the social construction of art and the coordination of artistic activity, Howard S. Becker (1982, pp.314-350) details a fascinating comparison between the development of two similar technologies with somewhat promising artistic applications in the late 19th and early 20th century. The first of these technologies is the stereoscope, the second is the photograph. In his account, he describes how the latter went on to be accepted by most of society from the amateur to international high art institutions as a tool for artistic creation while the former became a temporary fad that was quickly relegated to obscure collections of curiosities. Although Becker is unable to give the reader a definitive answer as to why one succeeded while the other did not, it is fascinating to extend his reflections to new media technologies and to ask how some of these may one day contribute to the work of a contemporary Stieglitz. One approach could involve documenting how these digital technologies are currently employed by artists. This way, it may be possible to contribute to a wider understanding of the dialogical relationships between artists, the tools they use, and the wider art world that surrounds them.

With this is mind, a case study has been put together to observe how artists apply the network metaphor to information and communication technologies (ICTs), specifically high-bandwidth academic networks, in order to coordinate the collaborative online production of new media artworks. The two main research questions are:

1) How does the network metaphor enable media artists to coordinate the online collaborative production of art works?

2) How does it enable the coordination of consumption/use of ICTs?

This paper will briefly set out the theoretical and methodological framework guiding this case study as well as provide some observations stemming from preliminary

fieldwork. It will then attempt to demonstrate how the research structure and its findings might benefit those working in the field of e-science.

2. Theoretical framework

Research in the social sciences pertaining to artists often focuses on their role as producers. The production of culture perspective (Peterson & Anand 2004), for example, applies organisational sociology to the study of how producers and distributors organise in order to better understand the dynamics of power that structure the meanings of cultural products before they eventually make it to the general public. But where the paint of a painting comes from is rarely of interest to the viewer. Some, such as those in the field of audience studies (Livingston 1998), would argue that the viewers themselves can generate all sorts of meanings independently of the producers intentions. But in the case of this research, the relationship between the artists, their support personnel and the new media technologies remains unstable. The research must focus instead on the "dynamics of *uses*" (Martin-Barbero 1993), the shifts between strategies and tactics (de Certeau 1984) and between the role of producer and consumer among actors and organisations that are trying to acquire or maintain creative power. We are observing what some in the 19th or 20th century might have called an "avant-garde movement" where experimentation with conventions is taking place. But these days it is called a "network" of "project managers" and "web developers". In order to this, the research calls on a theoretical framework that combines the model of conventions as developed by Becker (1982) with Martin-Barbero's (1993) model of mediation in order to allow the researcher to follow the articulations between institutional traditions (such as the contemporary art world, new media, and academia), the actors that maintain these traditions, and the technologies that support their activities.

Employing Lammers and Barbour's (2006) model of institutions it will be possible to identify institutional discourses and practices relating to the art world and to new media and to see how they are reproduced by individual actors and organisations in order to mediate the conventions pertaining to the use and production of ICTs. This may in turn

provide a glimpse into the ways ICTs are selected and used by individual artists and arts organisations. It may also provide us with the tools to observe whether institutional power relations enable the social construction of the new media artist as an empowered actor in the field of new media.

3. Research design and methodology

The case study selected for the research is the MARCEL Network. It is first conceived during a series of conferences in the late-1990s in Souillac (Foresta & Barton 1998). Following this, a number of artists and other new media art practitioners set out to build "a permanent broadband interactive network and web site dedicated to artistic, educational and cultural experimentation, exchange between art and science and collaboration between art and industry" (MARCEL 2004). By 2001 experimentations with high-speed academic networks between Le Fresnoy (France), the Wimbledon School of Art (UK), and Ryerson (Canada) are attempted. The Public in West Bromwich and other new media centres and academic institutions across Europe and North-America (Ibid) soon follow suit. Most of the experimentations centre on the realtime collaborative potential of video-teleconferencing software such as Access Grid.

The research design consists of a multi-sited ethnographic case study (Marcus 1998) of the MARCEL Network's activities. The researcher will follow the network's activities over a two and a half year period, visiting locations across Europe and North America with the objective of generating:

1) Multiple "career threads" through document analysis and interviews. It is hoped that, by not only documenting the careers of individual artists (Peterson 2004) but also documenting the careers of the technologies used (Kopytoff 1986, Silverstone et al. 1991) and the organisation as a whole, it will be possible to produce a sufficiently clear historical context for the Network's activities.

2) Field notes and audio recordings of participant observation of Networks activities as well as interviews of key actors for ethnographic and discourse analysis.

These two empirical objectives respectively constitute a moment of socio-historical analysis and a moment of formal analysis which will then be combined in a moment of interpretation-reinterpretation (Phillips and Brown 1993) as part of a complete critical hermeneutic methodology. Using this approach allows the researcher to attempt to triangulate the observations in order to distinguish particular conventions as well as go further into an analysis of how these conventions are mediated when communicated between different actors or organisations in the field. The principal research objective is to come to a critical analytical understanding of artists' application of the network metaphor to ICTs in order to coordinate the production and use of new media artworks. Although it will be impossible to reach a definitive objective conclusion, it is hoped that such an imprecise description can eventually lead to a better shared understanding of creative activity in new media.

4. Preliminary findings

The case study has passed its mid-point phase. Although there remains much work to be done before presenting any compelling findings, it is possible to present a few preliminary observations and hypotheses. Most new media artists encountered over the course of the research seem to face considerable challenges due to limited resources, both in terms of financial and institutional support. Although the word "network" is somewhat inconsistently applied by actors, it does seem to preserve a certain particular characteristic across most of the organisation's members which allow them to deal with this limited support. The characteristic can be summarized as an implicit understanding of the importance of disseminating and maintaining particular kinds of "ownership" (Strathern 1996) of conventions relating to the use of ICTs, one of which I will call "squatting". Although it is impossible to adequately develop and qualify this research's use of Strathern's notion of ownership and its theoretical influence on the the notion of "squatting", it is possible to briefly define "squatting" as the tactical use of institutional power in order to share and develop conventions relating to the use/consumption of ICTs in new media art.

5. Conclusion

Although it is certainly impossible to predict whether these artists' choices of technologies will one day lead to as successful an art world as photography, it is my hope that this research will at least lead to a greater appreciation of the dynamics of art world activity. The two principle objectives of this paper have been to present:
1) Conceptual tools for the study of organisational networks using ICTs, particularly how certain types of mediation of conventions, like ownership, might lead to a better understanding of the diffusion of creative practices and discourses relating to new media.
2) An introduction to the activities of the MARCEL Network and a wider understanding of artists' challenges and interests when working with ICTs.

References

[1] Becker, H. S. (1982) *Art Worlds*, University of California Press, Berkeley and Los Angeles.

[2] de Certeau, M. (1984) *The Practice of Everyday Life*, University of California Press, Berkeley and Los Angeles.

[3] Foresta, D. and Barton, J. (1998) *Leonardo*, **31,** 225-230.

[4] Kopytoff, I. (1986) In *The Social Life of Things: Commodities in cultural perspective*(Ed, Appadurai, A.) Cambridge University Press, Cambridge, pp. 64-91.

[5] Lammers, J. C. and Barbour, J. B. (2006) *Communication Theory*, **16,** 356-377.

[6] Livingston, S. M. (1998) *Making Sense of Television – The psychology of audience interpretation*, Routledge, London.

[7] MARCEL Network (2004), *About MARCEL* (Ed, Foresta, D.) MARCEL Network, London. www.mmmarcel.org .

[8] Marcus, G. (1998) *Ethnography Through Thick and Thin*, Princeton University Press, Princeton, New Jersey.

[9] Martin-Barbero, J. (1993) *Communication, Culture and Hegemony: From the media to mediations*, Sage, London.

[10] Peterson, R. A. and Anand, N. (2004) *Annual Review of Sociology*, **30,** 311-334.

[11] Phillips, N. (1993) *Academy of Management Journal*, **36,** 1547-1576.

[12] Silverstone, R., Hirsch, E. and Morley, D. (1991) *Cultural Studies*, **5,** 204-227.

[13] Strathern, M (1996) *Social Anthropology*, **4**, 1, 17-32.

PROCEEDINGS OF SCIENCE

Always the Bridesmaid and never the Bride! Arts, Archaeology and the E-Science Agenda

Prof. Vincent Gaffney
Insitute of Archaeology and Antiquity
University of Birmingham, Edgbaston, Birmingham, B15 2TT, United Kingdom
E-mail: v.l.gaffney@bham.ac.uk

Dr. R. P. Fletcher[1]
Computing Service
University of York, Heslington, York, YO10 5DD, United Kingdom
E-mail: R.Fletcher@york.ac.uk

There is, without doubt, a strong tradition amongst the Arts and Humanities community of the gifted individuals: academics who can, and do, labour long and hard alone in libraries or museums, to provide significant scholarly works. The creation and organisation of large data sets, the desire for enhanced accessibility to data held in disparate locations and the increasing complexity of our theoretical and methodological aspirations inevitably push us towards greater use of technology and a reliance on interdisciplinary teams to facilitate their use. How far such a process has become established, however, is a moot point. As the director of one Arts-based Visualisation laboratory[1] that possesses an UKlight connection, I would probably observe that the Arts and Humanity community has, largely, remained aloof from many of the recent developments of large-scale, ubiquitous technologies, with the notable exception of those that permit resource discovery. It remains a fact that the emergence, for instance, of GRID technologies in other disciplines has not yet had the impact one might have expected on Arts and Humanities. It seems certain that reticence has not been the consequence of a lack of data within the Arts. Others, including archaeology, sit at the edge of the natural sciences and are prodigious generators of data in their own right, or consumers of digital data generated by other disciplines. Another assertion that may be considered is that Arts research is not amenable to large-scale distributed computing. To a certain extent, successful Grid applications are linked to the ability of researchers to agree methodologies that, to some extent, permit a “mechanistic” world view that is amenable to such analysis. However, in contrast it is not true that Arts research is either so individual, so chaotic or anarchic that it is impossible to demonstrate the value, at least, of e-science applications to our disciplines.

Lighting the Blue Touchpaper for UK e-Science - Closing Conference of ESLEA Project
The George Hotel, Edinburgh, UK
26-28 March, 200

[1] Speaker

PoS(ESLEA)031

1. Introduction

Potential Arts and Humanities E-research activities can be broken down, in general terms, as shown in Figure 1 and one suspects that most researchers would be comfortable with the assertion of one or more of these general themes as being central to their own work.

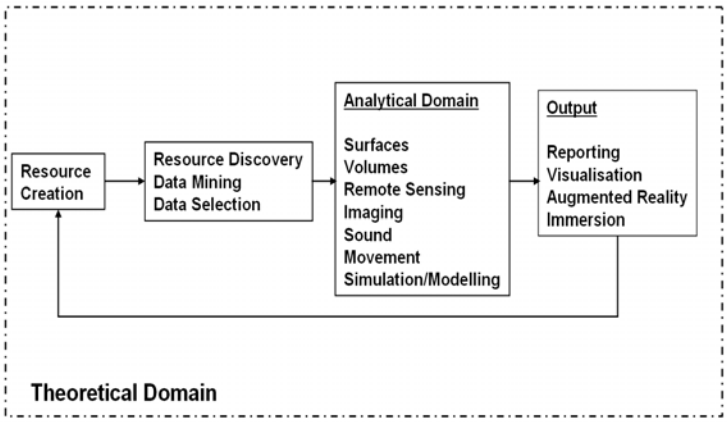


Figure 1. e-Research Activities

Visualisation, an area of specific interest for my own area of research (Landscape Archaeology), may be regarded as a useful point of departure in considering where the Arts stand in respect of forming an E-Science agenda and whether it may move towards a proportionate use of available technology, including high volume, low latency networks. Aside from the technical aspect of visualisation, visualisation carries the implication of being, in most senses, a final act within a larger research process that has involved the collection, selection and manipulation of data in some iterative manner.

2. Discussion

For the Arts the act of visualisation, therefore, carries considerable theoretical or methodological baggage. Intellectually the act of visualisation is a highly emotive act and, within archaeology at least, there has been considerable debate as to the legitimacy of visualisation as an isolated output. This is hardly surprising in a society in which dazzling imagery is ubiquitous and pervades everyday life. However, for historical disciplines it has particular significance. By definition, our reality is usually a degraded image of an original that has eroded or rotted away, or the proxy representation of past events through surviving records written, frequently, by people who never witnessed the events far from the places they lived. The impact that virtual reconstructions may have on the observer is inevitably more real than the heap of rocks and lines of postholes that we encounter daily on archaeological sites or the dry, detailed footnotes of a historical text. Media representations of historical events and archaeology in particular, seem particularly open to virtual representation and, perhaps, misrepresentation. In this context, we might consider the example of a digital model[2] of a Wild Goat style cup, see Figure 2, created from an original in the Museum of the Institute of

Archaeology at Birmingham. The cup itself is a relatively simple stemmed vessel and of interest mainly because of the suggestion that the painted design might indicate that the vessel was a skeuomorph whose metal prototype possessed incised angular designs that were repeated in paint on the clay copy. Some time ago, this cup was scanned and rendered with an image of the surface design. This model was used, with the surface of the design raised, to create a casting in metal but the digital model was also rendered haptically to permit the user to view the digital model either as metal or clay. The significance of the exercise was not so much the ingenuity of permitting visitors to touch, virtually, an object that would otherwise have only been experienced through the glass of a museum case: rather it was the fact that the visitor experienced an enhanced act of interpretation. It may be that there never was a metal original of the stemmed cup and its representation (digital or physical), remains an act of subtle imagination. The experience of interpretation is therefore a central role of visualisation rather than the simple representation of any particular reality.

Figure 2. Wild Goat style cup

The environment itself is a physical and viewable research asset that requires incorporation within interpretative schema. Imbued with meaning, we now study our physical and natural context in its own right and appreciate the ability of landscapes to manipulate human action as a consequence merely of its existence. There is an important point to be made here and that is, with respect to landscape, the significance of the new technologies is not simply power, but the application of the continuous measurement and analysis of space and the extension of the analytical sphere to virtually every part of a landscape may be equally significant. If so this means that the Arts will increasingly stray into the area of the environment, the natural sciences and, by implication, will therefore appropriate demands for substantial computing power.

Over the next 5 years or so one can predict that the emergent technologies for 3D landscape scanning from ground or air-base platforms[3] will further transform our capacity for control of space, filling the gap between ground-based geophysics and traditional aerial photography, through the reproduction of the surface of the landscape in an almost seamless fashion. Alternately, consider the potential of work piloted by Ian Haynes, with a Birmingham group at the Lateran where physical reconstruction of the superstructure of the church is

incorporated with volumetric rendering of data from ground penetrating radar to provide a wider interpretative context of this internationally important building.

However, infinitely finessing the resolution of data itself cannot satisfy our disciplinary aspirations to explain the past. Physical scale can become a driving force in its own right. The example here can be of route analysis which might vary in scale between local paths beside a farm to trade routes that span continents. Although analysis of any of these communication routes may be quite simplistic in itself the potential of scale, but also resolution, as a driver of computational need should be clear.

However, nothing demonstrates this more dramatically, perhaps than the Birmingham project on the Palaeolandscapes of the Southern North Sea.[4] This project seeks to explore the land inundated by the sea during the last great period of global warming. Between the end of the last glacial maximum and c. 6,000 BC an area larger than England was lost to the sea. This great plain was probably the heartland of the Mesolithic populations of North Western Europe. Man lived and walked the rivers and valleys of this country and the hills and the plains have since been lost beneath a remorseless onslaught of water. Today, we can barely trace the outline of this vast landscape. This month a research team in Birmingham, funded by the Aggregates Levy Sustainability Fund, is concluding the mapping of more than 20,000 km^2 of this landscape using 3-dimensional seismic data collected for the purposes of oil and gas exploration and provided by PGS. This is an archaeo-geophysical survey the size of the whole of Wales, see Figure 3. Using this data we have begun to trace, and even to name, the rivers, hills and valleys that have been lost to mankind for more than eight millennia, see Figure 4. We have not yet begun to finesse this relatively course data with the thousands of 2D lines, cores, borehole data, and the steps to recreate this lost world have begun to involve immersive VR, and intelligent life modelling. The past is not a foreign country it is an imaginary one, but that does not absolve us from the responsibility to study and learn from it.

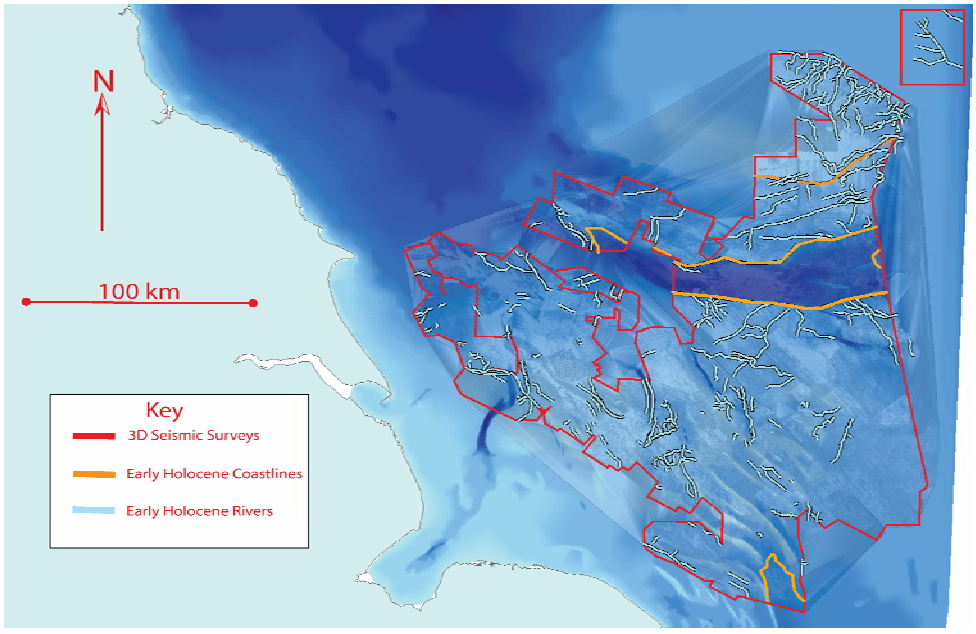


Figure 3. Area of Study

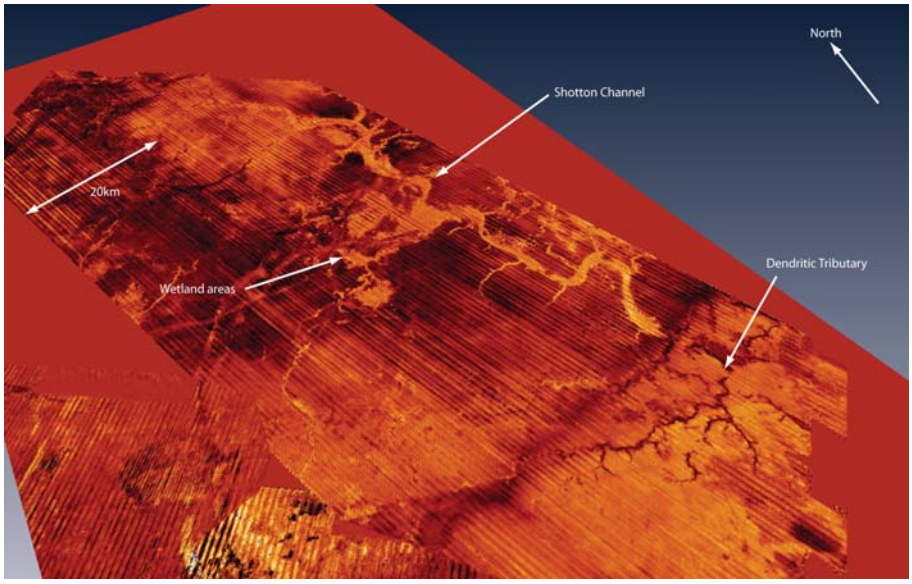


Figure 4 Rivers, Hills and Valleys

The intellectual freedom that is demanded for Arts projects, and the range of legitimate research routes to provide provisional statements, requires flexibility and the imaginative use of available technologies. However, whilst flexibility and imagination are desirable qualities in general, the intellectual and technological landscape relating to visualisation looks to have more in common with the Wild West rather than the Arcadian backdrop often more suitable for a considered national research strategy. Despite this, it is appropriate that some statement is made on both the practise and principles regarding ICT development in the Arts and Humanities. There are a number of specific points that I would also raise under these headings.

The number of groups that are currently involved in large-scale e-science programmes within the Arts is relatively small, nationally or internationally, and it is unlikely that any specific group will establish complete competence across the range of technologies likely to be used or be able to provide appropriate resource for all potential projects. However, it is critical that the United Kingdom possesses and develops the potential of existing expertise. Integrative technologies, most notably the Access Grid, perhaps linked with virtual network computing, will be central to linking the increasingly disparate groups that are required to study project which span the Arts and natural Sciences. Powerful low latency networks such as UKLight, with near seamless computing capabilities, may become central to development of distributed research programmes dealing with significant shared data sets. Integration with monuments, dress, and reception studies, particularly in Classics or Ancient History probably demand high quality photorealistic representation. The capacity for immersive replication of human movement is likely to become significant over time and appropriate data capture facilities and rendering capacity will follow from this.

Following this, I think it equally appropriate to highlight a number of strategic issues. It is essential for the Arts that technology is adopted on our terms. Even if the technology is shared, our requirements are not those of other disciplines. We must avoid, for instance, the situation relating to archaeological science, the responsibility for which was devolved to NERC.

External disciplines have advice to offer but they cannot be allowed to determine our research agendas. As a group the Arts must seek appropriate funding that allows infrastructural development. There has been a tendency in the past for "big science" to control dedicated computational facilities rather than the Arts. If we do not have appropriate funding, then we will fail to establish an appropriate pool of expertise within our disciplines and our research agendas will suffer as a consequence.

The nature of modelling behaviour is such that we require resources capable of modelling exponentially expanding data sets and the complexities of human action. In areas where we touch the physical sciences, the Arts must demand parity for our research with physical geography, geology and the Social Sciences. As a bottom line we have to be confident, and active, in our assertion of the significance of Arts and Arts visualisation. We are part of the larger visualisation market that feeds the cultural economy and this should be recognised.

3. Conclusion

In conclusion, we must acknowledge that large data and their use is a challenge for the Arts. However, this is not simply because visualisation is technically complex or even because it is expensive. Rather this reflects the fact that so much of our visual content is itself contentious. However, we cannot sidestep tension in interpretation as this may, in respect of multivocality and reception, be central to academic exploration. In such situations, sophisticated visualisation may well be the most appropriate route to investigate such phenomena. Following from this, if we do accept that complex visualisation is to be part of our academic strategy there can be no half measures in providing resource. Whilst it is good to share, and we have much to learn from other disciplines in relation to specific technologies, we must accept that responsibility regarding the relevance of visualisation technologies to the Arts, and the implementation of these technologies, is ours alone.

4. Acknowledgements

I would like to thank the following for their comments and suggestions when writing this paper - Paul Hatton, Meg Watters, Andy Howard, Keith Challis, Dr Henry Chapman, Ben Gearey, Helen Goodchild, Simon Fitch, Dr Ken Thomson, Dr Mark Jolly, Dr Lorna Hughes, Dr Sheila Anderson, Dr Stuart Dunn, Dr Roger White, Helen Gaffney, Professor Bob Stone, Eugene Ch'ng, Dr Ian Haynes.

References

[1] http://www.iaa.bham.ac.uk/Computing/HP_VISTA/HPindex.htm

[2] http://www.iaa.bham.ac.uk/research/opening/main.htm

[3] http://www.iaa.bham.ac.uk/bufau/services/3Dscan/Lidar.htm

[4] http://www.iaa.bham.ac.uk/research/filedwork_research_themes/projects/North_Sea_Palaeolandscapes/index.htm

PROCEEDINGS OF SCIENCE

Exploitation of Switched Lightpaths for e-Health: Constraints and Challenges

Lee Momtahan*
Oxford University Computing Laboratory
E-mail: Lee.Momtahan@comlab.ox.ac.uk

Andrew Simpson
Oxford University Computing Laboratory
E-mail: Andrew.Simpson@comlab.ox.ac.uk

The Exploitation of Switched Lightpaths for e-Science Applications (ESLEA) project is evaluating the feasibility of using switched lightpath networks to support various e-Science applications. This feasibility is being investigated via case studies, one of which pertains to the use of such networks to support distributed healthcare research and delivery. We consider some of the constraints and challenges associated with the use of switched lightpath networks in this context.

Lighting the Blue Touchpaper for UK e-Science - Closing Conference of ESLEA Project
March 26-28 2007
The George Hotel, Edinburgh, UK

*Speaker.

PoS(ESLEA)032

1. Introduction

The future of networking is moving towards switching in the optical domain, with the UK's effort being represented by the UKLight network. Within the UK, the national e-Science Programme [1] helped drive the development of infrastructures and applications to deliver 'large-scale' science. In this spirit, the Exploitation of Switched Lightpaths for e-Science Applications (ESLEA) project [2] aims to pioneer the use of UKLight to support a variety of e-Science applications. The concern of this paper is the e-Health 'mini-project' of ESLEA.

A two-pronged approach has been taken within the e-Health mini-project: the consideration of generic issues and the consideration of more specific issues pertaining to a particular application area have been split into two parallel tasks. Initial experiences with respect to the former were reported in [3]; initial experiences with respect to the latter were reported in [4].

The specific application associated with the e-Health mini-project pertains to typical use cases derived from the Integrative Biology project (IB) [5, 6], which is developing a customised virtual research environment to support researchers in the integration of biological information across multiple spatio-temporal scales. IB has established a community that is committed to sharing technologies and best practice with respect to heart and cancer modelling. The exploitation of networks such as UKLight has the potential to improve remote visualisation and to allow models to be run at more than one site—even when they are tightly coupled.

Current models of action-potential propagation within whole-heart models can involve finite element meshes with tens of millions of nodes, and up to 50 solution variables. Solutions in real-time for just a single heartbeat require (at least) millisecond temporal accuracy, resulting in data generation in the 10–100 TeraByte range from a single simulation. Establishing the feasibility of running a resource-intensive simulation model of electrical activity in the heart across a number of geographically distributed sites, using UKLight to provide the necessary low latency network, could offer significant benefits.

The structure of the remainder of this paper is as follows. In Section 2 we expand upon our use cases drawn from IB. In Section 3 we describe our experiences in trying to realise these use cases. Finally, Section 4 provides a brief summary of this contribution.

2. ESLEA and IB

IB is developing a customised virtual research environment that is capable of supporting research scientists in the integration of biological information across multiple spatiotemporal scales. Such integration could provide comprehensive descriptions at the system level, which can aid in the determination of biological function in both normal- and patho-physiology. It is intended that the tools developed by IB will support and accelerate the work of clinical and physiological researchers. The long-term beneficiaries of the work should be patients with heart disease and cancer, which together cause 60% of all deaths within the UK.

To this end, IB's end-users are drawn from the bio-mathematical modelling, cardiac physiology, and clinical oncology communities. These groups all wish to make use of mathematical modelling and HPC-enabled simulation to investigate multi-scale and multi-physics phenomena in the modelling of the heart or cancer. Much of this work is grounded in experimental data and involves

the development of complex mathematical models, usually involving the solution of non-linear systems of coupled ordinary and partial differential equations (or equivalent stochastic representations of the underlying processes on very small spatial scales) in complex, deformable three-dimensional geometries. In addition, very large quantities of simulation data are being generated which must be validated against experiments, curated, synthesised, analysed, and visualised.

IB is using, and has used, a variety of computational and storage resources distributed throughout the UK, including HPCx at Daresbury, CSAR at Manchester and NGS (National Grid Service) Data Clusters at the Rutherford Appleton Laboratory (RAL).

The activity pertaining to IB and ESLEA was split into three phases.

The intention of the first phase was to set up a link between Daresbury and RAL, with a view to transferring TeraBytes of simulation data between the main computation site at Daresbury and the main storage site at RAL.

The plan for the second phase was to utilise the link established in the first stage for the purposes of visualisation—with data being streamed for storage also being passed to visualisation servers based at RAL. The motivation for this was that such processing had the potential to reduce the high bandwidth stream of raw simulation results into a relatively low bandwidth stream of geometry information—which could then then be shipped to the end-user for rendering on their desktop over normal low bandwidth networks.

The third phase was more ambitious, and involved the execution of a heart model across heterogeneous compute resources. In modelling heart mechanics, the equations that model the mechanical pumping action are coupled to the electrical activity generating the muscular contractions via a large system of ordinary differential equations which model the ion flows through the cell membrane. A potentially attractive approach is to solve the ODE systems on a massively parallel processing (MPP) architecture whilst simultaneously solving the mechanical problem on an symmetric multi processing (SMP) architecture. To make this approach feasible would require very rapid communications between the HPC resources, with minimal latency, jitter and packet loss. The intention was to use UKLight to couple the MPP and SMP resources (respectively the HPCx and CSAR facilities), to provide an optimal architecture on which to execute the heart model.

3. Experiences

First, a large number of people external to the ESLEA and IB projects (such as systems administrators of the end systems and local networks at the relevant sites) were required to facilitate the first, relatively simple, use case. There was initially an understandable reluctance to connect the storage facility to UKLight in case the potentially very large workload from this connection caused a denial of service to other users. Eventually, however, the relevant parties agreed that the use case was feasible and would work on it.

To execute the first use case, it took on the order of a month to install the necessary Globus client software and obtain grid certificates to connect to the NGS Data Clusters at RAL. One of the factors with the software installation was that there were multiple sources of information, which were slightly different. For example, there is a generic installation guide from Globus, but this contains none of the specifics of the UK NGS. There was also a step-by-step instruction guide produced by the IB project, but this was in part out-of-date as the UK NGS was in a ‘state of

flux' at the time. In addition, advice involved reinstalling the Globus software as root—as non-root installs were not recommended; this, obviously, involved a significant amount of rework.

Eventually, these problems were overcome and it was possible to make some baseline measurements for the data transfer use case. The next task was to make the same measurements with the data routed over the UKLight network.

Even when UKLight was correctly provisioned, tests revealed that the end-to-end data transfer was an order of magnitude below what was expected. Diagnosing this problem proved very difficult. The ESLEA e-Health project did not initially have login access to the relevant nodes of the SRB storage facility, a prerequisite to performing the most basic of diagnostics. The security policy associated with the SRB prohibited external login, but after a period of time we successfully negotiated the login access we required.

Diagnostic tests could then be performed. The end-to-end network was tested using the iPerf program. iPerf needed to be compiled on HPCx which runs the AIX operating system, but this did not work out-of-the-box. The HPCx help desk could not resolve the issue, although it was at least responsive. Members of the wider ESLEA community were drawn upon for support with compiling iPerf and this proved extremely fruitful. The diagnostic tests revealed that the network was the bottleneck, rather than the end systems.

Unfortunately, drilling down into the network issue was not possible for the ESLEA e-Health project. The ESLEA project is not aware of the topology of the network between the end systems and the UKLight point of presence, and in any case has no administrative rights to the relevant network devices, which would normally be needed to run diagnostics. Therefore, the problem was passed to the local area network support. This problem was never resolved.

The second use case required modifications to the visualisation software being developed for IB. Instead of post-processing simulation results from a data file, it was required to work with a stream of data being received from the network. Understandably, those responsible for the visualisation software were not willing to make these changes, and it would have been difficult for anyone not already working on the software to make such changes; as such, the use case was dropped.

The software for the third use case was developed by the ESLEA e-Health project working closely with members of the IB project. The software took longer than expected to be developed. The scope of the software was reduced in that just an electrical model was developed. Nevertheless the software was capable of being run in parallel over the UKLight network, using MPICH-G2: a grid-enabled implementation of the Message Passing Interface for parallel computation.

CSAR went out of service before the end of the project, and therefore an attempt was made to run the software between the TeraGrid and HPCx. However, due to issues associated with the compilation of the large software stack, this was not achieved within the time-frame of the project.

Leveraging the IB experience, we see that collaboration within the heart modelling community stems from the development of a trust between scientists—and often these collaborations require the users to develop 'experiments' where disparate groups are required to run simulations, visualise results and contribute to joint papers. Scientists invite colleagues to participate in an experiment and require deployed systems and networks to ensure that these experiments, meshes, generated data and visualisation results are kept secure from competitors. The utilisation of high performance computing facilities requires users to manage accounts on multiple facilities and transport information across the world to secure data stores for real time visualisation or later analysis.

To realise the potential of utilising UKLight or similar networks for applications drawn from projects such as IB requires significant effort from many different parties—some of whom will never benefit (directly or indirectly) as a result of the activity. As often seems the case with large-scale projects, the political and social challenges have the potential to outweigh the technical ones. And if our experiences from ESLEA tell us one thing it is that if the potential of using switched lightpath networks for distributed heart modelling is to be realised fully, the political and social challenges will have to be tackled before any meaningful attempt can be made at tackling the technical ones.

4. Discussion

The e-Health 'mini-project' was an ambitious attempt to demonstrate the potential of switched lightpaths for e-Health applications. The chosen application was, in the authors' opinion, the correct one—the potential benefits for the heart modelling community are significant, especially if the resources provided by the IB infrastructure are being leveraged. Unfortunately, the hurdles faced—coupled with a lack of 'buy-in'—meant that progress was difficult. Our work has demonstrated that while other application areas have a pressing need for switched lightpaths *now*, the urgency of the need for them within the e-Health context is not quite as crucial—but we would argue that it is on its way.

References

[1] A. J. G. Hey and A. E. Trefethen. The UK e-Science programme and the grid. *Proceedings of Computational Science (ICCS 2002), Part I* 3–21, Springer-Verlag Lecture Notes in Computer Science volume 2329, 2002.

[2] C. Greenwood, V. Bartsch, P. Clarke, P. Coveney, C. Davenhall, B. Davies, M. Dunmore, B. Garrett, M. Handley, M. Harvey, R. Hughes-Jones, R. Jones, M. Lancaster, L. Momtahan, N. Pezzi, S. Pickles, R. Pinning, A. C. Simpson, R. Spencer, and R. Tasker. Exploitation of switched lightpaths for e-Science applications (ESLEA). *Proceedings of the 2005 UK e-Science All Hands Meeting*.

[3] L. Momtahan, S. Lloyd, and A. C. Simpson. Switched lightpaths for e-health applications: issues and challenges. *Proceedings of the IEEE Symposium on Computer Based Medical Systems 2007*.

[4] L. Momtahan and A. C. Simpson. Switched lightpaths for e-health applications: a feasibility study. *Proceedings of the IEEE Symposium on Computer Based Medical Systems 2006*.

[5] D. J. Gavaghan, A. C. Simpson, S. Lloyd, D. F. Mac Randal, and D. R. S. Boyd. Towards a grid infrastructure to support integrative approaches to biological research. *Philosophical Transactions of the Royal Society: Mathematical, Physical and Engineering Science* 363(1883):1829–1841, August 2005.

[6] S. Lloyd, D. J. Gavaghan, A. C. Simpson, M. Mascord, C. Sieunarine, G. Williams, J. Pitt-Francis, D. R. S. Boyd, D. Mac Randal, L. Sastry, S. Nagella, K. Weeks, R. Fowler, D. Hanlon, J. Handley, and G. de Fabritis. Integrative Biology: the challenges of developing a collaborative research environment for heart and cancer modelling. *Future Generation Computer Systems* 23(3):457–465, March 2007.

PROCEEDINGS OF SCIENCE

Monitoring the ESLEA UKLight network

Barney Garrett[1]
The University of Edinburgh
James Clerk Maxwell Building, Mayfield Road, Edinburgh. EH54 6TD. UK
E-mail: barney.garrett@ed.ac.uk

The eScience applications taking part in the ESLEA project aim to utilise the switched light paths offered by UKLight to deliver massive datasets and prove bandwidth intensive data streaming technologies. For debugging, analysis and reporting, the applications need to be able to monitor their bandwidth utilisation. This paper looks at the methods that could be used and the implementation of those technologies within the UKLight infrastructure.

Lighting the Blue Touchpaper for UK e-Science - Closing Conference of ESLEA Project
The George Hotel, Edinburgh, UK
26-28 March, 2007

[1] Speaker

PoS(ESLEA)037

1. Introduction

The ESLEA[1] project's objective is to utilise the switched lightpath network, UKLight[2], provided by UKERNA[3] to support the data transfer requirements of the eScience community. In order to assess the success of the project and to be able to provide reporting data the collection of traffic throughput and utilisation is essential. The project has UKLight links terminating in the UK at UCL, Lancaster, Manchester and Edinburgh, and due to the nature of UKLight it is only possible for us to monitor the network traffic at these points [Figure 1]. The monitoring has been implemented in several stages, with each stage attempting to improve the usefulness of the collected data.

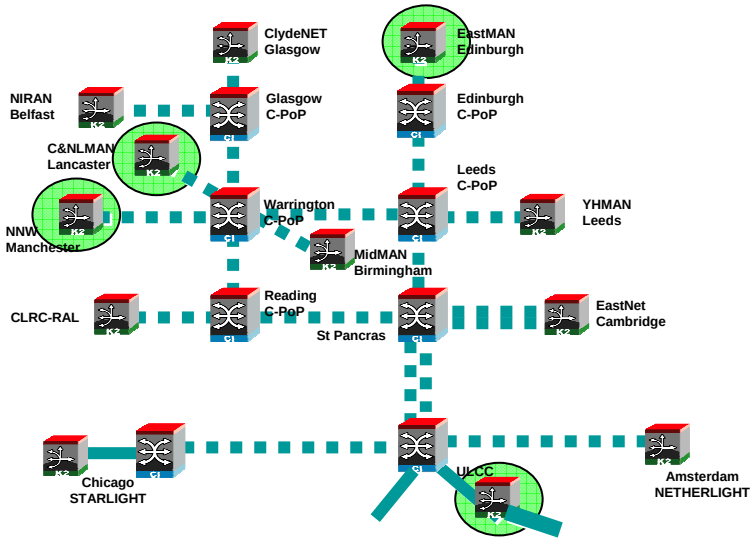


Figure 1

2. MRTG

The Multi Router Traffic Grapher (MRTG)[4] is a tool to monitor the traffic load on network links. MRTG generates HTML pages containing images which provide a visual representation of this traffic [Figure 2]. The first iteration of our "monitoring system" was to install an MRTG monitoring server attached to the project routers at each of the sites. This gave us basic utilisation monitoring capabilities for each of the UKLight links.

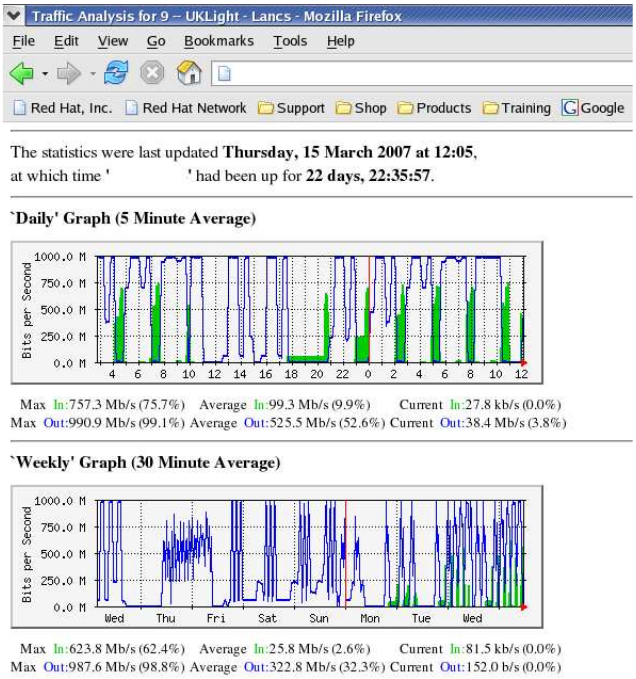


Figure 2

3. Weathermap

A "weathermap"[5] style of visualisation was made to bring all the MRTG resources together into a simple and easily understood interface. It acquires the information provided by the MRTG system and displays it on a map to represent the logical topology of the ESLEA project UKLight links [Figure 3][6].

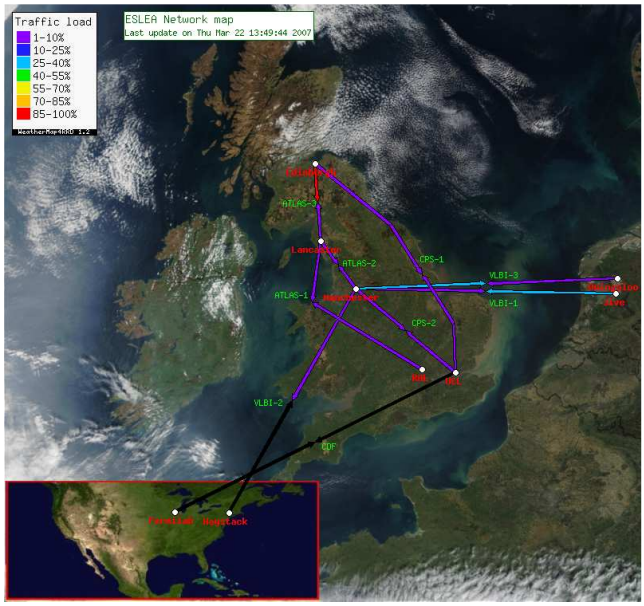


Figure 3

4. RRDTool and PerfSonar

The MRTG system is limited to taking samples every 5 minutes, and averages the data down into longer periods when looking at historic data which can lead to misleading figures. To overcome these limitations a Round Robin Database (RRD)[7] was then deployed. This is a tool set that allows the storage of time-series data in an efficient and systematic manner and provides tools for analysing that data. Using RRD we increased the sample collection rate to every 60 seconds, and increased the accuracy of the figures when looking at historic data.

The PerfSonar project[8], the result of a joint collaboration between ESnet, Geant2, Internet2 and RNP, is a web services infrastructure for network performance monitoring that provides services for publishing the data stored within these RRD files. By utilising the PerfSonarUI[9] it is possible to interrogate these web services to dynamically produce graphs similar to the ones produced by MRTG showing inbound and outbound throughput [Figure 4] but with more granularity and other additional features such as the ability to zoom in on areas of particular interest.

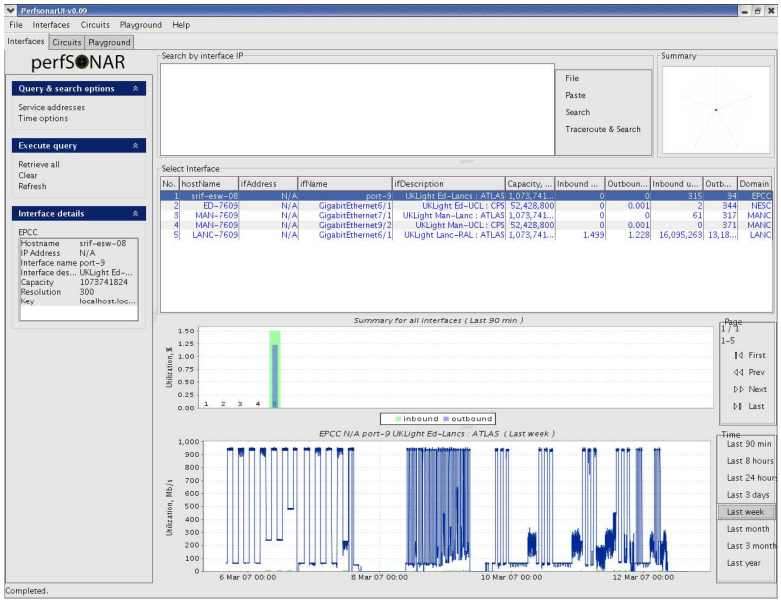


Figure 4

5. Summary

All three tools can provide useful utilisation metrics to a project assuming appropriate access to the networking equipment can be arranged. Both MRTG and the Weathermap tools were easy to install and configure but are limited in their resolution. The Weathermap in particular provides an easy to digest view where there are a larger number of connections being monitored. PerfSonar is more suited to being deployed by the network service providers rather than on a project level as is has a rather convoluted installation and configuration process, it is however by far the most advanced and flexible of the systems that we used.

References

[1] ESLEA http://www.eslea.uklight.ac.uk

[2] UKLight http://www.uklight.ac.uk

[3] UKERNA http://www.ukerna.ac.uk

[4] MRTG http://oss.oetiker.ch/mrtg

[5] Weathermap4php http://weathermap4rrd.tropicalex.net/index.php

[6] ESLEA Weathermap http://www.eslea.uklight.ac.uk/weathermap

[7] RRD Tool http://oss.oetiker.ch/rrdtool

[8] PerfSonar http://www.perfsonar.net

[9] PerfSonarUI http://perfsonar.acad.bg

PROCEEDINGS OF SCIENCE

VLBI_UDP

Simon Casey[1]
The University of Manchester, Oxford Road, Manchester, UK
E-mail: scasey@jb.man.ac.uk

Richard Hughes-Jones , Ralph E. Spencer, Matthew Strong
The University of Manchester, Oxford Road, Manchester, UK
E-mail: R.Hughes-Jones@manchester.ac.uk
E-mail: res@jb.man.ac.uk
E-mail: mstrong@jb.man.ac.uk

This paper describes the VLBI_UDP application, which has been designed to transport VLBI data using the UDP protocol. Modifications to provide additional features such as file access and packet dropping are described, and results obtained from tests conducted with the application are presented.

Lighting the Blue Touchpaper for UK e-Science - Closing Conference of ESLEA Project
The George Hotel, Edinburgh, UK
26-28 March, 200

[1] Speaker

PoS(ESLEA)038

1. Introduction

e-VLBI requires vast quantities of data to be sent from several outstations (telescopes) over high-speed computer networks and the Internet to a single correlator. Currently, VLBI data rates of 512 Mbit/s are achievable using the Transmission Control Protocol (TCP) [1]. An alternative to TCP is User Datagram Protocol (UDP), which is what VLBI_UDP relies upon.

2. The case for UDP

Whilst e-VLBI in the EVN can run at 512 Mbit/s with TCP, if longer baselines are used, for example across the Atlantic, TCP may struggle to sustain a constant 512Mbit/s should any packet loss occur. TCP guarantees all data sent will arrive and in the right order, but was designed with congestion control algorithms which reduce the transmission rate by half if any packet loss is detected. There are both software and hardware buffers in the Mark5A systems which can compensate for a reduced transmission rate for short periods of time, but extended slow periods would mean the buffers run empty. The higher the round trip time (RTT), proportional to the physical network distance, the longer it takes TCP to recover back to its previous rate after a packet loss event [2]. UDP, on the other hand, does not guarantee delivery of data, and the transmission rate is governed by the user.

3. VLBI_UDP architecture

VLBI_UDP was originally written as a monolothic application by Richard Hughes-Jones for iGrid 2002 as a simulation of the loads e-VLBI places on the networks. It has since undergone several revisions with extra features being added, and these are detailed below. The current architecture is represented graphically in *Figure 1*. There are 3 components to VLBI_UDP, the sending application, receiving application, and the control application. The send & receive components are run as console applications with no user input. The control is also a console application which drives the send & receive components, and is accessed via a variety of methods. It can either take user input from the console, commands via a webpage through a miniature http interface, or read commands from a file with no user interaction. A single instance of the control application controls multiple send/receive pairs.

3.1 Conversion to pthreads

As a monolithic application, VLBI_UDP was consuming almost all available CPU cycles due to constant polling, checking if there is data to be moved around. Clearly this is not an optimal situation, and so the send and receive programs were both split into 3 threads: control, data input and data output. Splitting the application into threads allows each thread to act with a reasonable amount of independence from the other threads, whilst still allowing communication between them.

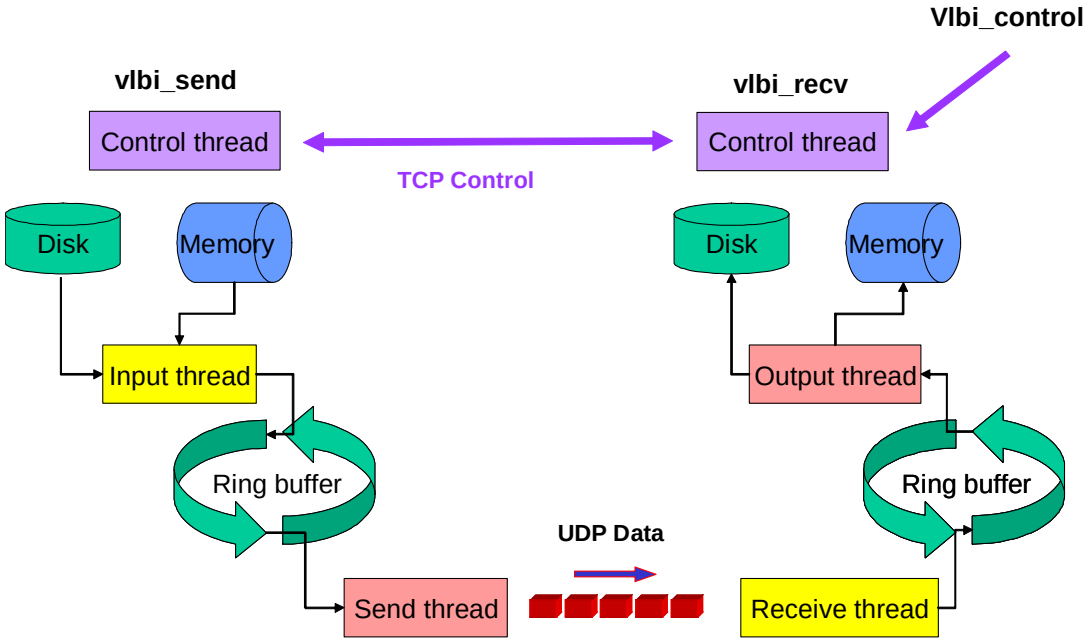


Figure 1: Flow diagram representing VLBI_UDP architecture

3.2 Ringbuffer

A ringbuffer was present in the original iGrid2002 application, but was incomplete so far as it didn't handle out of order packets correctly – not a problem for the demo but needed correcting for use with real data. As each UDP packet is received, it is written directly to the next usable location in the ring buffer. Each packet has a sequence number which allows missing and out of order packets to be caught. If there were one or more packets missing immediately previous to the received packet, then it would be in an incorrect position. A function RingMove() is called, which moves the last packet forward the required number of positions within the ring buffer such that it is then correctly placed. The next available location is now set to after the new location of the last packet.

Should the 'missing' packet(s) subsequently arrive, out of order, then they are first written to the next available location as before. The sequence number is checked, and RingMove() is called with a negative offset to place the packet back where it should have been. In this case, the next available location doesn't change and so the next packet will be written to where the last packet originally arrived. This process is illustrated in *Figure 2*.

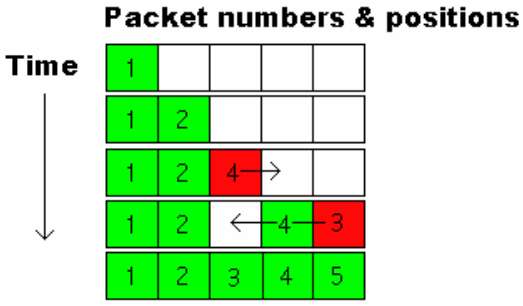


Figure 2: Simulation of packets being placed into ring buffer

3.3 File access

To allow for testing with real data, and as a precursor to interfacing with actual VLBI hardware, file access was implemented. Linux large file support is used, a necessity when dealing with VLBI data sets which are almost exclusively >2GB.

3.4 Packet dropping [3]

A packet dropping function has been added, which, when combined with the file access mode, facilitates the creation of data sets with missing packets under controlled conditions. This function is implemented only in the sender module. The send thread receives a pointer to a packet of data from the ring buffer, passes this to the dropping function as a parameter, along with a choice of dropping algorithm. The return value is a pointer, which will be the same as that passed to the function if the packet was not dropped, else will be a pointer to a randomly initialised portion of memory. Currently there are two algorithms available. The first drops single packets at a steady rate with no randomisation, the second can drop a bunch of between 1 and 10 consecutive packets, the value chosen randomly. To maintain a fractional loss rate f in the 2nd case, after a bunch of n packets are dropped, the subsequent $n(1/f - 1)$ packets are not dropped.

4. Results from tests with VLBI_UDP

VLBI_UDP has been used both as a demonstration tool at events such as iGrid2002, and more recently at the Geant2 network launch, as well as a tool to probe network conditions over extended periods of time. *Figure 3* demonstrates a 24 hour flow, transmitting data from a PC based in Jodrell Bank over a dedicated gigabit fibre connection into a PC based in Manchester University. Each point represents the average received bandwidth in a 30 second period, and it can be seen that rate stability is mostly maintained to 1 part in 1000

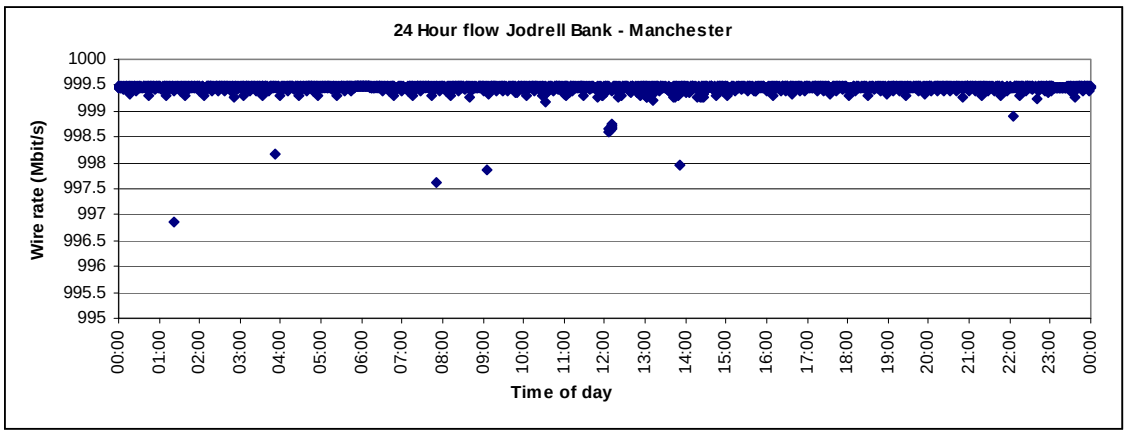


Figure 3: 24 Hour flow from Jodrell Bank Observatory to Manchester University

Figure 4 shows 3 simultaneous transfers into JIVE, one from Manchester travelling over a UKLight dedicated gigabit lightpath, another from Manchester but crossing the conventional

packet switched Internet, and the third from Bologna again over the conventional packet switched Internet. The lightpath performed as expected, with the transmit rate purposely capped at 800 Mbit/s and showing almost no packet loss. The second flow was capped at 600 Mbit/s, since this was travelling via the Manchester University campus access link and so would have swamped regular campus Internet traffic. Packet loss is present here due to contention, most likely over the campus access links, and can be seen to decrease through the test period, representing a decrease in campus traffic. The third plot was limited at 400 Mbit/s as the sending PC was underpowered and this was the maximum rate it could transmit at.

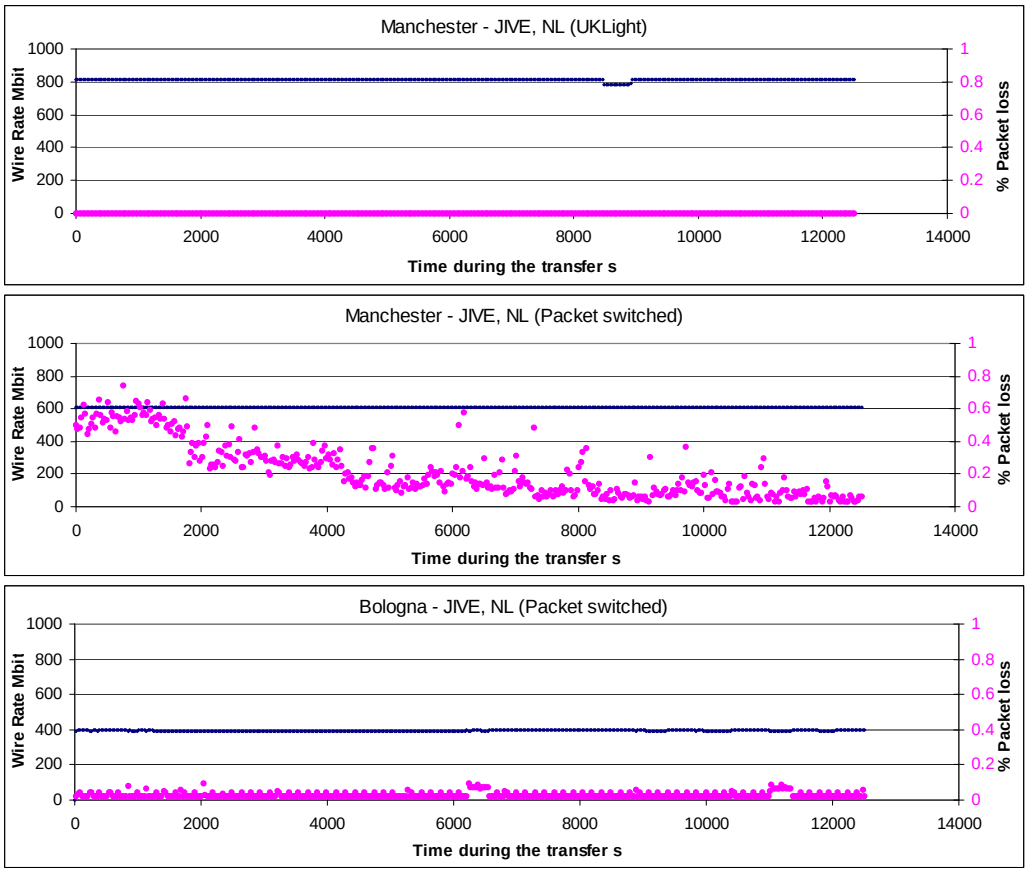


Figure 4: Multiple streams into JIVE

5. Conclusion

As transport protocols for VLBI, both TCP and UDP have their advantages and disadvantages. Currently TCP is a suitable transport protocol, but with the demand for higher data rates and longer baselines, it may be that TCP is unable to keep up and so this paper shows that UDP can provide a suitable alternative.

References

[1] M. Strong et al., *Investigating the e-VLBI Mark 5 end systems in order to optimise data transfer rates as part of the ESLEA Project.* In proceedings of Lighting the Blue Touchpaper for UKe-Science – Closing Conference of ESLEA Project, POS(ESLEA)024, 2007.

[2] R. Hughes-Jones et al., *Bringing High-Performance Networking to HEP users* in proceedings of CHEP04 Interlaken, 27 Sep - 1Oct 2004

[3] S. Casey et al., *Investigating the effects of missing data upon VLBI correlation using the VLBI_UDP application.* In proceedings of Lighting the Blue Touchpaper for UKe-Science – Closing Conference of ESLEA Project, POS(ESLEA)025, 2007.