The DataTAG Transatlantic Testbed

O. Martin[1*], J.P. Martin-Flatin[1], E. Martelli[1], P. Moroni[1],
H. Newman[2], S. Ravot[2] and D. Nae[2]

[1] CERN, Geneva, Switzerland
[2] California Institute of Technology, Pasadena, California, USA

Abstract

Wide area network testbeds allow researchers and engineers to test out new equipment, protocols and services in real-life situations, without jeopardizing the stability and reliability of production networks. The DataTAG testbed, deployed in 2002 between CERN, Geneva, Switzerland and StarLight, Chicago, Illinois, USA, is probably the largest testbed built to date. Jointly managed by CERN and Caltech, it is funded by the European Commission, the U.S. Department of Energy and the U.S. National Science Foundation. The main objectives of this testbed are to improve the Grid community's understanding of the networking issues posed by data-intensive Grid applications over transoceanic gigabit networks, design and develop new Grid middleware services, and improve the interoperability of European and U.S. Grid applications in High-Energy and Nuclear Physics. In this paper, we give an overview of this testbed, describe its various topologies over time, and summarize the main lessons learned after two years of operation.

Key words: Gigabit wide area networks, Grid networking, testbed.

1 Introduction

Testbeds in general, and Wide Area Network (WAN) testbeds in particular, are not a novel idea. The key motivation for using them is the complete freedom to deploy and test new hardware (network devices or end-systems), middleware and software to find the thresholds where they break, and the possibility to make frequent changes at short notice. Doing so is generally impossible, or at best very impractical, in production environments. Testbeds are there to be broken by pushing technologies and new ideas to their limits, whereas production infrastructures must provide stable and dependable services.

The experience gained with testbeds is very precious to shape and design future production infrastructures. At an early stage of a project's lifecycle, they give insight as to

[*] Corresponding author: Olivier Martin, IT Dept., CERN, 1211 Geneva 23, Switzerland
E-mail addresses: Olivier.Martin@cern.ch, jp.martin-flatin@ieee.org, Edoardo.Martelli@cern.ch, Paolo.Moroni@cern.ch, newman@hep.caltech.edu, ravot@caltech.edu, Dan.Nae@cern.ch

where the real technological and architectural problems lie. Correcting mistakes once a large infrastructure has been deployed and put into production can be immensely expensive; testbeds are a proven way to avoid making such mistakes in the first place. Network testbeds therefore complement test laboratories: once new equipment, protocols and services have been simulated, emulated, or tested in the laboratory, they can be tested out and thoroughly debugged in a testbed until they are ready for production. Although this cannot guarantee the safe deployment of new services across operational networks (because each network has its idiosyncrasies), it does definitely improve the confidence that engineers can place in them prior to large-scale deployment in a production environment.

The Data TransAtlantic Grid (DataTAG) testbed was jointly financed by European and U.S. government agencies. The European contribution was funded by the FP5/IST Program of the European Commission (DataTAG project, grant IST-2001-32459) [1]. This project ran from 1 January 2002 to 31 March 2004 and brought together five leading research organizations in the Grid networking community: the European Organization for Nuclear Research (CERN) in Switzerland, the National Institute for Nuclear Physics (INFN) in Italy, the National Institute for Research in Computer Science and Control (INRIA) in France, the Particle Physics and Astronomy Research Council (PPARC) in UK, and University of Amsterdam (UvA) in The Netherlands.

The U.S. contribution was funded by the U.S. Department of Energy (DoE)—grant DE-FG03-92-ER40701, won by the California Institute of Technology (Caltech)—and the U.S. National Science Foundation—grant ANI 9730202, won by the Electronic Visualization Lab (EVL) at University of Illinois in Chicago (UIC). The testbed was operated jointly by CERN and Caltech, with staff from both organizations located at CERN.

The main goal of the DataTAG testbed was to increase the Grid community's knowledge and understanding of how to leverage long-distance gigabit networks in data-intensive Grid environments. On the engineering side, this included deployment aspects, day-to-day operation, reservations of specific equipment by limited groups of users, and frequent upgrades or configuration changes. On the research side, networking people worked on fast transport protocols (variants of or alternatives to TCP, the Transmission Control Protocol), Quality of Service (QoS), advance reservation and network monitoring; software people designed and developed new Grid middleware services, and considerably improved the interoperability of Grid applications developed on both sides of the Atlantic in the field of High-Energy and Nuclear Physics (HENP). These applications are primarily destined for analyzing the PetaBytes of data that will be generated by the Large Hadron Collider (LHC) under construction at CERN, trying to discover the Higgs boson [2].

In order to meet this objective, a flexible multi-vendor testbed was made available to all project members as well as a number of partner organizations. This testbed offered various layer-1, layer-2 and layer-3 network topologies. The results gathered during the

DataTAG project were immense, as demonstrated by this special issue, and the need to fund such a testbed was demonstrated.

The rest of this paper is organized as follows. In Section 2, we define the terminology used in this paper and make a case for WAN testbeds. In Section 3, we review the main characteristics of the DataTAG testbed. In Section 4, we describe the different topologies of the testbed during its lifetime. In Section 5, we outline how the DataTAG testbed is interconnected with other research and education networks. In Section 6, we present the Internet2 land-speed records that were beaten several times using this testbed. In Section 7, we sum up the main lessons learned during the DataTAG project. In Sections 8 and 9, we give future prospects and study related work. Finally, we present concluding remarks and directions for future work in Section 10.

2 Terminology and Requirements

Before delving into the details of the DataTAG testbed, let us define a terminology and identify some requirements for WAN testbeds.

2.1 Terminology

Throughout this paper, testbed *users* are researchers or engineers using the testbed as a facility, whereas testbed *administrators* are people in charge of operating the testbed.

When we talk about layers 1, 2 and 3, we refer to the 7-layer OSI (Open Systems Interconnection) reference model [3]. In a *layer-1 testbed*, devices are interconnected by multiplexers; in a *layer-2 testbed*, by switches or bridges; in a *layer-3 testbed*, by routers. The distinction between different flavors of testbeds is not always straightforward, however, as a layer-1 testbed typically also supports Gigabit Ethernet attachments, in much the same way as a layer-2 switch or a layer-3 router. So, the difference between a layer-1, a layer-2 and a layer-3 testbed primarily has to do with (i) the way in which Ethernet frames are forwarded and (ii) the resulting delay, jitter and packet loss rate. SONET/SDH (Synchronous Optical Network / Synchronous Digital Hierarchy) multiplexers are characterized by a low latency: Ethernet frames are transmitted on-the-fly. In a layer-2 switch, conversely, Ethernet frames are stored and forwarded, so latency is higher. In a layer-3 router, the processing that needs to be done on IP headers increases latency even more.

A testbed can have a specific focus, e.g. Quality of Service (QoS) or Bandwidth on Demand (BoD). Alternatively, it can have a very general purpose such as *"enabling work on advanced networking"*, which includes IPv4, IPv6, Virtual Private Networks (VPNs), etc.

A testbed can be *native* (e.g., transparent SONET/SDH circuits, also known as *optical waves* or *lambdas*) or *emulated* (e.g., a layer-2 VPN over a layer-3 transport network).

Testbeds can encompass single or multiple technologies (in the latter case, they are called *hybrid*). They can also be *concatenated* in order to extend their reach as well as their capabilities. For example, via its extensions across Abilene, GÉANT and national research and education networks such as GARR in Italy, the DataTAG testbed evolved over time from a native testbed to a concatenated hybrid testbed.

2.2 Requirements

Among all the requests that we have received from users since the DataTAG testbed became operational, five seem to be important requirements for future WAN testbeds. It should be noted that there is no single truth in this respect: a requirement in one user's perspective may simply be a desirable feature in another's.

Requirement 1: A WAN testbed should be *flexible*. It should introduce as few technical restrictions as possible in order to allow users and administrators to test out the maximum number of topologies and scenarios. This is particularly important when people have limited *a priori* knowledge when they start a new project.

Requirement 2: A WAN testbed should be *dynamic*. It should be able to meet the fast changing requirements of users and evolve quickly in order to stay at the forefront of commercially available technologies (sometimes even using equipment still in beta-test, prior to large-scale distribution).

Requirement 3: It should be possible to partition a large testbed into meaningful smaller parts, giving simultaneous and independent access to different users. This is of particular interest in multi-vendor testbeds, where users often do not require access to all the equipment at the same time.

Requirement 4: A testbed should provide exclusive access to independently managed parts of it via an advance reservation application.

Requirement 5: Ideally, a WAN testbed should provide a layer-1 interface (i.e., at the optical layer) in order to allow the connection of any layer-2 (switch) or layer-3 (router) devices. In case a layer-1 testbed is not practical, a layer-2 testbed is still an acceptable compromise as layer-2 switches are usually fairly transparent. However, if possible, layer-3 testbeds should be avoided because they can introduce undesirable behaviors; e.g., the Juniper M160 routers in use across the 10Gbit/s GÉANT backbone have the unpleasant "feature" of re-ordering packets under heavy load (above 1Gbit/s).

3 Main Characteristics of the DataTAG Testbed

In this section, we review the main technical and organizational characteristics of the DataTAG testbed.

3.1 Key Technical Characteristics

The DataTAG testbed has been at the forefront of WAN technologies for two years. During this period, its main technical characteristics have been the following:

- High-speed 2.5Gbit/s transatlantic optical wavelength (lambda) between September 2002 and August 2003, upgraded to 10Gbit/s in September 2003.
- Between March 2003 and August 2003, transparent transport of 1Gbit/s Ethernet over a 2.5Gbit/s optical circuit (Ethernet over SONET/SDH).
- Since September 2003, transparent transport of 10Gbit/s Ethernet frames using a layer-2 emulation solution based on Juniper layer-2 VPN technology, which established DataTAG as the first transoceanic testbed with native 10Gbit/s Ethernet access capabilities.
- Open[1] multi-vendor layer-2 and layer-3 testbed with equipment from Alcatel (1670 and 7770 RCP), Chiaro/Enstara, Cisco (6506, 7606 and 7609), Extreme (Summit 5i), Juniper (M10 and T320) and Procket (8801).
- Given the requirement to provide native 10Gbit/s Ethernet capabilities and the lack of proven commercial layer-1 or layer-2 products allowing the transport of 10Gbit/s Ethernet frames over 10Gbit/s long-distance optical wavelengths, a layer-2 emulation solution based on Juniper T320 routers has been deployed to transport 10Gbit/s Ethernet.
- Thanks to a non-disclosure agreement between Caltech, CERN and Intel, the DataTAG project has been able to pioneer the use of 10Gbit/s Ethernet Network Interface Cards (NICs) since January 2003. Subsequently, agreements between AMD, Caltech, CERN, Microsoft, Newisys and S2io allowed us to have early access to the latest server technologies and to beat Internet2 land-speed records (see Section 6).
- To the best of our knowledge, the DataTAG testbed was the first transoceanic testbed with native 10Gbit/s Ethernet access capabilities.

3.2 Ethernet over SONET/SDH vs. G.709

The selection of Ethernet over SONET/SDH was not obvious and deserves a few explanations.

Next generation optical transport networks with up to 40Gbit/s capabilities are expected to be based on the ITU-T's G.709 recommendation [4], often known as *digital wrapper*. Unlike today's long-distance telecommunication networks, which can only transport SONET/SDH frames, these new WANs should also be able to transport 1Gbit/s Ethernet, 10Gbit/s Ethernet and several other types of frames transparently.

At the outset, the DataTAG stakeholders decided to build a multi-vendor testbed on top of a layer-2 Gigabit Ethernet transport network in order to get maximum flexibility and transparency. Unfortunately, as commercial deployment of G.709-enabled networks had

[1] The word "open" reflects two facts: the testbed had the potential to include equipment from any vendor, and it was available to a wide user community.

not yet happened (and still has not[2]), the only way to provide native transatlantic layer-2 services was to use Ethernet over SONET/SDH multiplexers.

The Generic Framing Procedure (GFP) [5], defined by the Telecommunication Standardization Sector of the International Telecommunication Union (ITU-T), specifies a standard low-latency method to transport a Gigabit Ethernet signal transparently across a SONET/SDH (Synchronous Optical Network/Synchronous Digital Hierarchy) network. We could not find a suitable GFP-capable multiplexer for phase 3 of the DataTAG testbed, so we selected instead the Alcatel 1670, a multiplexer that can encapsulate 1Gbit/s Ethernet frames over SONET/SDH frames using a proprietary pre-GFP encapsulation scheme. This enabled the DataTAG testbed to qualify as a layer-1 testbed between March 2003 and August 2003 (see requirement 5 in Section 2.2). Thanks to these multiplexers, a transparent bi-directional transatlantic 1Gbit/s Ethernet bridge was successfully built between the CERN Internet eXchange Point (CIXP) [6] in Geneva and the StarLight [7] Internet Exchange Point in Chicago. De facto, this provided users with a distributed transatlantic Internet exchange point.

Unfortunately, we could not find a similar solution during phase 4, when the testbed was operating at 10Gbit/s. Instead, we had to resort to layer-2 emulation over a layer-3 network (see Section 4.4).

3.3 Key Organizational Characteristics

In addition to state-of-the-art equipment, the DataTAG testbed has provided users with two features rarely found in large testbeds:

- It was the first time that a testbed of this size, with such a variety of network equipment, high-end CPUs and disk servers[3], was made available to a large community of researchers in such a rigorous and systematic manner.
- Access to the testbed was controlled by a sophisticated reservation application that allowed users to reserve all or part of the testbed in advance. This allowed them to make "clean" measurements when testing new protocols or services, without having to worry about other users generating extraneous traffic that could "pollute" their measurements.

These features were greatly appreciated by DataTAG users and made it easier to conduct experiments and performance evaluations in a scientific way.

---

[2] Because of the crisis of the telecommunication industry, commercial deployment of G.709 networks seems unlikely to happen before 2007. Deployment across the Atlantic is probably even further away, given the excess of bandwidth there.

[3] Dual Intel Xeon processors, 3.06GHz, 2 GigaBytes (GB) of RAM, SuperMicro X5DPE Motherboard (Intel E7501 chipset), HP RX2600, Dual Itanium2 1.5GHz, 4GB of RAM. 10 GbE interfaces: Intel Pro/10GbE-LR.

4 Chronology of the DataTAG Testbed

The DataTAG testbed became operational at the end of August 2002, four months ahead of schedule, just in time for the iGRID 2002 [8] demonstrations in Amsterdam, The Netherlands. In total, four major phases can be identified.

4.1 Phase 1: layers 1 and 2 (Sep. 2002)

During the first phase, which lasted only one month, the 2.5Gbit/s DataTAG circuit was integrated into the Amsterdam (NetherLight) – Chicago (StarLight) – Geneva (CERN) 2.5Gbit/s layer-1 triangle[4]. This triangle used SDH multiplexers (three Cisco ONS 15454s) owned and operated by SURFnet (the National Research and Education Network in The Netherlands). In this layer-2 configuration, 1Gbit/s Ethernet paths could easily be created across the Atlantic and also extended through CANARIE's infrastructure (CANARIE is the National Research and Education Network in Canada).

This topology allowed us to establish a 12,000km lightpath between TRIUMF, Canada's National Laboratory for Particle and Nuclear Physics in Vancouver, and CERN via Chicago and Amsterdam, and to demonstrate transfers of TeraBytes (TB) of real physics data at nearly 1 Gbit/s, a throughput never achieved before then over such long distances.

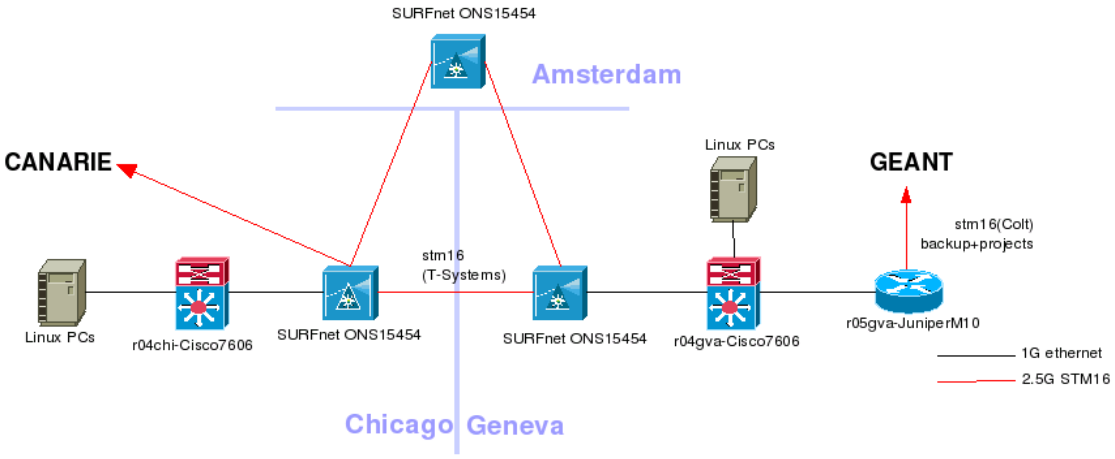

Fig. 1: DataTAG testbed, phase 1

4.2 Phase 2: layers 2 and 3 (Oct. 2002 – Feb. 2003)

During the second phase, a fairly conservative configuration was deployed using Cisco 7600 OSR routers at both ends of the circuit. Nonetheless, this phase allowed researchers to improve several variants of TCP by testing out development versions of their transport protocols at a throughput well above 1Gbit/s. In addition, for the first time at such as

[4] In 2003, this triangle was upgraded to 10Gbit/s.

large scale, a native 2.5Gbit/s circuit could be connected directly to high-end servers[5] each equipped with Intel's brand new 10Gbit/s Ethernet NICs: the Intel Pro/10GbE-LR.

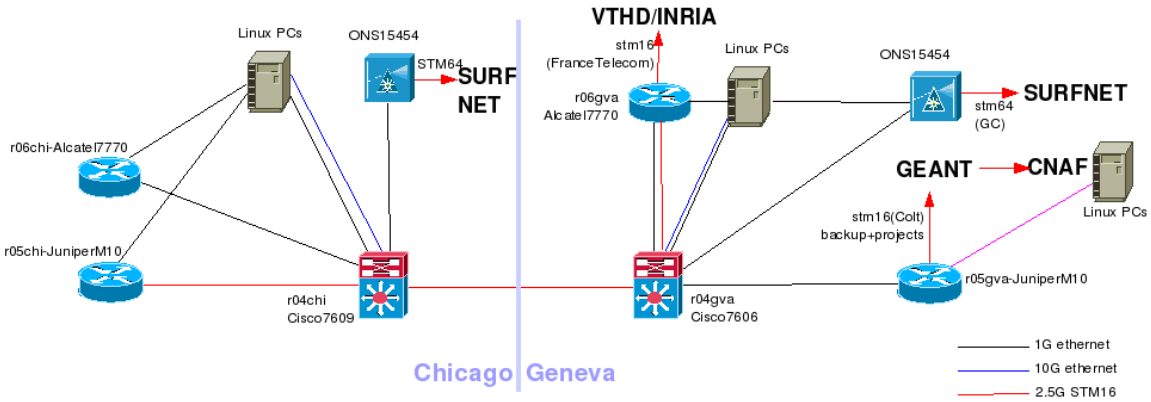

Fig. 2: DataTAG testbed, phase 2 (simplified)

4.3 Phase 3: layers 1,2 and 3 (Mar. 2003 – Aug. 2003)

During the third phase, we deployed layer-1 and layer-2 capabilities similar to those available across the SURFnet circuits during phase 1. We also improved the positioning of the DataTAG testbed as an open multi-vendor testbed, with equipment from Alcatel, Cisco, Extreme and Juniper now fully used.

Having a feature-rich testbed, not restricted to the limitations of a single supplier at any point in time, proved to be a very fruitful and successful concept. In particular, it allowed testbed users to compare the QoS (Quality of Service) and IPv6 capabilities of various vendors, and verify the interoperability between different Multi-Protocol Label Switching (MPLS) [9] implementations.

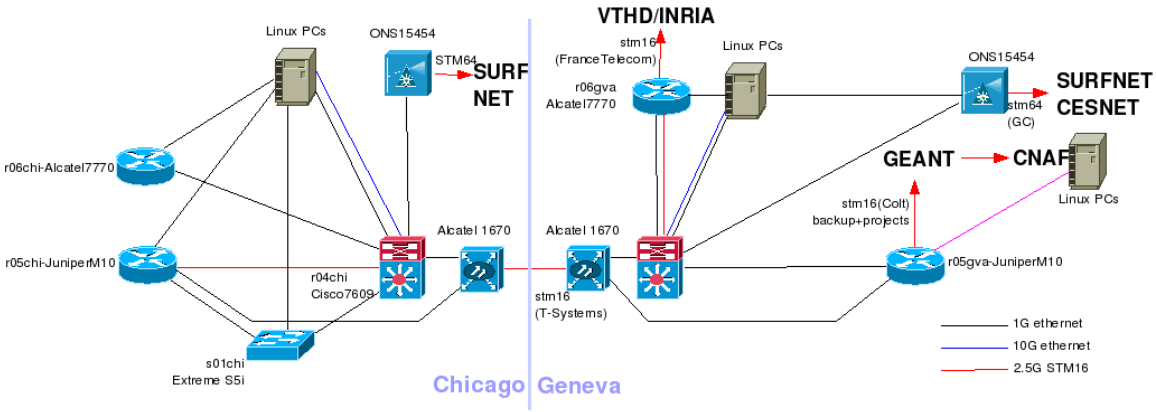

Fig. 3: DataTAG testbed, phase 3 (simplified)

[5] These high end servers have Dual Intel Xeon processors (2.4GHz with 512k level-2 cache), SuperMicro P4DP8-G2 motherboards with Intel E7500 chipsets, and 2GB of RAM.

4.4 Phase 4: emulated layer 2 and layer 3 (Sep. 2003 – Dec. 2004)

The DataTAG circuit was upgraded to 10Gbit/s in September 2003. We did not initially expect this to happen so quickly, but new market conditions created an opportunity that we could not miss. Given the lack of commercial layer-1 products supporting 10Gbit/s Ethernet access at that time, a difficult technological choice had to be made between layer-3 (router-based) or layer-2 (switch-based) solutions.

In the summer 2003, Force10 was the only vendor offering 10Gbit/s Ethernet over WANs and supporting the WAN-PHY option of the IEEE 802.3ae standard [10]. However, interoperability with commercial Wave Division Multiplexing (WDM) systems across long-distance optical circuits had not yet been demonstrated[6]. Based on the successful layer-2 VPN tests performed between CERN and INFN-CNAF across the GÉANT and GARR networks, we decided to choose a solution relying on Juniper T320 routers at both ends of the DataTAG circuit. This provided the required functionality at layers 2 and 3. In this configuration, the multi-vendor DataTAG testbed included equipment from Alcatel, Chiaro, Cisco, Extreme, Juniper and Procket.

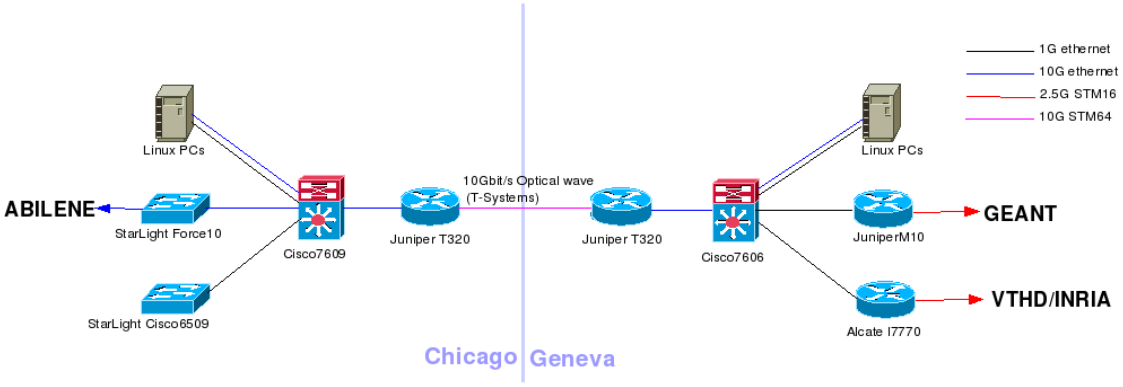

Fig. 4: DataTAG testbed, phase 4 (simplified)

5 Interconnections and Collaborations with Other Networks

Since October 2002, the DataTAG testbed has had *direct* broadband connections to the following testbeds and research networks:

- GÉANT, the European research and education backbone;
- SURFnet, the national research and education network in The Netherlands, and NetherLight, an Internet exchange point located in Amsterdam;
- VTHD, an advanced next generation Internet testbed in France based on IP over WDM;
- Abilene, a national backbone in the USA supporting high-performance connectivity and Internet innovation within the academic research community;
- CANARIE, the national research and education network in Canada;

[6] The proof of concept was subsequently established by a joint team with CERN, SURFnet and University of Amsterdam staff

- ESnet (Energy Sciences Network), a high-speed network in the U.S. used by thousands of DoE-funded scientists and collaborators worldwide.

During several events (e.g., IST 2003, Telecom 2003, SC 2003 and WSIS 2003), the DataTAG testbed was also temporarily connected to other networks and facilities, including TeraGrid (a very large distributed computing infrastructure) and OMNInet (the Advanced Optical Metro Network Initiative in Chicago).

Since its inception, the DataTAG testbed has had *indirect* physical connections to:

- INFN-CNAF, the INFN center of expertise in information technology and telecommunications located in Bologna, Italy;
- GARR, the Italian Research and Education Network;
- the Managed Bandwidth Next Generation (MB-NG) project in UK, which aims at creating a networking and Grid testbed focusing on advanced networking issues and interoperability of administrative domains.
- Caltech via Abilene and CENIC (Californian Research and Education Network).

6 Internet2 Land-Speed Records

In order to stimulate research and experimentation on high-speed high-latency TCP transfers, the Internet2 project [11] created an international competition for the largest bulk data transfers in four categories: single or multiple TCP streams over IPv4 or IPv6. This contest is known as the *Internet2 Land-Speed Record*, or I2-LSR for short [12]. It was created in March 2000, is still ongoing and is open (i.e., it is not limited to Internet2 members). To take into account the difficulty of achieving high throughput when sending data over long distances, the unit of measurement for this contest is the Petabit-meter per second (Pbit.m/s), that is, the product of the end-to-end network distance by the achieved throughput. Because of its relevance for Grid applications, this contest has rapidly become of major importance in the Grid networking community.

On seven occasions, the DataTAG testbed was used by different teams to beat I2-LSRs in different categories. Fig. 5 shows the evolution of the Internet2 land-speed record since its inception.

Record 1: 27 February 2003

During phase 2 of the DataTAG testbed (see Section 4.2), the IPv4 Single and Multiple Streams records were beaten with 23,888Tbit.m/s. The end-to-end network distance was 10,037km (between CERN in Geneva and Level3's Point of Presence in Sunnyvale, California through StarLight in Chicago), and 1.1TB of data were transferred in 3700s. Both end-hosts were running the RedHat 7.3 Linux distribution and Linux kernel 2.4.19. The data was sent using Iperf 1.6.5 with Jumbo frames (9,000 Bytes). The team members were Caltech, CERN, Los Alamos National Laboratory (LANL) and Stanford Linear Accelerator Center (SLAC). The two end-hosts had Dual Xeon CPUs clocked at 2.20GHz and SysKonnect 1Gbit/s Ethernet NICs with patched drivers.

Record 2: 6 May 2003

During phase 3 of the DataTAG testbed (see Section 4.3), the IPv6 Single and Multiple Streams records were beaten with 6,947Tbit.m/s. The end-to-end network distance was 7,067km (between CERN in Geneva and StarLight in Chicago), and 412GB of data were transferred in 3600s. The team members were Caltech and CERN. Both end-hosts were running the RedHat 7.3 Linux distribution and Linux kernel 2.4.20. The data was sent using Iperf 1.7.0. with Jumbo frames. The two end-hosts had Dual Xeon CPUs clocked at 2.20GHz and SysKonnect 1Gbit/s Ethernet NICs.

Record 3: 1 October 2003

During phase 4 of the DataTAG testbed (see Section 4.4), the IPv4 Single and Multiple Streams records were beaten with 38,420Tbit.m/s. The end-to-end network distance was 7,067km (between CERN in Geneva and StarLight in Chicago) and 1.1TB of data were transferred in 1620.5s. The team members were Caltech and CERN. Both end-hosts were running the RedHat 7.3 Linux distribution and Linux kernel 2.4.20. The data was sent using Iperf 1.7.0. with Jumbo frames. The sender end-host was an HP RX2600 workstation with Dual Itanium2 CPUs clocked at 1.5GHz, 4GB of RAM, and an Intel PRO/10GbE LR NIC. The receiver end-host had Dual Xeon CPUs clocked at 3.06GHz, 2GB of RAM, a SuperMicro X5DPE motherboard with an E7501 chipset, and an Intel PRO/10GbE LR NIC.

Record 4: 11 November 2003

During phase 4 of the DataTAG testbed (see Section 4.4), the IPv4 Single and Multiple Streams records were beaten with 61,752Tbit.m/s. The end-to-end network distance was 10,949km (between CERN in Geneva and Los Angeles in California via StarLight, Abilene and CENIC), and 2.3TB of data were transferred in 3600s. The team members were Caltech and CERN. Both end-hosts were running the RedHat 7.3 Linux distribution and the Linux kernels 2.6.0. The data was sent using Iperf 1.7.0. with Jumbo frames. The sender end-host was an HP RX2600 workstation with Dual Itanium2 CPUs clocked at 1.5GHz, 4GB of RAM, and an Intel PRO/10GbE LR NIC. The receiver end-host had Dual Xeon CPUs clocked at 3.06GHz, 2GB of RAM, a SuperMicro X5DPE motherboard with an E7501 chipset, and an Intel PRO/10GbE LR NIC.

Record 5: 18 November 2003

During phase 4 of the DataTAG testbed (see Section 4.4), the IPv6 Single and Multiple Streams records were beaten with 46,156Tbit.m/s. The end-to-end network distance was 11,539km (between CERN in Geneva and a booth at SC 2003 in Phoenix via StarLight and Abilene), and 560GB of data were transferred in 1200s. The team members were Caltech and CERN. Both end-hosts were running the RedHat 7.3 Linux distribution and the Linux kernel 2.6.0. The data was sent using Iperf 1.7.0. with Jumbo frames. The sender end-host was an HP RX2600 workstation with Dual Itanium2 CPUs clocked at 1.5GHz, 4GB of RAM, and an Intel PRO/10GbE LR NIC. The receiver end-host had

Dual Xeon CPUs clocked at 3.06GHz, 2GB of RAM, a SuperMicro X5DPE motherboard with an E7501 chipset, and an Intel PRO/10GbE LR NIC.

Record 6: 22 February 2004

During phase 4 of the DataTAG testbed (see Section 4.4), the IPv4 Single and Multiple Streams records were beaten with 68,431Tbit.m/s. The end-to-end network distance was 10,949km (between CERN in Geneva and Caltech in Los Angeles via StarLight, Abilene and CENIC), and 499GB of data were transferred in 600s. The team members were Caltech and CERN. Both end-hosts were running Windows Server 2003 64-Bit Edition. The data was sent using the NTttcp test tool (part of Windows 2000 DDK) with Jumbo frames. Both end-hosts were 4U Intel Quad Itanium2 SR870BN4 Servers with Intel E8870 chipsets and PCI buses clocked at 133MHz in 64-bit mode. On the sender, the NIC was a sender-s2io 10GE; on the receiver, the NIC was a receiver-Intel 10GE LR.

Record 7: 6 May 2004

During phase 4 of the DataTAG testbed (see Section 4.4), the IPv4 Single and Multiple Streams records were beaten with 77,699Tbit.m/s. The end-to-end network distance was 10,949km (between CERN in Geneva and Caltech in Los Angeles via StarLight, Abilene and CENIC), and 860GB of data were transferred in 970s. The team members were Caltech and CERN. Both end-hosts were running Windows Server 2003 64-Bit Extended Systems Edition. The data was sent using the NTttcp test tool (part of Windows 2000 DDK) with Jumbo frames. The sender was a Newisys 4300 Quad AMD Opteron Enterprise Server with AMD-8131 and an S2io 10GE NIC. The receiver was an Intel Quad Itanium2 SR870BN4 Server with the Intel E8870 chipset and an S2io 10GE NIC.

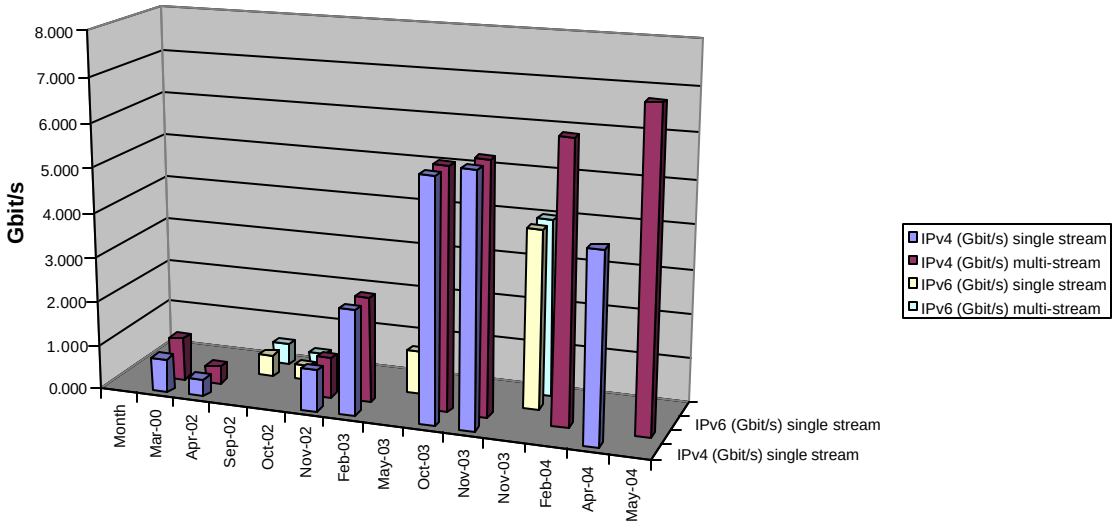

Fig. 5: Evolution of Internet2 land-speed records

Chasing such records may sound like a game, but the underlying goal is of great importance for the future of data-intensive Grids. In particular, for CERN and all the physicists in the world working on LHC experiments, the LHC Computing Grid (LCG) will depend critically on sustainable multi-gigabit per second throughputs between geographically dispersed sites.

Running after the I2-LSRs records also showed that against all expectations, and for the first time in the WAN history, performance is now limited by the end-systems and not by the network. In October 2003, Newman said [13]: *"This is a major milestone towards our goal of providing on-demand access to High Energy Physics data from around the world, using servers affordable to physicists from all regions. We have now reached the point where servers side by side have the same TCP performance as servers separated by 10,000 km. We also localized the current bottleneck to the I/O capability of the end-systems, and we expect that systems matching the full speed of a 10 Gbit/s link will be common-place in the relatively near future."*

7 Lessons Learned

The main lessons learned since the DataTAG testbed became operational in August 2002 are related to long-term collaborations, remote operations, advance reservations and cooperation between users.

Long-Term Collaborations

The success of the DataTAG testbed was due to a large extent to the quality of the pre-existing collaborative framework between CERN and Caltech, the long-term trust established between these two research partners, and CERN's experience in interacting with DoE and NSF. This was greatly beneficial to the DataTAG project as a whole, for it provided not only additional manpower but also funding for testbed equipment, which could not have been paid by the DataTAG project otherwise.

Remote Operations

We have learned how to operate a WAN testbed that frequently changes but needs to be as stable as possible. CERN and Caltech had a long experience in operating production networks, but we realized that operating fast-changing testbeds is quite different from operating a production network, which changes rarely. For instance, we learned that operating equipment located 7,067km away requires remote power cycling facilities, to recover equipment when it hangs. Initially, we thought that this would just be a nice feature and did not fully appreciate that it was mandatory. We have also learned how to store and retrieve remotely the information that Linux servers send to their console ports during the boot sequence or when they crash. This, too, was underestimated and proved to be necessary to debug several problems in the end-systems.

Advance Reservation Application

In Section 2.2, we mentioned that users wanted to be able to have exclusive access to independently managed parts of the testbed (for a maximum of 8 hours in a row, and up to several months in advance for major events) via an advance reservation application. The application that was developed at CERN was Web-based and written in Java, which allowed users to access it remotely and interactively via a nice Graphical User Interface (GUI). After a few months of regular use, we realized that GUI-based reservations were very useful in a testbed environment, but production and pre-production environments in fact require automated reservations. To be able to deal with both, all the interactions between our reservation application and the outside world were done using the eXtensible Markup Language (XML). However, elaborate locking and duplication mechanisms still need to be put in place, to make this reservation application more robust when it is used simultaneously (by different users) in interactive and unattended modes.

Another lesson learned with our reservation application was that it needs constant updates because of the frequent changes in the setup of the testbed. We expected to change it from time to time, but the overhead caused by these changes was grossly underestimated. More work needs to be done in the area of self-discovery and self-adaptation.

Cooperation between Users

The policy for using the DataTAG testbed was that all users should strive to share this facility in a cooperative manner. Dealing with last-minute changes and overlaps proved reasonable when people knew one another. But problems appeared when the testbed was opened to partner institutes: because these users never saw the physical machines or the people working on them, some were tempted to behave unreasonably. This attitude caused frictions with users and administrators. We have thus learned that the high availability of a testbed can mislead some users to behave as if they were dealing with a production infrastructure dedicated to them. When such situations arise, users need to be educated and reminded of a few principles governing the use of a shared test infrastructure.

Others

Many other lessons were learned but we cannot detail all of them here. They include the difficulty to share resources with users working in different time zones, switching to summer/winter time on different days; the need to schedule several months in advance high-visibility demonstrations at major events; the necessity to define policies and enforce them strictly; the difficulty to connect the DataTAG testbed to the LANs of high-visibility events, where administrators do not always fully appreciate the constraints posed by the use of state-of-the-art equipment; etc. A number of these aspects are detailed in the deliverables of Work Package 1 of the DataTAG project [14].

8 Future Prospects

The possibility to have access to 40Gbit/s WANs across the Atlantic appears to be rather slim for the next few years. Current estimates target 2009 for the wide deployment of this new technology in WANs. As a result, research and education backbones will probably have to remain several years at 10Gbit/s.

The suitability of Force10's 10Gbit/s Ethernet WAN-PHY solution has been successfully established by CERN, SURFnet and University of Amsterdam [15], first across the CERN-NetherLight circuit, and later across the NetherLight-StarLight circuit. As more manufacturers (e.g., Cisco and Foundry) have now decided to enter the 10Gbit/s WAN-PHY market, we expect that 10Gbit/s Ethernet WANs will often provide a cheaper alternative to 10Gbit/s SONET/SDH-based networks in the near future, and will therefore become popular in the academic and research community. We expect major cost savings to be achieved using layer-2 rather than layer-3 equipment in environments that do not require the complexity of layer 3. This may be the case, for instance, with *lambda Grids*, when statically or dynamically provisioned high-speed interconnections are provided at the site, cluster, server, or even flow level to fulfill the requirements of data-intensive Grids.

Regarding the prospects for pushing the performance of single stream TCP over IPv4 or IPv6 beyond 7Gbit/s, it has been proved that the limitations are currently due to the end-systems currently available on the market [16]. With the expected advent of PCI Express chips (i.e., fast chips for the Peripheral Component Interconnect bus), better 10Gbit/s Ethernet network adapters, faster CPUs and improved motherboards in the near future, there is little doubt that it will soon be feasible to push the performance of single stream TCP streams closer to 10Gbit/s, and thus the I2-LSR above 100,000Pbit.m/s.

Finally, although the possibility to partition the 2.5Gbit/s circuit into two independent 1Gbit/s Ethernet circuits proved very useful during phase 3 (e.g., to visualize the dynamics of different TCP stacks under artificially generated packet-loss conditions), we were unable, with the routers and switches at our disposal, to configure these two 1Gbit/s Ethernet circuits as a single 2Gbit/s circuit using *inverse Time Division Multiplexing (TDM)* or *Gig-Etherchannel* techniques. As a result, single TCP flows were always mapped to the same 1Gbit/s Ethernet circuit, in effect limiting the performance to 1Gbit/s per flow. Given these intrinsic limitations, and apart from the special cases of dynamic Virtual LANs (VLANs) and Gigabit Ethernet on demand, it is rather unclear what the use cases for Gigabit Ethernet-based Time Division Multiplexers are in the Grid community. The MPLS technology seems to offer a more flexible alternative.

9 Related Work

For the sake of conserving space, we focus on Europe in this section. Although the testbeds and networks were/are different in other geographical areas, similar conclusions can be drawn as to the availability of large high-speed WAN testbeds for research purposes [17].

Because of the high telecommunication costs usually associated with the deployment of WANs, especially during the monopoly era (which lasted until 1998 in Europe), most network testbeds were deployed only in Local Area Networks (LANs) or Metropolitan Area Networks (MANs) for many years. There were few exceptions (e.g., in the late 1980s and early 1990s, the network infrastructure of the Berkom project spanned an entire region in Germany).

In Europe, the first international network testbed was deployed in 1994 in the framework of the BETEL (Broadband Exchange for Trans-European Links) project [18]. During this one-year trial, four 34Mbit/s ATM links were deployed and operated between France and Switzerland. This enabled, for instance, the distribution of services (SHIFT project) between CERN in Geneva, Switzerland and IN2P3 in Lyon, France. This infrastructure was also used for teleteaching between the Swiss Federal Institute of Technology in Lausanne (EPFL), Switzerland and Institut Eurécom in Sophia Antipolis, France. This testbed demonstrated the possibility of having broadband communications in Europe.

Next, in 1995-1996, the JAMES network [19] was provided by a consortium of telecommunication operators. It consisted of a 34Mbit/s ATM infrastructure similar to the one used in BETEL, but on a much wider scale (many European countries were connected). This testbed was notably used to support the DANTE/TERENA TF-TANT experiments [20].

In 1997, the JAMES network was incorporated into the TEN-34 (Trans-European Network) [21] project and became the European research and education backbone, interconnecting most countries of the European Union at 34Mbit/s. This ATM infrastructure was later upgraded to 155Mbit/s in the TEN-155 project [22], which ran from 1999 to 2001. TEN-34 and TEN-155 both provided production networks with strict availability and stability constraints; none of them offered WAN testbed facilities.

In 2002, TEN-155/JAMES was superseded by GÉANT [23], the current European research and education backbone, which operates at 10Gbit/s. This, too, is a production network that cannot be used as a testbed: we cannot blast 10TB of data over GÉANT to see if a new transport protocol prevents other applications from working; we cannot test creative QoS settings in the production routers to measure their effect on traffic; we cannot change the network topology several times per week to try out different scenarios; etc. It should be noted that these limitations are, to a certain extent, compensated by the fact that a wealth of new services have been made available to users during the lifetime of this project: Premium IP, Less than Best Effort (LBE, known as *Scavenger* in Abilene), layer-2 VPNs, layer-3 VPNs, etc. GÉANT is, and has always been, a production network offering state-of-the-art facilities.

The need for large-scale testbeds that can be broken and dedicated temporarily to only a few advanced users (thereby allowing controlled usage and "unpolluted" performance measurements) has now been recognized by the European Commission, largely thanks to the success encountered by the DataTAG testbed. In the framework of the GÉANT2

project, the European research and education backbone that will replace GÉANT from 2005, a Europe-wide testbed is expected to be made available to networking researchers.

10 Conclusion

The DataTAG testbed has been at the forefront of high-speed WAN technologies for over two years. To the best of our knowledge, it was the first transoceanic testbed with native 10Gbit/s Ethernet access capabilities. The papers published in this special issue and elsewhere [24] [25] show how precious this testbed has been to researchers testing out new hardware and software. Beating Internet2 land-speed records on many occasions proved that very high throughputs are sustainable over gigabit WANs. This is very good news for the Grid community, particularly for the data-intensive Grid applications under development in High-Energy and Nuclear Physics.

More research is needed in this field, however. The Grid networking research community still needs to improve its understanding of how to structure and configure the gigabit WANs that underpin Grids (e.g., we are still far from self-configuration and self-adaptation). We also need to increase the efficiency of higher-layer protocols (e.g., QoS-based routing or transport protocols) to allow applications to exploit most of the theoretically available bandwidth without hampering non-Grid traffic. Better middleware services have to be designed and developed so as to hide the network complexity (e.g., on-demand optical path setup in lambda Grids) from Grid applications. All of this work would be facilitated if the DataTAG testbed, or similar high-speed high-latency transoceanic testbeds, could be funded in the future.

References

[1] DataTAG Project, http://www.datatag.org/
[2] H. Newman, M. Ellisman and J. Orcutt, "Data-Intensive E-Science Frontier Research", *Communications of the ACM*, Vol. 46, No. 11, pp. 68-75, November 2003.
[3] H.G. Hegering, S. Abeck and B. Neumair, *Integrated Management of Networked Systems: Concepts, Architecture, and Their Application*, Morgan Kaufmann, 1998.
[4] ITU-T, *Interfaces for the Optical Transport Network*, Recommendation G.709, March 2003.
[5] ITU-T, *Generic Framing Procedure*, Recommendation G.7041/Y.1303, December 2003.
[6] CERN Internet eXchange Point (CIXP), http://www.cixp.ch/
[7] StarLight, http://www.startap.net/starlight/
[8] iGrid 2002 Conference, http://www.igrid2002.org/
[9] E. Rosen, A. Viswanathan, R. Callon, *Multiprotocol Label Switching Architecture*, RFC 3031, IETF, January 2001.
[10] IEEE, *Media Access Control (MAC) Parameters, Physical Layers and Management for 10Gb/s Operation*, IEEE Standard 802.3ae-2002 Amendment, August 2002.
[11] Internet2 Project, http://www.internet2.edu/
[12] Internet2 Land-Speed Records, http://lsr.internet2.edu/
[13] CERN Press Release, *CERN and Caltech Join Forces to Smash Internet Speed Record*, October 2003, http://info.web.cern.ch/Press/PressReleases/Releases2003/PR15.03ESpeedrecord.html
[14] DataTAG Project, Work Package 1, http://www.datatag.org/wp1/

[15] C. Meirosu, P. Golonka, A. Hirstius *et al.*, "Native 10 Gigabit Ethernet Experiments over Long Distances", *Future Generation Computer Systems*, Dec 2004.
[16] R. Hughes-Jones, P. Clarke and S. Dallison, "Performance of 1 and 10 Gigabit Ethernet Cards with Server Quality Motherboards", *Future Generation Computer Systems*, Dec 2004.
[17] M. Brown (Ed.), Special Issue on "Blueprint for the Future of High-Performance Networking", *Communications of the ACM*, Vol. 46, No. 11, pp. 30–77, November 2003.
[18] BETEL Project, http://archive.dante.net/ten-34/tf-ten/papers/martin.ps
[19] The Commission's Role in the Implementation of the Information Society, http://www.cordis.lu/infowin/acts/analysys/nathosts/euhost/ch2.htm1
[20] TF-TANT: A Joint Activity Between TERENA and DANTE, http://archive.dante.net/tf-tant/
[21] TEN-34 Project, http://archive.dante.net/ten-34.html
[22] TEN-155 Project, http://archive.dante.net/ten-155.html
[23] GÉANT Website, http://www.dante.net/server/show/nav.007
[24] M. Rio, A. di Donato, F. Saka, N. Pezzi, R. Smith, S. Bhatti and P. Clarke, "Quality of Service Networking for High Performance Grid Applications", *Journal of Grid Computing*, Vol. 1, No. 4, pp. 329-343, 2003.
[25] DataTAG Publications, http://www.datatag.org/papers/

Biographies

Olivier Martin is the Project Leader of the DataTAG project. He received an M.Sc. degree in EE from École Supérieure d'Électricité (Supélec), Paris, France in 1962. He joined CERN in 1971, held various positions in the Software Group of the Data Handling Division, and then moved to the Communications Group of the Computing & Networks Division in 1984, where he has been Head of the External Networking Section since 1989. Prior to the DataTAG project, he was involved in several European projects (including BETEL, BETEUS and STEN) in the framework of the RACE, ACTS and TEN programs. His research interests include high-speed networking, transport protocols and Grids.

Jean-Philippe Martin-Flatin is the Technical Manager of the DataTAG Project. Based at CERN, he coordinates the research activities of 50 people in Grid networking, middleware and applications. Prior to that, he worked at AT&T Labs Research in New Jersey and ECMWF in UK. His research interests include distributed systems management, UML modeling, Web Services, self-organized systems and software architecture. He holds a Ph.D. degree in CS from the Swiss Federal Institute of Technology in Lausanne (EPFL). He is a co-chair of the GGF Data Transport Research Group and a member of the IRTF Network Management Research Group. He was a co-chair of GNEW 2004 and PFLDnet 2003. He serves on the Editorial Boards of eTNSM, JNSM and JWSR.

Edoardo Martelli currently works at CERN as an Internet expert and is responsible for the IPv6 deployment. He previously worked in Italy for Cineca (www.cineca.it) and Nextra (a Norwegian ISP). He joined CERN for the DataTAG project. He received an M.Sc. degree in Computer Science from University of Bologna, Italy in 1994.

Paolo Moroni received an M.Sc. degree in Mathematics from University of Pisa, Italy in 1982. After working three years as an MVS systems programmer in a financial institute, he moved in 1987 to the Scuola Normale Superiore of Pisa, where he was in charge of the local IBM VM service. In 1988, he joined CERN, where he initially worked as a systems engineer on IBM mainframes, then as a security engineer. Since 1998, he has been working on the design and operation of CERN's external networks (IP engineering, DNS, firewall, etc.). He manages Work Package 1 (infrastructure management) of the DataTAG project.

Harvey Newman (Sc. D, MIT 1974) is a Caltech faculty member since 1982. He co-led the MARK J Collaboration that discovered the gluon, the carrier of the strong force, at the DESY laboratory in Hamburg in 1979. He has had a leading role in the development of international networks and collaborative systems serving the High-Energy and Nuclear Physics communities since 1982, and served on the NSFNet Technical Advisory Group in 1986. He originated the Data Grid Hierarchy concept adopted by the high energy physics collaborations of the Large Hadron Collider based at CERN in Geneva. He is the PI of the LHCNet project and the DOE Particle Physics Data Grid Project (PPDG), and a Co-PI of the NSF International Virtual Data Grid Laboratory (iVDGL). He is a member of the Internet2 Applications Strategy Council, and chairs the Standing Committee on Inter-Regional Connectivity of the International Committee on Future Accelerators). He has led the US part of the CMS Collaboration (450 physicists at 40 US Institutions) as Collaboration Board Chair since 1998. He leads the Caltech team that has won the Internet2 Land Speed records several times in 2003 and 2004.

Sylvain Ravot is a Senior Network Engineer with Caltech, Division of Physics, Mathematics and Astronomy. Currently based at CERN in Geneva, he is one of the engineers responsible for the operation of the CERN/US-HENP transatlantic network. He is also a member of the Internet2 Hybrid Optical and Packet Infrastructure (HOPI) design team and is active in the Internet2 High-Energy and Nuclear Physics Working Group. He holds an M.Sc. degree in Communication Systems from the Swiss Federal Institute of Technology in Lausanne.

Dan Nae is a Network Engineer with Caltech, Division of Physics, Mathematics and Astronomy. Currently based at CERN in Geneva, he is one of the engineers responsible for the operation of the CERN/US-HENP transatlantic network. He is also involved in the UltraLight project. He received an M.Sc. degree in CS from Politehnica University of Bucharest, Romania. Prior to joining Caltech, he worked for Politehnica University of Bucharest, Cisco CATC and the Romanian Education Network. His research interests include network protocol performance, gigabit end-system performance and high-speed data transfers.