A new look at Ethernet: Experiences from 10 Gigabit Ethernet End-to-End network between Amsterdam and Geneva

Antony Antony[a] Johan Blom[b] Freek Dijkstra[b] Cees de Laat[b]

[a]NIKHEF,Amsterdam, The Netherlands, antony@nikhef.nl
[b]University of Amsterdam, Amsterdam, The Netherlands

October 23, 2003

ABSTRACT

We had a unique opportunity to build and evaluate the first Trans-European network using 10 Gbps Ethernet WAN PHY technology. Ethernet frames were transported over the SONET and DWDM infrastructure. The path was between NIKHEF, Amsterdam and CERN, Geneva. The Round trip time was about 17 msec. This report covers our experience of building the test bed and some results using TCP and UDP to transfer data. We used two workstations on each side equipped with a 10 Gbps network interface card. The maximum throughput obtained from memory-to-memory using single stream TCP was about 5.22 Gbps. From our experience we think this technology is a good alternative to build high bandwidth wide area networks for bulk data transfer.

1. INTRODUCTION

Motivations for these tests are validation of 10 Gbps Wide Area Network (WAN) Physical Layer (PHY) technology to extend Local Area Network (LAN) beyond a campus and study scalability of transport protocols over LANs interconnected using WAN PHY. The GRID community is actively trying to build application specific high bandwidth networks interconnecting few locations. Current technology used for wide area interconnection is either routers with Packet over SONET (POS) interfaces or encapsulate Ethernet into a SONET using a device such as ONS 15454.[13] The ability to send Ethernet directly from an Ethernet switch over a SONET link will will eliminate the need for expensive POS interfaces and in some cases need for expensive routers. The challenge of transporting data at high speed between LANs interconnected using WAN is an active research area. We try to understand scalability issues of network architectures and transport protocols. Especially we tested TCP and UDP based applications from work stations equipped with 10 Gbps Network Interface Cards (NIC). To be able to send and receive data at several Gigabits per seconds workstations need a high speed bus and interconnection between bus and memory. We used Itanium 2 based systems. During the initial tests these systems clearly showed advantage over Xeon based systems. We believe it is due to the high bandwidth available between PCI-X bus and main memory.

1.1. Ethernet WAN PHY Technology

By far the most widely used local area network technology is Ethernet. Similarly SONET/SDH is by far the most widely used wide area network technology. SONET has its roots in voice technology while Ethernet has that in computer networks. Both these technologies has advantages of large installed base, stability and they are easy to manage. If we can extend Ethernet over SONET that will be a major step towards convergence and extending the reach.

An Ethernet standard defines Ethernet Media Access Control (MAC) that maps to layer 2 in OSI reference model, Physical Layer to Layer 1 (Copper or fiber). The ten Gigabit Ethernet standard 802.3ae-2002 architecture defines more layers than previous Ethernet standards. Important layers are Physical Media Dependent (PMD) and a Physical Coding Sub-layer (PCS). An example of a PMD is an optical transceiver. The PCS sub-layer performs coding functions such as 8b/10b (10GBASE-X) or 64b/66b and scrambling (10GBASE-R).

The maximum allowed distance between two Ethernet interfaces is medium dependentmuch, a a few hundreds of meters with copper as medium and a few tens of kilometers when dark fiber has been used as medium. When SONET has been used for interconnection there is a difference in the way links and interfaces are presented

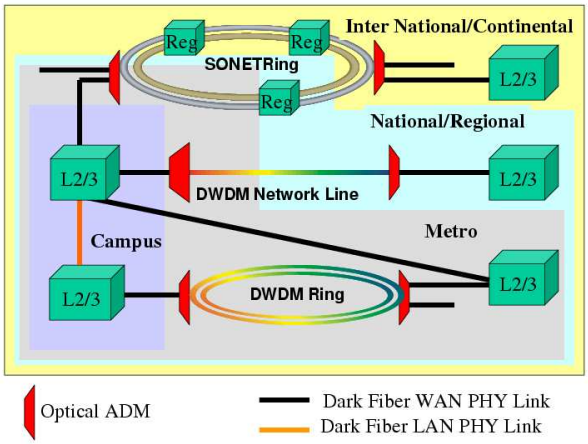


Figure 1. A possible scenario to use WAN PHY technology. 10 Gbps Ethernet can be used to interconnect networks over 2, 20 and 200 msec round trip in other words networks that span Metropolitan, National and International areas

in routers. Most IP routers using SONET links see them as POS interfaces. So there is another layer of encapsulation between the POS and SONET. There are signal regenerators for a SONET/SDH links to extend its range. So effectively the distance between two POS interfaces, IP routers is large, thousands of kilometers. If we can send Ethernet over the same link (with regenerators) then we can extend distance between two Ethernet switches or routers unlimitedly.

Ethernet standard has evolved form 10 Mbps to 10 Gbps. At 10 Gbps both Ethernet and SONET OC192/SDH STM16 has become compatible bit rates. There is a very minor difference in payload rates but that is negligible. So it is possible to send 10 GbE over a SONET link, a "WAN-compatible 10GbE is significantly simpler than the full SDH implementation required to carry POS". This is a great opportunity to extend the well known LAN technology over WAN. Last year IEEE 802.3ae-2002 standard defined two Ethernet PHYs: LAN PHY (approximately 10.00 Gbps) and WAN PHY (approximately 9.29 Gbps). The WAN PHY is simply an optional extended operating feature added to a LAN PHY. Because the difference is very small there are MAC layer implementations that inter-operate between LAN PHY and WAN PHY. According to the architecture of 10 Gbps Ethernet these PHYs are distinguished in the PCS layer. The major differences between LAN PHY and WAN PHY are SONET/SDH framer in the WAN Interface Sub-layer (WIS). Ethernet terminations will not fulfill all requirements of SONET sections such as stringent grid laser specifications, jitter requirements and a stratum clock. The clock specification of WAN PHY is less stringent. It is also important to note that on WAN PHY, Ethernet remains an asynchronous link protocol. As in every Ethernet link, 10 Gigabit Ethernets timing and synchronization must be maintained within each character in the bit stream of data, but the receiving switch, or router may re-time and re-synchronize the data. In contrast, synchronous protocols, including SONET/SDH, require that each device share the same system clock to avoid timing drift between transmission and reception equipment and subsequent increases in network errors where timed delivery is critical. With the relaxed standard it is possible to send Ethernet frames over SONET. SONET interfaces provide detailed and accurate link status reports which is very helpful for trouble shooting. On the other hand classical Ethernet interfaces do not have such good status reports. Ethernet WAN PHY interfaces do provide status reports but their accuracy and reliability are yet to be proven in the field. Advantages of using Ethernet WAN PHY is that most of the networking technology get unified for both local area networks and wide area networks. Simple inexpensive Ethernet switches with WAN PHY interfaces can be used to interconnect LANs over wide area links. It can be used over the existing dark fibers, SONET/SDH infrastructure with regenerators and DWDM infrastructure with SONET ports. By using WAN PHY, the reachability of Ethernet increases, or in short: "Ethernet every where".

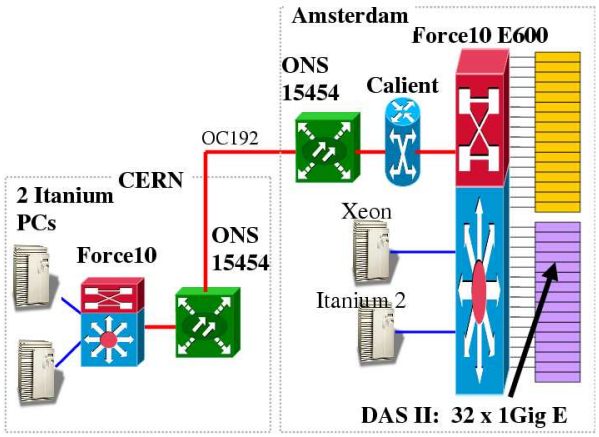


Figure 2. Setup using ONS

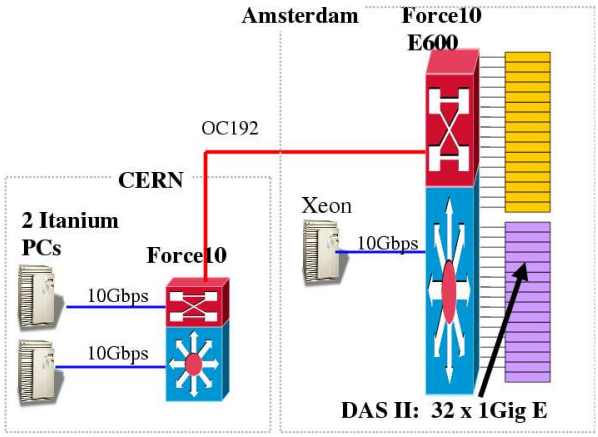


Figure 3. Setup Without ONS

The 10 Gbps switched network between Amsterdam and CERN, Geneva. The Lambda, SONET SONET OC192 Circuit, was terminated using 10 Gbps WAN PHY interface on the Ethernet Switch. At both ends Force10 switches were used.

Switched networks using Ethernet WAN PHY will complement for many ISPs their routed networks. The Possibility to create an all Ethernet based network is an important step to build networks for high bandwidth application users, such as GRID users. Figure 1 shows a typical interconnection scenario of a high bandwidth GRID like network using 10 Gbps Ethernet. Unlike a high speed ISP network which connects to hundreds of networks, a GRID like network connects to very few sites for selected high bandwidth applications. In such cases switched network using Ethernet WAN PHY is very useful.

2. SETUP

The setup we used in this trial was evolved after initial tests conducted by CERN at CANARIE labs to verify the concept of interconnecting two 10 Gbps Ethernet WAN PHY using SONET ports on an add devision multiplexer. The first wide area trial of this concept was between Amsterdam and CERN. The initial setup used IXIA traffic generators with 10 Gbps WAN PHY modules connected to Cisco ONS? OC192 SONET ports. It proved that this technology is capable of transporting Ethernet over SONET between Amsterdam and CERN with bit error rates better than 10^{-14}. The next setup was to interconnect two Force10[14] switches with 10 Gbps WAN PHY interfaces over SONET. Initially we had some troubles getting this to work. The problem was consisted of connecting the WAN PHY interface with the ONS's SONET interface. At the CERN side it went smoothly while at the Amsterdam side there were some problems. We discovered that it had been caused by a broken interface. Once we had replaced it everything worked fine. We got the 10 Gbps switched Ethernet path between Amsterdam and CERN operational as shown in Figure 2. From the small problem we learned that the SONET errors were apparently correctly interpreted by the Force10 switch. After running several tests for a week we removed the ONS' from the path and connected the Force10 directly to the DWDM equipment of Global Crossing as shown in Figure 3. It is worth noting that the receiver side power at the Global Crossing POP in Amsterdam was just within the tolerable limit. Compared to the ONS interface (LR, with launch power +7 dB) the Force10 interface(-3.9 dB) launch power is low. They used 5 dB attenuators with LR interface. May be in the future there will be WAN PHY cards available with higher launch power. The change from ONS to Force10 was transparent to workstations and application layer protocols. The path was about 1700 Km long with three regenerators. Figure 2 shows the setup used for testing the performance of transport protocols. We started our tests using two Itanium systems one system in Amsterdam and the other at CERN. The RTT between two workstations was about 17.3 msec. Figure 4 shows the VLAN and IP address configurations. We choose two VLANs over the link. This made it possible to be able to run two types of tests: one using switched path of 17 msec, and another

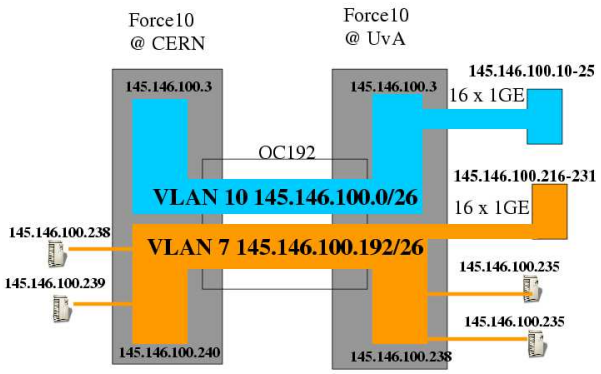


Figure 4. IP Addressing plan and VLAN configuration. This setup allows two type of paths. A switched 17 msec path and a 34 msec routed path.

routed path of 34 msec. This setup could also support the sending of raw Ethernet frames e.g using the IXIA traffic generators.

2.1. Important Specifications of the Force10 WAN PHY Card

In the Force10, E600, a "2-port 10-Gigabit Ethernet WAN PHY Line Card" (LC-EE-1GEW-2Y)[6] was used. To the best of our knowledge Force10 is the only vendor which sell such a card. An important point to note is that according to the ONS specification the wavelength is 1550 nm, while the Force10 interfaces are 1310 nm. Lab tests have shown that if the power levels are a bit above the minimum required level that even with a mismatch in wavelengths they can inter-operate. This is purely based on experience. Below follows a listing of important specifications of the used card.
Physical:
Two full-duplex 10-Gigabit Ethernet WAN PHY ports with LC connectors
IEEE 802.3ae 10GBase-LW compliant

Optical (10Base-LW)
10 Km reach typical on ITU-T G.652 (IEC 60793-2 B1.1/B1.3) SMF
Wavelength: 1290 nm to 1330 nm (1310 nm nominal)
Average launch power: -4 dBm to -1 dBm
Receiver saturation: -1 dBm
Receiver sensitivity: -12 dBm

SONET/SDH:
Local (internal) or loop timed (recovered from network)
Stratum 3 clock accuracy over full operating temperature range
Performance monitoring
Fault management (alarms and status reporting)

Standards and Protocols:
IEEE 802.3ae
IEEE 802.3x Flow Control
IEEE 802.1Q VLAN Tagging
IEEE 802.1p Prioritization

IEEE 802.3ad Link Aggregation
IEEE 802.1D Spanning Tree
MTUs of 9252 bytes supported

2.2. SONET errors reported by the Force10 WAN PHY

The switch reported receiving side SONET errors for path and section. We found that these errors have very similar meanings as explained in a CISCO technical document.[3] In our case errors were caused by two reasons: first by a broken interface, second when there was too much attenuation of the signal in the fiber. The SONET controller output with error is shown in Figure 5 and without error in Figure 6. These errors may help to debug SONET/WDM related problems.

3. PCS USED

In total we used three Itanium 2 systems, one Xeon based system, and 32 Pentium-III systems. The Itanium 2 and Xeon systems were equiped with the 10 Gbps NICs. In Amsterdam site a Xeon system, an Itanium 2 system and 32 DAS-2 nodes. A DAS-2[11] node is essentially a dual Pentium-III processor based system with 1.5 GByte RAM, Gigabit, Fast Ethernet and Myrinet NIC. At the CERN side there were two Itanium 2 systems with 10 Gbps NIC.

3.1. Configuring Itanium 2

Three HP Integrity rx2600 Itanium 2[8] based systems had been used: one in Amsterdam and two in CERN. The workstation in Amsterdam had dual Itanium 2 1.3 GHz Processors and 4 GB main memory. While the CERN hosts had dual 1.5 GHz processors. These systems were equipped with 64 bit 133 MHz PCI-X slots connected memory with a maximum bandwidth of 4 GBps.

The systems were delivered by HP. They were pre-installed with Red Hat Linux. Red Hat Linux Advanced Server release 2.1AS (Derry). We updated few packages to install a new modutils so that we can test 2.6 kernels.The updated packages were: `binutils-2.14.90.0.5-7.ia64.rpm`, `glibc-devel-2.3.2-78.ia64.rpm`, `glibc-2.3.2-78.ia64.rpm`, `libgcc-3.3.1-3.ia64.rpm`, `glibc-common-2.3.2-78.ia64.rpm`, `tzdata-2003a-2.noarch.rpm`.

After the update we compiled the Linux 2.4.21 kernel with Net100 extensions and Linux 2.6t5 kernel. Both appeared to be working fine. Note that LILO configuration is different from the x86 architecture. The configuration file is `/boot/efi/efi/redhat/elilo.conf`. The configuration file is read by the program`elilo`. The directory `/boot/efi` is a vfat partition. The new kernel and the system map file should be copied to the directory `/boot/efi/efi/redhat/` and to make the initrd image in the same directory using the command `mkinitrd --omit-scsi-modules -v -f initrd-${version} ${version}`. To be able to boot with the new kernel is should be added to the file `elilo.conf` in the directory. There no need to run a command as like `lilo` to be able to boot with the new kernel.

Iperf V. 1.6.5 has been used in stead of version 1.7.0, where the shaped UDP bandwidth is defined with a 32 bit Integer which is not easy to change in the code to a required 64 bit Integer. The usual modifications to V. 1.6.5 has been applied from which the most important is the usage of 64 bit Integers in stead of the default 32 bit Integers. All modifications are described in the UvA DataTAG Web site.[10]

Especially for the Itanium architecture, in the configure file `./cfg/config.sub` below the installation root the RE `ia64-*` has to be add to the OR'ed RE's in the `case` statement of line 174.

3.2. Intel 10 Gbps NIC

We used the Intel 10 GbE NIC.[9] To be able to get maximum performance the NIC was connected to the PC via a 64 bit 133 MHz PCI-X slot. The card configurations can be checked by reading the `/proc` entry which has been created by the driver. The used driver was is V. 1.0.47 compiled with NAPI enabled. NAPI is an option in src/Makefile `ENABLE_NAPI := 1`. Figure 7 shows the important parameters from a system we used. The card supports upto 16 KByte MTU. Due to limitations of the switch during our tests we used MTU of 9000 bytes.

```
Force10#show controllers  tengigabitethernet 4/0
Interface is TenGigabitEthernet 4/0

 SECTION
  LOF = 1    LOS = 1                              BIP(B1)  = 124

LINE
 AIS = 0    RDI = 0               FEBE = 0    BIP(B2) = 9813

PATH
 AIS = 0    RDI = 0    LOP = 0    FEBE = 700    BIP(B3) = 117


Active Defects:  SLOS SLOF

Active Alarms:   SLOS


Alarm reporting enabled for: NONE

 Framing is SONET, AIS-shut is enabled
 Scramble-ATM is enabled, Down-when-looped is enabled
 Loopback is disabled, Clock source is line, Speed is Oc192
 CRC is 32-bits, Flag C2 is 0x1a, Flag J0 is 0xcc, Flag S1S0 is 0x0
```

Figure 5. SONET Controller error shown by Force10 when there is no signal on RX

```
Force10#show  controllers tengigabitethernet 4/0
Interface is TenGigabitEthernet 4/0

SECTION
 LOF = 0    LOS = 0                              BIP(B1)  = 3497

LINE
 AIS = 0    RDI = 0               FEBE = 0    BIP(B2) = 134475

PATH
 AIS = 0    RDI = 0    LOP = 0    FEBE = 725    BIP(B3) = 2081

Active Defects: NONE

Active Alarms:  NONE

Alarm reporting enabled for: NONE

 Framing is SONET, AIS-shut is enabled
 Scramble-ATM is enabled, Down-when-looped is enabled
 Loopback is disabled, Clock source is line, Speed is Oc192
 CRC is 32-bits, Flag C2 is 0x1a, Flag J0 is 0xcc, Flag S1S0 is 0x0
```

Figure 6. SONET Controller output show by Force10 when there is no error

```
[root@pcepatr28 root]# cat /proc/net/PRO_LAN_Adapters/eth2.info
Description                     Intel(R) PRO/10GbE Network Connection
Part_Number                     a82505-005
Driver_Name                     ixgb
Driver_Version                  1.0.47-k1
PCI_Vendor                      0x8086
PCI_Device_ID                   0x1048
PCI_Subsystem_Vendor            0x8086
PCI_Subsystem_ID                0xa11f
PCI_Revision_ID                 0x01
PCI_Bus                         128
PCI_Slot                        1
PCI_Bus_Type                    PCI-X
PCI_Bus_Speed                   133MHz
PCI_Bus_Width                   64-bit
IRQ                             57
System_Device_Name              eth2
Current_HWaddr                  00:07:E9:0D:41:73
Permanent_HWaddr                00:07:E9:0D:41:73
```

Figure 7. 10 Gbps NIC reported by Linux /proc variable

4. TRANSPORT PROTOCOL TESTS

The objective of the following tests is to understand the capabilities of transport protocols over the test bed. We are interested in single stream, multi-stream between a pair of hosts and multiple flows between multiple hosts. The used transport protocols are stock TCP, High Speed TCP (HSTCP)[2] and UDP based protocol UDT.[12] The motivations for TCP tests are understand scalability of the protocol when 10 Gbps NIC's are used, to compare reactive response of stock TCP and HSTCP.[2]

4.1. Single Stream TCP

These tests are aimed at understanding the TCP buffer size required to support single stream. The duration of each test was 60 seconds. We varied the buffer size, `iperf -w` option between `1-32` Megabytes. It is interesting to note that the bandwidth delay product of 5.4 Gbps times 17 msec is about 12 MBytes while actually the socket size required to obtain this speed is about twice the bandwidth delay product. The `iperf -l` seems to be quite sensitive. We used `-l 8000`. The results of single stream TCP is shown in Figure 9 on the left side. Since there was no congestion on the link it is possible to sustain this throughput for long periods. Long tests were conducted by the CERN group. In the multi-stream experiment, the first series of tests show some anomalies. The right side of Figure 9 shows the results of multi-stream tests with variable buffer size. We have noticed that such occasional difference in the results are mostly related to free memory and system load. A drastic measure which helped within the given time constraint was rebooting the system.

4.2. TCP responsiveness with HSTCP

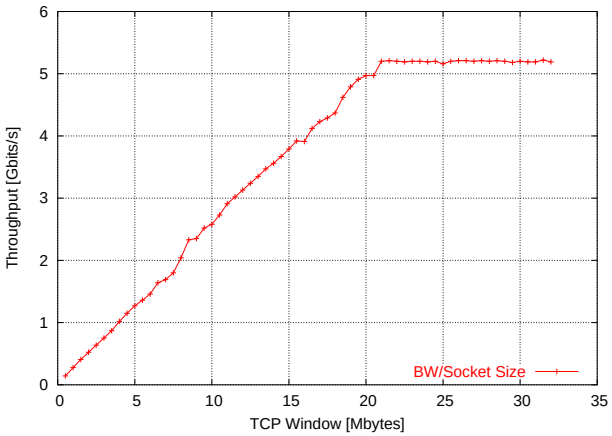
The objective is here to observe TCP responsiveness to a total congestion event. First we start a TCP `iperf` session between two hosts and send large UDP stream from a third host to the receiver. Large UDP stream causes multiple packet loss at the receiver. TCP responsiveness is the time it takes to recover after the congestion event. Figure 10 shows the comparison between the stock TCP implementation and Net100 implementation of HSCTP in the 2.4.21 kernel. The responses are as expected. HSTCP has clear advantage over TCPNewReno.

4.3. UDT

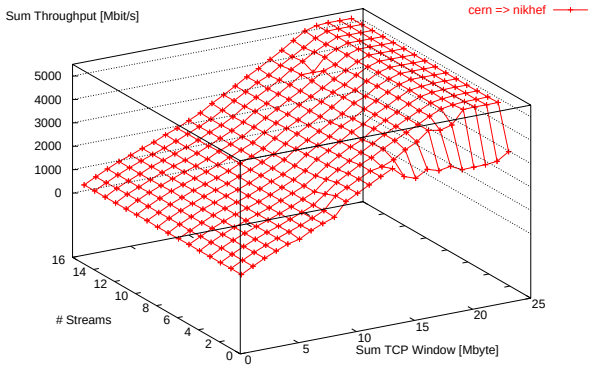
UDT is an UDP based Data Transfer protocol. It has been developed at the University of Illinois, Chicago. The major advantage of UDT is more effective link utilization on high bandwidth delay product networks. The parameter to control the maximum throughput is Flow Control (FC). The slow start part is also controlled by FC. It limits the number of unacknowledged packets. The flow window size is the product of packet arrival speed

```
# to modify the MMRBC field in PCI-X configuration space to
# increase  transmit burst lengths on the bus.
/sbin/setpci -d 8086:1048 e6.b=2e

#increase the MTU
/sbin/ifconfig eth2 mtu 9000

# Following sysctl variables are changed using sysctl command
# for really big buffers
net.core.rmem_max = 33554432
net.core.wmem_max = 33554432
# leave the default low
net.core.rmem_default = 65536
net.core.wmem_default = 65536
# increase Linux autotuning TCP buffer limits
# 32M
net.ipv4.tcp_rmem = 4096 87380 33554432
net.ipv4.tcp_wmem = 4096 65536 33554432
net.ipv4.tcp_mem = 33554432 33554432 33554432
# suggested by tk for high speed flows
# probably need a *little* tweaking
net.core.mod_cong = 2800
net.core.lo_cong = 1000
net.core.no_cong = 200
net.core.no_cong_thresh = 2900
net.core.netdev_max_backlog = 3000
# some web100 stuff
# should have no effect on non-web100 hosts
net.ipv4.web100_default_wscale = 8
net.ipv4.web100_no_metrics_save = 1
# turn on floyd's AIMD
net.ipv4.WAD_FloydAIMD = 1
# or ifconfig txqueuelen 10000
net.ipv4.WAD_IFQ = 1
# don't trust these, and they get in the way of testing
net.ipv4.web100_sbufmode = 0
net.ipv4.web100_rbufmode = 0
### ATLAS Canada guys igrid paper recommendations.
vm.bdflush = "30 500 0 0 500 3000 60 20 0"
vm.min-readahead=127
vm.max-readahead=256
```

Figure 8. Host variables tuned to get better performance.

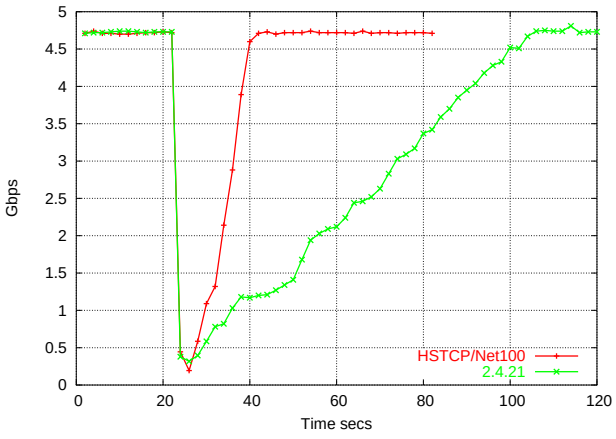


throughput vs socket size between Amsterdam and CERN. Maximum throughput is is 5.22 Gbps



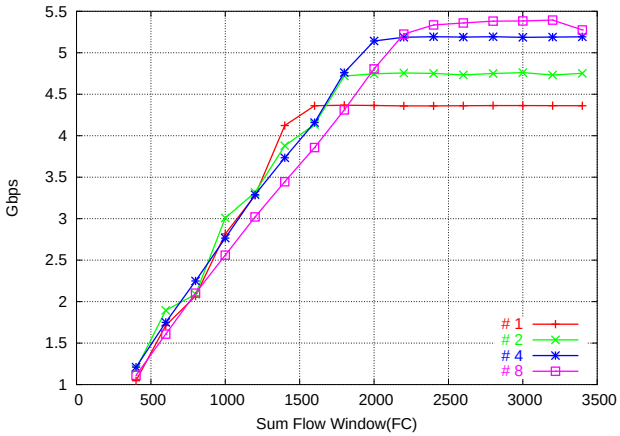
Throughput vs socket size and number of streams between CERN and Amsterdam

Figure 9. TCP Avg throughput of 60 second tests using iperf between a pair of hosts. RTT 17 msec

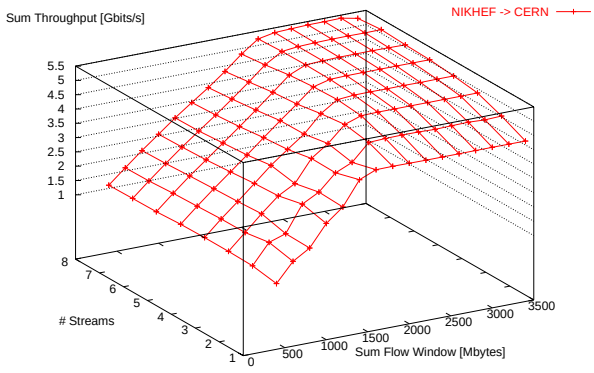


TCP congestion avoidance response. Throughput vs time for HSTCP and stock with 2.4.21 kernel

Figure 10. TCP response



with 1,2,4,8 flows



1-8 flows

Figure 11. UDT memory to memory throughput vs FC and number of streams between Amsterdam and CERN, RTT = 17 msec

and sum of RTT and constant ACK time. The fairness control is similar to AIMD based algorithms (TCP). The tunable parameter in UDT, FC, is given by the relation $FC = Bandwidth * (RTT + 0.01)/MTU$. For example, on a 10 Gbps link, 100 ms RTT, and 1500 byte MTU, FC is 91666.

Figure 11 shows the results between Amsterdam and CERN using two Itanium 2 systems. The maximum performance is about 5.4 Gbps. It is similar to the performance of TCP using the same set of machines. FC has been varied from 400 to 3400 Mbytes (sum over all flows) an MTU of 9000 bytes and 1 - 8 flows.

4.4. Multiple TCP One Gbps Flows

In these tests we used the DAS-2 clusters as shown in Figure 2. Also half of the nodes can reach the other half via the router at CERN. Sixteen nodes are routed to the other sixteen nodes via the Force10 switch at CERN acting as a router. The path is a 34 msec RTT. We have also set the MTUs on the DAS-2 nodes to 8192 Bytes. Even though the gigabit NICs support 9000 Bytes we choose lower value to be able to compare the results with other tests. The ideal expected behavior is that each flow can get an equal and maximum bandwidth of about one Gbps. There may be small variations due to system overhead which can be tuned by carefully adjusting the host parameters. Figure 12 shows the results of 1-10 Flows in one direction. For single flow we get about 760-800 Mbps. When there are multiple flows some of them get lower throughput than other flows. This behavior is rather unexpected; there is no full explanation yet. But in our understanding it may be either that TCP flows can't equally share an uncongested link or host parameters are not well tuned. It is interesting to note that multi-stream (two hosts, parallel TCP flows) TCP tests show a higher bandwidth utilization.

In the next tests half of the flows were send in one direction while the other half was send in the opposite direction. In this way we expect to introduce traffic in both direction on the duplex OC192 link to be able to observe the behavior of multiple TCP flows. Since the link is not saturated there is no bottleneck so the expected behavior is that each flow gets an equal and maximum bandwidth of about 1 Gbps. Figure 13 shows the results of this test. The distribution is not uniform which needs further investigation. We also do not observe any pattern with higher number of flows. May be the test could be improved by tuning host and application parameters.

5. FUTURE WORK

Building a Trans-Atlantic path using WAN PHY technology will enable us to investigate the scalability of the technology. The dynamics of multiple flows using multiple workstations and single flow with background traffic need to be investigated further. Mixing TCP and UDP based protocols is another interesting area. In future tests we could also explore disk-to-disk transfer. Another possible area is to explore the shaping and rate-limiting

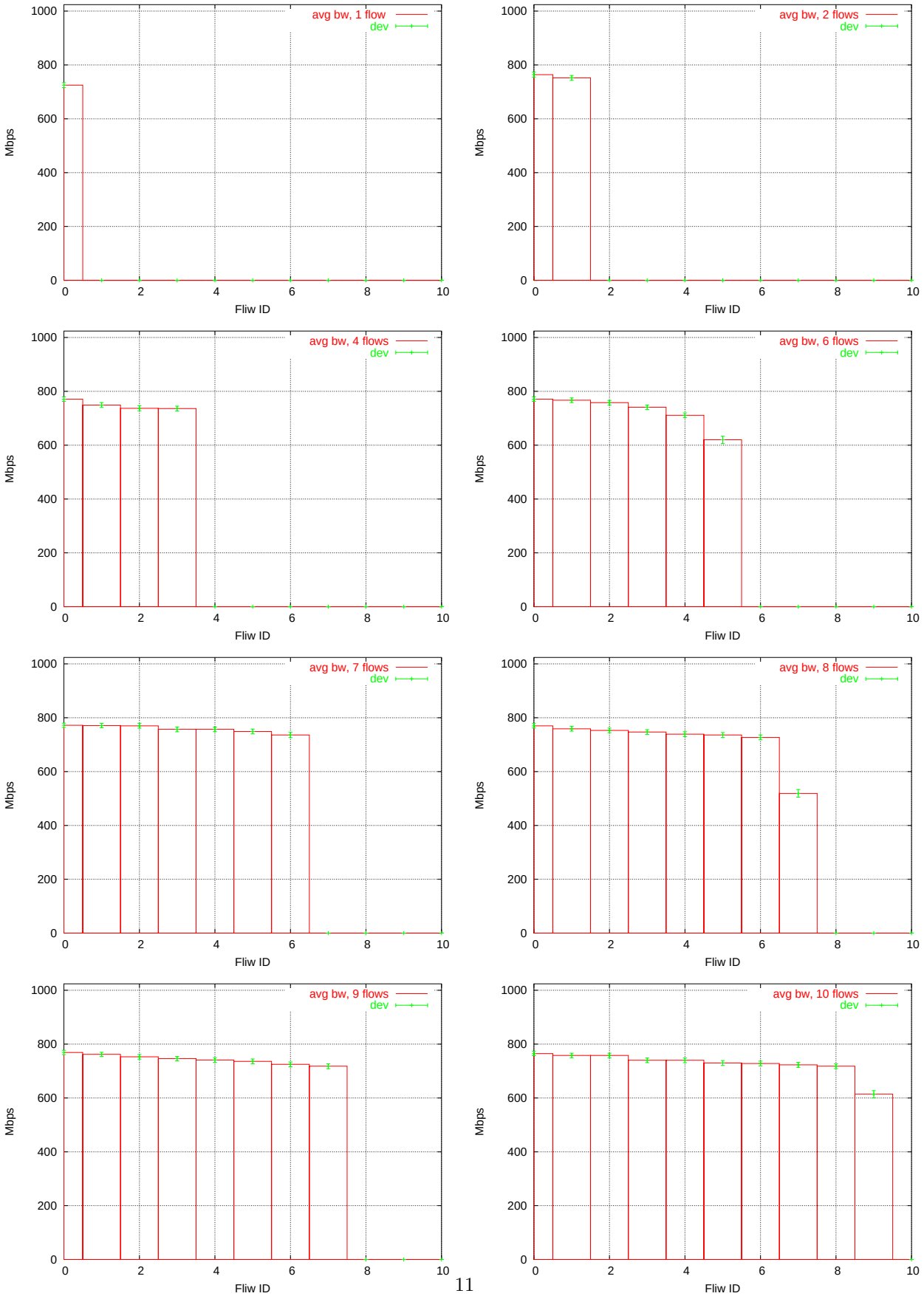


Figure 12. Average throughput of 40 minutes and standard deviation. Both two DAS-2 node groups were connected via a routed loop with an RTT of 34 msec. Sender and receiver in Amsterdam

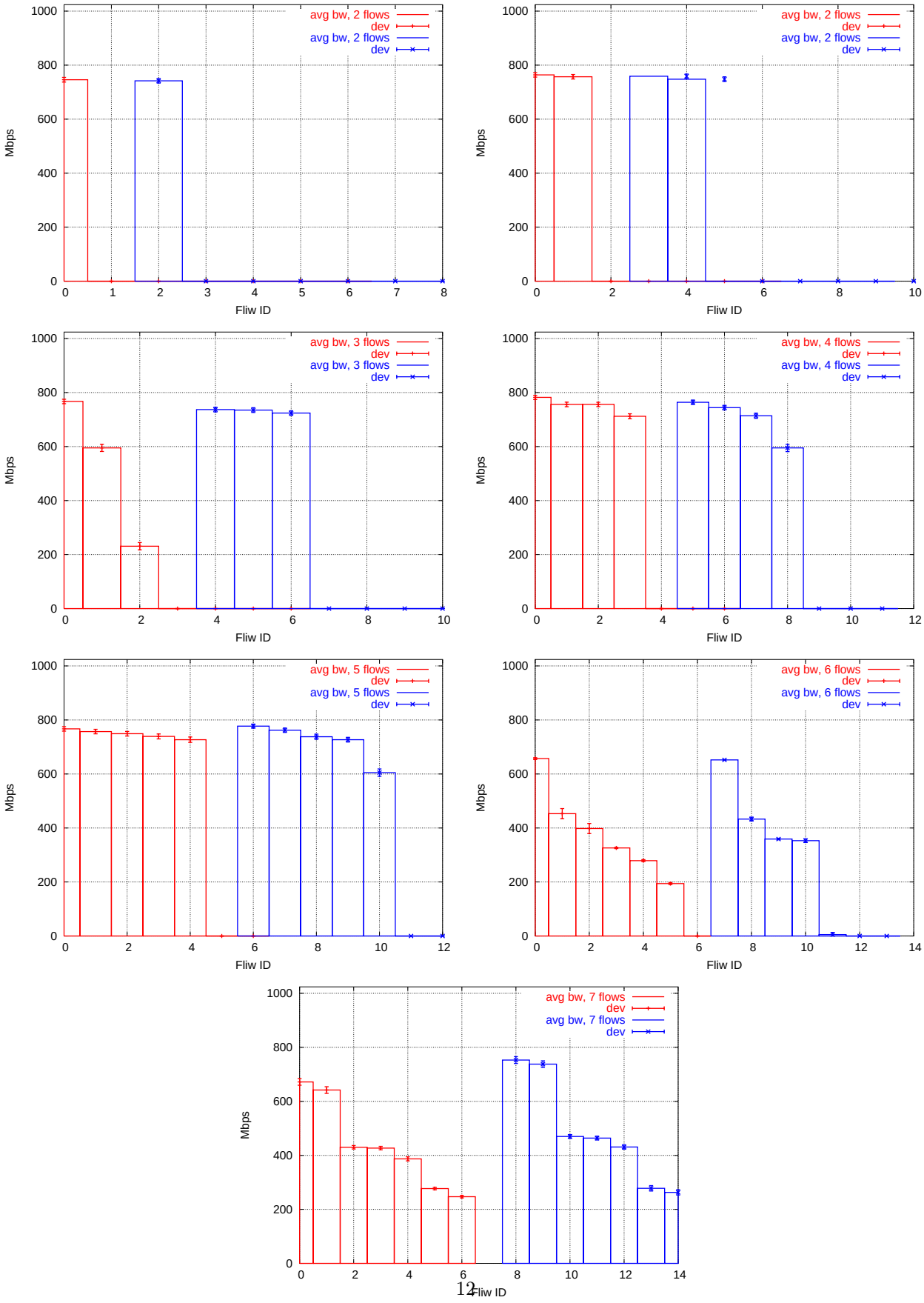


Figure 13. Average throughput of 25 minutes and standard deviation. The red lines denote the flows in one direction. The flows in the opposite direction are marked with blue lines. In this test the RTT between the sender and receiver node was 34 ms.

features of Force10 switch. If we need to use the rate limiting on switches that may have major influence on TCP flows.

6. CONCLUSION

As speed of networking increases Ethernet is keeping pace with the need to create faster research networks and commodity internet services. Transport protocols are also becoming better and better at supporting demands of applications to transport bulk data over long distances.

We have demonstrated that 10 Gbps Ethernet WAN PHY interfaces on switches can successfully inter-operate with OC192 SONET terminations as well as SONET/WDM add devision multiplexers. A SONET or a DWDM network infrastructure can transport Ethernet frames. This technology can be used to create switched networks over existing SONET and/or DWDM infrastructure. With respect to workstations, currently PCI-X interconnections to CPU and memory bus appear to be the bottleneck. Single stream TCP with HSTCP modification and UDP based protocols are scalable for Trans Europe distances (17 msec) at high speeds. Maximum performance using single stream TCP is about 5.22 Gbps.

7. ACKNOWLEDGMENTS

This work was made possible by a large collaboration. We had lot support from vendors network operator and researchers. We wish thanks the following colleagues for their valuable support.

CANARIE, Rene Hatem. CERN EP division, Bob Dobinson, Stefan Stancu, Piotr Golonka, Mihai Ivanovici. CERN IT division, Stan Cannon. CERN Openlab, Andreas Hirstius. Cisco Europe. Cortex Networks. Ottawa, Ray Belleville. ESTA EU Project (IST-2001-33182), Catalin Meirosu. Force10 Canada and Europe. Global Crossing. HP Europe. IST-2001-32459 DataTAG EU Project, Johan Blom, Antony Antony. Intel. Ixia Canada and Europe. SARA, Pieter de Boer. SURFnet, Erik Radius. University of Amsterdam, Cees de Laat, Freek Dijkstra, University of Carleton ,Wade Hong.

We would also like to thank the Net100/WEB100 collaboration for instrumenting TCP with modifications.

REFERENCES

1. DataTag Project webpage: http://www.datatag.org
2. Floyd, Sally, "HighSpeed TCP for Large Congestion Windows", IETF Internet Draft, http://wwwietf.org/inter-drafrts/draftfloyd-highspeed-02.txt
3. CISCO, Troubleshooting Physical Layer Alarms on SONET and SDH Links, `http://www.cisco.com/warp/public/127/sonetalarms_16154.html`
4. Installing Cisco ONS 15454 OC192LR/STM64 LH 1550 Cards `http://www.cisco.com/en/US/products/hw/optical/ps2006/prod_module_installation_guide09186a008007ec45.h`
5. 10 Gigabit Ethernet Interconnection with Wide Area Networks : `http://www.10gea.org/10GbE%20Interconnection%20with%20WAN_0302.pdf`
6. 10-Gigabit Ethernet WAN PHY Line Card : http://www.force10networks.com/products/lcee-10gew-2y.asp
7. WAN PHY Definitions http://grouper.ieee.org/groups/802/3/ae/public/terminology.pdf
8. HP Integrity rx2600 server specifications: `http://www.hp.com/products1/servers/integrity/entry_level/rx2600/specifications.html`
9. Intel PRO/10GbE LR Server Adapter Specifications http://www.intel.com/support/network/adapter/pro10gbe/pro10gbelr/specifications.htm
10. Modifications to iperf: `http://carol.science.uva.nl/~jblom/datatag/wp3_1/tools/test_tools.html#Iperf-Mods-Sect`
11. DAS-2 architecture : http://www.cs.vu.nl/das2/das2-machine.html
12. UDT http://www.dataspaceweb.net/
13. Cisco ONS 15454. http://www.cisco.com/en/US/products/hw/optical/ps2006/ps2010/index.html
14. Force10 Switch specifications. http://www.force10networks.com/products/products6.asp