A Simulation model for e-VLBI traffic on network links in the Netherlands.

Julianne Sansa
Kapteyn Astronomical Institute
Postbus 800
9700 AV Groningen
The Netherlands
sansa@astro.rug.nl

Arpad Szomoru
Joint Institute for VLBI in Europe
Oude Hoogeveensedijk 4
7991 PD Dwingeloo
The Netherlands
szomoru@jive.nl

J.M. van der Hulst
Kapteyn Astronomical Institute
Postbus 800
9700 AV Groningen
The Netherlands
vdhulst@astro.rug.nl

September 10, 2006

Abstract

This paper presents an ns-2 [1, 2] based simulation model for e-VLBI traffic and the bottlenecks currently limiting the transfer speeds of astronomical data from radio telescopes across the world over high speed Internet links to the central processing centre in the Netherlands. This model is useful for studying the transport characteristics of the e-VLBI data as it traverses the networks. This model accentuates that a combination of large idle times between data bursts caused by application limitation, inefficient receiver hardware and excessive background traffic negatively affect the performance of e-VLBI data transfers.

Keywords: e-VLBI, radio astronomy, simulation, transport protocols

1 Introduction

The telescopes used in the European VLBI Network (EVN) produce data at rates of up to 1 Gbps each. Traditionally, these data streams were recorded on tapes (nowadays hard disk drives) and shipped to the correlator located at JIVE, the Joint Institute for VLBI in Europe, in Dwingeloo, the Netherlands. During the last few years JIVE, in collaboration with the European National Research Networks and the pan-European Research Network GEANT, have worked on a proof-of-concept (PoC) project to connect several telescopes across Europe in real-time to the correlator via the Internet (electronic VLBI or e-VLBI). This project has led to an EC sponsored project called EXPReS, which over the next few years will transform the EVN to a fully functional real-time e-VLBI network.
During the PoC project it became clear that in spite of the vast capacity of the connecting networks, the actual transport of large amounts of data poses quite a challenge especially for real time correlation and in utilizing all the physically available bandwidth.
The Mark5 [3] application that handles e-VLBI data uses the Transport Control Protocol (TCP). By the nature of e-VLBI, huge amounts of data have to be transported via the Internet over long distances from geographically dispersed telescopes to one central correlator.
In order to investigate the e-VLBI data transport characteristics and scrutinise the data flow limitations we simulated an e-VLBI data flow similar to the one reported in [4], while gathering TCP statistics on the congestion window (CWND), receive widow (RWND), Round Trip Time (RTT), packet loss and the resulting throughput. In this paper we report on the simulation setup and the extent to which each identified bottleneck limits the e-VLBI flow as well as the combined bottlenecks effect.

1.1 Background and Motivation

A TCP connection (used to transport application data such as e-VLBI data) [5, 6] keeps track of a set of variables namely CWND, RWND, RTT and packet loss that impact its throughput. The CWND is the amount of data that a sender can send before receiving any feedback from the receiver. A TCP receiver maintains a RWND for the purpose of informing the sender how much data it is willing to accept

in one go without needing to generate acknowledgements. It is thus in the best interest of the TCP connection that CWND and RWND are close to one another, otherwise the lesser value will be the effective value for both. The CWND is controlled by TCP congestion control which may exist in either one of two states, the slowstart or the congestion avoidance. The RWND on the other hand is set to the minimum between the size of the receive buffer and the bandwidth delay product (BDP) shown in by *equation (3)* below. The receive buffer is one of the high performance setting referred to in *subsection 2.3* and may also be changed by the application. On highspeed links, the like of which are used for e-VLBI the bulk of the TCP connection's life time is spent in the congestion avoidance state in which the CWND update is characterised by equations (1) and (2).
On acknowledgement (ACK):

$$CWND_{new} = CWND_{old} + a/CWND_{old} \quad (1)$$

On packet loss:

$$CWND_{new} = CWND_{old} - b \times CWND_{old} \quad (2)$$

$$\text{where } a = 1 \ \& \ b = 0.5$$

From Floyd [7] the following relationships between the TCP variables are established:

$$CWND_{optimal} = Bandwidth \times RTT \quad (3)$$

$$Throughput = \frac{CWND_{average}}{RTT} \quad (4)$$

$$CWND_{average} = \frac{1.2}{\sqrt{p}} \quad (5)$$

Where RTT is the *Round Trip Time*, and p is the *packet loss rate*.
From equations (4) and (5), we note that the TCP throughput is directly proportional to the CWND and inversely proportional to the RTT and that the CWND in turn is inversely proportional to the square root of packet loss rate. Making long distance connections over high speed links as is the case for e-VLBI, implies that the optimal CWND required to ensure high throughput has to be large, as suggested by equation (3), which is also referred to as the BDP. The TCP congestion avoidance algorithm, in equations (1) and (2), however exhibits a weakness in that it can not maintain a large CWND in the presence of packet loss.
From a previous study [4] on the network links of interest with an e-VLBI data flow, we concluded that the data transfer rate being below optimal could be explained as caused by application limitation (large bursts separated by large idle times), receiver hardware inefficiencies and some background traffic. We therefore set out to design an e-VLBI application model, which highlights the traffic generation patterns, background traffic and network limitations on such a network. This model can be used to test suggested improvements of the underlying transport protocols.

1.2 Related Work

Some of the initial important efforts to design application-specific traffic models are reported in [8, 9, 10, 11, 12]. With the growth of the web, general traffic generation models that characterise the web traffic have been developed [13, 14, 15, 16, 17, 18]. Other, aplication-specific models developed are for the File Transfer Protocol and Simple Mail Transfer Protocol and are reported in [19]. For networking studies these models provide critical data characterising the TCP connections between the sender and receiver in terms of connection establishment rates and the sizes and timing of exchanges of request and response data. In developing these models the various methods used to trace the data include embedding instrumentation software in the client, installing specialised software and hardware in the network or installing publicly available packet capture tools on off-the-shelf hardware. In developing an e-VLBI traffic model we used the most viable option of publicly available software on our existing hardware. In addition to characterising the traffic generation patterns of the e-VLBI disk2net - net2disk data flow, our model also describes the network limitations faced by such a flow as observed in [4].

2 Methodology

In this section we describe the network topology over which we conducted the network tests from which data was used to develop the model as well the hardware configuration and the tools we used to gather network statistics and to simulate the model.

2.1 Setup

Tests were conducted between Mark5 units located at JIVE interconnected via Amsterdam through Netherlight [20] as shown in Figure 1. The end-to-end path has a minimum of 1 Gbps.

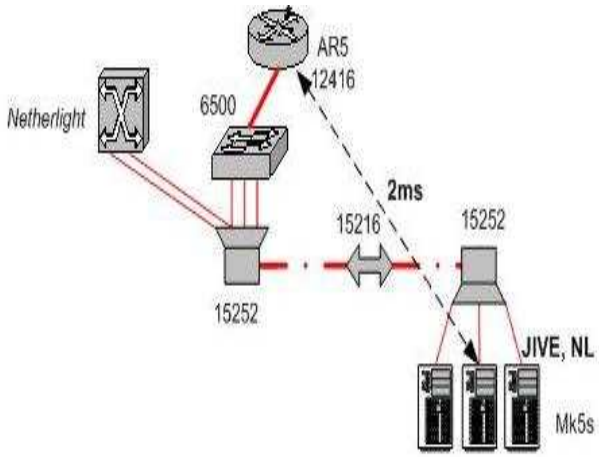

Figure 1: *Network topology over which tests were conducted*

Each of the links from the router AR5 to each of the Mark5's is 1 Gbps. Statistics were gathered during Mark5 disk2net-net2disk transfers from one Mark5 unit through the router AR5 to another Mark5 unit. The Mark5 application has several data transfer modes. The disk2net-net2disk mode which we used gets already recorded VLBI data on a disk and sends it to the network on the sending side and on the receiving side data is obtained from the network and written to the disk.

2.2 Hardware configuration

The hardware configuration of the Mark5 units is as follows:

- Motherboard: Supermicro P3TDLE, FSB 133 MHz
- Processor: Pentium III 1.26 GHz, FSB 133 MHz, 512K L2
- Chipset: ServerWorks Serverset III LE
- Memory: 256M PC133 SDRAM
- Operating System: RedHat Release 9.0, linux kernel 2.4.20-6

2.3 High Performance Networking Options

In order to ensure maximum throughput, the following well known high performance networking options [21] were tuned. Maximum Transmission Unit (MTU) of 8192 bytes was supported for the tests compared to the default 1500 bytes. TCP buffers were set to 4 Mbytes compared to the 64 Kbytes default while txqueuelen was set from a default of 100 to 20000, which has been proved to offer good performance [22]. The default values are too small to support high speed data transfers.

2.4 Measurements with TCPdump

During e-VLBI transfers we gather statistics using TCPdump [23], a publicly available tool. TCPdump generates no traffic, it merely keeps track of all the traffic going through a particular network interface, making it a passive tool.

2.5 Simulating with ns-2

Based on the data (on behaviour of CWND, RWND, RTT and packet loss) we obtained from measurements with TCPdump, we simulate an e-VLBI data flow in the ns-2 simulation environment. This simulation assists us to construct a data generation model that approximates the e-VLBI flow behaviour. ns-2 is a publicly available network simulator that supports extensive network transport simulation in both the wired and wireless environments. It returns results faster than other network simulators [24] but is also very memory intensive.

3 Tests & Simulations

In this section we present the results of our tests and simulations in four subsections: observed CWND & RWND, packet loss, RTT and TCP throughput.

3.1 CWND & RWND

As mentioned in Section 1.1, CWND and RWND should be nearly equal and large enough to allow full bandwidth utilisation. The links have a bandwidth of 1 Gbps, the RTT is ~ 4 ms. From equation (3) in Section 1.1 the optimal CWND and RWND for our tests is 0.5 Mbyte.

Figure 2A shows the CWND/RWND results for a real e-VLBI data flow between two MarkV units, both located at JIVE. Both CWND and RWND are far below the optimum expected. The flow yields a CWND average of 0.06×10^6 bytes and a RWND average of 0.05×10^6 bytes, which is 10% of the optimal value . RWND is steadily below CWND. *Figure 2B* on the other hand shows the CWND/RWND results for a simulated e-VLBI data flow also with the same averages.

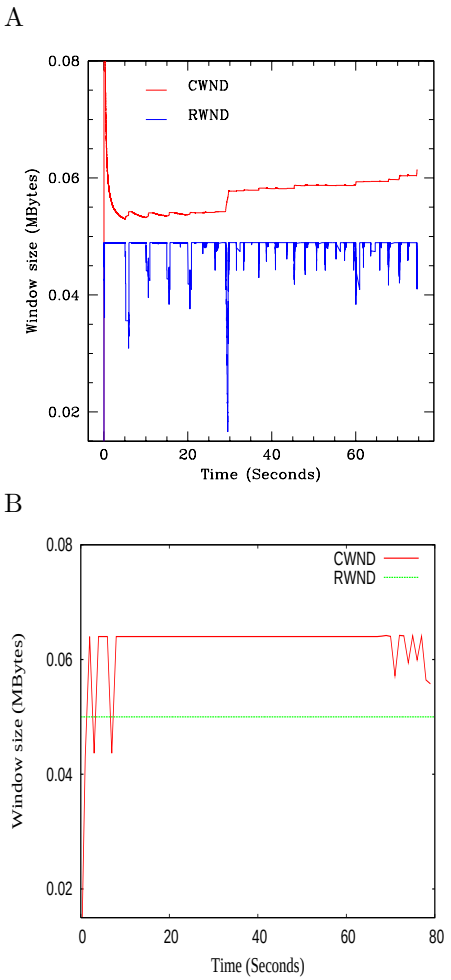

Figure 2: *Congestion Window and Receive Window for a disk2net-net2disk e-VLBI data transfer (RWND steadily lower than CWND): A - Real flow, B - Simulated flow.*

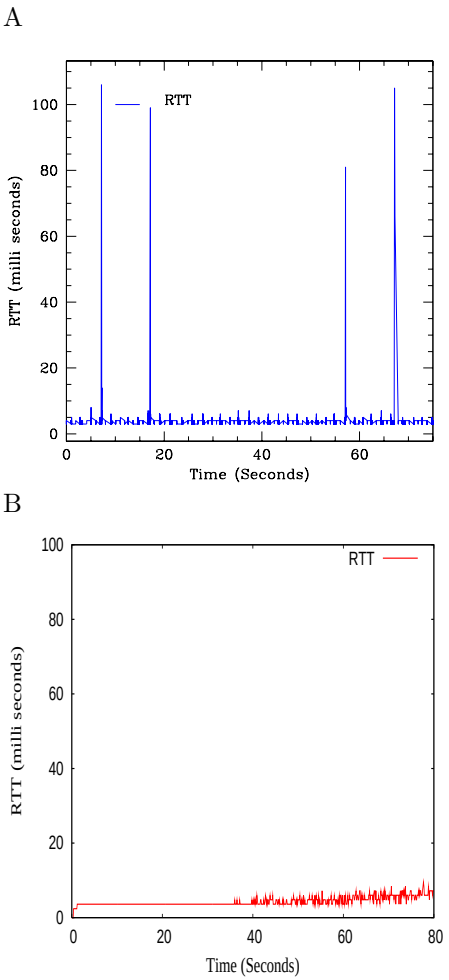

Figure 3: *Round trip time for a disk2net-net2disk e-VLBI data flow between two MarkV units both in the Netherlands: A - Real flow, B - Simulated flow.*

3.2 Packet loss

From our measured maximum CWND we calculate a steady state packet loss of 7.49×10^{-6} packets /second [7]. A rate larger than this steady state rate will cause a decrease of the CWND, while a smaller rate will cause an increase. We observed zero packet loss rate during the e-VLBI data flow. As the observed packet loss rates are much smaller than the steady state packet loss rate, the CWND should be increasing. This is however not the case. We will return to this in Section 4.

3.3 RTT

During e-VLBI data transfers between two hosts both located at JIVE in the Netherlands, we noted an average RTT of 3.8 ms with a few spikes to approximately 100 ms. These results are shown in *Figure 3*. The general observation is that RTT values are quite stable making RTT ideal to use as a sign of congestion.

3.4 TCP throughput

Figure 4A illustrates the achieved TCP throughput during a real disk2net-net2disk e-VLBI transfer in which we observed an average of 366.9 Mbps. The performance was affected by the start of tcpdump at the same time making the throughput to fluctuate at the beginning of the flow and stabilising a short while later. The effect is by a factor of between 20% and 40%, which implies this average throughput would have been a value between 440.3 Mbps and 513.7 Mbps without TCPdump. *Figure 4B* shows the throughput of simulated flow with an average of 374.5 Mbps. The throughput pattern and value is close enough to the real flow.

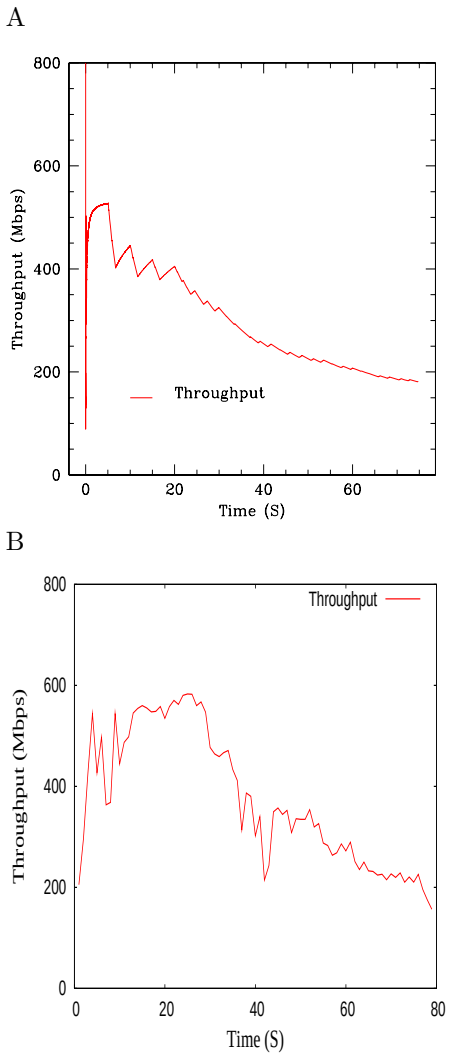

Figure 4: *TCP throughput for a disk2net-net2disk e-VLBI data flow: A - Real flow, B - Simulated flow.*

4 Discussion of Results

In this section we discuss the results obtained from the tests and simulations presented in the Section 3 as well as the model that estimates the e-VLBI data flow.

4.1 e-VLBI Flow behaviour

For both the real and simulated e-VLBI flows, shown in *Figure 2* we see the same CWND sustained for a period of time and yet with both slow-start and congestion avoidance states of TCP we should be seeing CWND either increase or decrease. This can be explained as idle connection window validation [25], in which the same constant CWND is maintained during an application limited period, which means the CWND is not increased merely by reception of ACKs, as long as during the previous RTT the flow did not fully use the available CWND. This seems to imply that the flow experiences severe application limitation on these high speed links. Application limitation happens when the application does not produce data fast enough [26] for two reasons. Either the application is transfering small amounts of data at a relatively constant rate to the TCP layer or the application is producing data in bursts separated from each other by idle periods. Based on the fact that e-VLBI generates huge amounts of data, we conclude that the flow is application limited due to the latter.

In addition having set the TCP buffers (both send and receive buffers) to 4 Mbytes as mentioned in *subsection 2.3* and having computed the BDP to be 0.5 Mbytes as in *subsection 3.1*, we would expect the RWND to be 0.5 Mbytes but having obtained an average RWND of 0.05 Mbytes we conjecture that the application process altered the RWND based on its perception of the receiver's capacity. This then suggests limitation in the receiver.

4.2 An e-VLBI Traffic Model

4.2.1 Data Generation Characteristics

Data generation based on "on/off" traffic patterns, the like of which we simulate for the e-VLBI data flow can be varied based on the burst time and idle time between the bursts. During "on" periods, packets are generated at a constant burst rate. During "off" periods, no traffic is generated. Burst times and idle times are taken from exponential distributions. Having observed that the e-VLBI data flow is composed of large bursts separated by large idle times, we set the initial burst time to a constant of 500 ms and begin to vary the idle time to find the most suitable value. *Figure 5* indicates the achieved throughput for each of the idle periods simulated. It is shown that the larger the idle period the less throughput will be attained by the flow.

4.2.2 Background Traffic Effect

Background traffic can be varied depending the applications and users in the network. The starting time as well as the duration of the background traffic can also be significant effect. Generally the web being the most widely used application on the Internet implies that web traffic contributes substantially to background traffic. A number of applications such as those used for remote access are also common and have the characteristic of using small window sizes. In addition since TCP is the commonest transport protocol on the Internet, the majority of background traffic is carried over TCP connections. Based on these observations we simulated the e-VLBI data flow amidst changing background traffic, which consists of varying number of web sessions, small TCP flows (some going in the same direction as the e-VLBI flow and others in the opposite direction) and normal sized TCP flows (going in the reverse direction as the e-VLBI flow). *Figure 5* indicates the achieved throughput for each of the background traffic combinations simulated. From

the plot the more the back ground traffic the less throughput will be attained by the flow.

4.2.3 Receiver Limitation

To simualate the receiver limitation we set the maximum CWND and RWND to values estimated by the real flow, however we also simulate the flow with varying values of maximum CWND. This is shown in *Figure 5*, which indicates that initially the larger the maximum CWND the higher the achieved throughput. However beyond a maximum CWND of 256 packets (0.2 Mbytes), the same throughput is achieved owing to the bursty limitation in the application.

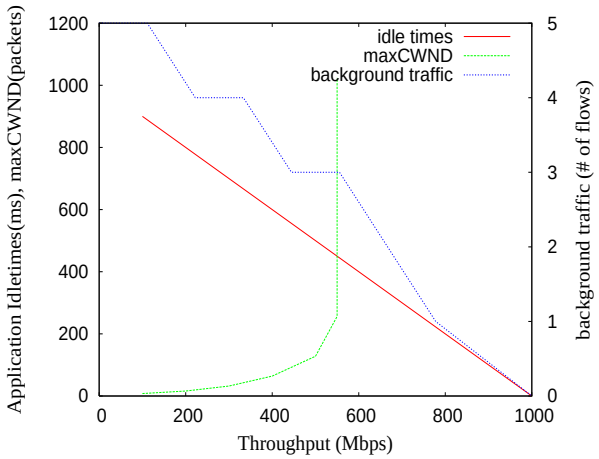

Figure 5: *Impact of each limitation on the achieved throughput*

4.2.4 Combined Bottleneck Effect

The traffic generation model used to estimate the e-VLBI flow has the following characteristics: "on/off" bursty data generation, initially with data bursts of 500 ms and idle times of 500 ms. To simulate the receiver limitation, we set the maximum CWND to 64 packets (0.06 Mbytes) and RWND to the 50 packets (0.05 Mbytes). The background traffic used is characterised by ten normal sized TCP flows from the reverse direction as the e-VLBI flow starting randomly with in the first 5 seconds as the flow, twenty five small TCP flows starting after the first 25 seconds flowing in the same direction as the e-VLBI flow, five small TCP flows starting after the 25 seconds and flowing in the opposite direction, one hundred and ten web sessions starting randomly during the flow, one hundred of them in the same direction and ten in the opposite direction as the e-VLBI flow.

5 Conclusions & Future Work

In this paper we present a data generation model that approximates an e-VLBI data flow by comparing results of a real flow against those of a simulation. Our model shows that the e-VLBI flow generated data in a bursty pattern i.e. large bursts separated by large idle periods. Other factors seen to affect the flow include receiver limitation and background traffic.
Future work will include designing data generation models for the other commonly used Mark5 transfer models such as In2Net-Net2Out, In2Net-Net2Disk, e.t.c... There is also need to lookout for wide spread data traces on which to base data generation models in order to avoid biasness in models due to local network conditions such as hardware and local usage patterns. Finally since the main goal of this work is to improve e-VLBI transport, models that eliminate or shorten the idle time between data bursts during the lifetime of an e-VLBI data flow are to be explored.

References

[1] L. Breslau, D. Estrin, S. Fall, J. Heidemann, P. Helmy, H. S. McCanne, K. Varadhan, Y. Xu, and H. Yu. Advances in network simulation. *IEEE Computer*, 55(5):59–67, 2000.

[2] The ns2 simulation. www.isi.edu/nsnam/ns.

[3] Mark5 vlbi data system. Haystack observatory Website. http://web.haystack.mit.edu/mark5/.

[4] J. Sansa, A. Szomoru, and J.M. van der Hulst. On network measurement and monitoring of end-to-end paths used for e-vlbi. *2nd Annual International Conference on Sustainable ICT Capacity in Developing Countries*, 2006.

[5] Gary R. Wright and W. Richard Stevens. *Tcp/Ip Illustrated: The Protocols*, volume 1. Addison-Wesley, 1994.

[6] M. Allman, Paxson V., and W. Stevens. Tcp congestion control. RFC 2581, Internet Engineering Task Force, 1999.

[7] S. Floyd. Highspeed tcp for large congestion windows. RFC 3649, Internet Engineering Task Force, 2003.

[8] R. Caceres, P. Danzig, S. Jamin, and D. Mitzel. Characteristics of wide-area tcp/ip conversations. *Proceedings of ACM SIGCOMM*, pages 101–112, 1991.

[9] P. Danzig and S. Jamin. tcplib: A library of tcp internetwork traffic characteristics. *USC Technical Report USC-CS-91-495*, 1991.

[10] P. Danzig, S. Jamin, R. Caceres, D. Mitzel, and D. Estrin. An empirical workload model for driving wide-area tcp/ip network simulations. *Internetworking: Research and Experience*, 3(1):1–26, 1992.

[11] V. Paxson. Empirically derived analytical models of wide-area tcp connections. *IEEE/ACM Transactions on Networking*, 2(4):316–336, 1994.

[12] V. Paxson and S. Floyd. Wide-area traffic: The failure of poisson modelling. *IEEE/ACM Transactions on Networking*, 3(3):226–244, 1995.

[13] B. Mah. An empirical model of http network traffic. *Proceedings of IEEE INFOCOM*, 1997.

[14] P. Barford and M. E. Crovella. Generating representative web workloads for network and server performance evaluation. *Proceedings of ACM SIGMETRICS*, pages 151–160, 1998.

[15] P. Barford, A. Bestavros, A. Bradley, and M. E. Crovella. Changes in web client access patterns: Characteristics and caching implication. *World Wide Web, Special Issue of Characterisation and Performance Evaluation*, 2:15–28, 1999.

[16] M. Crovella and A. Bestavros. Self-similarity in world wide web traffic: Evidence and possible causes. *IEEE/ACM Transactions on Networking*, 5(6):835–846, 1997.

[17] Cunha C. R., A. Bestavros, and M. E. Crovella. Characteristics of www client-based traces, technical report. *Technical Report TR-95-010 Boston University Computer Science Department*, 1995.

[18] F. D. Smith, F. H. Campos, K. Jeffay, and D. Ott. What tcp/ip protocols headers can tell us about the web. *Proceedings of ACM SIGMETRICS*, 2001.

[19] F. D. Smith, F. H. Campos, K. Jeffay, and D. Ott. Methodology for developing empirical models of tcp-based applications. *Work In Progress*, 2001. www.cs.unc.edu/ fhernand/publications.html.

[20] The open optical internet exchange in amsterdam. Netherlight Website. www.netherlight.net.

[21] J. Mahdavi, M. Mathis, and R. Reddy. Enabling high performace data transfers (system specific notes for system administrators and privileged users). Pittsburgh Supercomputing Center Website. www.psc.edu/networking/projects/tcptune.

[22] Effect of txqueuelen on high bandwidth delay product network (datatag). Website. www.hep.ucl.ac.uk/ ytl/tcpip/linux/txqueuelen/datatag-tcp/.

[23] The tcpdump public repository. www.tcpdump.org.

[24] D. M. Nicol. Scalablity of network simulators revisited. *Proceedings of the Communication Networks and Distributed Systems Modeling and Simulation Conference*, February 2003.

[25] M. Handley, J. Padhye, and S. Floyd. Tcp congestion window validation. RFC 2861, Internet Engineering Task Force, 2000.

[26] M. Siekkinen, G. Urvoy-Keller, E. W. Biersack, and T. En-Najjary. Root cause analysis for long-lived tcp connections. *Proceedings of the ACM Conference on Emerging Networking Experiments and technologies*, 2005.

Biography

Julianne Sansa holds a BSc. (Maths, Computer Science) and MSc. (Computer Science) from Makerere University. Since 2001 she has worked with the Faculty of Computing and IT of Makerere University in various capacities. She is currently registered as a Ph.D. Student at the University of Groningen, in the Netherlands and her research interests are Quality of Service and Network protocols.

Arpad Szomoru obtained a PhD in astronomy at the University of Groningen, the Netherlands. He has worked at JIVE, the Joint Institute for VLBI in Europe, for several years as a senior software scientist. During this period he was heavily involved in the deployment and integration of disk-based recording systems and the early development of e-VLBI. Since 2006 he is the head of data processor research and development at JIVE.

Thijs van der Hulst got his PhD in Groningen in 1977 on a study of neutral hydrogen emission from interacting galaxies. He spent 5 years in the USA at the National Radio Astronomy Observatory and the University of Minnesota before returning to the Netherlands, where he joined the staff of ASTRON, the Netherlands Foundation for Radio Astronomy. In 1982 he moved to the Kapteyn Astronomical Institute of the University of Groningen, of which he now is the director. His main field of interest is structure and evolution of galaxies