Measuring Transient Performance of a Multistage Interconnection Network Using Ethernet Networking Equipment*

Marcus Brenner, Dietmar Tutsch and Günter Hommel
Technische Universität Berlin
Real–Time Systems and Robotics
D–10587 Berlin, Germany

Keywords: Multistage interconnection networks, transient measurement, self-similar network traffic, Ethernet

Abstract

Multistage interconnection networks (Banyan networks, MINs) are frequently proposed as connections in multiprocessor systems or in high-bandwidth network switches.

Using off-the-shelf Ethernet networking components, a small multistage interconnection network was re-created and several measurements concerning delay times and throughput were conducted on this setup using Poisson as well as self-similar input traffic. The analysis of the transient measurements reveals a dependency of the delays' variance on the offered input load.

1 Introduction

Multistage Interconnection Networks (MINs) are an efficient implementation of packet switching networks. Because MINs require less switching elements compared to a fully meshed crossbar switch, it is possible to create very large networks (e. g. 1024×1024) at low cost.

Areas of application for MINs lie in multiprocessor systems [1] or high-bandwidth communication networks. Internal buffering greatly increases the performance of a MIN. In [5] Dias and Jump developed an analytical model to cope with buffered MINs. Multiple extensions have been made to this model: Yoon, Lee and Liu [17] allowed the MIN to have arbitrary buffer length as well as arbitrarily sized switching elements. Theimer, Rathgeb, and Huber [10] refined Jenq's model to take into account the dependency between two successive clock cycles, thereby increasing the model's accuracy. Cut-through switching was taken into account by Widjaja, Leon–Garcia, and Mouftah [15], and by Boura and Das [2]. Tutsch and Hommel [13, 12] extended this model to include packet multicasting (while routing). Transient analysis of MINs was also performed by this group [11] using multifractal input traffic [14] showing that MINs retain the multifractal nature of the offered traffic if the network is not saturated. In this paper, measurements are only conducted using unicast traffic due to restrictions in the experimental setup.

In many modern telecommunication system scenarios, self-similar traffic models better reflect real-world measurements than Poisson-based models do. This is especially true in the context of packet oriented network traffic both in

*this research is supported by Deutsche Forschungsgemeinschaft (DFG) under Grant Numbers *Ho 1257/14-2* and *Ho 1257/16-2*

LAN (Ethernet [6]) or WAN (Internet) environments. Measurements in these areas clearly show a very high degree of traffic "burstiness", even on a medium to large-scale aggregation level. This lack of smoothing out over an increasing timescale is a property also found in heavy-tailed distributions, which were utilized in this work to generate the input traffic for the multistage interconnection network.

The remainder of this paper is structured as follows: In section 2 the kind of multistage interconnection networks we examine are described in further detail. Section 3 presents a short introduction to self-similar traffic and methods of its examination. Section 4 shows the measurement setup that was used to re-create a small MIN using an Ethernet switch in a VLAN configuration as the main component. In section 5 we present the results obtained by measuring delay times and throughputs for two different traffic models. Finally, section 6 presents a conclusion and shows directions for further studies.

2 Multistage Interconnection Networks

Multistage interconnection networks are composed of several small-sized switching elements (SEs) that are arranged in stages. Figure 1 shows a (Banyan property) $N \times N$ MIN consisting of $c \times c$ switching elements that are arranged in $n = \log_c N$ stages. All SEs forward packets synchronously according to an internal clock. Each switching element has a buffer (FIFO discipline) attached to each of its c input ports in order to store packets that cannot be sent forward due to an output port conflict or a full target buffer.

Packets entering the network at an input port are routed according to their target address information.

3 Self-similar Traffic

An object or a phenomenon is being described as *self-similar* if it appears roughly the same at different levels of magnification. Fractals are a well-known example of self-similar objects in geometry. This very informal definition is usually

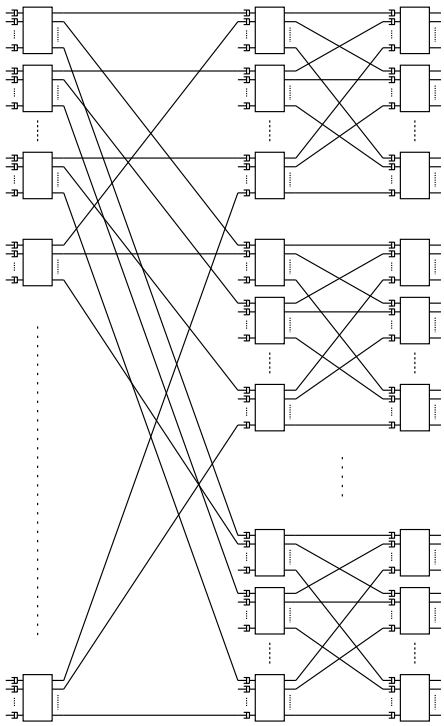

Figure 1: structure of a MIN consisting of $c \times c$ SEs

formalized as follows to describe a self-similar signal $\boldsymbol{x} = \{x_0, x_1, \ldots\}$, e. g. network traffic (time-discrete case).

The m-aggregated time series $\boldsymbol{x}^{(m)} = \{x_0^{(m)}, x_1^{(m)}, \ldots\}$ is the summation of the original series x over non-overlapping intervals of block size m:

$$x_k^{(m)} = \frac{\sum_{i=km-(m-1)}^{km} x_i}{m}$$

With m becoming larger, $\boldsymbol{x}^m$ represents an increasing compression of the original signal's timescale. Comparing the variances of both $\boldsymbol{x}$ and $\boldsymbol{x}^{(m)}$ for different values of m it can be shown that for a self-similar process, $\mathrm{Var}[\boldsymbol{x}^{(m)}]$ decays more slowly than at a rate of $\frac{1}{m}$ (see [9] for a more detailed explanation):

$$\mathrm{Var}\left[x^{(m)}\right] = \frac{\mathrm{Var}[x]}{m^\beta}$$

Thus, β is a measure for the degree of self-similarity of the signal. To detect self-

similar properties in data samples, a commonly used method is to plot the variance of the m-aggregated time-series over various levels of aggregation m on a logarithmic scale. By estimating the slope of the line of regression, β and therefore the degree of self-similarity is determined.

Self-similarity in network data traffic was observed by a number of researchers. Leland, Taqqu, Willinger and Wilson discovered this characteristic behavior when studying Ethernet local networking traces [6]. In [16] the same authors presented a detailed explanation of long-range dependence in LAN traffic and suggested a generation method for self-similar data traffic using ON/OFF sources. A similar study for wide area network traffic was conducted by Crovella and Bestavros [3] who analyzed WWW traces which were gathered over a period of several months.

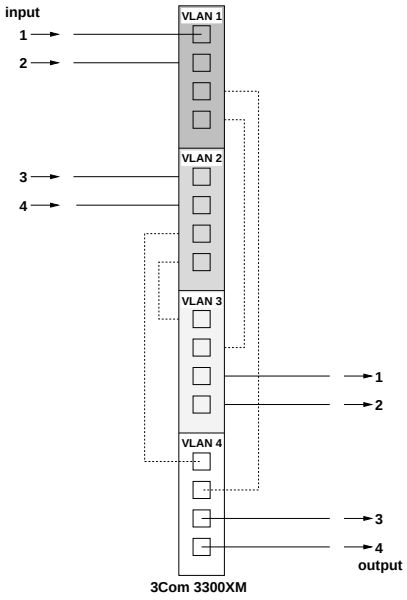

Figure 2: re-creating a MIN-structure using an Ethernet switch

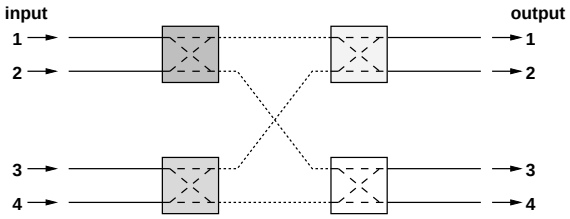

Figure 3: 4×4 MIN consisting of 2×2 SEs

4 Measurement Setup

A single, manageable 24-port Ethernet switch was used to re-create the multistage structure of a 4×4 multistage interconnection network (MIN), which is composed of 2×2 switching elements. Several additional steps had to be taken to emulate the behavior of a MIN as closely as possible:

By partitioning the switch into virtual LANs (VLANs) the required four 2×2 switching elements were represented. VLANs limit the broadcast domain in a similar way switches limit the collision domain for Ethernet frames: no traffic can cross a VLAN boundary within the switch. The four VLANs were connected externally (using patch cables) to implement the inner connections of the switching elements as illustrated in Figures 2 and 3.

To eliminate any ambiguities in path finding, the switch's automatic learning mode for Ethernet addresses (MAC addresses) was disabled, instead it was programmed manually with the MAC addresses of all network interfaces. Furthermore, the computers representing the end stations had permanent entries put in their ARP (address resolution protocol) tables to ensure that no discovery of Ethernet (MAC) addresses (via ARP) will occur. Allowing ARP traffic would involve sending Ethernet broadcasts to all communication endpoints and thus would interfere with the measurements.[†]

To avoid any interference between data traffic and acknowledgment packets, which are generated when using the TCP/IP protocol, sending and receiving was performed by separate network interfaces. In order to reduce protocol stack overhead, a C-Library was developed that allows to send Ethernet frames directly into the network, without being restricted to only use the TCP/IP

[†]this only applies to measurements using the Internet Protocol (IP), which requires IP address to MAC address translation

stack. The measurements presented in this paper were conducted exclusively using this library.

Four standard PCs performed the role of the sending and receiving stations. The measurements were conducted using the Linux operating system with the KURT real-time extension developed at the University of Kansas [4]. KURT contains a component *utime* that increases the Linux kernel's temporal granularity (normally 10 ms with Linux on Intel CPUs) to a level necessary for generating self-similar traffic over a large time scale.

The major drawback of this approach is the inability to control the individual switching elements' (each SE is represented by a virtual LAN) buffer distribution (shared buffer for all SEs vs. individual buffers), size and frame forwarding strategy. The latter was pre-set to store and forward routing in this case, because of the switch's frame forwarding capabilities.

To evaluate the network's performance, transient delay measurements were conducted using Poisson traffic as well as Pareto traffic generators. The latter allowed to study the behavior under self-similar traffic conditions. The Pareto distribution (with probability density function $f(x) = \frac{\alpha}{c}\left(\frac{c}{x}\right)^{\alpha+1}$, depicted in Figure 4) has the property of slowly decaying variances over time aggregation and was used to provide the input traffic for the self-similar traffic measurements.

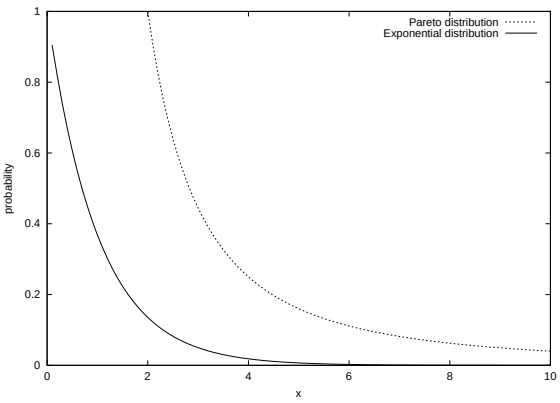

Figure 4: probability density functions for Pareto and exponential distributions

After synchronizing the senders' and receivers' clocks in a manner similar to the Network Time Protocol NTP [8] (measuring the propagation delay between client and server and assuming statistically equal propagation in each direction), traffic was generated on the sender's network interface and the transit time for each Ethernet frame received was recorded at the receiver for offline measurement evaluation. Experiments showed that clock skew of the PCs timers was low enough to allow for measurement runs of approximately 30 minutes without disturbing the delay time measurements.

Realization of the different traffic models (heavy-tailed and memoryless distributed network input load) was performed by modulating the time between sending of two successive Ethernet frames as well as their size. To examine if the output traffic of the multistage network shows again self-similar characteristics, variance-time plots (using the sample variance) were conducted to estimate the Hurst parameter.

5 Results

To evaluate the delay time measurements, variance was plotted over increasing intervals of aggregation m on a logarithmic scale. Self-similarity can be perceived as non- (or slowly) decaying variance over time aggregation.

One can see from Figure 5, that the multistage structure does not change the self-similar characteristics of the input traffic as the slope β of the line estimates to approximately -0.4.

β is directly related to the Hurst parameter H, which is given by $H = 1 - \frac{\beta}{2}$.

For comparison, Figure 6 shows the variance-time plot when Poisson generated traffic is offered to the network. Here, the output traffic measurement does not exhibit self-similar properties as the Hurst parameter was estimated to about 0.5.

In both cases, there is a considerable decrease in the absolute amout of the delays' variances as the load offered to the network increases. This is due to the filling up of the queues in front of the switching elements and therefore the decreasing probability of a packet to traverse a SE with little

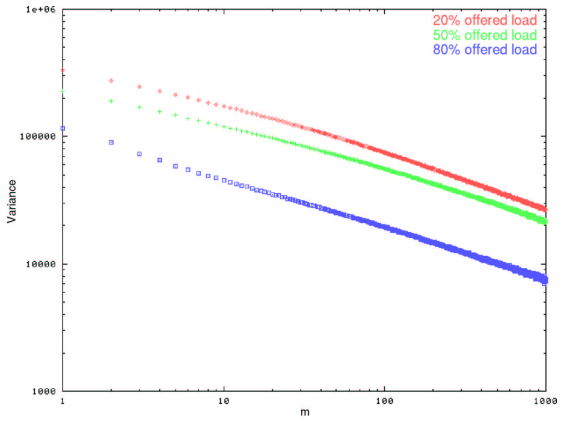

Figure 5: output traffic variance (Pareto case)

or no waiting time. All packets are delayed by roughly the same amount of time. Since the measurement setup cannot provide support for multicast traffic, the self-similarity does not decay completely because network saturation does not provide for all buffers to fill up entirely.

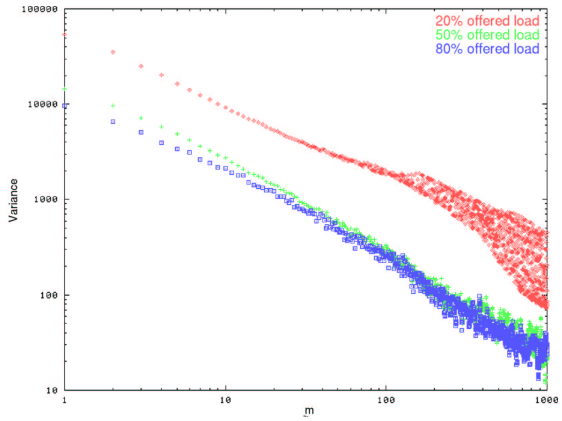

Figure 6: output traffic variance (Poisson case)

When comparing the output traffic for the different packet generators (Pareto and Poisson distributed) it should be noted that the variance in case of Poisson input traffic is about one order of magnitude less than in case of self-similar traffic due to the heavy-tailed property of the Pareto distribution.

6 Conclusion and Outlook

In this paper we have described an experimental setup to re-create a 4×4 multistage interconnection network using standard Ethernet equipment. The VLAN feature of a single Ethernet switch was used to provide for several small switching elements without requiring additional hardware. The experiments conducted consisted of measuring packet delay times and throughputs. Using this setup, one can obtain results for arbitrarily distributed input traffic in a short amount of time, thus allowing for validation of results obtained by analytical or simulative means.

In addition to the traditional approach of Poisson traffic modeling self-similar input traffic was considered as well. Even under high load conditions, the examined multistage network retains the input traffic's self-similar properties, while reducing the absolute amount of variance due to saturation.

The first step in expanding this research would be enlarging the number of input and output ports of the network and to examine differently sized switching elements (3×3, 4×4 SEs). The approach presented here appears scalable to re-create larger networks.

When dealing with MINs, another important area of interest is packet multicasting while routing, i. e. copying a packet destined for multiple output ports inside the switching elements instead of inserting it repeatedly into the network. This aspect was not considered in the work presented here because it would have involved changing the switch's internal programming. Whether or not the registration of multicast groups (GARP[‡] Multicast Registration Protocol, GMRP) in a VLAN context as defined in [7] can be used to emulate the kind of multicasting described above is yet to be determined.

References

[1] Gheith A. Abandah and Edward S. Davidson. Modeling the communication performance of the IBM SP2. In *Proceedings of the 10th International Parallel Processing*

[‡]Generic Attribute Registration Protocol

Symposium (IPPS'96); Hawaii. IEEE Computer Society Press, 1996.

[2] Younes M. Boura and Chita R. Das. Performance analysis of buffering schemes in wormhole routers. *IEEE Transactions on Computers*, 46(6):687–694, June 1997.

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

[4] Sean B. House D. Niehaus, Winliam Dinkel. Effective real-time system implementation with KURT linux. Technical report, Information and Telecommunication Technology Center, Electrical Engineering and Computer Science Department, University of Kansas, 1999.

[5] Daniel M. Dias and J. Robert Jump. Analysis and simulation of buffered delta networks. *IEEE Transactions on Computers*, C–30(4):273–282, April 1981.

[6] Will Leland, Murad Taqqu, Walter Willinger, and Daniel Wilson. On the self-similar nature of ethernet traffic (extended version). *IEEE/ACM Transactions on Networking*, 2(1):1–15, February 1994.

[7] W. Lidinsky. IEEE draft standard P802.1Q/D11: Virtual bridged local area networks, July 1998.

[8] David L. Mills. RFC 1305: Network time protocol (version 3), specification, implementation and analysis, March 1992.

[9] William Stallings. *High-Speed Networks*. Prentice Hall, 1998.

[10] Thomas H. Theimer, Erwin P. Rathgeb, and Manfred N. Huber. Performance analysis of buffered banyan networks. *IEEE Transactions on Communications*, 39(2):269–277, February 1991.

[11] Dietmar Tutsch. Performance analysis of transient network behavior in case of packet multicasting. In *Proceedings of the 11th European Simulation Symposium 1999 (ESS'99)*, pages 630–634. SCS, 1999.

[12] Dietmar Tutsch and Günter Hommel. Analysis of multicasting in buffered multistage interconnection networks. Technical Report FB Informatik 1997–20, Technische Universität Berlin, 1997.

[13] Dietmar Tutsch and Günter Hommel. Performance of buffered multistage interconnection networks in case of packet multicasting. In *Proceedings of the 1997 Conference on Advances in Parallel and Distributed Computing (APDC'97); Shanghai*, pages 50–57. IEEE Computer Society Press, March 1997.

[14] Dietmar Tutsch and Günter Hommel. Multifractal multicast traffic in multistage interconnection networks. In *Proceedings of the Proceedings of the High Performance Computing Symposium 2001 (HPC 2001)*, pages 257–262. SCS, 2001.

[15] Indra Widjaja, Alberto Leon-Garcia, and H.T. Mouftah. The effect of cut-through switching on the performance of buffered banyan networks. *Computer Networks and ISDN Systems*, 26:139–159, 1993.

[16] Walter Willinger, Murad Taqqu, Robert Sherman, and Daniel Wilson. Self-similarity through high-variability: Statistical analysis of ethernet LAN traffic at the source level. *IEEE/ACM Transactions on Networking*, 5, 1997.

[17] Hyunsoo Yoon, Kyungsook Y. Lee, and Ming T. Liu. Performance analysis of multibuffered packet–switching networks in multiprocessor systems. *IEEE Transactions on Computers*, 39(3):319–327, March 1990.