Realization of a stable network flow with high performance communication in high bandwidth-delay product network

Y. Kodama*, T. Kudoh, O. Tatebe, S. Sekiguchi
Grid Technology Research Center,
National Institute of Advanced Industrial Science and Technology (AIST)
Tsukuba, Ibaraki 3058568, JAPAN

Abstract

It is important that the total bandwidth of the multiple streams should not exceed the network bandwidth in order to achieve a stable network flow with high performance in high bandwidth-delay product networks. Software pacing of TCP/IP for each stream sometimes exceeds the specified bandwidth, especially at the beginning of the stream or when buffer was overflow at the network router. We proposed the hardware control technique for total bandwidth of multiple streams with high accuracy.

GNET-1 is the hardware gigabit network testbed that we developed. It provides functions such as wide area network emulation, network instrumentation, and traffic generation at gigabit Ethernet wire speeds. GNET-1 is a powerful tool for developing network-aware grid software. It can control the total bandwidth of the multiple streams with high accuracy by adjusting the inter frame gap (IFG).

To see the effect of the highly accurate bandwidth control by GNET-1, the file exchange of large-scale data was done on a Trans-pacific Grid Datafarm testbed between Japan-U.S.. We used three trans-pacific networks, APAN/TransPAC Los Angels line and its Chicago line and SuperSINET New York line. Its total bandwidth that can be used was 3.9 Gbps. In this feasible study, GNET-1 controlled five gigabit Ethernet ports, and achieved the total bandwidth of 3.78 Gbps in stable for about one hour. The bandwidth was 97 % of the peak bandwidth of used networks.

INTRODUCTION

Multi gigabit networks have come to be used as wide area networks (WAN) in recent years. However, since there is a large delay on a WAN, it is difficult to effectively use such high bandwidth-delay product network from a single application. One reason is that the several parameters of standard TCP/IP are not adequate for such large delay and high bandwidth network. Several protocols, for example HighSpeed TCP[1], have been developed to solve this problem. But even in the HighSpeed TCP, if packets are lost, the bandwidth is reduced. It is the most important to make no packets lose on such long fat network.

We will describe the reason of the packet loss using a simple model. It assumes that three 500 Mbps streams are transferred over a 2.4 Gbps network in Figure 1. Since

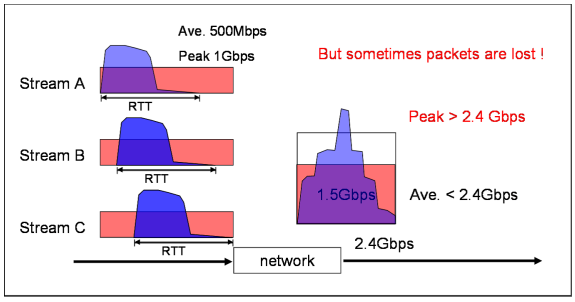

Figure 1: packet loss on long-fat network

the simple total bandwidth of streams is 1.5 Gbps and it is lower than the network capacity, packets would not be lost. However, sometimes packets are lost. It is because the stream bandwidth is not stable and it has the peak of 1 Gbps. When the peaks of streams conflict with each other, the total bandwidth becomes larger than the network capacity. It causes packet loss.

TCP/IP controls the output bandwidth by self clocking. In the self clocking, packets of TCP/IP are transmitted at the timing of receiving acknowledgments of the previous packets. Since the previous packets and the acknowledgement packets are transfered over the bottleneck line, they are paced by the bottleneck bandwidth. But self clocking is not always effective. For example, it is not effective at the beginning of streams or at buffer overflows on network routers.

We proposed the hardware smooth pacing method that always controls the bandwidth. Following this introduction, the remainder of the paper is organized as follows. Section 2 introduced the proposed method of smooth traffic shaping. We also described the hardware network emulator GNET-1, by which we implemented the technique. Section 3 described the experimental results in two cases. One is a network emulated by GNET-1, and another is the real network on transpacific. Section 4 summarized the paper.

SMOOTH TRAFFIC SHAPING

It is important that the total bandwidth of the multiple streams should not exceed the network bandwidth in order to achieve a stable network flow with high performance in high bandwidth-delay product networks. Software pacing of TCP/IP for each stream sometimes exceeds the specified bandwidth, especially at the beginning of the stream or

* y-kodama@aist.go.jp

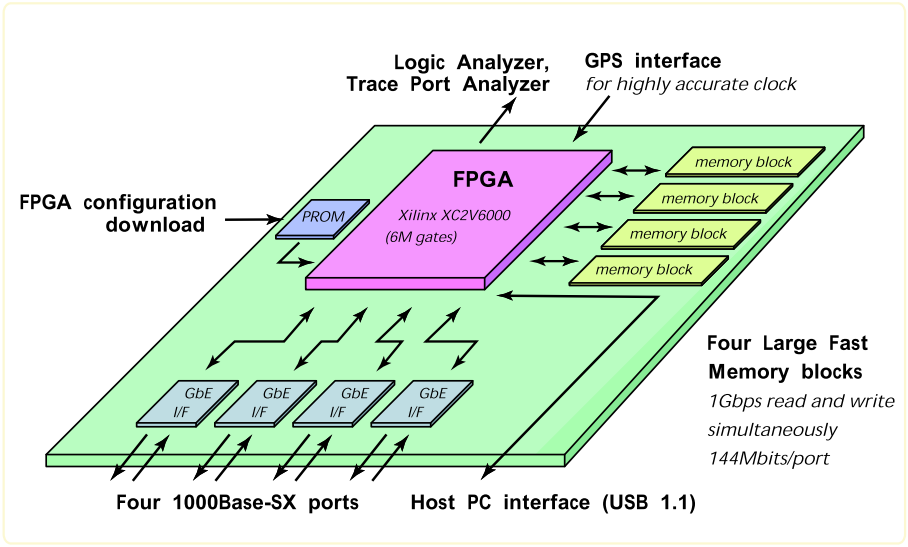

Figure 3: Block diagram of GNET-1

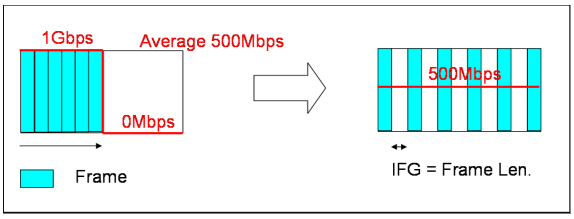

Figure 2: Smooth traffic shaping by adjusting inter frame gap(IFG)

when buffer overflow at the network router. We proposed the hardware pacing method that always control the bandwidth of each stream with high accuracy.

GNET-1 [3, 4] is a hardware network testbed we developed, and it can control the output bandwidth by adjusting the inter frame gap (IFG) to the specified bandwidth. If we want to limit the output bandwidth lower than m Mbps, the size of IFG is formulated by $len(1000/m - 1) - 8$, where len is the size of previous frame. Since GNET-1 calculates the size of IFG for every frame, it can control the output bandwidth even if the stream includes frames those lengths are different. It inserts the adjusted IFG between frames, the stream becomes smooth and the output bandwidth does not exceed specified bandwidth in any time interval longer than the frame interval.

In the simple model of the introduction, if GNET-1 controls the bandwidth of each stream to 500 Mbps, the total bandwidth is rigidly controlled to 1.5 Mbps. Since GNET-1 can guarantee that total bandwidth does not exceed specified bandwidth using smooth shaping, the transfer of each stream becomes stable.

GNET-1

GNET-1 is a fully programmable network testbed. It provides functions such as wide area network (WAN) emulation, network instrumentation, traffic shaping, and traffic generation at gigabit Ethernet wire speeds by programming the core FPGA. GNET-1 is a powerful tool for developing network-aware grid software. It is also a network monitoring and traffic-shaping tool that provides high-performance communication over WAN.

Figure 3 shows a block diagram of GNET-1. GNET-1 has four GBIC ports. Each port is connected to a large-scale FPGA. The FPGA is a Xilinx XC2V6000, which includes 76K logic cell, 2.5 Mbit memory, and 824 user I/O pins.

GNET-1 has four SRAM ports. Each SRAM has 144 Mbit capacity and can be simultaneously read-and-written with 1 Gbps speed. GNET-1 is connected to the control PC by USB 1.1. The control PC sets and gets the parameters of GNET-1. It also has two MICTOR ports for connecting a logic analyzer in order to observe internal signals, and a GPS (Global Positioning System) port in order to get the micro-second accurate absolute time.

The most attractive feature of GNET-1 is that its functions are reconfigurable. If you design the control circuit for the desired function and load it into the FPGA, you can use any function easily.

The circuits in the FPGA consist mainly of four GbE MACs, four FIFO memory controllers, and other control circuits. The GbE MACs run at a 125 MHz clock, the FIFO memory controllers run at a 62.5 MHz clock, and other circuits run at a 31.25 MHz clock. Since most functions are implemented to run at a 31.25 MHz clock, the timing of the user circuits is not tight. We designed the circuit in verilog hardware description language, and synthesized it using Xilinx ISE and XST. It takes about 1 hour to create the configuration data.

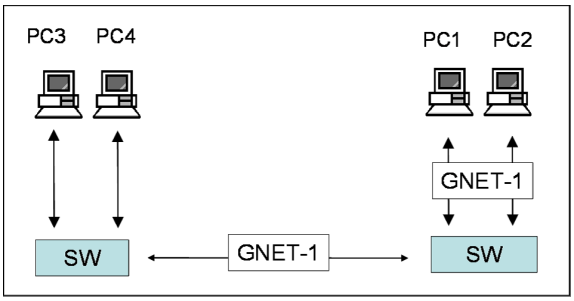

Figure 4: Environment of experiment

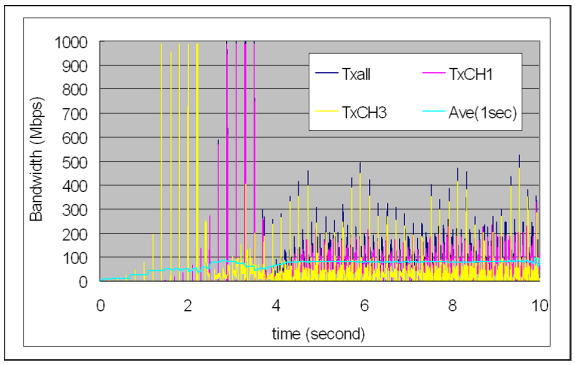

Figure 5: Bandwidth with no traffic shaping

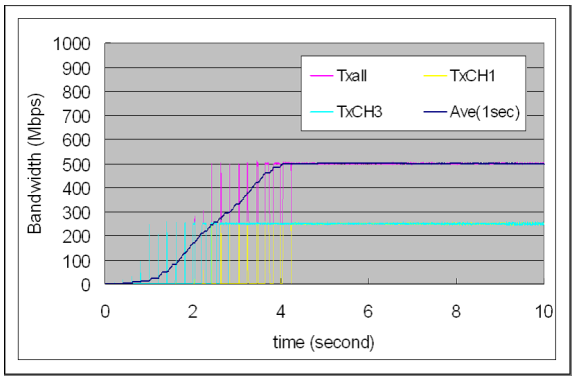

Figure 6: Bandwidth with traffic shaping

EXPERIMENTS

results on network emulated by GNET-1

Figure 4 shows the environment of the experiments. 2 PCs are connected to a network switch. GNET-1 emulates a network between two switches to evaluate the bandwidth of multiple streams on the bottleneck line. We compared the streams with no shaping and the streams with shaping by GNET-1. The bottleneck line has 100 millisecond one-way latency, 500 Mbps bandwidth, and 512 Kbytes buffer for each port.

We used desktop PCs which consists of Intel Pentium 4 2.8GHz, 1GB memory (DDR400), Intel D865GLC mother board and the on-board Intel 82547 GbE NIC. All PCs are running the Redhat Linux 9.0 and the Linux kernel 2.4.26 with the Web100[2] 2.3.8 patch. We used the iperf program with 8 Mbyte socket buffer to evaluate the bandwidth. We used the normal TCP/IP and enabled *WAD_IFQ* in Web100 in order to ignore send stalls at IFQ where it is the same effect with increasing IFQ size.

Figure 5 shows the results of two streams with no traffic shaping. The x-axis is the time, and y-axis is the bandwidth. The yellow and the pink line is the bandwidth of each stream, and dense blue line is a total bandwidth. These bandwidths are measured in 2 millisecond interval by GNET-1. The light blue line is the average bandwidth in 200 millisecond interval. The CH3 stream starts at time 0, and the CH1 stream starts at 1.3 second. The total bandwidth of two streams was less than 100 Mbps. Peaks of two streams conflicted at the bottleneck line, and packet buffer overflowed on it. It decreased the bandwidth of both streams.

Figure 6 shows the results of two streams with traffic shaping. The x-axis and the y-axis are the same as Figure 5. Each stream is limited to 250 Mbps by GNET-1. The CH3 stream starts at time 0, and it becomes stable at 2.8 seconds. The CH1 stream starts at 1.4 second, and became stable at 4.2 seconds. After 4.2 seconds, total bandwidth of two streams is stable and it is 500 Mbps.

results on Transpacific networks

We evaluated the effect of smooth traffic shaping on an actual wide area network. Figure 7 is the network that we used. It is the Trans-Pacific Grid Datafarm testbed between Japan and the U.S. constructed for Bandwidth Challenge of SC2003, an international conference held in Phoenix in November, 2003. Gfarm [5] is a grid file system we developed. It shares the distributed disks on a grid, and managements them. Gfarm can assign processes to the computers that have the data to be computed for scalability and locality. It also maintains the replica, the copy of data, for load balancing and dependability.

We used three transpacific networks. The APAN/TransPAC Los Angeles line (LA) is 2.4 Gbps, and the round trip latency is 141 milliseconds. The APAN/TransPAC Chicago line (Chicago) is a 644 Mbps ATM line, its peak bandwidth on Ethernet is about 500 Mbps, and the round trip latency is 250 milliseconds. The SuperSINET New York line (NY) is 4.8 Gbps. We used only 1 Gbps on it, and the round trip latency is 285 milliseconds. Total network bandwidth is 3.9 Gbps. We used three Gigabit Ethernet for the LA line, and called them LA1, LA2, and LA3, respectively. We used a Gigabit Ethernet each for the Chicago and NY lines. We controlled five Gigabit Ethernet by GNET-1.

We used 11 PCs on both ends. Each PC consists of dual Xeon 2.8GHz. We used HighSpeed TCP and jumbo packets whose MTU size is 6000 and enabled *WAD_IFQ* in Web100. We transferred the data from disks to disks from the U.S. site to the Japan sites using replica copy of Gfarm.

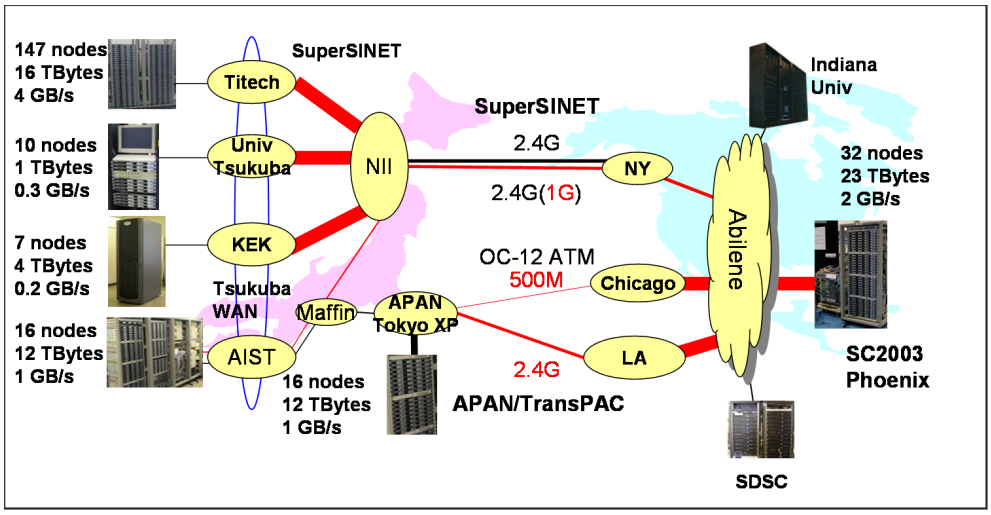

Figure 7: Transpacific Gfarm network testbed

Figure 8 shows the data transfer performance using Gfarm when the files were transferred using multiple streams between Japan and the U.S.. First, GNET-1 controlled the output bandwidth of LA1, LA2, LA3, Chicago and NY to 800 Mbps, 750 Mbps, 800 Mbps, 500 Mbps and 930 Mbps, respectively. But the LA line caused packet loss, and the bandwidth was unstable. So we decreased the LA1 bandwidth to 780 Mbps, and increased NY to 950 Mbps at 4 minutes. After that, the bandwidth was stable. We achieved 3.78 Gbps total bandwidth, that is, 97 % of the peak bandwidth, in a stable manner.

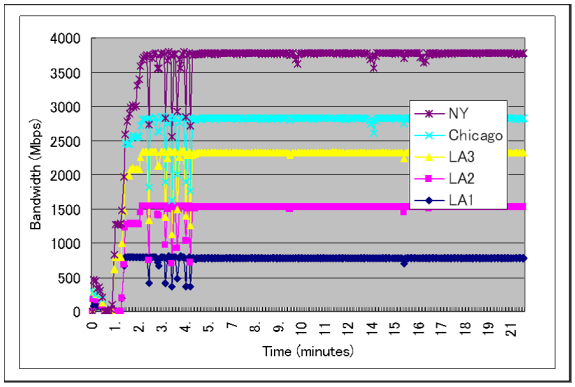

Figure 8: Total bandwidth on BWC'03

CONCLUSION

Smooth traffic shaping of GNET-1 realizes stable network traffic on a high bandwidth-delay product network. Currently the bandwidth value of shaping has to be defined and maintained by user manually. Automatic tuning of the bandwidth of each stream is a future challenge. We are also developing a software pacing method in network driver. We are now developing a new tool for 10GbE.

ACKNOWLEDGMENTS

This research was partially supported by an NEDO Grant-in-Aid for private sector fundamental technology "Research on Large-Scale and Reliable Servers."

REFERENCES

[1] S. Floyd. "HighSpeed TCP for large congestion windows," In Internet draft, draft-floyd-tcp-highspeed-02.txt, 2003. http://www.icir.org/ floyd/hstcp.html.

[2] M. Mathis, J. Heffner and R. Reddy, "Web100: Extended TCP Instrumentation for Research, Education and Diagnosis," ACM Computer Communications Review, Vol.33, No.3, July 2003.

[3] http://www.gtrc.aist.go.jp/gnet/.

[4] Y. Kodama, T. Kudoh, R. Takano, H. Sato, O. Tatebe and S. Sekiguchi, "GNET-1: Gigabit Ethernet Network Testbed," Proceedings of 2004 IEEE International Conference on Cluster Computing (Cluster2004), pp.185-192, 2004.

[5] O. Tatebe, Y. Morita, S. Matsuoka, N. Soda and S. Sekiguchi, "Grid Datafarm Architecture for Petascale Data Intensive Computing," Proceedings of the 2nd IEEE/ACM International Symposium on Cluster Computing and the Grid (CCGrid 2002), pp.102-110, 2002.