The Research Internet Gateways

David J. Iannucci and John Lekashman

Report RND-94-008 August 1994

NAS Systems Development Branch

NASA Ames Research Center
Moffett Field, California 94035
lekash@nas.nasa.gov

Abstract

(RIG) in experimental local and wide area network testbeds is reported. These RIGs are pre-production Internet routers, capable of forwarding data among multiple networks. Experimental tools for metering data transfers are described. This report documents delivered throughput, performance under traffic load, and failure modes of the units. Behavior of TCP/IP (Transmission Control Protocol/Internet Protocol) file transfers under terrestrial, satellite and multiple satellite data transfers are shown. Throughput for the multiple satellite case drops off radically, far more than to be expected from the delay. Further tests are conducted with production equipment, revealing that defects in the pre-production units are responsible for the loss. The pre-production units are suitable for encouraging industry in the development of technology, and networking algorithm experiments. They are not suitable for long term service use. This experience provides support for exercising caution when dealing with new technologies, and performing thorough metering of systems prior to service use.

1 Introduction

The Research Internet Gateway Program is a joint project between the Defense Advanced Projects Agency (DARPA), National Aeronautics and Space Administration (NASA), and Department of Energy (DoE). It was initiated in 1988. The following paragraph from the statement of work describes the expected results.

> DARPA's objective in this effort is to obtain a high performance packet switch/gateway (Research Internet Gateway (RIG)) that can be evaluated as the basis for a Defense Research Internet and can provide a highly flexible platform for experiments into routing, congestion management and network management issues that are critical to DARPA's long term networking goals. [1]

Three vendors were selected to produce units for testbed evaluation. Each supplied four RIG units. The three agencies each took responsibility for one testbed. The pairings were as follows:

- DARPA : Bolt, Beranek and Newman
- NASA : GTE Government Systems/Proteon Inc.
- DoE : SRI International/Cisco Systems

The contract was initiated in late 1988, and the units were delivered eighteen months later, in early 1990. This document addresses the NASA experience with the GTE/Proteon RIGs. This testbed was deployed at the Numerical Aerodynamic Simulation (NAS) Program at NASA Ames Research Center.

2 The Evaluation

There are two important considerations in the evaluation of the RIG units. First is their applicability for use in a wide area network service environment. There are a number of aspects to this, both technical and non-technical. The hardware and software must be stable, i.e., perform gateway functions for extended time without constant attention. The units must deliver performance, forwarding data traffic at rates matched to the attached networks.

The company producing the equipment must continue to develop the technology, repairing software problems as well as enhancing functionality and performance.

Second is their applicability for use in network experiments. Different factors come into play. A complete software source code build environment is necessary. Stability is a factor, but not as critical. The vendor future commitment is less important, since interest is in the experimental results, rather than technology advances.

The approach to this evaluation consisted of three steps. First, initial lab tests were conducted to verify functionality. Second, the RIGs were deployed in a live wide area network testbed. An experiment in satellite file transfer was conducted, desgined to stress test the units, as well as advance the utility of file transfer capability. Third, the RIGs were returned to the lab for close examination of the problems that occurred during the live testbed.

2.1 RIG Hardware Requirements

- Provide interfaces to IEEE 802.3 (ethernet) and DS1 (serial line) networks.
- Provide from 0 to 8 ethernet interfaces and 0 to 8 serial lines simultaneously.
- Provide an RS232 physical console access port.
- Use modular hardware design. New processors, memory or interface upgrades can be installed without a major overhaul.

The RIG is composed of a VMEbus backplane in a standard 19 inch rackmount cabinet. It has an Ironics IV-9001 Single Board CPU processor with an AMD29000 RISC CPU, and a one megabyte daughter memory board (IV-9102).

The chassis can hold up to eight serial interfaces, SBE VCOM4 cards. It can hold eight ethernet interfaces, SBE VLAN-E ethernet cards. It is also capable of using Proteon Pronet token ring cards, however, these were not evaluated in the testbed. These are all off-the-shelf products. There is one RS232 serial console port.

The delivered units met all the hardware requirements.

2.2 RIG Software Requirements

- Comply with Requirements for Internet Gateways (RFC 1009).
- Interoperate with current versions of TCP/IP (RFCs 791 and 793).
- Configure without recompiling software or changing any hardware switches.
- Provide an "inter-RIG" protocol for coordinating routing information.
- Provide a method for downloading the software via the network.
- Provide access controls (packet filters).
- Provide traffic and error statistics.
- Provide a logical (software) console access port for telnet access over a network.

The RIG software is a port of Proteon's P4200 C gateway source to the AMD29000. It is compliant with RFC 1009. It correctly meets the TCP and IP protocol specification. All configuration can be done via remote login to the units. The Open Shortest Path First (OSPF) routing protocol is used for inter-RIG routing. The units download their software via the standard Trivial File Transfer Protocol (TFTP). Access controls and statistics are available from the units. This software met all the requirements, with two minor exceptions. Two network statistics provided by the RIG were incorrect:

- The number of bad network and subnet addresses observed
- The number of packets discarded through filters.

The value remained 0 when it should have incremented.

2.3 Laboratory Testing

First stage testing took place in the NAS Facility Long Haul Communications laboratory. Verification of functionality and measurement of performance ability were conducted. The original RIG design and fabrication requirements for RIG units is described in the Statement of Work (SOW) [1]. That document is a procurement specification, and as such contains much information not relevant to the system functionality. The salient characteristics of interest in this evaluation are listed below.

2.4 Lab Equipment

The following pieces of equipment were used:

- Two Sun 3/260 host workstations running SunOS UNIX
- One TTC satellite digital delay circuit simulator (T1)
- One TTC satellite digital bit-error circuit simulator (T1)
- One 56 Kbps local circuit
- Two Network General Sniffer protocol analyzers
- One Excelan LANalyzer protocol analyzer
- Multi-port ethernet transceiver boxes (TCL, DELNI)
- Ethernet, V.35, and RS232C cables

2.5 Performance

The following are the performance requirements for the RIGs:

- Provide throughput of 1500 user data packets/second
- Provide throughput up to 10 Mbits/second
- Be capable of evolving to the level of 10,000 packets/second

Table 1 shows the performance measurements made in the Long Haul Communications Laboratory. The table shows the maximum number of packets per second forwarded by the RIG without loss. The testbed configuration consisted of one RIG passing packets between two ethernets. Sniffer[1] network analyzers were used to source the packet stream and count the packets forwarded. Inter Packet Gap (*IPG*) is the time period between packets that the RIG could handle without loss. It is measured in microseconds. Forty μs is the smallest IPG that our test equipment could generate. The test data indicate that for packets of size greater than 512 the RIG could sustain a 100% forwarding rate at IPG of much less than $40\mu s$. The fourth column shows the same performance figures when several filters (access controls) are enabled.

Size	IPG	PPS no filters	PPS w/filters	Maximum
64	290	2940	2857	14880
128	240	2940	2778	8445
512	40	2222	2222	2349
1024	40	1162	1162	1197
1500	40	806	806	812

Table 1: Protocol Analyzer Throughput (Packets/sec)

Note: While the RIG was busy forwarding packets at its fastest rate, all console access was frozen. This behavior will present some difficulty in managing a system under load.

2.6 Wide-Area Testbed Activities

For the evaluation, the units were deployed in a live wide area network. Host computer systems were attached, and an experiment in file transfer conducted. The following reports the results.

2.6.1 The Experiment

Network technology is pushing a limit in TCP. TCP uses a flow control window that has a maximum value of sixty-four thousand octets for the product of bandwidth and delay over the network. One solution adopted by NAS is

[1]Sniffer is a trademark of Network General Corporation

outlined below; it involves doing data striping with multiple simultaneous TCP streams. More information and background may be found in [2]. Jacobson and Braden describe another method, with modifications to the TCP protocol to handle this problem, in RFCs 1072 and 1185 [3,4].

The goals of the wide-area RIG experiment are:

1. To determine the effectiveness of the RIG units when placed into a live service environment.

2. To determine the effectiveness of using multiple transport connections per data transfer to utilize a high bandwidth delay product network.

3. To compare the effectiveness of the above scheme (multiple transport connections with relatively small window sizes) with a single connection using a large window for achieving the same data transfer.

4. To determine the effectiveness of "type-of-service" routing in enhancing this utility. Specifically, to measure the performance effects of sending data acknowledgments via an out-of-band, low-delay path.

2.6.2 The Testbed

See Figure 1 for the wide-area testbed. A RIG was installed at each of four NASA centers. They were joined in a circular topology, with each pair connected by two circuits in parallel: a 1.544 Mbps (T1) satellite and a 56 Kbps terrestrial line. At each NASA center were one or two Sun workstations on an ethernet. The ethernet connection is labelled MPT. The workstations were running SunOS 4.0 or higher, including slow-start TCP. Each of the RIG units had an out-of-band console access connection through the X.25-based NASA Packet Switching System (NPSS).

The concept of type of service explored here is to route large data packets over the high bandwidth, high delay connection. Data acknowledgement flows back over the low bandwidth, low delay path. Data acknowledgement is relatively low bandwidth, so can therefore exploit the better delay characteristics, and speed the effective transfer.

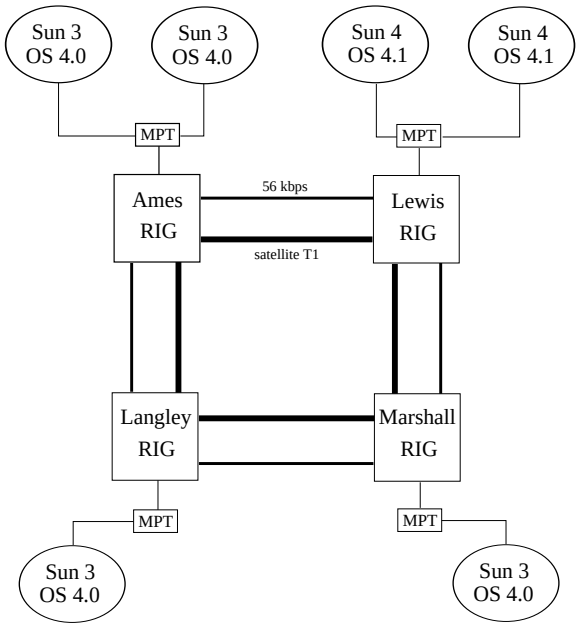


Figure 1

2.6.3 Tools

The primary tool used in the experiment was a program called *mftp*. It is a variation of the standard ftp[5]. It can open multiple parallel TCP connections for data transfer, rather than being limited to one. The number of connections is specified by the user. This program can be used to adjust the effective transport window size for the entire transfer by adjusting the *number* of individual fixed-size windows. There is one such window per connection.

The basic paradigm for the experiment consists of a file transfer between two endpoints separated by a network consisting of one or more serial "hops" through the RIG routers. All data-carrying packets passed over the satellite circuits. In the type-of-service-routed cases, acknowledgement packets were returned via the low-delay terrestial 56 Kbps circuit. The parameters varied in the experiment were:

1. Number of mftp connections
2. TCP send/receive window size
3. End-to-end delay.

The transfer file size was held constant at ten megabytes. This value is large enough to eliminate startup and internal computer memory buffering effects. The quantity of interest in all cases was delivered throughput. (See figures 2, 3, 6, and 8) Further data were taken in the form of packet traces using Van Jacobson's *tcpdump* program. The traces allowed to visualization of TCP behavior by charting the number of packets sent and acknowledged as a function of time. (See figures 4,5,7, and 9).

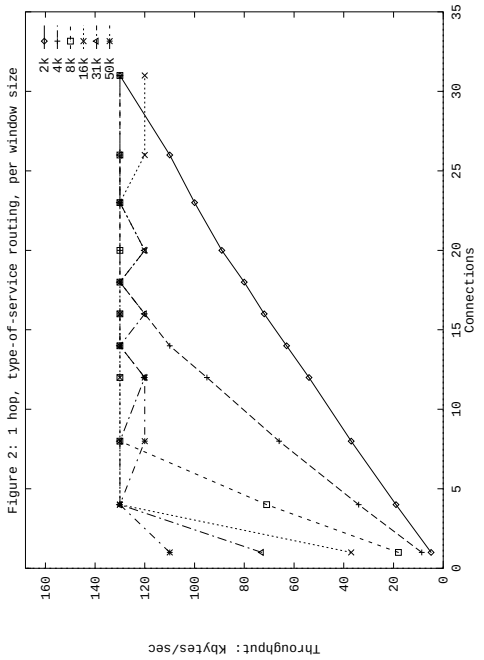
2.7 Results

Live metering of the testbed produced some expected and unexpected results. Increasing the number of connections did lead to increases in delivered throughput, as expected. However, there are some unstable areas, where the delivered throughput does not match the expected result. These are traced to faults in the RIG hardware used in the testbed.

2.7.1 Expected Results

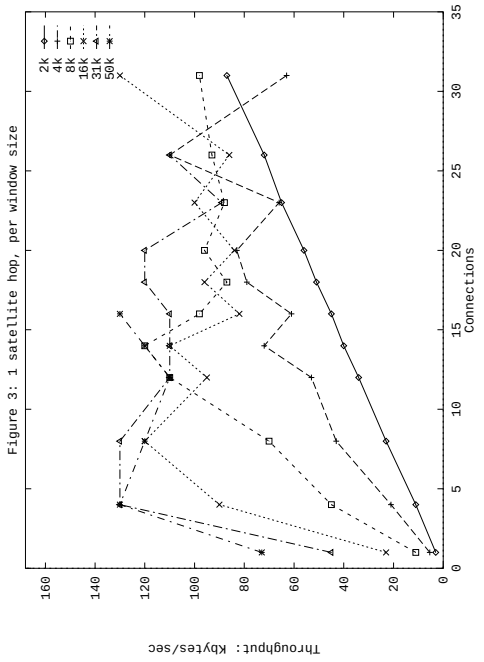
1. Figure 2 presents a family of curves, each line representing the window size of an individual connection. The number of connections at each window size is increased, and the corresponding delivered throughput measured.

   Throughput increases almost linearly in the number of connections used as long as the total effective window size (TEWS) is less than the bandwidth-delay product. This total size is the product of window size and the number of connections. This is seen in the behavior of the curves in Figure 2 for small numbers of connections and window sizes. It eventually reaches a ceiling, caused by the overall line capacity. The graph slope increases with window size. This is the expected result, as an individual larger window would provide a larger increase in bandwidth.



2. Throughput values are similar whether few large windows are used or many small ones, if the TEWS is the same. Compare the data points for 1 connection at 31 Kb, 4 connections at 8 Kb, 8 connections at 4 Kb, and 16 connections at 2 Kb in Figure 2. All produce a delivered throughput of about 75 Kbytes per second. This is the expected result.

3. The use of type-of-service routing as in Goal 4 gives a significant throughput benefit with fewer connections over routing data transfer and acknowledgement on the same path. This is seen by comparing Figures 2 and 3, noting that curves in Figure 2 reach throughput values with far fewer connections. This is an expected result. Data acknowledgement can flow over the lower capacity paths without problem, realizing the benefit of lower delay paths. This produces an effective reduced Bandwidth Delay product.



2.7.2 Unexpected Results

1. There is erratic behavior as the number of connections increases, particularily with larger individual window sizes. This is shown in Figure 3. Problems appear at eight or more simultaneous connections.

   Analysis of the data flowing during a single transfer is shown by the *tcpdump* traces in Figures 4, 5, 7 and 9. A single transfer is made up of data flowing over multiple connections simultaneously. These graphs consist of a pair of curves. The "sent" curve shows the number of bytes transmitted or retransmitted since the last point on the graph. The "acked" graph shows the number of bytes *newly* acknowledged since the last point on the graph.

   Figure 4 represents the expected case. Data sent rates rise quickly, then fall back. Further transmissions are clocked by the acknowledgements, and the data transfer stabilizes at the available line capacity, producing a steady flow of data.

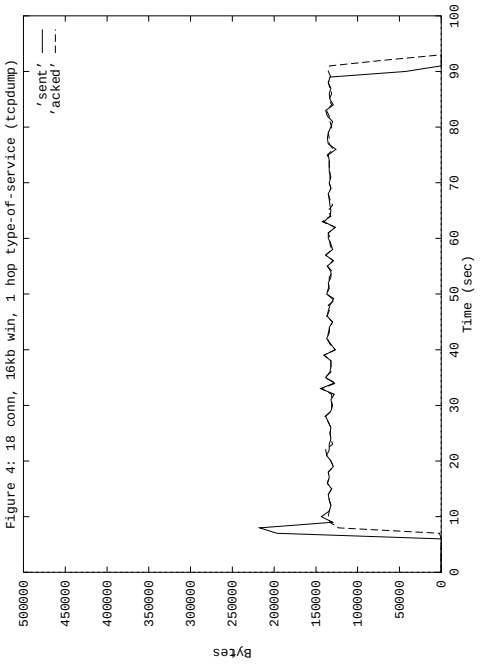
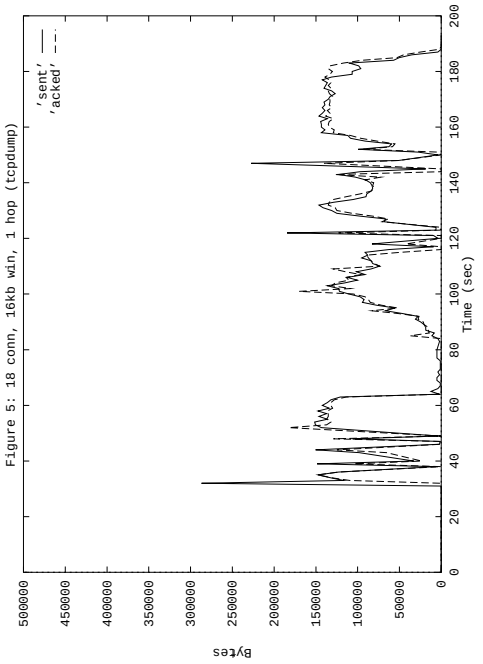


Figure 4 contrasts quite radically with Figure 5. Graphically, the anomolies appear as tall spikes and deep valleys. There are periods of silence, when no connection is transmitting for ten to twenty seconds, resulting in poor line utilization, and increased transfer time. A packet by packet examination of these blackouts reveals that during this interval retransmission of a single packet is taking place. Sometimes four or five retransmissions take place before it is acknowledged. Meanwhile, the data path is quiet.

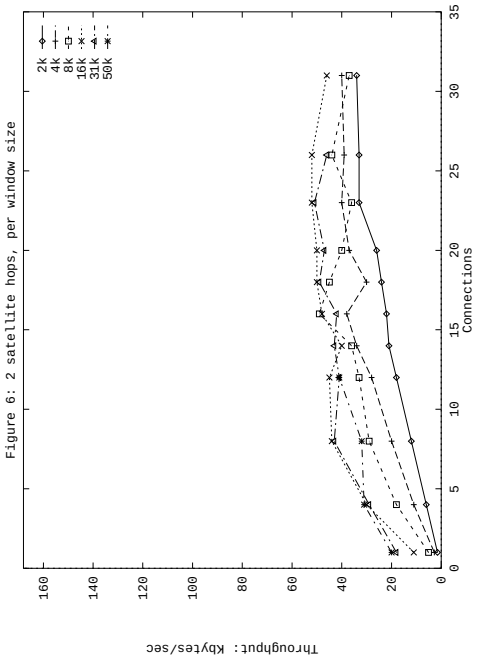
Congestion is not the reason for this. There are no (or very few) other packets flowing. No source quench messages are being sent, indicating that the destination resources are not being overloaded. The problem is not limited to an individual connection. Almost all of the connections are affected simultaneously. Such behavior is non-existent in all of the type-of-service routed cases (c.f. Figure 4).



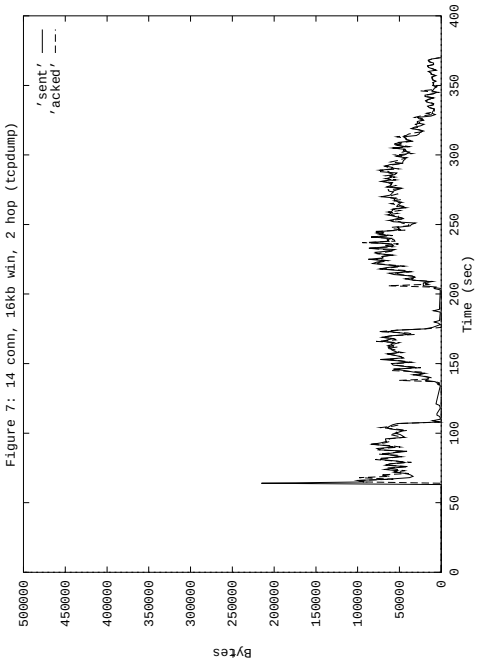
2. A second unusual effect appears in the data taken in the double satellite-hop case. (See Figures 6 and 7).

   Throughput increases in a relatively linear fashion with the number of connections, but at an incremental rate far lower than expected. The bandwidth delay product of this path is 210 Kbytes.

   Many of the data points in Figure 6 represent more than this. For example, twenty connections with a sixteen Kbyte window should fully drive a path with a bandwidth delay product of 320 Kbytes. In this experiment, this configuration was only able to produce 50 Kbytes/second, of the 130 Kbytes/second reached in other scenarios. (Figure 3.)



Close examination of an single transfer over a double satellite-hop is shown in Figure 7. The erratic retransmission effects seen in the one-hop case are less pronounced, but still present. The data transfer is again interrupted by repeated packet loss. This causes all the TCP connections to back off on retransmission, again resulting in reduced channel utilization, and increased transfer time.



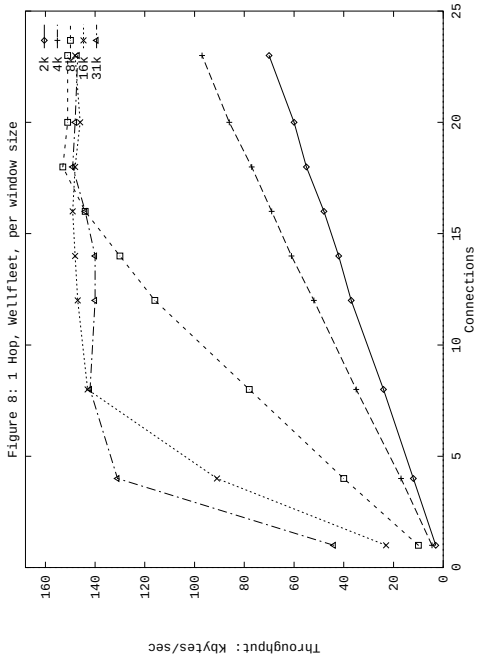
2.8 Post Experiment Analysis

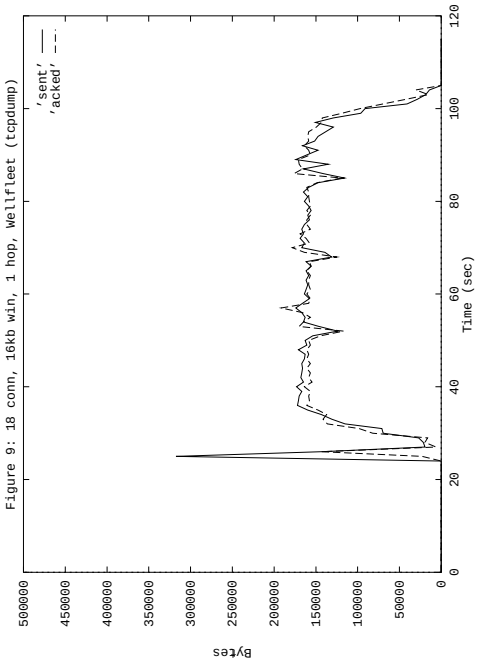
Upon completion of the above wide area experiment, the RIG units were returned to the Communications Lab for further study, to determine the cause of the black-out periods.

Two parallel testbeds were set up in the Lab. Each consisted of a pair of routers, connected by a T1 and 56 kbps line, as in the live testbed. Satellite delay were simulated with a delay line. One testbed consisted of the RIG units, the second of a pair of Wellfleet FN routers. These are a commercial product of similar construction to the RIG units. They reached market by the time of completion of the various testbed experiments described earlier.

The RIG units exhibited the same behavior as in the wide area testbed. The Wellfleet results are shown in Figures 8 and 9. These are comparable to the RIG results shown in Figures 2 and 4. The significant result is that the problem does not exist with Wellfleet units in place in the network. After further analysis of the RIG hardware and software, it was determined that the units would drop packets scheduled to be output on a particular physical T1 interface, if there were many queued to come in over that

same interface. Thus, the RIGs had no problems in the Type of Service case, where the acknowledgement traffic returned over an alternate path.





2.9 Other Factors

Additional factors play a strong role in the evaluation.

1. Proteon, the original manufacturer of the product, allowed the product line to remain dormant for two years following production of the initial units. When they finally released a product, it was based upon the RIG architecture, but of completely different manufacture. The RIG units became a dead product line.

2. During the experimental phase described above, competitive procurement activity also took place, for acquisition of additional units to place in service. Other vendors saw the potential market, and contracts were let with Wellfleet Communications and Cisco Systems. In December 1988, when the original RIG contract let, there were no vendors producing commercial equipment. By June 1991, when the contracts were awarded, units with all the capabilities of the RIG units were available from multiple sources for five to ten thousand dollars.

3. Equipment maintenance became an issue. The units were assembled completely from commercial off-the-shelf parts. Because of this, one could continue hardware maintenance completely without the original builder. This had been a serious issue at the time, because DARPA previous experience with internetwork gear had led to excessive equipment maintenance costs associated with proprietary one time board designs. However, with the advent of the large commercial market for these devices, whole routers could be obtained as described above, at the cost of a single board to be replaced in the RIG.

3 Conclusion

The bottom line for the RIG units is that they quickly became unsuitable for continued use. One of the major objectives of the project was to encourage the existence of a strong commercial market. This occurred far beyond our expectations. It cannot be known to what extent this would have occurred without the RIG project. Our data is quite sketchy. Prior to the program, potential vendors expressed their belief that there was little or no market for high performance internetwork gateways. The RIG program created such a market. By the end of the program, all three RIG vendors (as well as several others) were actively offering such products. This high performance internetwork routing paradigm has become the primary architecture for packet-switched networks. This widespread commercial support and infrastructure has made the design and deployment of the current and future generations of Data Network Systems much faster and more effective.

The experimental results were useful as well. The prediction that data striping multiple TCP connections would overcome file transfer limitations in a satellite network environment was verified in a live testbed. Further, the initial use of type of service data classification was also shown to function. Data transfer flowed over the high capacity path, and acknowledgement traffic returned over the low delay path, resulting in greater efficiency and delivered throughput. We are capable of deploying file transfer software such that a satellite path will no longer impact the delivered performance.

There is an important systems lesson in the RIG program as well. Deploying and using pre-production units has its pitfalls. As commercial technology matures, it may quickly outstrip such a unit. At the program start, the RIGs

had higher performance than any product on the market. After a year, they were out performed by two vendors. By this writing, the maintenance issues and lack of vendor support far outweighed the cost of simply using new commercial equipment. The RIG units are now surplus equipment. One must be prepared to move forward when such dynamic changes occur in a marketplace.

4 References

1. Statement of Work for SNIPE/DARPA Research Internet Gateway, PR No. B-8-3563, Rome Air Development Center, 12 Apr 1988

2. Iannucci, D. and Lekashman, J., *MFTP: Virtual TCP window scaling using multiple connections*, RND-92-002, NASA Ames Research Center, January 1992.

3. Jacobson, V. and Braden, R., *TCP Extensions for Long Delay Paths*, RFC 1072, NSFnet Service Center, nnsc.nsf.net, October 1988.

4. Jacobson, V., Braden, R. and Zhang, L., *TCP Extensions for High Speed Paths*, RFC 1185, October 1990.

5. Postel, J. and Reynolds, J., *File Transfer Protocol (FTP)*, RFC 959, NSFnet Service Center, nnsc.nsf.net, October 1985.