NASA/TM—1999-209445

ACTS 118x Final Report High-Speed TCP Interoperability Testing

William D. Ivancic, Mike Zernic, and Douglas J. Hoder
Glenn Research Center, Cleveland, Ohio

David E. Brooks, Dave R. Beering, and Arun Welch
Infinite Global Infrastructures, LCC, Chicago, Illinois

Prepared for the
Fifth Ka-Band Utilization Conference
sponsored by the Instituto Internazionale delle Comunicazioni
Taormina, Sicily Island, Italy, October 18–20, 1999

National Aeronautics and
Space Administration

Glenn Research Center

November 1999

Trade names or manufacturers' names are used in this report for identification only. This usage does not constitute an official endorsement, either expressed or implied, by the National Aeronautics and Space Administration.

Available from

NASA Center for Aerospace Information
7121 Standard Drive
Hanover, MD 21076
Price Code: A03

National Technical Information Service
5285 Port Royal Road
Springfield, VA 22100
Price Code: A03

ACTS 118x Final Report High-Speed TCP Interoperability Testing

William D. Ivancic, Mike Zernic and Douglas J. Hoder
National Aeronautics and Space Administration
Glenn Research Center
Cleveland, Ohio 44135
www.grc.nasa.gov

David E. Brooks, Dave R. Beering and Arun Welch
Infinite Global Infrastructures, LLC
Chicago, Illinois
www.igillc.com

Introduction

Background

With the recent explosion of the Internet and the enormous business opportunities available to communication system providers, great interest has developed in improving the efficiency of data transfer using the Transmission Control Protocol (TCP) of the Internet Protocol (IP) suite. The satellite system providers are interested in solving TCP efficiency problems associated with long delays and error-prone links. Similarly, the terrestrial community is interested in solving TCP problems over high-bandwidth links. Whereas the wireless community is interested in improving TCP performance over bandwidth constrained, error-prone links.

NASA realized that solutions had already been proposed for most of the problems associated with efficient data transfer over large bandwidth-delay links (which include satellite links). The solutions are detailed in various Internet Engineering Task Force (IETF) Request for Comments (RFCs). Unfortunately, most of these solutions had not been tested at high-speed (155+ Mbps). Therefore, the NASA's ACTS experiments program initiated a series of TCP experiments to demonstrate scalability of TCP/IP and determine how far the protocol can be optimized over a 622 Mbps satellite link. These experiments were known as the 118i and 118j experiments. During the 118i and 118j experiments, NASA worked closely with SUN Microsystems and FORE Systems to improve the operating system, TCP stacks, and network interface cards and drivers. We were able to obtain *instantaneous* data throughput rates of greater than 520 Mbps and *average* throughput rates of 470 Mbps using TCP over Asynchronous Transfer Mode (ATM) over a 622 Mbps Synchronous Optical Network (SONET) OC12 link. Following the success of these experiments and the successful government/industry collaboration, a new series of experiments, the 118x experiments, were developed.

Goals

The overall goals of the 118x experiments were:

1) to work in partnership with the government technology oriented labs, computer, telecommunication, and satellite industries to promote the development of interoperable, high-performance TCP/IP implementations across multiple computing / operating platforms;
2) to work with the satellite industry to answer outstanding questions regarding the use of standard protocols (TCP/IP and ATM) for the delivery of advanced data services, and for use in spacecraft architectures; and
3) to conduct a series of TCP/IP interoperability tests over OC12 ATM over a satellite network in a multi-vendor environment using ACTS.

Conditions Which Affect TCP Efficiency

Three issues needed to be addressed when considering TCP performance: congestion, the bandwidth-delay product, and bit errors.

Beginning around the fall of 1986, the Internet began showing signs of congestion collapse. To alleviate this problem, congestion control algorithms such as the slow start algorithm were adopted into the TCP standard implementations [Ref. 1 and 2]. These algorithms have been continually enhanced and provide an elegant solution to congestion control in an environment consisting of multi-faceted users operating on a variety of interconnected networks, the Internet. Many congestion control algorithms – slow start in particular – may result in inefficient bandwidth utilization for end-to-end communications where a moderated amount of data is being transferred over a link exhibiting large bandwidth-delay characteristics.

Networks with bandwidth-delay products greater than 65535 bytes are referred to as long fat networks (LFNs). The 16 bit Window field in standard TCP results in this 65535 bytes Window limitation – approximately 60% percent of a T1, 1.544 Mbps, over a geosynchronous satellite link. Also, there is a possibility that packet sequence numbers could be used more than once in a LFN. Adding extensions to TCP for scaled windows and timestamps solves these problems. The specification that defines these extensions is found in RFC 1323, TCP Extensions for High Performance [Ref. 3].

Currently, any loss of TCP data is considered to be caused by congestion. As such, congestion control algorithms may be triggered for congestion when data experiences corruption. The TCP fast-retransmit and fast-recovery [Ref. 4], and the selective acknowledgment options [Ref. 5] improve TCP performance in many situations where congestion and/or corruption may occur.

Participants

A consortium was established to perform the 118x experiments. Participants included government agencies, the communication, computer, and satellite industries and academia. Participation took place in a variety of forms including engineering support, in-kind equipment loans, software support, communication links, and consulting.

Connectivity Models

There are three types of connectivity models we planned to work with in 118x: the communication satellite model, the relay satellite model, and the direct broadcast satellite model [Figures 1, 2 and 3 respectively].

For the majority of our testing, we used the communications satellite model. The communications satellite was modeled with symmetrical, balanced links between ground terminals representing a fully meshed satellite network or a trunking system. A communication satellite network may also exhibit moderate asymmetry – particularly for hub-spoke star networks.

The relay satellite model has a highly asymmetric link. The return channel bandwidth is only a small percentage of the data-path bandwidth. This model was put together at the request of NASA's Space Operations Management Office to investigate use of TCP for bulk data transfer over the Tracking and Data Relay Satellite System (TDRSS).

The direct broadcast satellite (DBS) model represents a hybrid satellite/terrestrial network where data is distributed using a high-bandwidth satellite channel and acknowledged using a low-bandwidth terrestrial link. This model closely represents a commercial system such as the Hughes Direct PC product. During 118x experiments, the DBS model was not exercised due to time limitations.

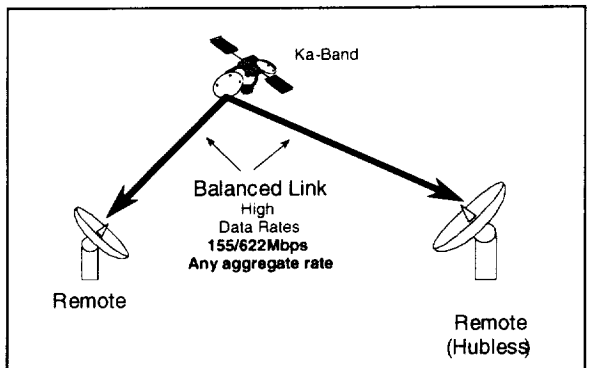

Figure 1 - Communication Satellite

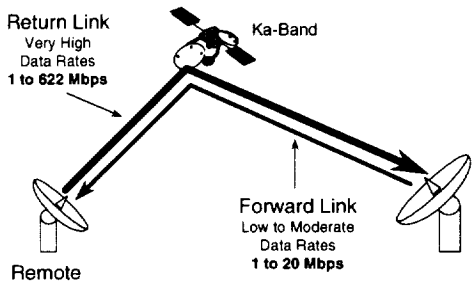

Figure 2 - Relay Satellite

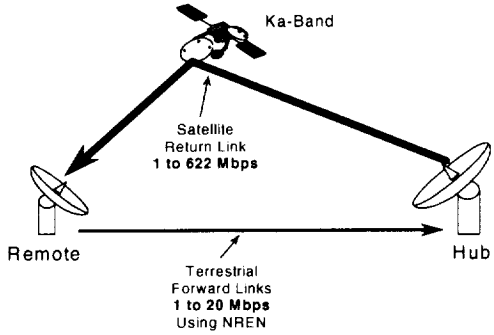

Figure 3 - Direct Broadcast Satellite

Network Configuration

Figure 4 shows the overall network configuration for the 118x experiments. Two High-Data-Rate terminals (HDR) and the Advance Communication Technology Satellite (ACTS) provided the satellite link with an effective bi-directional data throughput of 622 Mbps. The interfaces to the ACTS ground terminals is 622 Mbps using the SONET physical link protocol. The HDR located at NASA's Glenn Research Center (GRC) in Cleveland, Ohio was connected directly to a FORE ASX-1000 ATM switch. The workstations and personal computers were also connected directly to this switch. No routers were used for these experiments. The second HDR was located at Lawrence Livermore National Laboratory. The HDR was connected to an ATM switch at Sprint's Advanced Technology Laboratory in Burlingame, California through the National Transparent Optical Network Consortium (NTONC) using Dense Wave Division Multiplexing (DWDM) technologies. At Sprint's Advance Technology Laboratory, there was a mirror image of the GRC site. Again, no routers were used at the Sprint site.

Workstations, ground stations, and switches in this network were also connected via the Internet in order to configure, control, and monitor all workstations remotely. This was critical in order to operate and maintain the network. The ACTS satellite has many experiments scheduled; therefore, the satellite link was only available during scheduled experiment times.

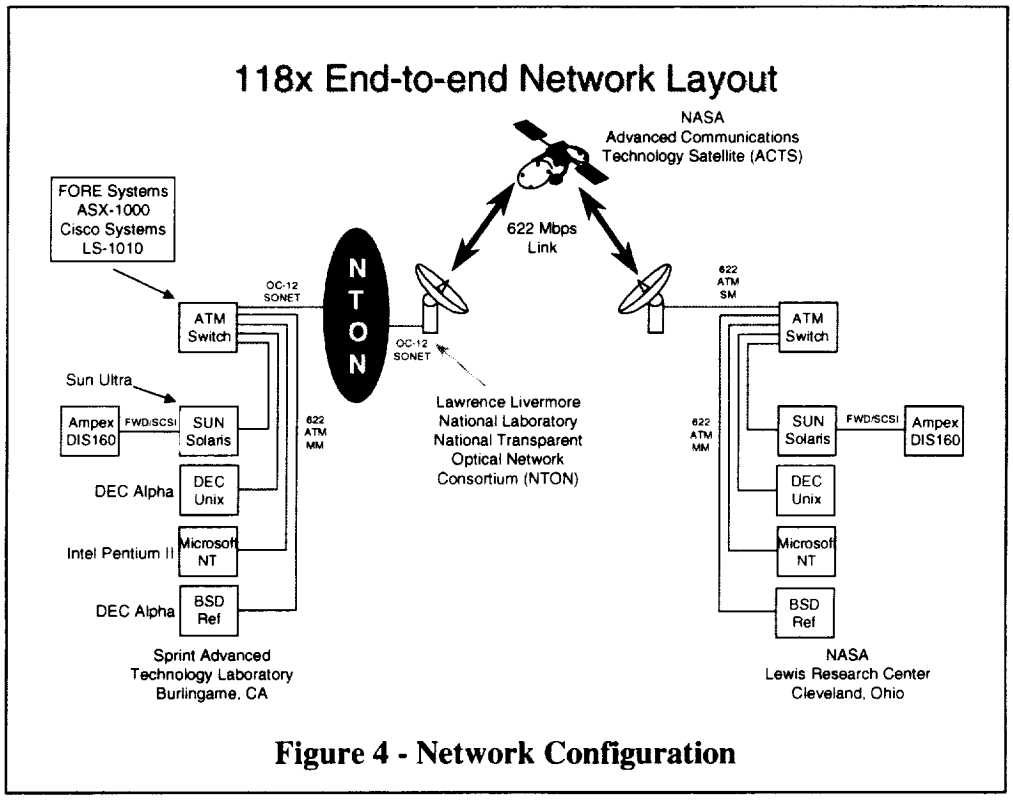

Figure 4 - Network Configuration

Symmetric Interoperability Tests

Configuration

For the symmetric tests, the network configuration represented in Figure 4 was setup to represent a duplex trunking satellite network as shown in Figure 1. These tests were designed to determine TCP interoperability. It is important to note that only the large-window, time-stamp, and protect against sequence number wrap-around portions of the TCP stack should be exercised during these tests as the system should have been operating error-free and congestion-free.

Results

The maximum theoretical throughput for TCP over classical Asynchronous Transfer Mode (ATM) over a Synchronous Optical NETwork (SONET) is approximately 134 Mbps for a 155 Mbps link and 537 Mbps for a 622 Mbps link when taking ATM and SONET overhead into consideration.

We were able to test – to varying degrees – interoperability on four operating systems: SUN's Solaris, Microsoft's NT4 and NT5, Silicon Graphics IRIX, and Compaq's OSF1. Tables 1 and 2 highlight the sustained average throughputs we were able to obtain for symmetric links. Table 1 shows result when operating in a local area network (LAN) environment with 10's of milliseconds of delay. Table 2 shows results when operating over a satellite link with 570 milliseconds of delay. The variation of results can be due to TCP and operating system implementations, network interface cards, line drivers, and/or workstation processor speeds and resources. The details and all raw data is available from the 118x Web Site, http://mrpink.grc.nasa.gov/118x. All tests were run with optimal buffer size allocations and used classical IP over ATM, the IP packet set to 9180 bytes. It should be noted that these results show the state-of-the-system as of November 1, 1998. One should expect that the systems will become more stable and eventually obtain near theoretical throughput within the next few years – particularly for OC3 links.

Operating System	Data Link Rate (Mbps)	Acknowledgment Link Rate (Mbps)	Average Data Throughput (Mbps)
OSF1	155	155	133
IRIX	622	155	500+
NT4/NT5	622	622	357
Solaris	622	622	400-500

Table 1 - Data Throughput for TCP over ATM over SONET in A LAN Environment

Operating System	Data Link Rate (Mbps)	Acknowledgment Link Rate (Mbps)	Average Data Throughput (Mbps)
OSF1	155	155	120[1]
IRIX	622	155	465[2]
NT4/NT5	622	622	Unstable
Solaris	622	622	473

[1] OSF1 transmitting data, Solaris Acknowledging
[2] Solaris transmitting data, IRIX Acknowledging

Table 2 - Data Throughput for TCP over ATM over SONET over a 570 msec Delay

Dual-Hop

During the early stages of the 118x experiments, we experienced fiber outages in the NTON. While these were being corrected the network was put into a loopback configuration as shown in Figure 5. This created a dual hop situation effectively doubling end-to-end delay and thereby doubling the bandwidth-delay product. Thus, we were able to demonstrate high-speed TCP operation in a dual-hop environment. As far as the TCP stack is concerned, demonstrating TCP performance over this bandwidth-delay product (622 Mbps x 1080 msec) effectively demonstrates a Gbps transmission over a single hop geostationary satellite relative to the protocol.

The link in each direction supports approximately 539 Mbps of user data throughput after overhead is subtracted (ATM, SONET, TCP/IP). Given that the acknowledgements occupy approximately 1/30 as much bandwidth as the forward data stream, we need to size our TCP/IP windows such that the data transfer bandwidth is less than about 97% of the available bandwidth. For our 539 Mbps forward link, the BW-delay product is 72,293,375 Bytes, given a delay of 1.073 seconds. We allowed four percent of the link to be allocated for acknowledgements. Therefore, we set our TCP/IP socket buffer to 69,401,640 bits. With these setting we were able to achieve an average TCP throughput of approximately 344.16 Mbps with server CPU usage at 83 percent.

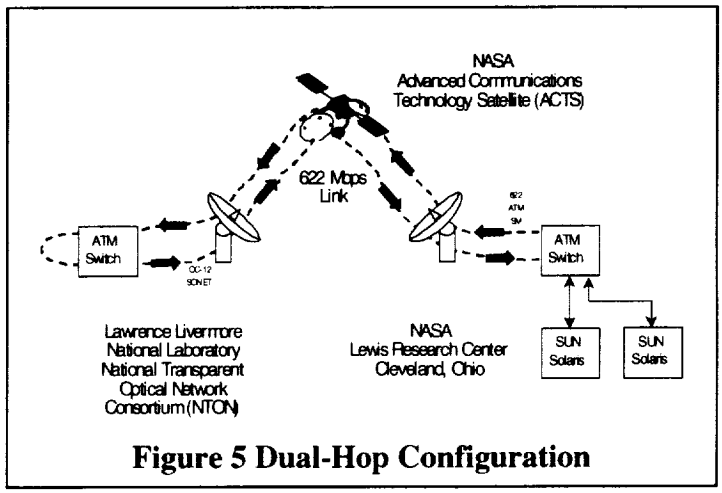

Figure 5 Dual-Hop Configuration

Asymmetric Tests

Configuration

For the asymmetric tests, the network configuration represented in Figure 4 was setup to represent NASA's Tracking and Data Relay Satellite System (TDRSS) or any other relay satellite network [Fig. 2]. These tests were performed to empirically determine "for a given forward satellite channel capacity, how much return capacity can be supported using TCP/IP?" In other words, what is the ratio of data path bandwidth to acknowledgment path bandwidth for bulk data transfers? For these experiments, the Sun workstations were used. The operating system was OS version 5.7 with Kernel version SunOS Release 5.7 version kcpoon.

For NASA's TDRSS, White Sands, New Mexico is the center of reference. A remote spacecraft sending data to White Sands is "returning" data, and the link is called a Return Link. Conversely, if a link is mapped from White out to the remote spacecraft, they call it a Forward Link. For the 118x tests, we established a Forward Link on the 118x test network operating at 1 Megabit per second. That link speed was enforced at the Sun workstation's ATM interface by a setting the rate queue, and also in the FORE Systems ATM switch by implementing a Constant Bit Rate traffic contract. The Return Link was not constrained, so theoretically it could go up as high as 622 Megabits per second. The Sun workstations set up to support TCP congestion windows as large as 32 Megabytes. When we ran the tests between the two workstations without the traffic contracts enforced, the workstations were able to support sustained data rates of approximately 500 Megabits per second. We then ran the tests with the traffic contract enforced for the forward channel at one Megabit per second and then again at two Megabits per second.

Results

The tests were run with a 1-, 2-, and 3-Megabit traffic contracts. For the 1 and 2 Mbps Forward Link Cases we were able to obtain a sustained return link data rates of 177.14 Mbps and 309.45 Mbps as measured by ttcp. The ttcp measurements are average values and included the slow-start throughput in the calculation whereas the PDUs per second measurement is a windowed value. Using the PDUs per second, one can calculate the throughput as:

PDUs per second x 9180 Bytes per PDU x 8 bits per Byte.

UPC Mbps	Peak PDU	Sustained PDU (PDUs/sec)@1 sec poll	Avg.TPut Mbps	Traced?	atmspeed Reported
1	3476	2885	177.14	Y	1 Mbps
1.5	4343 (?)	2885	182.53	Y	1 Mbps
2.0	5911	5776	309.45	Y	2 Mbps
2.5	N/A	N/A		N/A	N/A
3.0	6987	6800	357.60	Y	3 Mbps

Table 3 - Asymmetric Link Throughput Results

From the PDU data, for the 1 and 2 Mbps Forward Link Cases we were able to obtain a sustained return link data rates of 211 Mbps and 422 Mbps. Thus, the ratio of return-to-forward link is approximately 200:1 using the PDU data. The throughput for the 3 Mbps Forward Link Case is no longer linear as the maximum return link throughput had been reached. We were unable to obtain good data for Constant Bit Rate traffic contracts of 1.5 Mbps and 2.5 Mbps because the FORE ATM switches setup the traffic contacts in integer steps. Setting up a contract for 1.5 Mbps resulted in a contract for 1 Mbps.

Testing Tools

TCP Analysis

During the 118x experiments, it became evident that better tools were needed to debug and analyze the TCP stack – particularly with multiple vendors involved. Many tools are publicly available to perform TCP capture and analysis. Those chosen to use were: *ttcp*, *tcpdump*, *tcptrace*, *atmsnoop* and *xplot*. We had to develop special testing techniques and modify some of these to operate in the high-speed environment we were operating in. The details of this work are available in a paper entitled, "High-Speed TCP Testing" [Ref. 6].

Scripts

In order to obtain all of the pertinent information from each TCP run, a number of script files were generated. Information that we considered pertinent included: general workstation statistics, SONET ATM layer statistics from the switches at both the HDRs' and the workstations' ATM ports, TCP/IP statistics, TCP/IP settings, workstation driver information, type of ATM interface, and special workstation settings. The script files were combined into a single script that starts a bulk transfer between two workstations with the entire run recorded under a common directory on the initiating workstation. After the test, the receiver information is automatically copied to the transmitter side.

Where do we go from Here?

A) Investigate Other TCP Options

The ACTS HDRs use Reed-Solomon forward-error-correction (FEC) coding. This results in an error-free link. In addition, we only had a single TCP connection running over the link at any time – no congestion. Therefore, only the large-windows portion of the TCP stacks was exercised. The fast-retransmit, fast-recovery and selective acknowledgement (SACK) options should not have been. If these option were exercised, either something was wrong with the link or the TCP stack (including the hardware and drivers). The fast-retransmit, fast-recovery and SACK options need to be exercised.

B) Perform LAN Benchmarks

The 118x TCP interoperability experiments were run over ACTS with one set of equipment located in Cleveland, Ohio and a second set in Burlingame, California. Our desire was to originally have all the equipment shipped to GRC first for stringent baselining and LAN testing. This did not happen. Therefore, we did not have baseline information of the various stacks in a LAN environment. The TCP stacks need to be baselined and include baselining for the fast-retransmit, fast-recovery and SACK options as well as for large windows. This can be accomplished using at OC3 rates using a HP BSTS and/or an Adtech SX/14 to impair the channel. This cannot be readily accomplished at OC 12 rates, as we are not aware of any channel impairment equipment that operates at OC12 rates.

C) Build Upon these Findings and Experiences
A series of experiments is needed to evaluate various commercial and research TCP implementations in a controlled environment and determine the following:

1) Does the TCP implementation function properly relative to the specifications found in RFCs 1323 and 2018 (TCP Enhancements Compliance)?
2) If so, at what bandwidth-delay can the stack operate before breaking down and what is the cause of the breakdown (TCP Breakdown)?
3) How well does this TCP stack interoperate with other TCP stacks (TCP Interoperability)?

TCP Enhancements compliance testing is necessary to ensure we are using fully operational TCP implementations when evaluating TCP interoperability. Determining TCP breakdown will allow us to know to what bandwidth-delay we can properly evaluated TCP interoperability for various vendor's TCP implementations. In addition, we will be able to provide this information to the commercial vendors thereby enabling them to improve their products.

Conclusion

We believe the overall goals of the 118x experiments were met. We were able to establish partnerships with the computer, telecommunication, and satellite industries to promote the development of interoperable, high-performance TCP/IP implementations across multiple computing / operating platforms. We also were able to answer many outstanding questions regarding the use of standard protocols (TCP/IP and ATM) for the delivery of advanced data services and for use in spacecraft architectures. Of particular importance was the validation of TCP implementation, allowing hundreds of Mbps of throughput in a symmetric scenario as well as establishing a return-to-forward link ratio of 200:1 in an asymmetric scenario. However, there is still much work that was not completed. In particular, the TCP stacks need to be baselined and tested for interoperability with all aspects of TCP exercised including large-windows, fast-retransmit, fast-recovery and SACK. Follow-on experiment activities are being formulated to address these areas. The final report documenting this work in its entirety is now available [Ref. 7].

References

[1] Van Jacobson, Michael Karels: Congestion Avoidance and Control, ACM SIGCOMM, August 1988. Available from ftp://ftp.ee.lbl.gov/papers/congavoid.ps.Z as of July 1999.
[2] Van Jacobson: Note to the IEFT end2end-interest working group, April 1990.
[3] V. Jacobson, R. Braden, D. Borman TCP Extensions for High Performance. RFC 1323 May 1992. (Obsoletes RFC1072, RFC1185).
[4] W. Stevens: Slow Start, Congestion Avoidance, Fast Retransmit, and Fast Recovery Algorithms, RFC 2001 January 1997.
[5] Matt Mathis, Jamshid Mahdavi, Sally Floyd, and Allyn Romanow. TCP Selective Acknowledgment Options, October 1996. RFC 2018.
[6] D.E. Brooks, H. Gassman, D.R. Beering, A. Welch, D.J. Hoder, W.D. Ivancic: High-Speed TCP Testing November 1998, http://ctd.grc.nasa.gov/5610/inetprotocols.html.
[7] D.E. Brooks, C. Buffinton, D.R. Beering, A. Welch, W.D. Ivancic, M. Zernic, and D.J. Hoder: ACTS 118x Final Report High-Speed TCP Interoperability Testing, NASA TM-1999-209272, July 1999. Available from http://mrpink.grc.nasa.gov/118x/ as of July 1999.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (*Leave blank*)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	November 1999	Technical Memorandum

4. TITLE AND SUBTITLE

ACTS 118x Final Report High-Speed TCP Interoperability Testing

5. FUNDING NUMBERS

WU–632–50–5A–00

6. AUTHOR(S)

William D. Ivancic, Mike Zernic, Douglas J. Hoder, David E. Brooks, Dave R. Beering, and Arun Welch

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

National Aeronautics and Space Administration
John H. Glenn Research Center at Lewis Field
Cleveland, Ohio 44135–3191

8. PERFORMING ORGANIZATION REPORT NUMBER

E–11931

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

National Aeronautics and Space Administration
Washington, DC 20546–0001

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

NASA TM—1999-209445

11. SUPPLEMENTARY NOTES

Prepared for the Fifth Ka-Band Utilization Conference sponsored by the Instituto Internazionale delle Comunicazioni Taormina, Sicily Island, Italy, October 18–20, 1999. William D. Ivancic, Mike Zernic, and Douglas J. Hoder, NASA Glenn Research Center; David E. Brooks, Dave R. Beering, and Arun Welch, Infinite Global Infrastructures, LCC, Chicago, Illinois. Responsible person, William D. Ivancic, organization code 5610, (216) 433–3494.

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Unclassified - Unlimited
Subject Category: 17 Distribution: Nonstandard

This publication is available from the NASA Center for AeroSpace Information, (301) 621–0390.

12b. DISTRIBUTION CODE

13. ABSTRACT ***(Maximum 200 words)***

With the recent explosion of the Internet and the enormous business opportunities available to communication system providers, great interest has developed in improving the efficiency of data transfer using the Transmission Control Protocol (TCP) of the Internet Protocol (IP) suite. The satellite system providers are interested in solving TCP efficiency problems associated with long delays and error-prone links. Similarly, the terrestrial community is interested in solving TCP problems over high-bandwidth links. Whereas the wireless community is interested in improving TCP performance over bandwidth constrained, error-prone links. NASA realized that solutions had already been proposed for most of the problems associated with efficient data transfer over large bandwidth-delay links (which include satellite links). The solutions are detailed in various Internet Engineering Task Force (IETF) Request for Comments (RFCs). Unfortunately, most of these solutions had not been tested at high-speed (155+ Mbps). Therefore, the NASA's ACTS experiments program initiated a series of TCP experiments to demonstrate scalability of TCP/IP and determine how far the protocol can be optimized over a 622 Mbps satellite link. These experiments were known as the 118i and 118j experiments. During the 118i and 118j experiments, NASA worked closely with SUN Microsystems and FORE Systems to improve the operating system, TCP stacks, and network interface cards and drivers. We were able to obtain *instantaneous* data throughput rates of greater than 520 Mbps and *average* throughput rates of 470 Mbps using TCP over Asynchronous Transfer Mode (ATM) over a 622 Mbps Synchronous Optical Network (SONET) OC12 link. Following the success of these experiments and the successful government/industry collaboration, a new series of experiments, the 118x experiments, were developed.

14. SUBJECT TERMS	15. NUMBER OF PAGES
Networking; Satellites; Protocols; TCP	16
	16. PRICE CODE: A03

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	