1. Introduction

The electronic transmission of VLBI data (dubbed e-VLBI) presents a special challenge to the use of high-speed global networks. With long-term requirements for simultaneous or near-simultaneous Gbps data streams from antennas worldwide converging in a single processing center, e-VLBI is both a useful and highly synergetic application for global high-speed networks. Though the ultimate goal of a worldwide web of radio antennas connected at Gbps speeds is a few years away, the work described in this report is a first step in this direction.

In the work described here, funded by DARPA and NASA and executed in collaboration with several institutions, two antennas, one in Massachusetts and one in Maryland, were connected to a correlator in Massachusetts at sustained streaming rates approaching 1 Gbps over a network with several shared segments. This is believed to be the first connection of this kind, using shared networks, to demonstrate e-VLBI at these rates.

2. What is VLBI?

Very-Long-Baseline Interferometry (VLBI) has been used by radio astronomers for more than 30 years as one of the most powerful techniques for studying objects in the universe at ultra-high resolution and measuring earth motions with ultra-high accuracy [VLBA]. For astronomical studies, VLBI allows images of distant radio sources to be made with resolutions of tens of microarcseconds, far better than any optical telescope; the highest resolution VLBI images can resolve the dimples on a golf ball in Los Angeles as viewed from Boston! For Earth science studies, VLBI provides direct measurement of the vector between globally-separated telescopes with an accuracy of a few millimeters and, by measuring the changes in baseline length over time, yields motions of the tectonic plates with a precision of 0.1 mm/yr. Because VLBI is tied to the reference frame defined by very distant radio objects, and because VLBI can measure the sky position of these objects to sub-milliarcsecond precision, it is the most accurate technique to fundamentally measure the orientation of the earth in space. Paradoxically, VLBI is currently the only way to study the shape of the Earth's liquid iron core and to determine the magnetic connections of the liquid core to the mantle above it and to the solid core at the center of the Earth [MHB].

Basically, VLBI combines simultaneously acquired data from a global array of up to ~20 radio telescopes to create a single coherent instrument, as illustrated in Figure 1 for a simple 2-element VLBI array. Traditionally, VLBI data are collected on magnetic tapes that are shipped to a central site for correlation processing. This laborious and expensive data-collection and transport process now has the possibility of being replaced by modern global high-speed networks, potentially opening important new capabilities and scientific returns.
By nature, VLBI data are digital representations of the analog signal arriving at a radiotelescope. Almost always this is white Gaussian noise sampled at the Nyquist rate so that each sample is independent and the data are fundamentally incompressible.

Figure 2 and Figure 3 show examples of VLBI observations for both astronomy and geodesy.

Figure 2: Example illustrating how VLBI can produce ultra-high resolution images of the core of a radio source. This example show how VLBI can zoom into the very core of the giant radio galaxy NGC6251, at a distance of \(~100\) megaparsecs from Earth; the lower zoomed image is a VLBI image of the core where 1 parsec corresponds to \(~2\) milliarcseconds (1 parsec=3.26 light years).
3. What is e-VLBI

The transmission of VLBI data via high-speed networks is dubbed ‘e-VLBI’, the development of which is the main thrust of this demonstration project. e-VLBI can be exercised in either one of two ways: 1) real-time transfer via a direct connection from telescope to correlator or 2) quasi-real-time transmission by first buffering the data through a buffer memory (semiconductor or magnetic disc) capable of storing seconds to hours of data before transmission. The latter is generally preferred for a host of practical reasons, but both are workable and both provide many of the same advantages.

3.1 e-VLBI: The Advantages

The potential advantages for scientific productivity and technical operations of e-VLBI over traditional VLBI are:

1. **Faster turnaround of results**: Typical time to completion of processing of traditional VLBI is delayed until tapes or disks can be shipped, often from distant locations, and is usually measured in at least days and sometimes weeks. This makes it almost impossible to use ‘fast-response’ observations of important transient events as a means of guiding further critical observations. This is especially important as transient events such as extragalactic supernova or gamma-ray-burst events are becoming increasingly important in understanding various aspects of our universe.

2. **Higher sensitivity**: The potential to extend e-VLBI to multi-Gbps (Gigabit per second) data rates will allow an increase in the sensitivity of observations beyond those possible with traditional recorded media. For most observations, VLBI sensitivity increases as the square root of the data rate; larger antennae and quieter receivers are generally the only other methods to increase sensitivity. However, larger antennae are hugely expensive and many modern receivers are already near theoretical quietness limits.

3. **Lower costs**: e-VLBI will eliminate the need for expensive tape or disk pools while at the same time allowing full automation of observations, all towards the goal of lowering cost.

4. **Quick diagnostics and tests**: Some aspects of VLBI equipment are very difficult to test and diagnose without actually taking data with another station and processing it. Unfortunately, this characteristic
Both faster turnaround and higher sensitivity will open doors to better science while lower costs and easy diagnostics will lead to more science impact per dollar.

3.2 e-VLBI: The Challenges and Promises

3.2.1 e-VLBI on Shared Networks

e-VLBI has the potential to use a significant amount of the currently unused capacity on existing research network data pipes. Based on published usage statistics of networks like Internet2/Abilene [I2STAT], for example, the average usage over long periods is often below 20%, so there is potentially much available capacity for an application such as e-VLBI.

3.2.2 The ‘Last Mile’ Problem

Few radio observatories in the world are currently connected to high-speed networks, though many are near high-speed nodes. Completing the direct connection of a telescope to a high-speed node is known as the ‘last-mile’ problem [HINT]. e-VLBI will work most efficiently if all telescopes have a direct high-speed network connection, but that is not required in the short term; e-VLBI data may be temporarily buffered onto, for example, disks, which may then be transported to a nearby high-speed node for transmission to a correlator. This potential large-scale buffering of data is one of the characteristics of e-VLBI that can be exploited to design a workable and robust data transmission system that interfaces well with existing network traffic.

Nevertheless, some major global observatories, particularly in Europe and Japan, either are, or soon will be, connected to high-speed data links. In the U.S., Haystack Observatory is one of the few currently well connected, but links to other major observatories such as Arecibo in Puerto Rico and the Green Bank Telescope in West Virginia are being actively investigated. The development of a protocol to make the transmission of e-VLBI data over shared networks efficient and practical will certainly further spur these already ongoing activities.

4. History of e-VLBI

Electronic transmission of VLBI data from antenna to correlator has been an obvious but difficult goal of VLBI practitioners since the origin of the VLBI technique in the late 1960’s. A pioneering experiment in 1977 linked the signals from two antennas over a real-time satellite link [YEN]. Transmission of small amounts of data (~1 Mb/station) over ordinary telephone lines to a software correlator was successfully accomplished at Haystack in 1979 [LEVINE]. Since that time, the dominant activity in e-VLBI has been in Japan, with the Keystone project in ~1995 [YOSHINO] linking four antennas in real-time at 256 Mbps (megabits per second) and, more recently, dedicated Gbps networks. Furthermore, high-speed radio links have been used to transmit data from orbiting antennas, notably the TDRSS satellite, in 1986 [LEVY] and the Japanese dedicated orbiting VLBI satellite HALCA in 1997 [HIRA]; both of these satellites transmitted data to the ground for recording, but did not transmit data in real-time or near-real-time for correlation.

Until very recently, however, it has not been possible to take significant advantage of the rapidly developing public networks, most notably the Internet, which links nearly all worldwide VLBI antenna sites with network connections of varying speed and quality, but which is rapidly expanding in bandwidth and quality.

5. Haystack Observatory’s Role in the Development VLBI and e-VLBI

MIT Haystack Observatory has been at the forefront of VLBI instrumentation and technique development since the first VLBI experiments in the late 1960’s. Starting with modest tape recording capabilities of <1 Mbps, Haystack Observatory has developed four generations of special-purpose recording equipment, now extending to 1 Gbps. For most VLBI observations, a higher data rate translates directly to higher
signal-to-noise ratio, so there is a continual demand for higher data rates. With the advent of modern national and international networks, the potential exists not only to return data in real-time or near-real-time to a processing center, but to significantly extend data rates beyond those economically achievable by conventional recording techniques.

6. Development of Mark 5 e-VLBI Data System

With funding support from seven international institutions, Haystack Observatory has developed a next-generation VLBI system that is compatible with both traditional VLBI and e-VLBI and replaces the existing magnetic-tape-based data systems. Dubbed ‘Mark 5’, this system is based primarily on PC-based COTS components and subsystems [WHIT]. VLBI data up to 1 Gbps may be recorded and played back onto an array of 16 disks for traditional ship-and-process operation, or may alternately be transmitted electronically over standard network interfaces either in real-time or using disks to buffer and transmit in quasi-real-time.

The Mark 5 system, shown in Figure 4, has been adopted by both global geodetic and astronomical VLBI communities as the primary next-generation VLBI data system and is now being deployed worldwide. The Mark 5 system is a critical element in supporting the work detailed in this report.

![Mark 5 VLBI Data System](image)

Figure 4: Mark 5 VLBI Data System

7. Goals and Outline of the Gbps e-VLBI Experiment

The primary goal of the e-VLBI experiment is to perform an e-VLBI experiment under real-world conditions, using existing networks, transferring data from each of two antennas to a correlator at a rate of ~1 Gbps.

Figure 5 outlines the experiment in graphical form. Data are collected at the 20m (meter) diameter Westford antenna and transmitted to the correlator at Haystack Observatory, about 1.5 km (kilometer) distance. Data are simultaneously collected at the GGAO 10m-diameter antenna a NASA/GSFC in Greenbelt, MD and transmitted to the correlator at Haystack Observatory, a direct distance of ~650km.

---

1 NASA, U.S. Naval Observatory, National Radio Astronomy Observatory, Max Planck Institute (Germany), Bundesamt für Kartographie und Geodäsie (Germany), Joint Institute for VLBI in Europe (Netherlands), and Korean Astronomy Observatory.
2 Commercial Off-the-Shelf (COTS)
Data from the Westford antenna are transmitted on a local dedicated fiber link to the correlator. But data from the NASA/GSFC antenna are transmitted over a combination of dedicated and shared network segments, including NASA/HECN, MAX, ISI, Bossnet and Glownet; of these, the Bossnet and Glownet segments are dedicated and the balance are shared. Figure 6 shows a detailed diagram of the NASA/GSFC to Haystack link, including many intervening routers and switches.

Figure 5: Schematic of Gbps e-VLBI demonstration experiment

Figure 6: Details of the e-VLBI network path
A multi-step process was instituted to develop the necessary capability, as follows:

1. Development and testing of a Gbps e-VLBI data system.
2. Survey and document the details of the network.
3. Put in place a set of performance testing nodes along the path in order to test various parts of the network in isolation as well as in whole.
4. Regular and coordinated performance testing and tuning of various network components, as necessary.
5. Once all of the above was in place, experiments were performed with increasing demands on real-time performance:
   a. Record data locally at each of the two antennae onto magnetic disks files, then transmit the files to disks at the correlator, followed by correlation from disks at the correlator.
   b. Direct transmission of e-VLBI data in real-time from each antenna to disks at the correlator, followed by correlation.
   c. Direct transmission of e-VLBI in real-time from each antenna directly to the correlator at Haystack Observatory.

During the ~1 year duration of the experiment development, regular telephone conferences were held between all participants, which was a crucial element in coordinating all of the elements and actions necessary to carry out a successful set of experiments.

8. Elements in the e-VLBI Path

The network path used for the Gbps e-VLBI experiment spanned the domains of several organizations. In this section we describe each of the elements in the e-VLBI data path.

8.1 Glownet and Bossnet

Glownet is part of the MIT Lincoln Laboratory network infrastructure in eastern Massachusetts [GLOWNET]. It connects Millstone Hill (including the Haystack Observatory) in Westford, the Lincoln Laboratory facilities in Lexington; the MIT campus in Cambridge; and the northern terminus of Bossnet in Boston (see Figure 7). This metropolitan area network consists of high quality (low loss and dispersion) "dark" fiber leased from a number of providers and is "lit" with commercial optical network equipment. ("Dark" fiber is the fiber alone. It is "lit" with light from optical networking equipment.) For e-VLBI experiments IP data is conveyed in Gigabit Ethernet (GigE) format. Figure 8 shows a top-level diagram of the data flow. The GigE data flows through network switches within the Millstone complex. The switches are chosen for their ability to process GigE jumbo frames, one of the necessary factors in being able to sustain high processor-to-processor throughput. The data within the Millstone complex uses a wavelength of 1310 nm (nanometer) and is transponded to the 1550 nm band for minimum loss through optical fiber. This data is sent via Boston (56 miles away), where it is amplified and forwarded, to Lincoln Laboratory in Lexington (another 26 miles).

At Lincoln Laboratory the data stream is passed through Gigabit Ethernet switches to a Bossnet transponder which is designed to transmit and receive high speed data over long distances without the need for intermediate electronic data regeneration.
Bossnet is a DARPA-sponsored research program at MIT Lincoln Laboratory [BOSSNET]. The Bossnet program has been used for research in high rate (10's of gigabit/second per wavelength), long range (approximately 1000 km) all-optical transmission. It has been used by MIT-campus researchers as well as MIT Lincoln Laboratory. The portion of Bossnet used by e-VLBI consists of a pair of dark fibers that follow a coastal route from Boston to Washington DC (see Figure 9). The optical signal is amplified at fifteen "huts" along the route, but there is no intermediate electronic signal regeneration. (Hence the need for specially designed transponders.) At the Washington DC terminus of Bossnet there is an additional pair of dark fibers, which extends the reach to ISI East in Arlington, VA.
8.2 Information Sciences Institute East (ISI-E)

The Information Sciences Institute (ISI) is a research facility within the University of Southern California (USC) and includes facilities in Marina del Ray California and Arlington Virginia. The ISI facility in Arlington Virginia is known as ISI East (ISI-E) and host nodes for several research networks including:

- Mid-Atlantic Crossroads (MAX)
- BoSSNET
- ATDNet

Access to Internet2 Abilene is also available via this infrastructure. Figure 10 shows a diagram of the ISI-E network infrastructure.
Personnel at ISI-E and other institutions utilize these facilities to conduct a wide range of network research activities. Test and research for the electronic Very Long Baseline Interferometry (e-VLBI) is one such example.

### 8.3 Mid-Atlantic Crossroads (MAX)

The Mid-Atlantic Crossroads (MAX) network is a consortium of four founder universities: University of Maryland, Virginia Polytechnic Institute, Georgetown University and The George Washington University. Architecturally the MAX network forms a high-speed network ring around the Washington, D.C. area and provides the bridge between ISI-E and NASA/GSFC, which both connect to MAX. Figure 11 shows the core optical network of the MAX ring; and Figure 12 provides more detail on the physical connections of individual MAX participants. (Note: ISI-E is designated ‘ARLG’ in Figure 11 and Figure 12.)

As noted in a later section, data traffic between GGAO and MIT/Haystack normally is routed across the MAX core via its direct link between the MAX’s Juniper M160 routers at UMCP and ISI-E; but occasionally for various network debugging or performance tuning reasons it has been rerouted the longer way around the MAX ring between UMCP and ISI-E through the MAX Juniper M160 router at Eckington in DC.

![Core Optical Network](image)

**Figure 11: Simple schematic of MAX network**

### 8.4 MAX-GGAO Link

The e-VLBI path between the MAX Juniper M160 router at UMCP and the GGAO at NASA/GSFC includes both a shared part and an e-VLBI dedicated part, both engineered by GSFC’s High End Computer Network (HECN) team.
Installed about three years before the e-VLBI dedicated part but subsequently upgraded to also support e-VLBI experiments, the shared part supports a number of high end computer experiments between hosts on GSFC’s internal near-1-Gbps Scientific and Engineering Network (SEN) and the MAX and Abilene networks. Continuing with the general “north-to-south” description of the e-VLBI path already provided, the shared part starts with a MAX-provided jumbo-frame-capable GE port on the MAX Juniper M160 router at MAX/UMCP, and then includes a jumbo-frame-capable GE link through a HECN-provided Extreme Network 5i GE switch at MAX/UMCP, and use of an approximate 15 km jumbo-frame-capable GE connection through one wavelength of a coarse wave division multiplexed (CWDM) link over dark fiber (leased from Fibergate, Inc) between HECN-provided LuxN CWDM equipment at the MAX/UMCP and the HECN team’s lab in GSFC’s building 28, to another HECN-provided Extreme Network 5i GE switch in the HECN team’s lab in GSFC’s building 28. This latter GE switch supports local connections with several network testing computers in the HECN team’s lab, with a HECN-managed Cisco GSR 12016 router which is the default gateway for the hosts in GSFC’s SEN, and with the following described e-VLBI dedicated path part.
The e-VLBI dedicated part extends on dedicated dark fiber from the shared HECN-provided Extreme Network 5i GE switch in the HECN team’s lab in GSFC’s building 28, approximately 7.25 km through various optical fiber patch panels including the one in GSFC’s building 201 which serves to meet several other GGAO requirements, and then out approximately 500m to a dedicated Extreme Network 5i GE switch in the GGAO’s antenna trailer. The HECN-managed “pluto” 867-MHz (megaHz) Macintosh G4 workstation is locally attached to the latter GE switch and has been used extensively in partial and end-to-end checkout of the e-VLBI path, including those tests described in the following section. Also locally attached to the latter GE switch are two MIT/Haystack-provided Mark 5 units.

9. Testing and Performance

In preparation for the e-VLBI demonstration network performance testing was conducted. The purpose of the network performance testing was to determine the maximum end system to end system performance possible across the network path to be used for the e-VLBI demonstration.

This testing demonstrated the many issues associated with the conduct of high speed end to end data communications across high bandwidth-delay product networks. The combination of high bandwidth and long round trip times can be problematic for reliable transport protocols such as TCP. These issues revolve around hosts and TCP congestion control responses to packet loss and reordering. The parameters across the WAN are significantly different compared to that observed in lab testing. The result is that applications that perform well in the lab sometimes have less success when operated across the wide area.

9.1 Required Network Tuning

TCP flows on the order of 900 Mbits/s were required to support the e-VLBI application. With the proper host tuning, this was achieved. The host and network tuning required to achieve these results are summarized below:

- Path MTU Discovery. Host systems must utilize Path MTU Discovery per RFC1191. This allows hosts to discover the optimum MTU size to use for the current end to end path. For the e-VLBI testing the MTU size available was 9000 bytes.

- Large Window Extensions. TCP includes definition of a window parameter which defines how much data may be in the network pipe prior to receipt of specific acknowledgement. The bandwidth-delay product of the network path determines the optimum value of this parameter. For the e-VLBI testing this value was high enough that host system support of RFC1323 "Large Windows" extensions to TCP was required.

- Large Socket Buffers. The host system must support large enough socket buffers for reading and writing data to the network. These values needed to be in line with the bandwidth-delay product. For the e-VLBI testing a conservative value of 3-8 MB (megabytes) was used. Typical host default values are 64 kB (kilobytes).

- TCP Selective Acknowledgments (SACK). TCP SACK (RFC2018) allows for greater tolerance to packet reordering. Data communications across the wide area network results in increased packet reorder events so this feature is more important as compared to local area communications.
9.2 Example End-to-End Traceroute

The end to end path consisted of hosts connected with gigabit ethernet interface cards, gigabit ethernet switches, and routers connected to an OC48 network. A typical traceroute is shown below:

```
Traceroute:
MIT/LL ---> NASA Goddard
[dartnoc@superglide dartnoc]$ /usr/sbin/traceroute 206.196.178.53
traceroute to 206.196.178.53(206.196.178.53),30 hopsmax,38 byte packets
1 snet_mitll1.cairn.net(140.173.174.1) 0.302 ms 0.210 ms 0.190 ms
2 snet_mitll24.cairn.net(140.173.174.24) 9.925 ms 9.897 ms 9.883 ms
3 arlg-supernet.maxgigapip.net(206.196.177.117)9.950ms 9.953ms 9.928ms
4 clpk-so3-0-0.maxgigapop.net(206.196.178.37)10.410ms 10.436ms 10.387ms
5 clpk-3508.maxgigapop.net(206.196.178.53)10.581ms 10.568ms 10.570 ms
```

9.3 Performance Testing and Tools

Testing was conducted to determine if the sustained end-to-end TCP throughput of the system was sufficient to support the e-VLBI demonstration requirement of ~900 Mb/s during a 30 to 60 minute window. Note that there were no significant requirements on latency and jitter as this particular demonstration did not utilize real-time correlation. Previous testing done during e-VLBI platform selection had shown that directly connected e-VLBI hosts were able to sustain a 990 Mb/s throughput rate using 9000 byte jumbo frames.

Testing of network performance was conducted using three tools:

ping – (short for Packet InterNet Groper) a standard Unix tool for measuring round-trip response time (RTT)

iperf - ([http://dast.nlanr.net/Projects/Iperf/](http://dast.nlanr.net/Projects/Iperf/)). This application provides a basic test of the end to end performance. The data transfer is from host memory to host memory, so minimum host resources (disk drives, memory copies, CPU processing) are exercised. Results achieved with iperf are generally best case, since real applications will require disk access, memory copies, and additional application CPU utilization. Below is a typical iperf output from a network performance test.

```
Performance Test Result:
MIT/LL (Boston, MA) ---> ISI East (Arlington, VA)
[root@kame root]# iperf -s -w 8M -i 5
------------------------------------------------------------
Server listening on TCP port 5001
TCP window size: 16.0 MByte (WARNING: requested 8.0 MByte)
------------------------------------------------------------
[  6] local 140.173.180.5 port 5001 connected with 140.173.174.10 port
32822
[ ID] Interval Transfer Bandwidth
[  6]  0.0- 5.0 sec  577 MBytes  922 Mbits/sec
[  6]  5.0-10.0 sec  584 MBytes  935 Mbits/sec
[  6] 10.0-15.0 sec  584 MBytes  935 Mbits/sec
[  6] 15.0-20.0 sec  584 MBytes  935 Mbits/sec
[  6] 20.0-25.0 sec  584 MBytes  935 Mbits/sec
[  6] 25.0-30.0 sec  584 MBytes  934 Mbits/sec
[  6]  0.0-30.1 sec  3.4 GBytes  933 Mbits/sec
```

interconnecting network to a destination system, either transferring a specified number of buffers or alternatively transferring data for a specified time interval. In addition to reporting the achieved network throughput in Mbps, nuttcp also provides additional useful information related to the data transfer such as user, system, and wall-clock time, transmitter and receiver CPU utilization, and loss percentage (for UDP transfers).

9.4 Summary of Test Results

The path traversed by the e-VLBI data passed through the following nodes:

MIT Haystack ↔ MIT/LL ↔ ISI/East ↔ MAX_GigaPOP ↔ NASA/GSFC ↔ GGAO

The infrastructure along the path, as shown in Figure 6, consisted of four Juniper routers, multiple Extreme GigE switches supporting jumbo frames, and OC-48 POS and GigE links.

Network performance testing was conducted in two stages. Initial tests determined performance between ISI-E and MIT Haystack. Subsequent tests then determined network performance between GSFC and MIT Haystack.

The testing involved using the test programs ping, iperf and nuttcp to conduct RTT (round-trip time) and throughput measurements between several workstations distributed along the e-VLBI path. The workstation specifications and their network locations are detailed in Table 1. All workstation OS network parameters were pre-configured for performance over high bandwidth-delay paths, as listed in Table 2. Test runs typically entailed measuring the RTT with ping, calculating the send and receive buffer sizes based on the BDP (Bandwidth-Delay Product), and then running the throughput test program using these buffer sizes over a 30 minute window with 5 sec averaging. The setup and results from four test runs (designated A through D) are shown in Table 3 and 4. These results show an asymmetry with respect to network direction but fortunately it is biased favorably in the e-VLBI data direction; the cause of the asymmetry is likely due to details of the transmit and receive optical paths, parameter tuning and CPU loading (see Section 9.5), but was not fully investigated. Careful characterization would certainly require further investigation but the motivation here was to determine sufficiency. The final test run, D, was performed between system endpoints. Measurements were made for 30 minutes in one direction using 30 sec averages and then the reverse direction was measured similarly. This pattern was repeated for approximately 10 hours. The results are plotted in Figure 13.

<table>
<thead>
<tr>
<th>Workstation Designator</th>
<th>Location</th>
<th>System Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>ws1 (evlbihay)</td>
<td>Haystack Obs</td>
<td>Pentium III 1.1GHz, ServerWorks HE-SL, 512MB</td>
</tr>
<tr>
<td></td>
<td>(Westford, MA)</td>
<td>SysKonnect SK-9834SX PCI 64b/66MHz GigE NIC Linux 2.4.9-12 (RedHat 7.1)</td>
</tr>
<tr>
<td>ws2 (evlbill)</td>
<td>Lincoln Lab</td>
<td>Pentium III 1.0GHz, ServerWorks LE 3.0, 256MB</td>
</tr>
<tr>
<td></td>
<td>(Lexington, MA)</td>
<td>SysKonnect SK-9843SX PCI 64b/66MHz GigE NIC Linux 2.4.9-12 (RedHat 7.1)</td>
</tr>
<tr>
<td>ws3 (superglide)</td>
<td>Lincoln Lab</td>
<td>Dual PIII Xeon 1.0GHz</td>
</tr>
<tr>
<td></td>
<td>(Lexington, MA)</td>
<td>SysKonnect SK-9843SX PCI 64b/66MHz GigE NIC</td>
</tr>
<tr>
<td>ws4 (kame)</td>
<td>ISI-E</td>
<td>Dual PIII Xeon 733MHz, 1GB</td>
</tr>
<tr>
<td></td>
<td>(Arlington, VA)</td>
<td>SysKonnect SK-9843SX PCI 64b/66MHz</td>
</tr>
<tr>
<td>ws5 (pluto)</td>
<td>NASA/GSFC</td>
<td>Apple Mac G4 PowerPC 867MHz, 640MB</td>
</tr>
<tr>
<td></td>
<td>GGAO antenna</td>
<td>NetGear GA620T PCI 64b/33MHz GigE NIC</td>
</tr>
<tr>
<td></td>
<td>(Greenbelt, MD)</td>
<td>Linux 2.4.17-pre2-ben0 (YellowDog 2.1)</td>
</tr>
</tbody>
</table>

Table 1: Test Workstation Specifications
### Table 2: Parameter Settings for Performance Tests

<table>
<thead>
<tr>
<th>Setup Ref</th>
<th>Parameter Settings</th>
<th>Setting</th>
</tr>
</thead>
<tbody>
<tr>
<td>TCP SACK</td>
<td>TCP Selective Acknowledgements (SACK, RFC 2018)</td>
<td>Enabled</td>
</tr>
<tr>
<td>TCP Extension</td>
<td>TCP Extensions for High Performance (RFC1323)</td>
<td>Enabled</td>
</tr>
<tr>
<td>Path MTU</td>
<td>Path MTU Discovery (RFC 1191)</td>
<td>Enabled</td>
</tr>
<tr>
<td>TCP Timestamps</td>
<td>TCP Timestamps</td>
<td>Enabled</td>
</tr>
<tr>
<td>Maximum socket buffer size (read and write)</td>
<td>Maximum socket buffer size (read and write)</td>
<td>8 MB</td>
</tr>
<tr>
<td>Maximum MTU</td>
<td>Maximum MTU</td>
<td>9000 bytes</td>
</tr>
</tbody>
</table>

### Table 3: Test Parameters for test runs A through D

<table>
<thead>
<tr>
<th>Setup Ref</th>
<th>Test</th>
<th>MTU (bytes)</th>
<th>MSS (bytes)</th>
<th>RTT (msec)</th>
<th>buffer size (MB)</th>
<th>duration (min)</th>
<th>avg period</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>iperf</td>
<td>4470</td>
<td>4418</td>
<td>13</td>
<td>2.5</td>
<td>30</td>
<td>5</td>
</tr>
<tr>
<td>B</td>
<td>iperf</td>
<td>4470</td>
<td>4418</td>
<td>15</td>
<td>4.0</td>
<td>30</td>
<td>5</td>
</tr>
<tr>
<td>C</td>
<td>iperf</td>
<td>4470</td>
<td></td>
<td></td>
<td>8.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>nuttcp</td>
<td>4448</td>
<td>4418</td>
<td>16</td>
<td>4.0</td>
<td>600</td>
<td>1</td>
</tr>
</tbody>
</table>

### Table 4: Test-run TCP Throughput (setup references in parenthesis)

<table>
<thead>
<tr>
<th>From</th>
<th>To</th>
<th>ws1</th>
<th>ws2</th>
<th>ws3</th>
<th>ws4</th>
<th>ws5</th>
</tr>
</thead>
<tbody>
<tr>
<td>ws1</td>
<td>980 (A)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ws2</td>
<td>980 (A)</td>
<td>940 (B)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ws3</td>
<td></td>
<td>933 (C)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ws4</td>
<td>965 (B)</td>
<td>932 (C)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ws5</td>
<td>960 (D)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 13: Throughput vs. time for test run D

Figure 14 shows further details of the test of Figure 13 as captured by an Extreme 5i switch at NASA/GSFC in the data path, which also includes two additional 2-hour tests (from roughly 10:30 to 14:30 on the time axis), as follows:

1. A sustained 2-hour transmission from pluto at GSFC/GGAO to superglide at MIT/LL specifying a TCP window size of 2 MB. The following nuttcp analysis shows an average sustained data rate of ~931 Mbps, with a 92% CPU utilization on the transmitter (pluto) and 86% CPU utilization on the receiver (superglide):
pluto% nuttcp -t -T120m -w2048 -b superglide
800984.938 MB / 7215.58 sec = 931.1997 Mbps 92 %TX 86 %RX

2. A sustained 2-hour transmission in the opposite direction, namely from superglide at MIT/LL to pluto at GSFC/GGAO, again with a TCP window size of 2 MB. The following nuttcp analysis shows an average sustained data rate of ~906 Mbps, with a 83% CPU utilization on the transmitter (superglide) and 36% CPU utilization on the receiver (pluto):

```plaintext
pluto% nuttcp -r -T120m -w2048 -b superglide
778104.688 MB / 7200.78 sec = 906.4593 Mbps 83 %TX 36 %RX
```

During each of these 2-hour tests, ~0.8 TB (terabytes) of data were moved across the network!

![Traffic analysis of sustained tests](image)

**Figure 14**: Traffic analysis of sustained tests (from Extreme 5i switch in data path)

### 9.5 Examples of Problems, Solutions, and Lessons Learned

A number of problems were encountered and corrected before it was possible to achieve the goal of near Gbps network performance. Initially, while the performance from MIT/LL to GGAO was excellent, the GGAO to MIT/LL link, which is actually the crucial direction for the e-VLBI scientific experiments, was non-functional when using large MTU transmissions. This was tracked down to an MTU mismatch between the two ends of one of the OC-48 POS links in the end-to-end network path.

Once this problem was solved, it was possible to transfer data successfully from GGAO to MIT/LL, but the network performance was still very poor, ranging from only 100 to 200 Mbps. This problem turned out to be much more interesting in terms of the lessons learned in trying to obtain very high network performance across wide area networks. In trying to isolate the performance problem, it was determined that excellent performance was possible from GGAO to ISI/East and also from ISI/East to MIT/LL, but yet the performance from GGAO to MIT/LL was poor. This illustrates that network performance is not transitive, i.e. just because the performance from A to B is good and the performance from B to C is good, performance from A to C will not necessarily be good.

Further investigation involved checking each link in the path from GGAO to MIT/LL, and revealed that there were many FCS (Frame Check Sequence) errors on the link from the MAX router at UMCP to the MAX router at ISI/East, and that these errors correlated extremely well with the number of TCP retransmissions seen when performing the network performance testing. This knowledge explained the non-transitive nature of the network performance results being observed. The path from ISI/East to MIT/LL did not include the problematic link and so had excellent performance. And while both the GGAO to ISI/East and GGAO to MIT/LL paths would encounter the problematic link with the resultant FCS errors and TCP retransmissions, the effect on the network performance in the two cases was quite different. This is because there is an order of magnitude difference in the round trip times for the two paths. The GGAO to ISI/East path is a metropolitan area connection with a RTT of about one millisecond, while the RTT for the GGAO to MIT/LL wide area network path is about ten milliseconds. Both paths experienced the FCS errors and resultant TCP retransmissions, but since TCP only increases
its send window linearly after a timeout/retransmission, the path with the ten millisecond RTT takes a far
greater performance hit than the path with the one millisecond RTT. Basically the path from GGAO to
ISI/East was obtaining good performance in spite of the problematic link in the path, with the small RTT
limiting the damage to the network performance.

Once the problem was isolated, a workaround was devised. The MAX core network is a ring consisting
of three nodes, one at UMCP, one at ISI/East, and one at Eckington in DC (where it connects to the
Abilene network). The workaround simply consisted of rerouting the GGAO to ISI/East and GGAO to
MIT/LL traffic the long way around the ring. Instead of taking the direct path from UMCP to ISI/East, it
took the path from UMCP to Eckington and then from Eckington to ISI/East. It was verified that this
path was clean, and the network performance tests were rerun from GGAO to MIT/LL, and they then
finally achieved excellent end-to-end network performance. Subsequent to this initial testing, the link
from the MAX UMCP router to the MAX ISI/East router was cleaned up, and the network paths switched
back to using that direct link.

The bottom line is that, though all components along the path from Haystack to GGAO were each capable
of near Gbps performance, getting them to work together to provide an end-to-end path at a Gbps was not
an easy task and consumed many hours of testing, configuring and re-testing.

Workstation Limitations

Some of the limitations that surprised us somewhat during the course of the testing were due to the end-
terminal workstations themselves. As indicated in Section 9.4, CPU utilization doing nothing more than
testing the data paths at speeds of ~900 Mbps were often above 80%, particularly on the transmitting
workstation. Though all test workstations were chosen to be quite fast, with care taken to choose chipsets
that were known for good performance, this situation was observed nearly across the board including
Pentium, Xeon, Alpha and G4 processors. Limitations of bus speeds, along with required software
management of the TCP/IP stack, are believed to be the primary reasons for these limitations, though no
in-depth investigations were performed.

As always, the march of technology will likely overcome this limitation. Faster bus and CPU speeds will
help to some degree, but hardware management of TCP/IP stacks on NIC cards will probably make the
largest difference. Such NIC cards are just beginning to become available, primarily for Windows
platforms, but Linux support is expected to follow in a few months.

10. The e-VLBI Experiments

Near-real-time e-VLBI Experiment on 4 October 2002

On 4 October 2002, after many months of preparation, the first high-data-rate near-real-time experiment
was successfully conducted using the antennas at Westford and GGAO. The experiment was conducted
as follows:

1. Each antenna simultaneously recorded data onto the disks of two Mark 5P VLBI data systems,
each operating at 576 Mbps, for a total aggregate data rate at each antenna of 1152 Mbps. Data
were simultaneously collected on 10 X-band (radio wavelength ~3.5 cm) channels and 6 S-band
channels (radio wavelength ~13 cm), where each channel received a bandwidth of 16 MHz. The
observations were taken on extra-galactic quasar 4C39.25. Hydrogen masers were used at each
site as a frequency reference, with timing synchronization derived from GPS.

2. The data from the two Mark 5 systems at the GGAO antenna were then simultaneously
transferred to Haystack Observatory at an average sustained data rate of 788 Mbps over the
~700km link from Maryland to Haystack, where they were re-recorded on disks. A standard
TCP/IP transfer was used.

3. Similarly, the data from the two Mark 5 systems at the Westford antenna were then
simultaneously transferred the short distance to Haystack Observatory at an average sustained
data rate of ~1 Gbps, where they were re-recorded on disks. A standard TCP/IP transfer was used.

4. The data were then correlated on the Mark 4 VLBI correlator system at Haystack Observatory. The results of this correlation showed the data to be entirely normal.

Figure 15 shows the results of the correlation of one X-band channel in the standard output format of the Mark 4 correlator. The upper panel with the narrow peak shows the ‘fringe rate’ spectra, which is a measure of the RF coherence of the observations and is a good indicator that the system, including data recording and processing, is working well; the peak of this plot shows that the correlation amplitude is ~2x10^-4, which is the expected amplitude for this source with these antennas. The other plots show the data in different dimensions; for example, the lower-left plot shows the signal strength as a function of delay and the lower-right plot shows the cross-power spectrum over a 32 MHz-wide band consisting of adjacent upper-sideband and lower sideband channels. The small correlation amplitude reflects the fact that most extragalactic radio sources are very weak compared to the noise in the receiver systems.

Figure 15: Correlation results of e-VLBI test observation

Direct-data-transfer e-VLBI experiment on 24 October 2002

On 24 October 2002 another e-VLBI experiment was performed with one major difference. In this experiment, data from the GGAO antenna was transferred in real time at 288 Mbps to Haystack Observatory, where it was recorded on disks. The data rate was constrained by the fact that the experiment had to be performed as an add-on to a pre-scheduled operational experiment in which the Westford antenna was participating at 288 Mbps. The GGAO data were correlated in the normal manner with the Westford data and normal fringes were obtained.

11. Summary and Suggestions for Future Work

After several months of quite concentrated work, we were able to execute a successful near-Gbps e-VLBI experiment with data transfer from both the Westford antenna in Massachusetts and the GGAO antenna in Maryland to the VLBI correlator at MIT Haystack Observatory. The path involved a combination of dedicated and shared network segments, crossing through many routers and switches, all of which were nominally capable of ~Gbps operation. However, much experimentation, analysis and tuning was actually required to achieve the desired data rate. We believe that this situation is probably the typical state of affairs for many similar networks in 2002. Extreme attention must be paid to detailed configurations, equipment must be capable of handling ‘jumbo frames’, and optical fiber links must be
maintained in top physical condition. As networks mature, no doubt operation at Gbps and higher will become commonplace, but our work illustrates that it cannot be taken for granted.

Though straight TCP was used for this work, it is clear that e-VLBI and other similar applications could benefit from protocols that are especially designed to take advantage of the special characteristics of the real-time and near-real-time scientific instrumentation applications. For example, the following characteristics apply to e-VLBI and may apply to other similar applications:

- High bandwidth
- High overall volume
- Continuous-data nature
- Relative insensitivity to small data losses
- Relatively high tolerance for latency
- Extreme cost sensitivity

One suggestion is to pursue new protocols which *scavenge* unused low-priority bandwidth, which is currently available in relative abundance on many research networks. Recently MIT Haystack Observatory has received a grant from the National Science Foundation to pursue such work in collaboration with the MIT Laboratory for Computer Science and MIT Lincoln Laboratory. Other proposals are pending.

Because e-VLBI is normally international in nature, it is critical that both national and international high-speed research networks are connected. Work is progressing in this area.

Because e-VLBI is so well suited as an application to high-speed networks, we look forward to its rapid development and deployment over the next few years. Other similar applications will also undoubtedly appear as well. The work accomplished in this small project is a first step in that direction.

12. Acknowledgements

The e-VLBI demonstration could not have been accomplished without the support and assistance of many. DARPA, NASA and the US Air Force provided financial support and resources in support of this effort, with special thanks to Dr. Mari Maeda of DARPA for providing seed support. MIT Lincoln Laboratory, NASA/GSFC, ISI-E and MAX all generously contributed much time and effort. Additional thanks is due to Jim Calvin and Leslie Weiner of MIT/LL, Dr. John LaBrecque of NASA/HQ, Herbert Durbeck, Jeff Martz and Jay Redmond of NASA/GSFC, Dan Magorian of Univ. of Maryland, and Dr. John Ball, Dr. Brian Corey, Richard Crowley, Anne Gorczyca, Mike Poirier and Mike Titus of MIT Haystack Observatory.
## Glossary

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>BDP</td>
<td>Bandwidth-Delay Product: an important quantitative measure of a particular network connection</td>
</tr>
<tr>
<td>Bosnet</td>
<td>‘BoSton-South NETwork’: a research network supported by DARPA and operated by MIT/LL</td>
</tr>
<tr>
<td>cm</td>
<td>centimeter</td>
</tr>
<tr>
<td>CPU</td>
<td>Central Processing Unit</td>
</tr>
<tr>
<td>CWDM</td>
<td>Coarse Wave Division Multiplex; a method of placing multiple independent wavelength bands on an optical fiber</td>
</tr>
<tr>
<td>DARPA</td>
<td>Defense Advanced Research Projects Agency</td>
</tr>
<tr>
<td>e-VLBI</td>
<td>Same as VLBI, but with electronic transmission of data</td>
</tr>
<tr>
<td>FCS</td>
<td>Frame Check Sequence: Part of the cyclic redundancy check in TCP data frames</td>
</tr>
<tr>
<td>Gbps</td>
<td>Gigabit per second</td>
</tr>
<tr>
<td>GGAO</td>
<td>Goddard Geophysical and Astronomical Observatory: a 10-m diameter antenna near NASA/GSFC</td>
</tr>
<tr>
<td>Glownet</td>
<td>‘Gigabit Lincoln Optical WDM Network’; a private fiber network connecting MIT Haystack Observatory and MIT Lincoln Laboratory</td>
</tr>
<tr>
<td>HECN</td>
<td>High End Computer Network: a research network operated by NASA/GSFC</td>
</tr>
<tr>
<td>IP</td>
<td>Internet Protocol</td>
</tr>
<tr>
<td>ISI</td>
<td>Information Sciences Institute; affiliated with Univ. of Southern California</td>
</tr>
<tr>
<td>kB</td>
<td>kilobytes</td>
</tr>
<tr>
<td>km</td>
<td>kilometers</td>
</tr>
<tr>
<td>m</td>
<td>meters</td>
</tr>
<tr>
<td>MAX</td>
<td>Mid-Atlantic Crossroads: a multi-state networking consortium connecting universities and research institutions in the greater Washington, D.C. area</td>
</tr>
<tr>
<td>Mbps</td>
<td>megabits per second</td>
</tr>
<tr>
<td>MB</td>
<td>megabytes</td>
</tr>
<tr>
<td>mm</td>
<td>millimeter</td>
</tr>
<tr>
<td>MHz</td>
<td>megaHz</td>
</tr>
<tr>
<td>MIT/LL</td>
<td>MIT Lincoln Laboratory: a research laboratory operated by MIT</td>
</tr>
<tr>
<td>MTU</td>
<td>Maximum Transmission Unit: maximum packet size, usually measured in bytes, that can be transmitted across a network link</td>
</tr>
<tr>
<td>NIC</td>
<td>Network Interface Card: a plug-in interface card to support networking protocols in personal computers</td>
</tr>
<tr>
<td>nm</td>
<td>nanometer</td>
</tr>
<tr>
<td>NRAO</td>
<td>National Radio Astronomy Observatory</td>
</tr>
<tr>
<td>OC-48</td>
<td>‘OC’ is a protocol associated with SONET networks; the base rate is (OC-1) is 51.84 Mbps. OC-48 corresponds to 48 times OC-1 or approximately 2.5 Gbps</td>
</tr>
<tr>
<td>RTT</td>
<td>Round-Trip Time: An important quantitative measure of a particular network connection</td>
</tr>
<tr>
<td>SEN</td>
<td>Scientific and Engineering Network: a network operated by NASA/GSFC</td>
</tr>
<tr>
<td>TB</td>
<td>Terabyte</td>
</tr>
<tr>
<td>TCP</td>
<td>Transport Control Protocol: a ISO layer 4 protocol, most often used on top of IP (layer 3 protocol)</td>
</tr>
<tr>
<td>UMCP</td>
<td>University of Maryland at College Park, MD</td>
</tr>
<tr>
<td>VLBI</td>
<td>Very-Long Baseline Interferometry</td>
</tr>
<tr>
<td>WAN</td>
<td>Wide-Area Network</td>
</tr>
<tr>
<td>WDM</td>
<td>Wavelength Division Multiplex: a method of placing multiple independent wavelength bands on an optical fiber</td>
</tr>
</tbody>
</table>
14. References

[BOSSNET] “Bossnet – BoSton-South NETwork”,
http://www.ll.mit.edu/AdvancedNetworks/bossnet.html, MIT Lincoln Laboratory.


