Verification of High Data Rate Bandwidth-on-Demand networks: User Based Test Equipment

Jimmy Cullen
Research Associate, University of Manchester
Contents

• NEXPReS
• Very Long Baseline Interferometry
• Test equipment
• BoD tests
• Conclusions
What is NEXPReS?

- NEXPReS = Novel EXplorations Pushing Robust e-VLBI Services
- Three years (start 1 July 2010)
- Funded through the European Community’s Seventh Framework Programme (FP7/2007-2013), Contract n°: RI-261525
- Budget: 5,745,000 € (EC contribution: 3,500,000 €)
- Objective: further improve the astronomy technique of electronic Very Long Baseline Interferometry (e-VLBI) with the objective of incorporating it into every experiment conducted by the European VLBI Network (EVN)
- Means: develop data caching and implement dynamically provisioned network resources to offer the best of both worlds: the data archiving and re-processing afforded by traditional disk-based VLBI and the speed and flexibility of e-VLBI
NEXPReS Partners

Coordinator
• Joint Institute for VLBI in Europe (JIVE), EU (The Netherlands)

National Astronomy Institutes
• The Netherlands Institute for Radio Astronomy (ASTRON), The Netherlands
• Istituto Nazionale di Astrofisica (INAF), Italy
• Max Planck Gesellschaft zur Foerderung der Wissenschaften E.V. (MPG), Germany
• The University of Manchester (UMAN), United Kingdom
• Chalmers Tekniska Hoegskola AB (OSO), Sweden
• Ventspils Augstskola (VENT), Latvia
• Fundación General de la Universidad de Alcalá, together with Instituto Geográfico Nacional (FG-IGN), Spain
• Aalto University Metsähovi Radio Observatory (AALTO), Finland
• Commonwealth Scientific and Industrial Research Organisation (CSIRO), Australia

NREN Providers and Advanced Computing Facilities
• NORDUnet A/S (NORDUnet), Denmark
• SURFnet bv (SURFnet), The Netherlands
• Poznan Supercomputing and Networking Center (PSNC), Poland
• Delivery of Advanced Network Technology to Europe Limited (DANTE), EU (United Kingdom)
• Technische Universität München (TUM), Germany
NEXPReS Infrastructure

- NEXPReS partner
  - radio telescope
- NREN partner
- Correlation facility
- Storage facility*
  - *also at all telescopes
<table>
<thead>
<tr>
<th>#</th>
<th>Description</th>
<th>Leader</th>
</tr>
</thead>
<tbody>
<tr>
<td>WP1</td>
<td>Management of the Consortium</td>
<td>T. Charles Yun, JIVE</td>
</tr>
<tr>
<td>WP2</td>
<td>EVN-NREN</td>
<td>Richard Hughes-Jones, DANTE</td>
</tr>
<tr>
<td>WP3</td>
<td>eVSAG</td>
<td>Francisco Colomer, FG</td>
</tr>
<tr>
<td>WP4</td>
<td>Communication</td>
<td>Kristine Yun, JIVE</td>
</tr>
<tr>
<td>WP5</td>
<td>Cloud Correlation</td>
<td>Arpad Szomoru, JIVE</td>
</tr>
<tr>
<td><strong>WP6</strong></td>
<td><strong>High Bandwidth on Demand</strong></td>
<td><strong>Paul Boven, JIVE</strong></td>
</tr>
<tr>
<td>WP7</td>
<td>Computing in a Shared Infrastructure</td>
<td>Mark Kettenis, JIVE</td>
</tr>
<tr>
<td>WP8</td>
<td>Provisioning High-Bandwidth, High-Capacity Networked Storage on Demand</td>
<td>Ari Mujunen, Aalto</td>
</tr>
</tbody>
</table>
Interferometry

- Angular resolution $\approx \frac{\text{wavelength}}{\text{collecting area diameter}}$
- For fixed wavelengths (radio), must increase collecting area (dish size) to gain higher resolution images
- Lovell Telescope (JBO) $D = 76m$
- Interferometry combines signals from two or more telescopes to generate images
- Resolution proportional to the largest separation of dishes, not individual dish size

$\sin \theta \approx 1.22 \frac{\lambda}{D}$
Very Long Baseline Interferometry (VLBI)

- Distance between individual telescopes described as baselines
- Baselines are often hundreds or thousands of km - transglobal
VLBI and eVLBI

- Large volumes of data recorded at each telescope, typically 1 Gbps or higher for several hours (8 – 12 hours)
- Data are stored on hard disks, then shipped to a central correlator for processing
- Observation results can take weeks to be processed

- Electronic VLBI (eVLBI) uses dedicated light paths from observatories to the correlator
  - Allows faster processing of data (⇒ observation results)
  - Realtime monitoring of telescopes and error detection and correction
  - Enabled new science – Target of Opportunity (ToS) observations

- VLBI observations infrequent however, therefore inefficient use of the light paths (~ once a month)
Bandwidth on Demand and eVLBI

• BoD offers an ideal solution for eVLBI:
  ▫ Resources needed only for limited periods of time
  ▫ Dedicated links required since using UDP

• NEXPReS, the NRENs and GÉANT aim to provide a BoD service for eVLBI in Europe

• Essential to verify connectivity, bandwidth and packet loss on BoD links prior to observation
Test equipment

- Tests to evaluate different hardware approaches to characterise networks

- Evaluated two solutions for characterising networks
  - FPGAs
  - PCs
Test Network

- A dedicated 10 Gbps light path between Jodrell Bank Observatory and main university campus (Schuster Building, physics)
- ~30km direct distance, but fibre is ~80km
- No other users, uncontended bandwidth
Test Hardware - FPGA

- Interconnect Break-out Board (IBOB, from CASPER, Berkley)
- Xilinx Virtex-II Pro 2VP50 Field Programmable Gate Array
- 200MHz PowerPC
- 2 x CX4 10 Gbps
- 1 x RJ45 100 Mbps interfaces
Test Hardware - PCs

- Asus Crosshair IV Formula motherboard
- AMD Phenom(tm) II X6 1090T Processor
- 4 x Hynix 4GB DDR3 PC3-10600 (1333)
- Chelsio N310E-CXA 10GbE NIC and Myricom 10G-PCIE-8A-C NIC.
Test Software

• UDPmon* is a network diagnostic program which uses UDP datagrams to test endhost and network performance

• Client/server model written in C

• Measures many aspects of network communication including packet loss, packet reordering and variation in interpacket arrival times (jitter)

• iNetTest is a control framework designed to control the “gateware” in the IBOB in a manor compatible with UDPmon

*http://www.hep.man.ac.uk/u/rich/net/index.html
Test setup

- Two PCs and two IBOBs used
  - Manchester – jbnexpres1 & iNetTest1
  - JBO – jbnexpres2 & iNetTest2
Received Wire Rate

- Received wire rates measured for 10 different packet sizes
- At small requested inter-packet separation, steady data rate
- As separation increases, data rate decreases (inverse power law)
Received Wire Rate (2)

- Both hardware achieved > 9.9 Gbps data rates
- At smaller packet sizes, the hardware must process more packets per second to achieve high data rates
- IBOBs were able to achieve higher data rates than PCs for small packet sizes
- PCs achieved higher data rates than IBOBs
  - IBOB has a 200 MHz processor $\Rightarrow$ 5 ns temporal resolution
  - PC has a 3.2 GHZ processor $\Rightarrow$ ~0.3 ns temporal resolution
# Received Wire Rate (3)

- **IBOB to IBOB**

<table>
<thead>
<tr>
<th>Packet Size (bytes)</th>
<th>Number of Packets</th>
<th>Time between sending packets (nanoseconds)</th>
<th>Requested inter-packet delay (microseconds)</th>
<th>Mean Time Between receiving successive packets (nanoseconds)</th>
<th>Received Wire Rate (Mbps)</th>
</tr>
</thead>
<tbody>
<tr>
<td>8232</td>
<td>100000</td>
<td>6643.05</td>
<td>0</td>
<td>6643.13</td>
<td>9944.7</td>
</tr>
<tr>
<td>8232</td>
<td>100000</td>
<td>6643.05</td>
<td>1</td>
<td>6643.13</td>
<td>9944.7</td>
</tr>
<tr>
<td>8232</td>
<td>100000</td>
<td>6643.05</td>
<td>2</td>
<td>6643.13</td>
<td>9944.7</td>
</tr>
<tr>
<td>8232</td>
<td>100000</td>
<td>6643.05</td>
<td>3</td>
<td>6643.13</td>
<td>9944.7</td>
</tr>
<tr>
<td>8232</td>
<td>100000</td>
<td>6643.06</td>
<td>4</td>
<td>6643.13</td>
<td>9944.7</td>
</tr>
<tr>
<td>8232</td>
<td>100000</td>
<td>6643.06</td>
<td>5</td>
<td>6643.13</td>
<td>9944.71</td>
</tr>
<tr>
<td>8232</td>
<td>100000</td>
<td>6643.06</td>
<td>6</td>
<td>6643.13</td>
<td>9944.7</td>
</tr>
<tr>
<td>8232</td>
<td>100000</td>
<td>6643.06</td>
<td>7</td>
<td>6643.13</td>
<td>9944.7</td>
</tr>
<tr>
<td>8232</td>
<td>100000</td>
<td>6643.06</td>
<td>8</td>
<td>6643.13</td>
<td>9944.7</td>
</tr>
<tr>
<td>8232</td>
<td>100000</td>
<td>6969.93</td>
<td>9</td>
<td>6969.95</td>
<td>9478.41</td>
</tr>
<tr>
<td>8232</td>
<td>100000</td>
<td>7969.92</td>
<td>10</td>
<td>7969.94</td>
<td>8289.15</td>
</tr>
</tbody>
</table>
## Received Wire Rate (4)

- **PC to PC**

<table>
<thead>
<tr>
<th>Packet Size (bytes)</th>
<th>Number of Packets</th>
<th>Time between sending packets (nanoseconds)</th>
<th>Requested inter-packet delay (microseconds)</th>
<th>Mean Time Between receiving successive packets (nanoseconds)</th>
<th>Received Wire Rate (Mbps)</th>
</tr>
</thead>
<tbody>
<tr>
<td>8232</td>
<td>100000</td>
<td>6001.6</td>
<td>0</td>
<td>6649.52</td>
<td>9983.28</td>
</tr>
<tr>
<td>8232</td>
<td>100000</td>
<td>6288.5</td>
<td>1</td>
<td>6655.61</td>
<td>9974.14</td>
</tr>
<tr>
<td>8232</td>
<td>100000</td>
<td>6314.3</td>
<td>2</td>
<td>6654.99</td>
<td>9975.07</td>
</tr>
<tr>
<td>8232</td>
<td>100000</td>
<td>6084.0</td>
<td>3</td>
<td>6652.61</td>
<td>9978.64</td>
</tr>
<tr>
<td>8232</td>
<td>100000</td>
<td>6255.2</td>
<td>4</td>
<td>6656.12</td>
<td>9973.38</td>
</tr>
<tr>
<td>8232</td>
<td>100000</td>
<td>6180.4</td>
<td>5</td>
<td>6655.21</td>
<td>9974.74</td>
</tr>
<tr>
<td>8232</td>
<td>100000</td>
<td>6117.9</td>
<td>6</td>
<td>6656.16</td>
<td>9973.32</td>
</tr>
<tr>
<td>8232</td>
<td>100000</td>
<td>7024.4</td>
<td>7</td>
<td>7027.83</td>
<td>9445.87</td>
</tr>
<tr>
<td>8232</td>
<td>100000</td>
<td>8035.6</td>
<td>8</td>
<td>8036.74</td>
<td>8260.07</td>
</tr>
<tr>
<td>8232</td>
<td>100000</td>
<td>9019.4</td>
<td>9</td>
<td>9020.85</td>
<td>7358.95</td>
</tr>
<tr>
<td>8232</td>
<td>100000</td>
<td>10015.2</td>
<td>10</td>
<td>10010.6</td>
<td>6631.37</td>
</tr>
</tbody>
</table>
Jitter

- Defined here as the variability in inter-packet reception times
- Measured jitter for requested inter-packet delays of zero and 50 μs
- Histogram bin widths:
  - IBOBs – 5 nanoseconds
  - PCs – 1 microsecond
Jitter (1)

- Zero inter-packet delay requested

Range ~ 150 nanoseconds
FWHM ~ 15 nanoseconds

Range ~ 60 microseconds
FWHM ~ 8 microseconds
Jitter (2)

- 50 μs inter-packet delay requested

Range ~ 400 nanoseconds
FWHM ~ 80 nanoseconds

Range ~ 70 microseconds
FWHM ~ 8 microseconds
## Jitter (3)

<table>
<thead>
<tr>
<th></th>
<th>Server</th>
<th>Histogram bin width (µs)</th>
<th>Requested inter-packet delay (µs)</th>
<th>Achieved inter-packet delay</th>
<th>Achieved inter-packet delay</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Min (µs)</td>
<td>Max (µs)</td>
</tr>
<tr>
<td><strong>PC - PC</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>jbnexpres1</td>
<td>jbnexpres2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>83</td>
</tr>
<tr>
<td>jbnexpres2</td>
<td>jbnexpres1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>50</td>
</tr>
<tr>
<td>jbnexpres1</td>
<td>jbnexpres2</td>
<td>1</td>
<td>50</td>
<td>3</td>
<td>89</td>
</tr>
<tr>
<td>jbnexpres2</td>
<td>jbnexpres1</td>
<td>1</td>
<td>50</td>
<td>2</td>
<td>83</td>
</tr>
<tr>
<td><strong>IBOB - IBOB</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JBO iBoB</td>
<td>Schu iBoB</td>
<td>0.005</td>
<td>0</td>
<td>6.620</td>
<td>6.805</td>
</tr>
<tr>
<td>Schu iBoB</td>
<td>JBO iBoB</td>
<td>0.005</td>
<td>0</td>
<td>6.620</td>
<td>6.815</td>
</tr>
<tr>
<td>JBO iBoB</td>
<td>Schu iBoB</td>
<td>0.005</td>
<td>50</td>
<td>48.370</td>
<td>48.775</td>
</tr>
<tr>
<td>Schu iBoB</td>
<td>JBO iBoB</td>
<td>0.005</td>
<td>50</td>
<td>48.365</td>
<td>48.805</td>
</tr>
</tbody>
</table>
Local Testing Conclusions

- Both IBOBs and PCs can transmit and receive UDP data at 10 Gbps without packet loss and packet reordering
  - PCs can send data marginally faster than IBOBs due to faster CPUs
  - IBOB data rates are more reproducible than PCs

- IBOBs have a very small jitter in inter-packet reception time
  - Deterministic behaviour

- PCs have larger jitter
  - General purpose machine that has other processes to control besides network communications
  - Data caching in socket buffers and on the NIC

- Standard deviation of PC jitter at 10 Gbps < length of time for packet delivery

- In conclusion, IBOBs are shown to be more accurate and reproducible in network testing, however they are expensive and hard to programme when compared to modern PCs. PCs, whilst less accurate, are of sufficient precision for our needs.
International BoD Links

- BoD service still experimental, not production
- GÉANT pilot BoD system uses Inter-Domain Controller (IDC) protocol
- Networks used in test a combination of BoD and static paths to connect endhosts
- Data rates of 4 Gbps used
International BoD Links

• Using the GÉANT core dynamic network, a BoD circuit was set up with endhosts in:
  ▫ JBO, UK
  ▫ London, UK
  ▫ Stockholm, Sweden
  ▫ Metsähovi, Finland

• Static light paths were provided by:
  ▫ JANET – JBO/Manchester to London
  ▫ GÉANT – Amsterdam to Copenhagen
  ▫ NORDUnet – Copenhagen to Stockholm
  ▫ NORDUnet & Funet – Copenhagen to Metsähovi
International BoD Links

NEXPReS BoD connectivity

Key
- STM-64/10GE
- 10GE port at
- Static 10GE - mapping
- NREN data path

NORDUnet conference - 18th September 2012 - Oslo, Norway
Network Test Website

- Graphical interface to run UDPmon tests between preconfigured endhosts
Network Test Website

Network Bandwidth Test

NORDUnet conference - 18th September 2012 - Oslo, Norway
AutoBAHN Reservation Interface

Start time  
End time  
Source  
Destination  
Bandwidth  
Packet size  

<table>
<thead>
<tr>
<th>State</th>
<th>Start time</th>
<th>End time</th>
<th>Start port</th>
<th>Start mode</th>
<th>Start VLAN</th>
<th>End port</th>
<th>End mode</th>
<th>End VLAN</th>
<th>Capacity [Mbps/s]</th>
<th>Pkt size [bytes]</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACTIVE (10)</td>
<td>Wed Mar 07 16:00:33 EET 2012</td>
<td>Fri Mar 09 17:20:00 EET 2012</td>
<td>GEANT London host at port 15/64 (GEANT.RIC.110)</td>
<td>VLAN 2002</td>
<td>VLAN</td>
<td>GEANT Connection to NORDUnet at Amsterdam port 15/91 (GEANT.RIC.90)</td>
<td>VLAN 2002</td>
<td>4000.5</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

Service: GEANT@1331129131500

Service: GEANT@1331128803316
Tests Performed

- Network test website used to verify connectivity of the BoD links, and measurement of available bandwidth and any packet loss

- Command line used to run longer running, more detailed tests
Received Wire Rate

Requested inter packet spacing against received wire rate

Received Wire Rate (Mbps)

Requested Spacing Between Frames (microseconds)

- 8972
- 8000
- 7000
- 6000
- 5000
- 4000
- 3000
- 2000
- 1472
- 1000
Packet Loss

Requested inter packet spacing against percentage packet loss

- 8972
- 8000
- 7000
- 6000
- 5000
- 4000
- 3000
- 2000
- 1472
- 1000
Jitter

- One million packets sent between hosts and jitter in inter-packet reception times recorded
- Three inter-packet delays used to generate different data rates
  - 72 microseconds – 1 Gbps
  - 36 microseconds – 2 Gbps
  - 18 microseconds – 4 Gbps

<table>
<thead>
<tr>
<th>Hosts</th>
<th>Requested inter packet delay (µs)</th>
<th>Mean inter packet delay (µs)</th>
<th>Mode inter packet delay (µs)</th>
<th>Standard deviation of inter packet delay (µs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stockholm – JBO</td>
<td>72</td>
<td>71.4832</td>
<td>72</td>
<td>2.14274</td>
</tr>
<tr>
<td></td>
<td>36</td>
<td>35.4732</td>
<td>34</td>
<td>2.60401</td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>17.4783</td>
<td>17</td>
<td>1.29231</td>
</tr>
<tr>
<td>JBO – Stockholm</td>
<td>72</td>
<td>71.4342</td>
<td>71</td>
<td>5.91037</td>
</tr>
<tr>
<td></td>
<td>36</td>
<td>35.4516</td>
<td>36</td>
<td>5.93888</td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>17.4664</td>
<td>17</td>
<td>5.35474</td>
</tr>
</tbody>
</table>
Jitter

Histogram of inter packet jitter
EXPReS2 - jbnexpres2

Histogram of inter packet jitter
jbnexpres2 - EXPReS2

Number of packets

Inter packet spacing (microseconds)

1 Gbps - 72 microsecond separation
2 Gbps - 36 microsecond separation
4 Gbps - 18 microsecond separation
24 Hour Tests

- UDPmon tests at 4 Gbps were run in each direction between JBO and Stockholm

<table>
<thead>
<tr>
<th>Client</th>
<th>Server</th>
<th>Number of packets Received</th>
<th>Number of packets lost</th>
<th>Number of packets Reordered</th>
<th>Bytes received</th>
<th>Mean received wire rate (Mbps)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stockholm</td>
<td>JBO</td>
<td>4794303918</td>
<td>0</td>
<td>18288</td>
<td>4.30145E+13</td>
<td>4012.810546</td>
</tr>
<tr>
<td>JBO</td>
<td>Stockholm</td>
<td>4794875087</td>
<td>2102</td>
<td>18292</td>
<td>4.30196E+13</td>
<td>4013.286133</td>
</tr>
</tbody>
</table>

- Stockholm → JBO bit error rate of less than 1 in $3.45 \times 10^{14}$
Conclusions

• BoD matches very closely the requirements of eVLBI observations

• PCs are suitable for network parameter verification up to 10 Gbps

• BoD links are:
  ▫ Stable
  ▫ Reliable
  ▫ Suitable for eVLBI operations

• However, important to verify links before experiments
Acknowledgements

- NetNorthWest – Anthony Ryan
- Metsähovi – Ari Mujunen, Tomi Salminen
- JANET – Dave Tinkler, David Salmon
- NORDUnet – Fredrik Pettai
- Funet - Jani Myyry
- DANTE/GÉANT
Questions/Comments

• Contact information
  Jimmy Cullen
  Research Associate
  The University of Manchester
  jcullen@jb.man.ac.uk

• Additional Information at http://nexpres.eu

• NEXPReS is an Integrated Infrastructure Initiative (I3), funded under the European Union’s Seventh Framework Programme (FP7/2007-2013) under grant agreement no RI-261525.
eVLBI

- Data has a constant bit rate
- Transport protocols
  - TCP
    - + Reliably transfer all data
    - - Large TCP buffer (window) required for long links
    - - Lost packets cause reduction window size and hence data rate
    - - Retransmission of lost packets and lower data rates cause large delay in data delivery
  - UDP
    - + No buffering
    - + Timely delivery of data
    - - Unreliable transfer
- UDP most suitable format for VLBI data since timely arrival is more important than small amounts of data loss
  - $SNR \propto \sqrt{(1 - \text{fractional loss})}$
- VLBI observations infrequent however, therefore inefficient use of the light paths (~ once a month)
GÉANT AutoBAHN

- Uses VLANs across the network
- Data streams are tagged with the appropriate VLAN tag
- VLAN tags dynamically altered based upon BoD path requests.
  - e.g. BoD Metsähovi to JBO:
  - Data are tagged with VLAN 2004 over Funet and NORDUnet, then altered to VLAN 2002 at GÉANT, which identifies the VLAN to JBO