

Achieving 98Gbps of Cross-country TCP traffic using 2.5 hosts, 10 x 10G NICs, and 10 TCP streams

Eric Pouyoul, Brian Tierney

ESnet

January 25, 2012

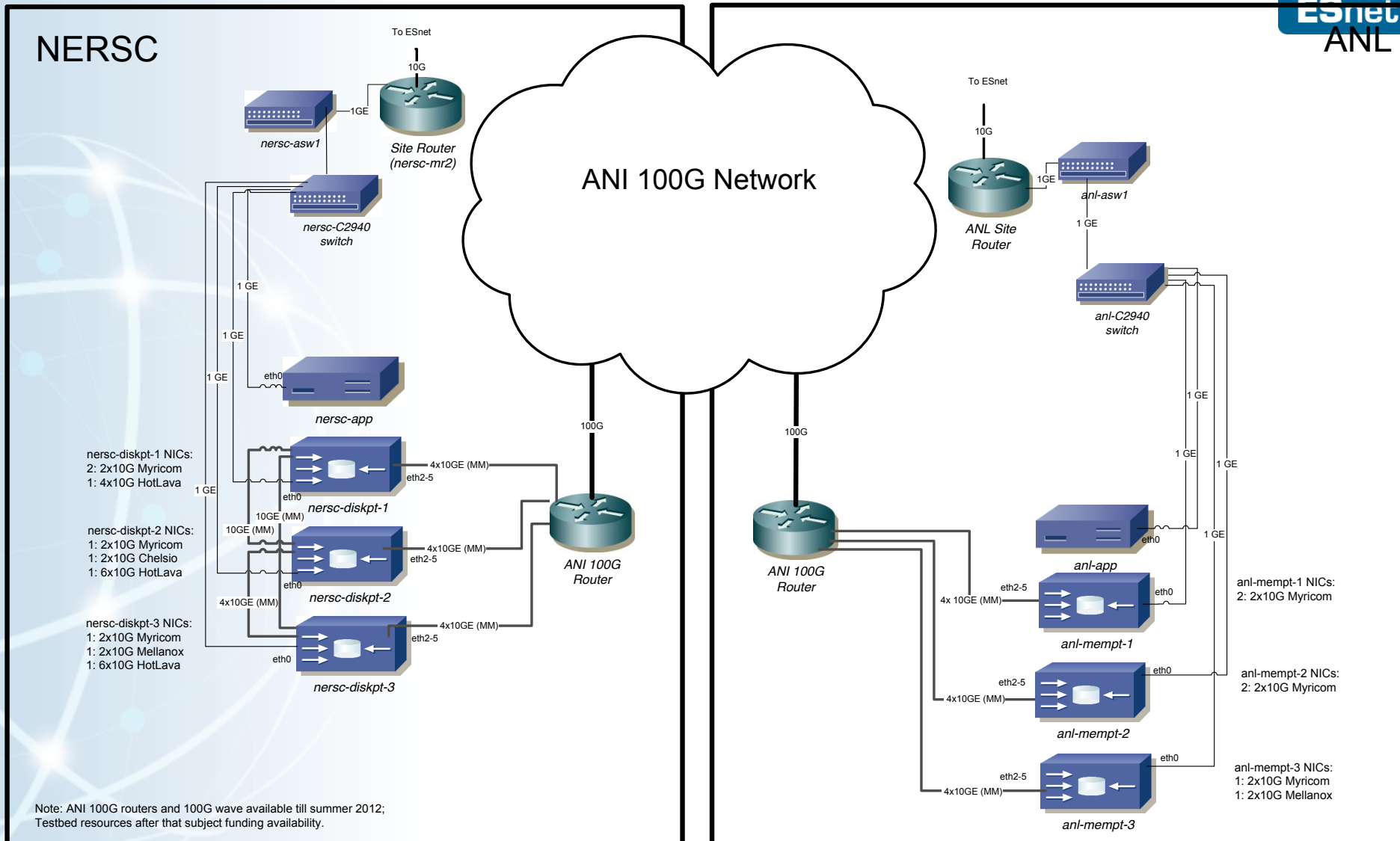


U.S. DEPARTMENT OF
ENERGY
Office of Science



ANI 100G Testbed

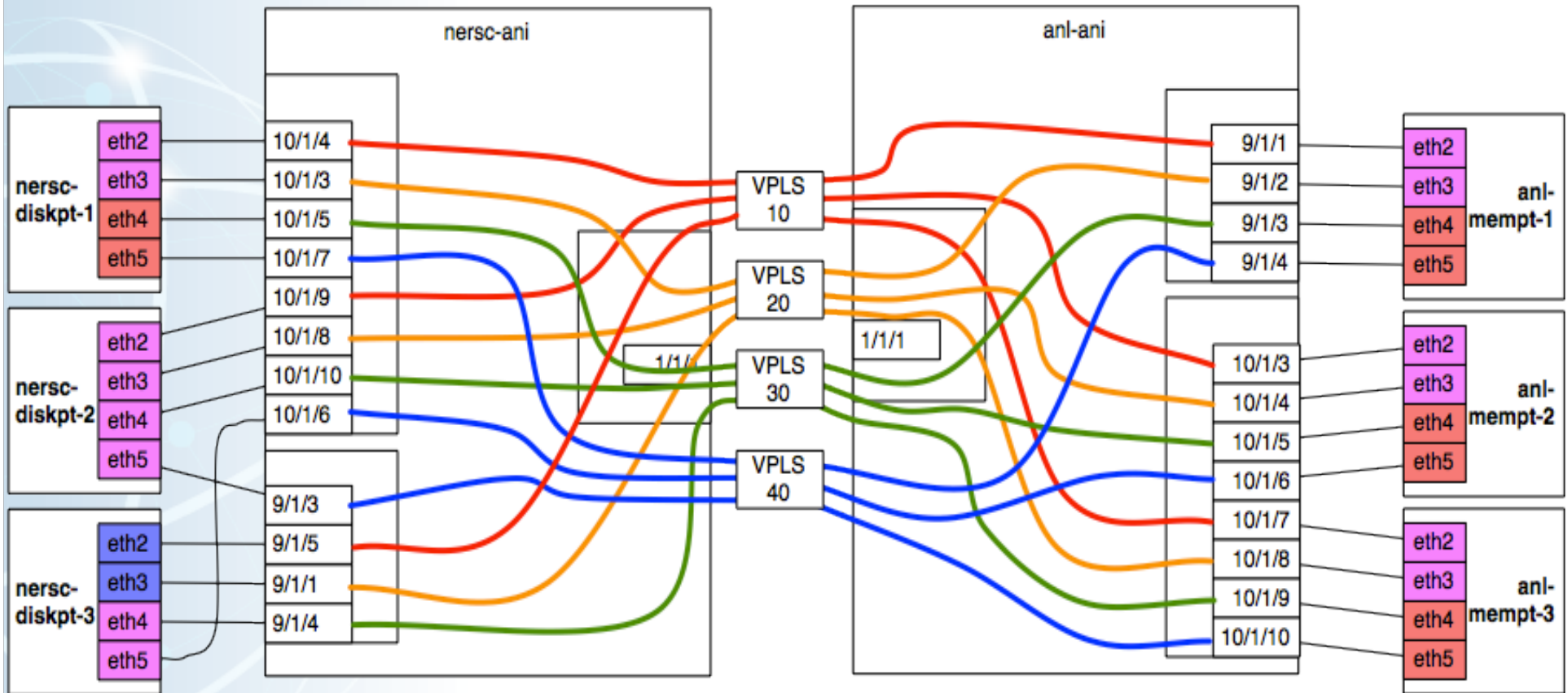
ANI Middleware Testbed



1/25/12

Updated December 11, 2011
2

Test Configuration Details



48.6ms RTT, 97.9Gbps aggregate TCP throughput* with 10 TCP streams

Per-Stream Results



nesc-diskpt-1-v4012:	1179.1875 MB /	1.00 sec =	9891.8010 Mbps	0 retrans
nesc-diskpt-1-v4013:	1179.2500 MB /	1.00 sec =	9888.4787 Mbps	0 retrans
nesc-diskpt-1-v4014:	1179.1875 MB /	1.00 sec =	9891.1482 Mbps	0 retrans
nesc-diskpt-1-v4015:	1179.1250 MB /	1.00 sec =	9891.1581 Mbps	0 retrans
nesc-diskpt-2-v4012:	1179.2500 MB /	1.00 sec =	9891.9494 Mbps	0 retrans
nesc-diskpt-2-v4013:	1179.0625 MB /	1.00 sec =	9891.1580 Mbps	0 retrans
nesc-diskpt-2-v4014:	1179.3750 MB /	1.00 sec =	9893.1365 Mbps	0 retrans
nesc-diskpt-2-v4015:	1179.1250 MB /	1.00 sec =	9891.0690 Mbps	0 retrans
nesc-diskpt-3-v4014:	1121.8750 MB /	1.00 sec =	9410.9602 Mbps	0 retrans
nesc-diskpt-3-v4015:	1121.8750 MB /	1.00 sec =	9410.9884 Mbps	0 retrans

	Input	Output
-----	-----	-----
Octets	18462079	12387383345
Packets	184615	1369129
Errors	0	0
Utilization (% of port capacity)	0.17	99.31

Hosts Must be Tuned



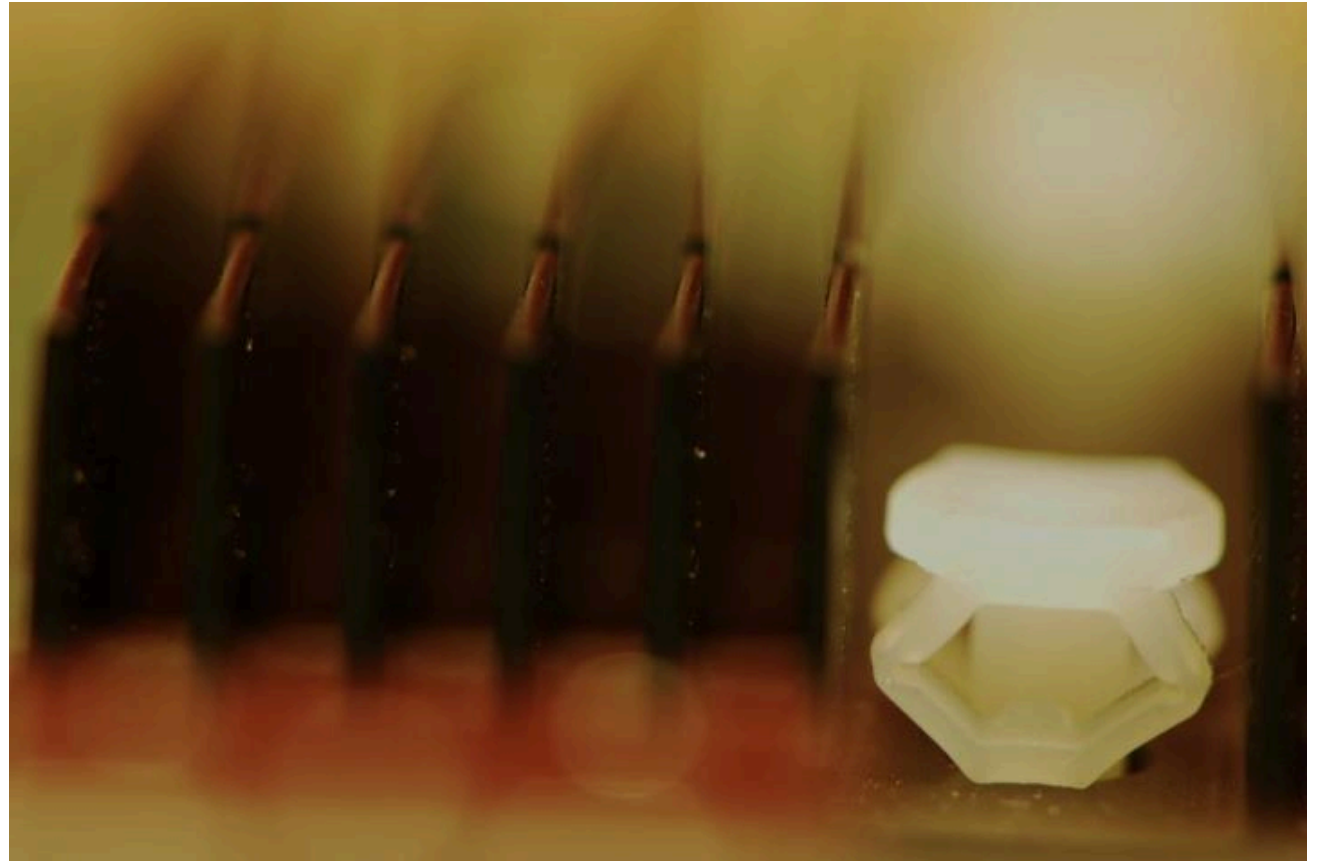
BIOS

Firmware

Device Drivers

NIC/TCP

Process Affinity



BIOS Tuning: Free the Hardware



Hyperthreading: disable, we want real cores.

CPU frequency scaling: disable, as well as all energy saving features: we want the full power all the time.

Check memory bus speed (force to max.)



NIC/TCP Tuning

We are using Myricom 10G NIC

- Download latest drive/firmware from vendor site
- Version of driver in RHEL/CentOS fairly old
- Enable MSI-X
- Increase txqueueelen
`/sbin/ifconfig eth2 txqueueelen 10000`
- Increase Interrupt coalescence
`/usr/sbin/ethtool -C eth2 rx-usecs 100`

Standard TCP Tuning:

```
net.core.rmem_max = 67108864  
net.core.wmem_max = 67108864  
net.core.netdev_max_backlog = 250000
```



100G = It's full of frames !

Problem:

- Interrupts are very expensive
- Even with jumbo frames and driver optimization, there is still too many interrupts.

Solution:

- Turn off Linux irqbalance (`chkconfig irqbalance off`)
- Use `/proc/interrupt` to get the list of interrupts
- Dedicate an entire processor core for each 10G interface
- Use `/proc/irq/<irq-number>/smp_affinity` to bind rx/tx queues to a specific core.

Application Tuning



nuttcp from NASA provides command line option to bind to a processor core.

<http://fasterdata.es.net/fasterdata/network-troubleshooting-tools/nuttcp/>

Server command:

```
nuttcp -S -P5200 -p5201 &; nuttcp -S -P5300 -p5301 &; nuttcp -S -P5400 -p5401 &; nuttcp -S -P5500 -p5501
```

Client command:

```
nuttcp -xc2/2 -l1 -P5200 -p5201 -i1 -T30 -w50M 192.168.2.11 &; nuttcp -xc3/3 -l2 -P5300 -p5301 -i1 -T30 -w50M 192.168.3.11 &; nuttcp -xc4/4 -l3 -P5400 -p5401 -i1 -T30 -w50M 192.168.4.11 &; nuttcp -xc5/5 -l4 -P5500 -p5501 -i1 -T30 -w50M 192.168.5.11 &
```

Note: hand tune TCP window to 50M



More Information

ANI 100G Testbed:

<http://sites.google.com/a/lbl.gov/ani-testbed/>

email: **ani-testbed-proposal@es.net**, BLTierney@es.net

Next round of proposals is due April 1st 2012

Host tuning:

<http://fasterdata.es.net/fasterdata/host-tuning/>

Even more information: lomax@es.net



Extra Slides



PCI optimization: MSI-X

- Extension to MSI (Message Signaled Interrupts)
- Increases the number of interrupt “pins” per card
- Associates rx/tx queues to a given core
- Allows to stitch together on the same core, the thread that runs the program and the asynchronous event it may receive (incoming network packets, asynchronous I/O...), resulting in maximizing L1 cache hit.
- Requires Chipset, card, and operating support.
- Optimized for Linux’ kernel > 2.6.26
- This is a major optimization: on a system with 4 x 10G ethernet, performance gain can be up to 20%



Testbed Access

Proposal process to gain access described at:

<https://sites.google.com/a/lbl.gov/ani-testbed/>

Testbed is available to anyone:

- DOE researchers
- Other government agencies
- Industry

Must submit a short proposal to the testbed review committee

- Committee is made up of members from the R&E community and industry

Goal is to accept roughly five proposals every 6 month review cycle

- Next round of proposals is due April 1, 2012