



## Initial End-to-End Performance Evaluation of 10-Gigabit Ethernet

Justin (Gus) Hurwitz  
Wu-chun Feng

Computer and Computational Sciences Division  
Voice: 505-665-4930  
Fax: 505-665-4934  
ghurwitz@lanl.gov  
http://www.ccs.lanl.gov



## 10GbE Evaluation

- 10 Gigabit Ethernet:

- 10.3 Gb/s bandwidth
- SONET compatible
- Standard Ethernet
- Only full duplex
- Only over fiber

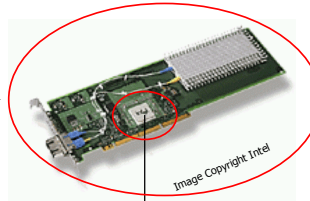
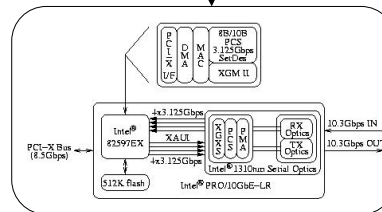


Image Copyright Intel

- Intel's PRO/10GbE LR

- 8.5 Gb/s PCI-X Bus
- Single Mode fiber
- Commercially Available
- Up to 16114 byte MTU (Maximum Transfer Unit)



## 10GbE Eval Outline

- Introduction
  - Outline & Results
- The "Meat"
  - Tests and Results
- Summary of Results and Analysis
  - Full analysis was beyond the scope of our paper
    - (though we have done it!)
- The Future
  - TOE? Scalability? What don't we know yet?
- Fin



## 10GbE Results

- 4.09 Gb/s
  - back-to-back between 2 Dell PE 2650s
  - Using 16000 byte MTU
  - 21- $\mu$ s latency
- 4.11 Gb/s
  - 8160 byte MTU (Jumboframe compatible)
  - Average performance below 16000 byte MTU's
- 2.47 Gb/s
  - 1500 byte MTU
  - CPU limited

COMPUTER & COMPUTATIONAL SCIENCES

## 10GbE Tests- Primary Systems Used


- Dell PE2650
  - 2x 2.2 GHz Intel Xeon CPUs, 400 MHz FSB
  - Serverworks GC-LE chipset
    - up to two 8.5 Gb/s, 133-MHz PCI-X slots
    - 25.6 Gb/s memory bandwidth
  - Available for ~\$1700
- Also used Dell PE4600s
  - Serverworks GC-HE chipset
    - 51.2 Gb/s memory bandwidth!

Images Copyright Dell

Justin (Gus) Hurwitz  
ghurwitz@lanl.gov

www.lanl.gov/radiant

5 

COMPUTER & COMPUTATIONAL SCIENCES

## 10GbE Tests- Network Topology

- 3 test configurations
  - back-to-back, single flow
  - indirect, single flow
  - indirect, multiple flow
    - Indirect tests run through a Foundry FastIron 1500 switch
    - Thanks Foundry!
  - All tests focus on throughput, not latency.


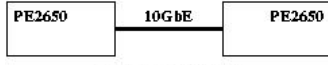
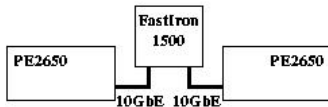


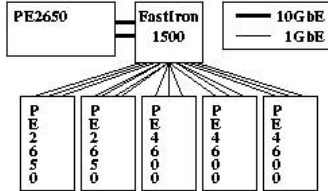
Image Copyright Foundry Networks



(a) Direct single flow




(b) Indirect single flow



(c) Multiple flows through the switch

Justin (Gus) Hurwitz  
ghurwitz@lanl.gov

www.lanl.gov/radiant

6 

COMPUTER & COMPUTATIONAL SCIENCES

## 10GbE Results- Baseline Results

- We start with stock TCP:
  - Default window size less than  $BW * Delay$  (BDP)
    - $\sim 21\text{-}\mu\text{s}$  latency \*  $2 * 10 \text{ Gb/s} = \sim 52 \text{ KB}$
    - Default = 64 KB
  - Common optimisations are not very helpful.
  - Optimisations are shown cumulatively.
- 1500 byte MTU
  - 1.8 Gb/s, 0.9 CPU load
- 9000 byte MTU
  - 2.7 Gb/s, 0.4 CPU load
  - What are those big dips?

Justin (Gus) Hurwitz  
ghurwitz@lanl.gov

www.lanl.gov/radiant

7 Los Alamos NATIONAL LABORATORY

COMPUTER & COMPUTATIONAL SCIENCES

## 10GbE Results- Better PCI-X Burst Size

- Maximum Memory Read Byte Count (MMRBC)
  - Controls PCI-X transmit burst sizes
  - Typically 512 bytes
  - 10GbE adapter supports up to 4096 bytes
- 1500 byte MTU
  - Marginal benefit
- 9000 byte MTU
  - Over 3.6 Gb/s
  - 33% Performance increase!
  - $8x \text{ MMRBC} \neq 1.3x \text{ BW}$ 
    - BW is likely not bus limited

Justin (Gus) Hurwitz  
ghurwitz@lanl.gov

www.lanl.gov/radiant

8 Los Alamos NATIONAL LABORATORY

COMPUTER & COMPUTATIONAL SCIENCES

## 10GbE Results- Uniprocessor Kernel

- Running a Uniprocessor is *faster* than SMP:
  - Interrupts are all processed by CPU 0.
  - SMP kernels have up to 20% extra overhead due to locking.
- 1500 byte MTU
  - 20% improvement
  - 2.15 Gb/s
- 9000 byte MTU
  - Similar *peak* performance
  - *Average* performance:
    - improves ~10%
    - improves ~20% for packets < 2 KB

Justin (Gus) Hurwitz  
ghurwitz@lanl.gov

www.lanl.gov/radiant

9 Los Alamos NATIONAL LABORATORY

COMPUTER & COMPUTATIONAL SCIENCES

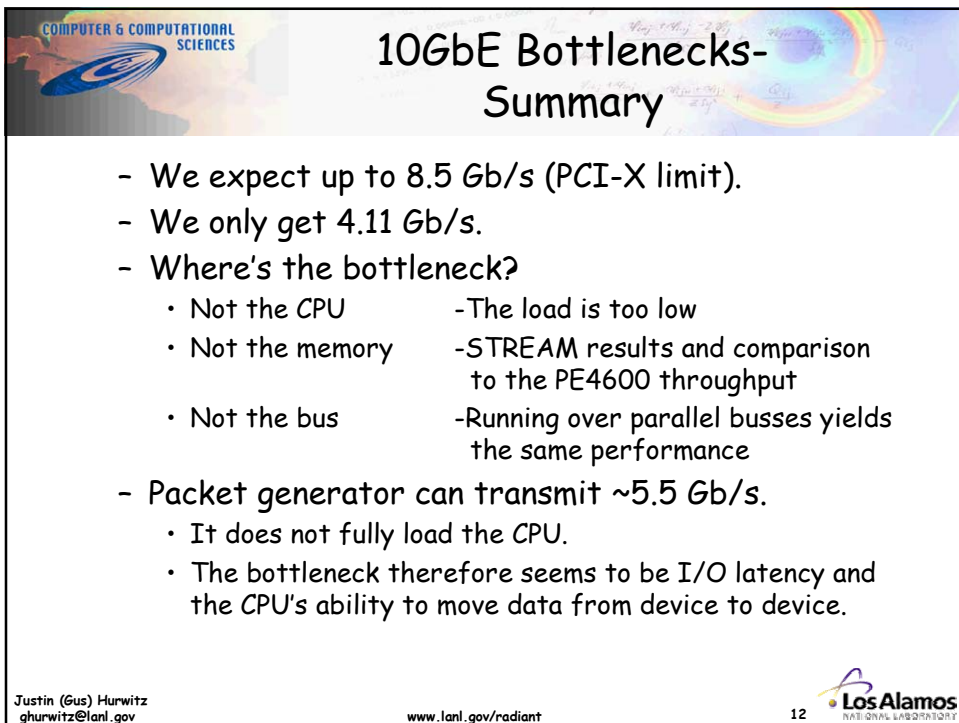
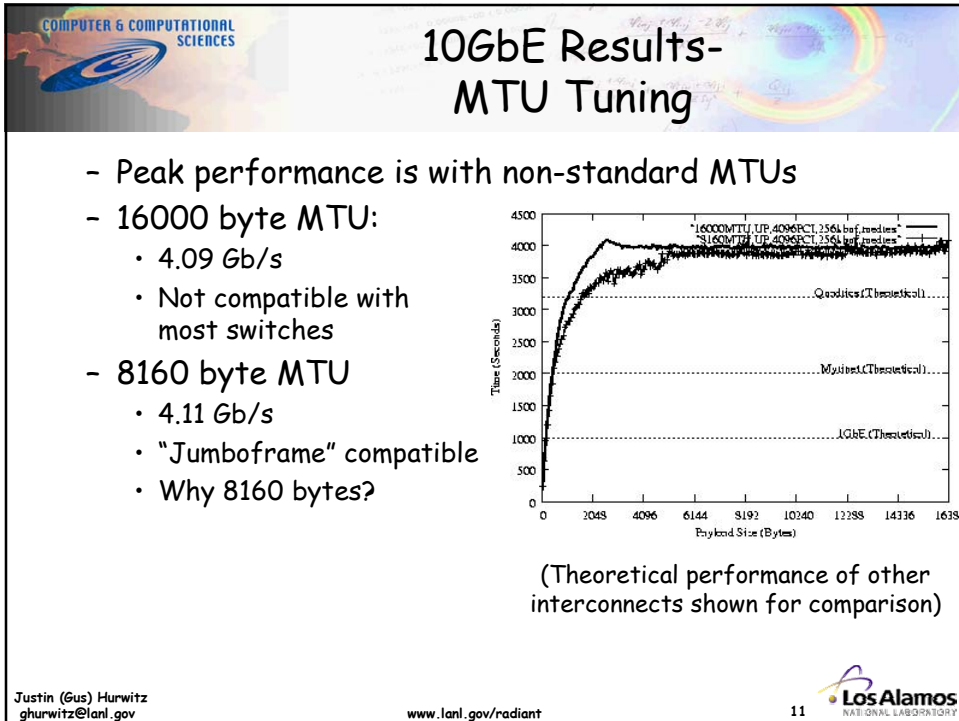
## 10GbE Results- "Too-Large" Windows

- Default window is larger than BDP
  - Larger windows should not improve performance!
  - Larger windows should *hurt* performance!
- Increasing the window improves performance:
  - 256 KB window
  - 1500 byte MTU:
    - 2.47 Gb/s, +15%
  - 9000 byte MTU:
    - 3.9 Gb/s, +8%
- Why?
- And where did the dips go?

Justin (Gus) Hurwitz  
ghurwitz@lanl.gov

www.lanl.gov/radiant

10 Los Alamos NATIONAL LABORATORY



COMPUTER & COMPUTATIONAL SCIENCES

## 10GbE Results-Summary

- Start with "stock" TCP stack      1500: 1.80 Gb/s, 9000: 2.7 Gb/s
- Increase MMRBC                      1500: 1.80 Gb/s, 9000: 3.6 Gb/s
- UP kernel instead of SMP          1500: 2.15 Gb/s, 9000: 3.6 Gb/s
- 256 KB large window              1500: 2.47 Gb/s, 9000: 3.9 Gb/s
- MTU tuning                            8160: 4.11 Gb/s, 16000: 4.09 Gb/s

- Questions:

- Why the 8160 byte MTU?
- Why do "too-large" windows help?
- What are those dips?

Justin (Gus) Hurwitz  
ghurwitz@lanl.gov

www.lanl.gov/radiant

13 Los Alamos NATIONAL LABORATORY

COMPUTER & COMPUTATIONAL SCIENCES

## 10GbE Analysis-Summary

- Don't waste allocated memory!
  - Memory allocation is expensive, especially for large chunks.
- LAN/SAN window optimisation is not as simple as it is in a WAN environment.
  - As the MSS increases relative to the TCP window, this problem will only increase.
  - A bigger MSS/MTU is not always better.
- The hardware throughput bottleneck seems to be intercomponent latency (i.e., I/O latency).
  - We've yet to reach the CPU, memory, or bus limits

Justin (Gus) Hurwitz  
ghurwitz@lanl.gov

www.lanl.gov/radiant

14 Los Alamos NATIONAL LABORATORY

## 10GbE Comparisons- LAN/SAN and WAN

- 10GbE isn't only good in the LAN/SAN
  - Designed to interoperate with WAN technologies (e.g., seamless integration into SONETs).
- Internet2 Land Speed Record:
  - 23,888,060,000,000,000 meters-bits/second.
  - Or, over 1 terabyte of data transferred in an hour.
  - 2.38 Gb/s sustained from Geneva to Sunnyvale over trans-Atlantic 2.5 Gb/s OC-48 connection.
  - Record set by CalTech, CERN, SLAC, and LANL collaboration.
  - Even certified by the Guinness Book of World Records!

## 10GbE The Future- Approaches

- Future approaches to high-speed Ethernet include
  - TCP Offload Engines
    - We're not a fan of them.
  - ST-like header parsing engines
  - RMDA over IP?
- Whatever the solution, checksumming must be done on the payload *after* it has reached main memory, or the bus must guarantee reliability!
  - Put the adapter on the Memory Controller Hub (MCH)?
  - *A la* AGP and Intel's CSA



## 10GbE The Future- Research

- Future research includes:
  - Path-oriented profiling of the TCP stack:\*
    - Quantifying which packets traverse which control path through the TCP stack,
    - Identifying what determines which control path a packet will take,
    - Profiling how long each step of each path takes.
  - TCP behaviour in large MSS/small window networks
    - For WAN performance to scale, the MSS needs to grow.
    - This conflicts with the needs of LANs and SANs.
    - A rift in TCP?
    - Not if the MSS can dynamically scale to fit the network.

\* This analysis is being done with MAGNET, a publicly available tool developed by our team at LANL.

## 10GbE Fin- Acknowledgements

- Many thanks go out to:
  - LANL: Eric Weigle and Adam Englehart
  - Intel: Caroline Larson, Peter Molnar, Patrick Connor, Marc Rillema
  - Foundry Networks: Peter Kersten and John Szewc
- And for the WAN tests/Internet2 LSR:
  - CalTech: Harvey Newman, Sylvain Ravot, Cheng Jin, Xiaoliang (David) Wei, Stephen Low, Suresh Singh, Julian Bunn
  - CERN: Olivier Martin, Paolo Moroni, Daniel Davids, Edoardo Martelli
  - SLAC: Les Cottrell, Fabrizio Coccetti, Gary Buhrmaster
  - ANL: Linda Winkler, Tom DeFanti
  - Level(3): Paul Fernes
  - Cisco: Doug Walsten

COMPUTER & COMPUTATIONAL SCIENCES

*This slide intentionally left blank.*

Justin (Gus) Hurwitz  
ghurwitz@lanl.gov

www.lanl.gov/radiant

19 Los Alamos NATIONAL LABORATORY

COMPUTER & COMPUTATIONAL SCIENCES

## 10GbE Tests- Testing Tools

- Tests run with:
  - Iperf (bulk data transfers)
    - <http://dast.nlanr.net/Projects/Iperf>
  - nttcp (bulk data transfers)
    - <http://www.leo.org/~elmar/nttcp/>
  - NetPipe (ping-pong bandwidth & latency)
    - <http://www.scl.ameslab.gov/netpipe>
  - STREAM (measures memory bandwidth)
    - <http://www.cs.virginia.edu/stream>

Justin (Gus) Hurwitz  
ghurwitz@lanl.gov

www.lanl.gov/radiant

20 Los Alamos NATIONAL LABORATORY

COMPUTER & COMPUTATIONAL SCIENCES

## 10GbE Analysis-Latency

- Latency shows a roughly linear increase with respect to the payload size.
- Disabling the interrupt delay shaves 5- $\mu$ s off of latency.
  - At little-to-no throughput cost when properly tuned...
- Higher performance systems show slightly better latency (as low as 12- $\mu$ s).

Justin (Gus) Hurwitz  
ghurwitz@lanl.gov

www.lanl.gov/radiant

21 Los Alamos NATIONAL LABORATORY

COMPUTER & COMPUTATIONAL SCIENCES

## 10GbE Analysis-Memory and MTUs

- Full analysis is beyond the scope of this paper.
- Why is 8160 byte MTU faster than 9000 bytes?
  - Memory is allocated in chunks of  $2^n$  bytes.
    - (i.e., 2, 4, ..., 8192, 16384, ...)
  - 9000 byte MTUs waste nearly 2 whole pages
    - This stresses the memory subsystem.
    - 8160 byte MTUs fit the *entire* packet (including headers) into 8192 bytes.
  - The kernel can more easily allocate smaller chunks.
    - Not only do 9000 byte MTUs waste a lot of memory,
    - they waste harder to allocate memory!

Justin (Gus) Hurwitz  
ghurwitz@lanl.gov

www.lanl.gov/radiant

22 Los Alamos NATIONAL LABORATORY

COMPUTER & COMPUTATIONAL SCIENCES

## 10GbE Analysis- Windows and Dips

- The "too-large" window concern is related to the unusual dips in throughput.
- The large MSS relative to the BDP limits the values that can be used for the window.
  - This artificially limits both the sender and the receiver windows.
  - This is a big problem for LANs/SANs...
  - And contradicts the general wisdom about windows in WANs.

Theoretical (~26KB) or advertised window

~9K MSS

~9K MSS

~9K MSS

Best possible window due to MSS

Justin (Gus) Hurwitz  
ghurwitz@lanl.gov

www.lanl.gov/radiant

23 Los Alamos NATIONAL LABORATORY

COMPUTER & COMPUTATIONAL SCIENCES

## 10GbE Analysis- Windows and Dips

- Regardless, we can "work around" the problem.
  - We set the window to be really gosh darn big.
- This is a bad solution-
  - It wastes memory.
  - It can significantly hurt performance
    - (e.g., it can halve performance of a WAN).
  - It doesn't address the cause of the problem.
- Nonetheless, even after we fix the software problem, we still only get 4.11 Gb/s
  - Where are the bottlenecks?

Justin (Gus) Hurwitz  
ghurwitz@lanl.gov

www.lanl.gov/radiant

24 Los Alamos NATIONAL LABORATORY

## 10GbE Better Results!

- Anecdotal, or non-rigorous
- 4.64 Gb/s
  - back-to-back, 2 Intel E7505 based systems
  - Using 16000 byte MTU
  - 12- $\mu$ s latency
- 7.2 Gb/s
  - Receiving multiple GbE flows
  - Through a switch
  - Quad 1GHz Itanium 2 CPUs

## 10GbE Comparisons- Interconnect Throughputs

- 4.11 Gb/s, 21- $\mu$ s latency presented in paper
  - More recently, we've reduced the latency to 14- $\mu$ s
- Myrinet/GM = 1.984 Gb/s, 6- to 7- $\mu$ s latency
  - Myrinet/IP = 1.853 Gb/s, ~30- $\mu$ s latency
  - Results published by Myricom
- QsNet/Elan 3 = 2.456 Gb/s, 4.9- $\mu$ s latency
  - QsNet/IP = 2.240 Gb/s, less than 30- $\mu$ s latency
- Gigabit Ethernet = 990 Mb/s, "high" latency
- 10GbE w/ High-end host system: 7.2 Gb/s