Gigabit Ethernet and NFS Server Performance

Mohan Srinivasan
mohan@netapp.com

Goals for Gigabit Ethernet

- IEEE 802.3z Gigabit Ethernet Task Force
  - Formed in early 1996
  - Goal - Complete Gigabit Ethernet spec by early 1998
- Complete Interoperability with (Fast) Ethernet
  - Leverage investment in NICs, hubs, switches, routers and network management
- Conformance to the Ethernet standard
  - Frame Format, Min and Max frame sizes
  - Half Duplex operation (IEEE 802.3 CSMA/CD)
  - Full duplex operation
  - 802.2 LLC specification
  - Simple forwarding mechanism between 10, 100, 1000 Mbps (no fragmentation needed)

Switch and NIC vendors

Switch vendors

Alteon Networks
Extreme Networks
Foundry Networks
Nbase Systems
3Com
Bay Networks, Inc.
Cisco Systems, Inc.
Cabletron
Lucent Technologies
Acacia Networks
Packet Engines (Buffered Dist)
XLNT Designs (Buffered Dist)

NIC vendors

Alteon Networks
Packet Engines
Essential Communications
Intel (Announced)

Gigabit Ethernet and Network Appliance

- Fall ‘96 Interop
  - Proof of concept demo using a prototype Alteon NIC
- Spring ‘97 (Las Vegas) and Fall ‘97 (Atlanta) Interops
  - NetApp among 25 vendors participating in the GEA GbE technology and interoperability demo
  - NetApp participated in the multi-vendor GbE interoperability event leading up to both the Interops.
- Gb Ethernet support in NetApp filers since 4Q‘97
- Full checksum offloading support in NetApp filers since 4Q ‘97.

Alteon AceNIC Server-centric Features

- Checksum Offloading
- Jumbo Frames (upto 9K MTU)
- Efficient PCI Utilization
- Highly Integrated ASIC
- Adaptive Interrupt Coalescing
- Ability to split off protocol headers from data
  - ☞ Could be leveraged to implement 0 copy NFS writes
- Scatter-gather DMA support with no alignment restrictions on xmit/recv buffers
- Dual homing for automatic NIC failover
  - ☞ 2nd NIC is a passive standby

Highly Integrated ASIC

- 2 (MIPS) 100MHz MIPS CPUs (AceNIC II)
- 1MB onboard memory
- 640MB/s memory bandwidth (AceNIC II)
- Dual DMA channels
  - Independent DMA channels for recv and xmit paths
  - Each DMA channel supports checksum offload
- 24KB on-chip scratchpad memory (AceNIC II)
- 10/100/1000 Ethernet Controller

Efficient PCI Utilization

- **32/64 bit data (and addressing) support**
- **Support for 33MHz/66MHz PCI clocks**
- **Supports long bursts (upto 64KB)**
- **Support for Mem Read Mult and Mem Write Invalidate**
- **No PCI reads for xmit or recv (except reading the ISR)**

Checksum Offloading

- TCP/UDP and IP header checksums are offloaded
- Checksum offloads controllable on per-packet basis
- Checksum offloads of fragmented packets supported
- Benefit
  - Laddis/SFS2.0 - 10% to 15% CPU savings. Results in near-linear thruput increase
- Drawback
  - Checksum Offload increases NIC CPU utilization
  - Depresses aggregate network thruput by about 20%
  - Expect this to be a non-issue with the new (2nd) generation Alteon NICs (with the 2x100MHz MIPS cores)

Checksum Offload - Fragmented UDP Datagrams

- Typically NICs implement checksum offload on an Ethernet Frame (IP Fragment) basis
- For Fragmented UDP datagrams
  - Receive - This does not present any problem. Host collapses checksums after reassembly
  - Transmit - More work needed (both stack and NIC). Most NICs don't offload in this case.
- My Solution (Now included in all Alteon NIC FW releases)
  - Stack marks first and last IP frags of a fragmented UDP datagram
  - Stack does not interleave IP frags from different datagrams while enqueueing IP frags
  - NIC FW accumulates checksums of IP frags and stuffs final checksum in UDP header

Jumbo Frames

- Alteon proposal to increase MTU on GbE to 9K
- NetApp support for Jumbo Frames this year (CY '98)
- 9K MTU over GbE Huge Advantage for NFS servers
  - Reduces CPU utilization in the driver and protocol stack, reduces interrupt overhead
  - On NetApp F630 under Laddis, I computed a 7% - 9% gain in CPU at the higher load points
  - Reduces PCI overhead
  - Increases network thruput dramatically (later)
- Disadvantages
  - No real technical disadvantages (GbE FD-only)
  - Discounted by competitors as "Alteon proprietary" or "non-standard". But I expect most vendors to support this sooner or later

Example - Peak NFS Thruput with GbE (on F630)

- Hardware and Traffic
  - Server - F630, 512MB,13 9GB drives, 1 GbE
  - Clients - 4 Sun Ultra2's
  - Switch - Alteon. 2xGbE, 8x10/100Mb/s
  - Load - NFSv3/UDP, 32KB r/wsize. Cached Sequential reads
  - Using the Alteon AceNIC I (40MHz NIC CPU)
- Measured (peak) NFS Thruput of 1 GbE 36MB/s => 300Mb/s
- In terms of frames per second/GbE interface
  - 36MB/s at 32KB xfers => ~1100 NFS RDs/s, and, 1100 NFS RDs/s => ~25,000 frames per second at a 1.5KB MTU. (An NFS RD is a NFSv3/UDP/32KB read)

Example (contd) - Peak NFS Thruput with Jumbo Frames

- FACT - The per-frame overhead on the NIC does not increase (measurably) with frame size
  - So we'd expect to get around a 5-6 fold (9K/1.5K) increase in thruput with Jumbo Frames, assuming servers scale
- Fixing the peak frames/second for the NIC at 25,000
  - With a 9K MTU, we would expect the NIC to sustain a peak of 5000 NFS RDs/s
  - 5000 NFS RDs/s => 160MB/s => well over 1Gb/s
  - The F630 maxes out on PCI (in the given test) at around 47MB/s
  - But with Jumbo Frames I can easily get 47+MB/s with1 GbE NIC (whereas with the 1.5K MTU, I need to slap in 2 NICs to reach that limit)

Summary

- **Ubiquity of Ethernet => Success of GbE is guaranteed**
- **GbE products shipping for well over 6 months, with**
  - **Good Performance**
  - **Surprisingly Few Interoperability Problems**
  - **Good Stability**
- **Opportunities for NFS server vendors**
  - **Aggregation of Server to LAN Backbone Links**
  - **Exploit Features of Smart NICs and Switches**