Latency Considerations for 10GBase-T PHYs

Shimon Muller

Sun Microsystems, Inc.

IEEE 802.3an Task Force

March 16, 2004

Orlando, FL

Sun microsystems

We make the net work.

Outline

- Introduction
- Issues and non-issues
- PHY Latency in The Big Picture
- Observations
- Summary and Recommendations

Why is PHY Latency an Issue for 10GBase-T

- In the past, PHY latency for Ethernet was driven by bit-budget requirements of CSMA/CD
  - Determines the physical span of the Ethernet network
  - Latency requirements are very tight
- 10-Gigabit Ethernet is the first standard that does not support CSMA/CD
  - Latency requirements can be substantially relaxed
    - Allows for useful implementation tradeoffs
- Caution needs to be exercised when selecting the maximum allowable latency for the 10GBase-T PHY
  - Some network applications that run over Ethernet are latency-sensitive
  - May suffer performance degradation if the PHY latency becomes significant

When is Latency Not a Problem

- Support for Pause flow control
  - Rarely used
    - Not a very popular (or useful) protocol
  - At 10Gb/s speeds the size of the flow control buffers is already large
    - If implemented, is probably already off-chip
- Network applications that mostly use bulk data transfers
  - Backups, file serving, etc.
- Network applications (bulk data or transactional) that use lots of low-throughput connections
  - Web servers, some databases, etc.
- Pipelining and parallelism hide the latency for above applications

When is Latency a Problem

- Applications that have a significant transactional network traffic profile
  - Message-based and/or request-response traffic patterns
    - Clustering, HPCC, OLTP, etc.
  - High-throughput connections where bulk data transfers are typically preceded by message exchanges
    - Most databases (Oracle), etc.
- For above applications latency directly affects performance
- Relatively few connections do not lend themselves well for pipelining

Additional Latency Requirements

- Ethernet has never been considered a low-latency interconnect
  - Mostly due to overheads incurred above the Ethernet sublayer
- However:
  - Physical layers tend to be leveraged between various interconnect technologies
    - Fiber Channel, InfiniBand, PCI-Express, etc.
- A low latency 10GBase-T PHY will have a broader market potential

Networked Systems' Latency Components

- Protocol stack and OS
  - In the lower 10s of microseconds in each direction
    - End-to-end: ~2x
  - Will continue to come down in the future
    - Used to be in the 100s of microseconds
    - Processors are getting faster
    - More efficient network traffic processing in the OS
    - Hardware hooks to speed up packet processing
- Server memory and I/O subsystem
  - Up to several microseconds per packet (multiple accesses)
    - NUMA effects, etc.
    - I/O bridge latencies
    - End-to-end: 2x
  - Will get much better in the future
    - Modern systems are already capable of minimizing this latency
    - New NIC architectures will be able to hide most of it

Networked Systems' Latency Components

- NICs and switches
  - Up to 1.2 microseconds per h/w component
    - Most implementations use store-and-forward
    - End-to-end: 3x
  - For latency sensitive applications cut-through is an option for both NICs and switches
- Cable delay
  - Up to 0.5 microseconds per hop
    - End-to-end: 2x
  - The vast majority of links in future datacenters will be shorter than 100m
    - Blade and rack systems

Observations

- **Goal:**
  - Pick a number for 10GBase-T PHY latency such that it is proportionally insignificant in the overall system in the foreseeable future

- **Latency consideration space:**
  - Ideally, the PHY latency should be on the order of 10s of nanoseconds
    - Will accommodate all Ethernet and non-Ethernet applications in the foreseeable future
  - PHY latency on the order of 100s of nanoseconds is acceptable
    - Will accommodate most Ethernet and some non-Ethernet applications
  - PHY latency should not exceed 1 microsecond
    - May start affecting Ethernet over TCP/IP application performance in the foreseeable future

Observations (Continued)

- Trade-offs:
  - From a system perspective, only end-to-end latency matters (Rx+Tx)
    - Can be budgeted asymmetrically
  - Given the choice between latency vs. complexity/power/cost, latency should take precedence
    - Moor's Law will eventually take care of the latter but not of the former
  - Given the choice between latency vs. UTP cable length, cable length should take precedence
    - In the short term support for installed cabling is more important

Summary and Recommendations

- Relaxing latency requirements for the 10GBase-T PHY does not come for free
  - Eventually may start affecting some application performance
  - May also reduce market potential
- Evaluate proposals in the context of observations and trade-offs presented
- Make final determination based on the “bang for the buck” trade-off

shimon.muller@sun.com

IEEE 802.3an Task Force

March 16, 2004

Orlando, FL

Sun microsystems

We make the net work.