Paper#1: Gigabit Ethernet – Design and Performance

Jonathan C.L. Liu, Ph.D.
Department of Computer, Information Science and Engineering (CISE),
University of Florida
Motivations

- Need for a faster Connectivity
- Increasing Multimedia content of Web Traffic, and Nature of traffic distribution
- Increased use of advanced applications
  - Streaming multimedia,
  - virtual reality,
  - distance learning
Gigabit Ethernet Network

- First Gigabit NIC in early 1997
- First Gigabit Hub in late 1997
- There are now at least 28 companies with Gigabit Ethernet products
- In late 1999, industry lab announced to prototype 10-Gbps Ethernet network
Major Components

- IEEE 802.3z Frame Format
- Gigabit Network Interface Cards
- Switching Hub for LAN
- Routing Switches for MAN
- Multimedia-Friendly Features
  - 802.1p and 802.1Q
  - Port-level and Box-level policy
Frame Format and PHY Layer

- **1000 BASE-SX standard**
  - Use 800-nm wavelength for multi-mode fiber
  - Maximum length ranges from 260 m to 440 m

- **1000 BASE-LX standard**
  - Use 1300-nm wavelength for single-mode fiber
  - Maximum length 3000 meters

- **1000 BASE-CX standard**
  - Up to 25 meters
Legend

1 octet = 1 byte
Minimum frame = 64 bytes
Maximum frame = 1518 bytes

IPG = Inter Packet Gap = 0.96μs = 12 bytes = 3 clocks = 31.25 MHz
IPG = Inter Frame Gap = 0.16μs = 20 bytes = 5 clocks = 31.25 MHz
Network Interface Design

- Based on 32-bit, 33-Mhz PCI bus
  - Theoretical limit is 1 Gbps
  - Practical bandwidth is 800 Mbps
- 64-bit, 66-Mhz GNIC
  - Theoretical limit is around 3-4 Gbps
- Block Diagram for GNIC Design
GNIC Design Issues

- Reduce host-CPU utilization
  - On-board independent DBDMA to stream RA and TX data without host CPU intervention
  - Dual-burst FIFO with 512 Kbyte to 2 Mbyte

- Achieved 800-Mbps at the board level
- The efficiency of CPU utilization is improved 6.8 times
FDR for LAN

- Operate at the data-link layer and process MAC and link-level frames
- Need to handle the following issues
  - Collisions
  - Link Congestion
  - Broadcast / Multicast
  - Flow Congestion
FDR Flow Control

- Full duplex for reducing the CSMA/CD collisions.
- Congestion control to avoid frame dropping
- Round-robin scheduling to prevent packet clumping
Approaches used in FDR

- Collisions
- Link Congestion
- Broadcast / Multicast
- Flow Congestion Schemes

- 802.3x Full Duplex
- 802.3x Flow Control
- Evolving VLAN & IP Switching Schemes
- Evolving QoS
Gigabit Routing Switch

- Architectural issues - Shared memory vs cross-bar.
  - Port-based memory.
  - Head-of-line- blocking.
  - Difficult to provide reliable QoS support.
- Novel parallel access shared memory architecture.
- Priority queue design.
Experiment Goals

- Measure the achieved throughput at the end-to-end application-level
  - Determine the system bottleneck
  - Identify the bounds
  - Peak throughput with netperf
  - Average sustained throughput
  - Jitter measurement
Application Natures

Two Categories of Dist. Multimedia Applications:
- Video Conferencing
- On-Demand Video Delivery

Major Challenges:
- Large Data Size
- Real-Time Constraint
- Supporting Concurrent Accesses
Peak Performance Measurement
Netperf Settings

- Works around the Client-Server model.
- Netperf uses a control connection (TCP using BSD sockets)
- Send the test parameters and collect results to/from the remote connection.
- Netperf usually gives the peak performance
Results for Peak Performance

![Graph showing peak performance results for different systems.](image)
Peak Performance Analysis

- Application level
  - Linux 186 Mbps vs. NT 90 Mbps
- TCP/UDP level
  - ??? Mbps vs. ??? Mbps
- IP level
  - ??? Mbps vs. ??? Mbps
- Device Driver level: 300~350 Mbps
- Hardware/firmware level: 700~800 Mbps
Sustained-throughput Experiment

- Measured in a peer-to-peer symmetric communication pattern.
- Round trip of a text message of a given size from the client to server and back.
- Varying sized messages from 128 bytes to 1 Mbytes
- AVERAGE Round trip delays and throughput values were computed (with 95% confidence interval)
**Client Process:**

*begin*

- Initialization
  - open socket;
  - bind SAP to host IP address;
  - send connect request;

- Set experiment parameters
  - buffer size;
  - TCP segment size;

*repeat N times*

*begin*

- mark time;
- send message of size m bytes;
- receive echoed message;
- mark time;
- compute the time difference;

*end*

- compute mean delay and throughput;

*end*

**Server Process:**

*begin*

- Initialization
  - open socket;
  - bind SAP to host IP address;
  - listen and accept connections;

- Set experiment parameters
  - buffer size;
  - TCP segment size;

*repeat forever*

*begin*

- receive message of size m bytes;
- send back the message;

*end*

*end*
Results on Sustained Throughput
Supporting on-demand streams

Table 3
Number of Concurrent Accesses Supported for the Gigabit GNICS

<table>
<thead>
<tr>
<th>Display Speed (frames/sec)</th>
<th>Rate</th>
<th>Frame Size</th>
<th>Requested Block Size (# frames)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4 Mbps</td>
<td>16 KBytes</td>
<td>2 4 8 16 32 32 N/A</td>
</tr>
<tr>
<td></td>
<td>8 Mbps</td>
<td>32 KBytes</td>
<td>16 16 17 18 19</td>
</tr>
<tr>
<td></td>
<td>16 Mbps</td>
<td>64 KBytes</td>
<td>6 7 8 9 9</td>
</tr>
<tr>
<td></td>
<td>32 Mbps</td>
<td>128 KBytes</td>
<td>2 3 3 3 3</td>
</tr>
</tbody>
</table>

Table 4
Number of Concurrent Accesses that Can Be Supported Using a Single 16 GByte Disk Array Source. From [3].

<table>
<thead>
<tr>
<th>Display Speed (frames/sec)</th>
<th>Rate</th>
<th>Frame Size</th>
<th>Requested Block Size (# frames)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4 Mbps</td>
<td>16 KBytes</td>
<td>4 8 16 23 29 33</td>
</tr>
<tr>
<td></td>
<td>8 Mbps</td>
<td>32 KBytes</td>
<td>4 7 11 14 15 17</td>
</tr>
<tr>
<td></td>
<td>16 Mbps</td>
<td>64 KBytes</td>
<td>3 5 6 8 8 NA</td>
</tr>
<tr>
<td></td>
<td>32 Mbps</td>
<td>128 KBytes</td>
<td>2 3 3 4 NA NA</td>
</tr>
</tbody>
</table>
Topics for Further Investigations

- IEEE 802.1p to expedite delivery of time-critical traffic and to avoid high-bandwidth multicast traffic
- IEEE 802.1Q for logically partitioning of a LAN
- Policy-based QoS
  - Port level - GNIC, FDR and Switch
  - Box level - FDR and 4884 Switch
  - End-to-End level - GNIC, FDR and Switch.
Summary

- Gigabit Ethernet is promising
- Short processing latency is needed
  - Better hardware/firmware co-design
  - Light-weight transport protocol stack
- To achieve the application-level gigabit throughput, we need ...