iWARP: It's Not Just For The LAN Anymore

Dennis Dalessandro
dennis@osc.edu

Commodity Cluster Symposium
July 25, 2006
Agenda

- Introduction to iWARP
- iWARP Hardware History
- Current iWARP Research
- iWARP Road Map
What is the problem with TCP/IP anyway?

- Network processing = lots of CPU
  - Costly at 1Gbps, imagine 10Gbps
- Why?
  - Complex protocol stack processed by CPU
  - Movement of data from memory to NIC by CPU
Why TCP/IP is so costly

OS Kernel

Application Memory

Read Data

Read Data

TCP

IP

Send Data

Send Data

NIC
**Possible solutions**

- **TCP Offload Engine (TOE)**
  - Offloads processing of TCP/IP stack
  - Good but not enough

- **Remote Direct Memory Access (RDMA)**
  - Offloads processing of TCP/IP stack
  - Also has Zero-Copy
TOE Card

OS Kernel

Application Memory

Read Data
Read Data
Send Data
Send Data

TOE
TCP
IP

Protocol work offloaded, but CPU still moves the data through OS
Protocol offloaded, and CPU does not move data.
Examples of RDMA

- InfiniBand, Myrinet, Quadrics
- Require special infrastructure
  - Do not work in the WAN
- Great performance
  - Latency can't be beat
- Tried and true
  - IB very common
iWARP - The new kid on the block

- iWARP = RDMA over Ethernet (TCP/IP)
  - Runs over existing network infrastructure
    - WAN Capable!
- IETF RFC specifications
  - RDMAP, DDP, MPA
- Downside
  - Switch cost for 10 Gigabit
  - New technology
Hardware History

- Ammasso Inc
  - First commercially available
  - Only 1 Gigabit
  - Blazed the trail
  - Allowed researchers to experiment with iWARP
  - Ceased operations late 2005

- Allowed researchers to continue iWARP work
  - Everything learned is still applicable

- Ammasso presence still felt
  - OpenIB - now OpenFabrics driver
New players on the scene

- **NetEffect**
  - 10 Gigabit iWARP adapter
  - Outperforms IB in terms of throughput
  - Boards are selling now
  - OSC leading the way
    - Paper to appear at RAIT'06 (IEEE Cluster 2006)
      - September 28 in Barcelona, Spain

- **Chelsio**
  - Has an adapter as well
    - Driver in OpenFabrics source tree

- **Broadcom**
  - ????????
NetEffect performance

Throughput

Message Size (bytes)

Throughput

NetEffect Verbs (PCI-X)
NetEffect MPI (PCI-X)
Infiniband MPI (PCIe)
Infiniband MPI (PCI-X)
NetEffect performance cont...
10 Gig iWARP

- Comparable (better?) in performance to IB
  - Higher throughput than standard 4X IB
  - Switch latency is comparable
  - A bit higher latency at small message sizes
  - Appropriate for cluster interconnect
  - Appropriate for high-end servers
  - Appropriate for storage (iSCSI)

- Just getting started with it
  - WAN tests
  - Interoperability with other iWARP HW
Current iWARP work

- iWARP in the WAN
  - Main point of this talk
- Interoperability of iWARP devices
  - Ammasso, NetEffect, Software iWARP
- RDMA enabled web server
  - Apache *mod_rdma* and proxy server
- RDMA enabled FTP client/server
- Real applications with NetEffect device
OSC iWARP resources

- NetRes Cluster
  - Up to 41 Ammasso
  - On TFN
- P4 Cluster
  - 17 Ammasso
  - On TFN
- NetEffect
  - 2 Servers
  - On TFN
Basic performance

- At 1 Gbps TCP about same as iWARP
  - Today's processors capable of 1Gbps
    - At high CPU utilization
- 10 Gbps will be a different story
- Things do not work the same in WAN
  - Tunable network parameters a must
    - Window Size
    - MTU?
Window size effect in WAN

*Note 113.1 KB not full BW for iWARP
iWARP FTP

- Demo at SC 2005
- Work in progress to create production version
- Written in OpenFabrics verbs API
  - Will work on iWARP or InfiniBand
- Intended use: Move large data sets in WAN
Basic FTP Performance

- Server: Springfield
- Client: Columbus
- Link: 10Gbps (TFN)
- About same perf

<table>
<thead>
<tr>
<th>Size</th>
<th>iWARP</th>
<th>TCP/IP</th>
</tr>
</thead>
<tbody>
<tr>
<td>200K</td>
<td>.010 s</td>
<td>.015 s</td>
</tr>
<tr>
<td>1M</td>
<td>.021 s</td>
<td>.031 s</td>
</tr>
<tr>
<td>10M</td>
<td>.117 s</td>
<td>.247 s</td>
</tr>
<tr>
<td>100M</td>
<td>1.05 s</td>
<td>2.59 s</td>
</tr>
</tbody>
</table>
The iWARP Benefit

*16 clients each time
iWARP in the WWW

- RDMA module for Apache
  - Add RDMA info in header
  - Request/Response in TCP
  - Data transfer with RDMA

- Downside
  - RDMA connection build up

- Why not full iWARP port?
  - Extensive changes to Apache code base
  - New feature not new web server
RDMA enabled Apache web server

- mod_rdma
  - Apache module to handle RDMA transfers
  - “Grab” out going data and ship it with RDMA
  - Manipulate headers for minor changes
    - Simple changes, nothing fundamental
  - All the benefits of Apache
    - No rewrite of Apache code needed
    - Utilizes Apache hooks
mod_rdma cont.

- **Server Writes**
  - Client has to guess size of file
  - RDMA connect takes time
    - Same as TCP connection (it is)
mod_rdma cont...

- **Client Reads**
  - Still have RDMA connect
  - Server replies with RDMA info
  - Client has to send an extra ACK to tell server RDMA read done
RDMA enabled Apache performance

- 1 page with 20 images
  - Stock wget
  - RDMA enabled wget

- CPU usage for 2, 4, 6 clients
  - RDMA
    - starts out low, stays low
  - TCP
    - starts out in middle goes and stays high

mod_rdma perf cont..
Example web based app

- Database of all US cities
  - Includes zip code, latitude, longitude, etc.
  - One fake person from each city
  - A little over 42,000 entries

- User: “give me all people within X miles of Zip”

- Server: responds with a variable number of results w/pictures per page
  - lots of trig for PHP to crunch on
  - lots of querying for MySQL database
  - pictures ensure lots of data to transfer

- Developed by Manu Mukerji
Two scenarios

1. Back end RDMA clients
   - iWARP
   - iWARP
   - iWARP
   - iWARP

2. Back end TCP clients
   - TCP
   - TCP
   - TCP
   - TCP
Sample app performance

Graphs showing performance comparison between back end iWARP and back end TCP.

- **iWARP wget**
  - Graph on the left:
    - X-axis: Number of backend clients
    - Y-axis: Time for iWARP wget (seconds)
    - Red line: back end iWARP
    - Blue line: back end TCP

- **TCP wget**
  - Graph on the right:
    - X-axis: Number of backend clients
    - Y-axis: Time for TCP wget (seconds)
Server performance

**iWARP wget**

**TCP wget**

Requests per second with iWARP wget

Requests per second with TCP wget

Number of backend clients

back end iWARP

back end TCP
Upcoming work...

- NetEffect interoperability
  - with Ammasso cards
  - with Software iWARP
- OpenFabrics port of mod_rdma
  - Including SSL support
- OpenFabrics port of wget
- Many 1Gig clients to single 10Gig server
  - http
  - ftp
iWARP road to adoption

Beginning
- Hardware iWARP in most high end of servers
- Software iWARP in clients

After time....
- HW iWARP clients will begin to appear
- SW iWARP will become common

In parallel.....
- Specialty clusters of iWARP

Eventually
- World will move beyond 1 Gig
- iWARP is one of the best answers for Ethernet
Conclusion

- iWARP is WAN capable
- iWARP is a viable cluster interconnect
- HW is now available
- Will make a difference in servers today
- Benefit all computing not just HPC
Questions?

Dennis Dalessandro
dennis@osc.edu
http://www.osc.edu/~dennis/iwarp