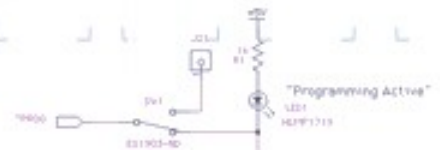
iWARP: Its Not Just For The LAN Anymore

Dennis Dalessandro
dennis@osc.edu

Commodity Cluster Symposium
July 25, 2006

Agenda

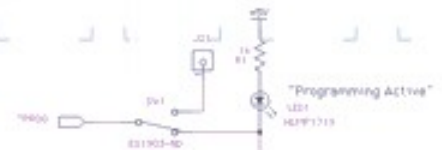
- Introduction to iWARP
- iWARP Hardware History
- Current iWARP Research
- iWARP Road Map



OSC

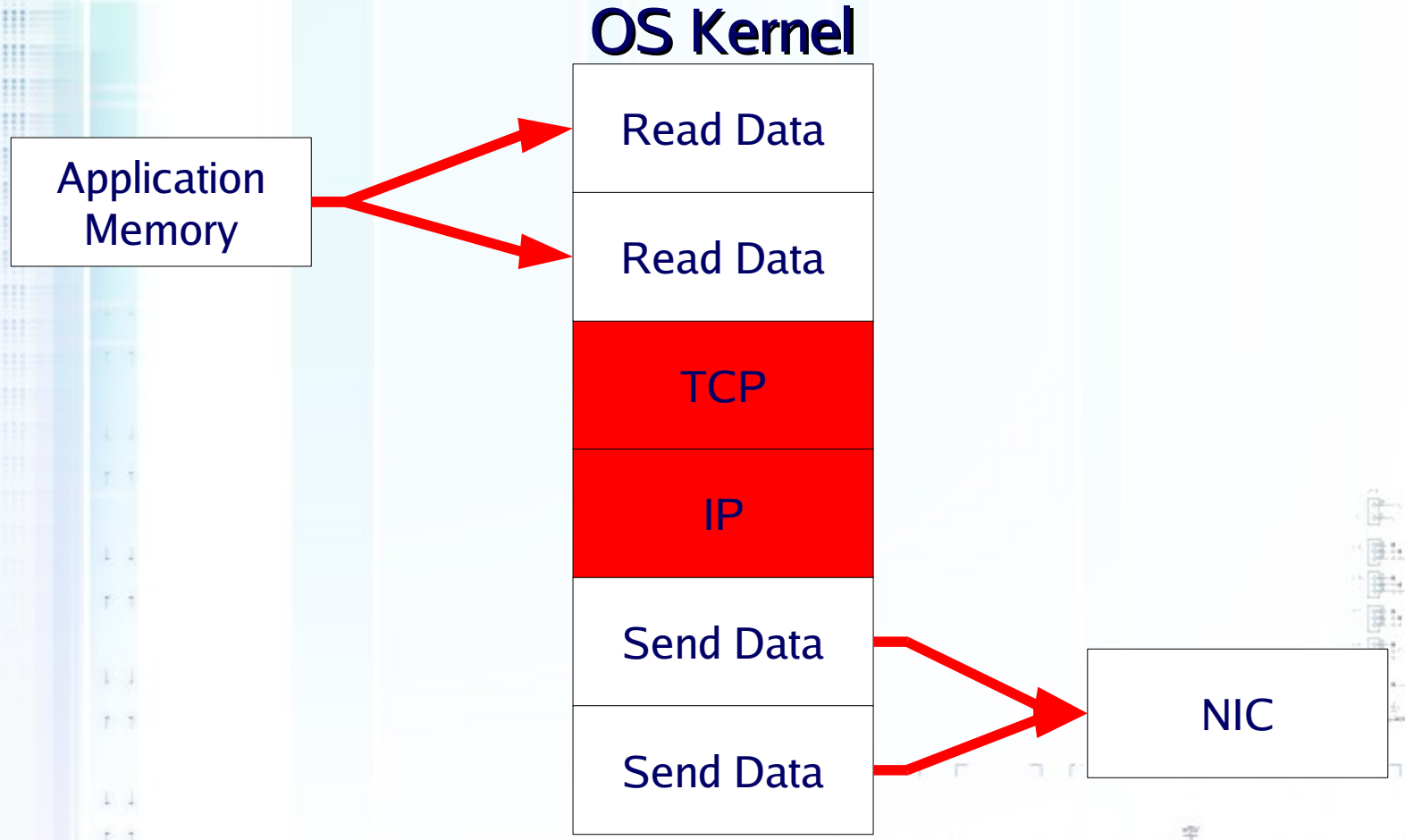
What is the problem with TCP/IP anyway?

- 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



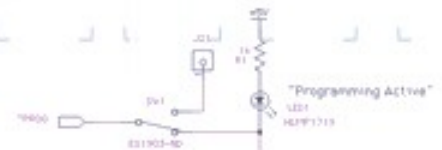
OSC

Why TCP/IP is so costly



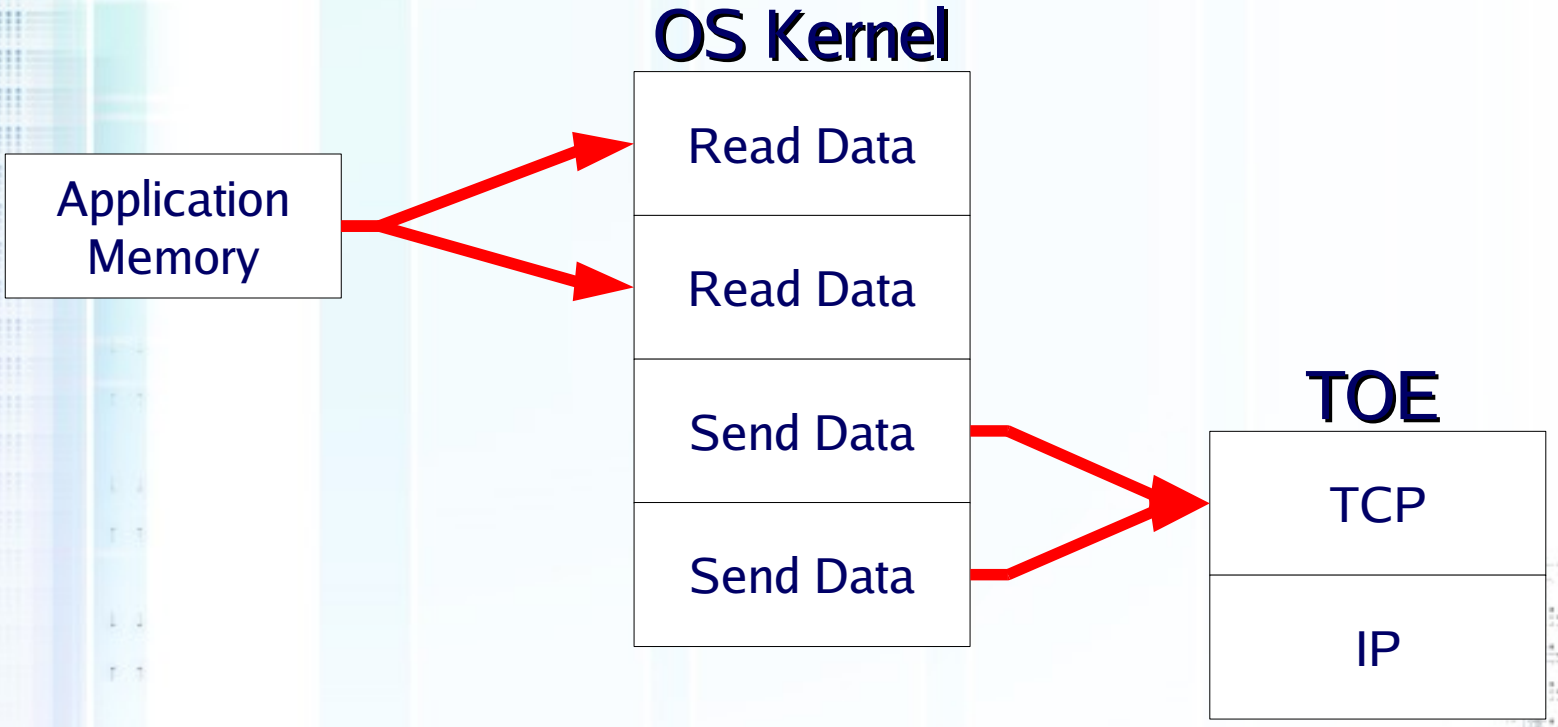
Possible solutions

- 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



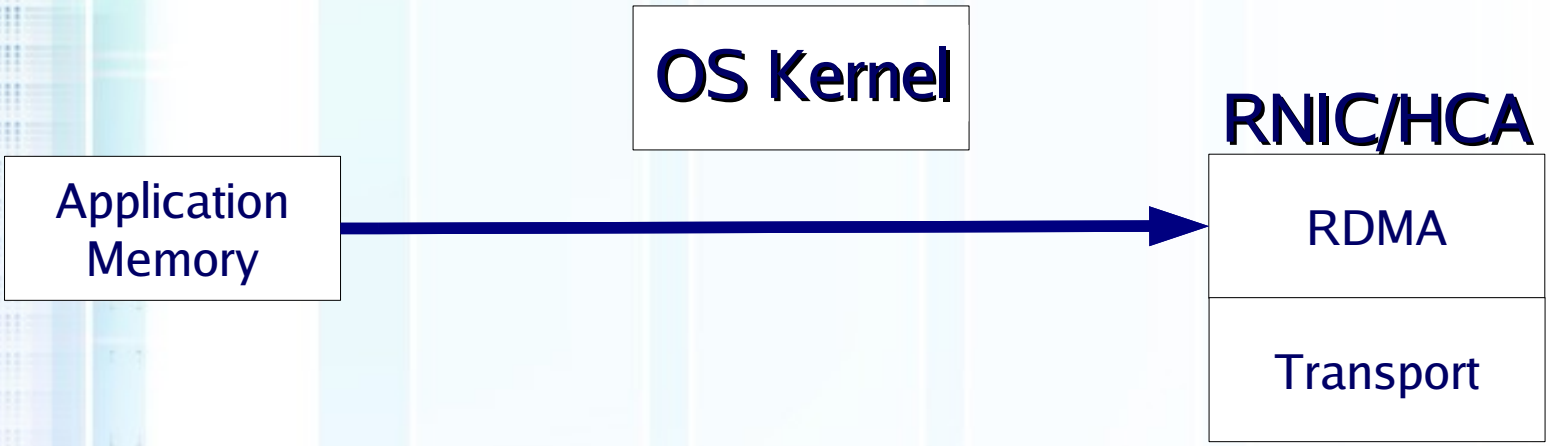
OSC

TOE Card



Protocol work offloaded, but CPU still moves the data through OS

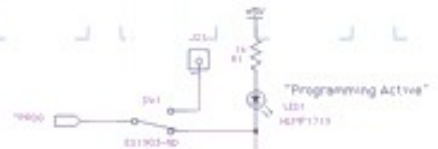
RDMA



Protocol offloaded, and CPU does not move data.

Examples of RDMA

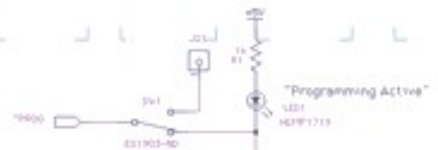
- **InfiniBand, Myrinet, Quadrics**
- Require special infrastructure
  - Do not work in the WAN
- Great performance
  - Latency can't be beat
- Tried and true
  - IB very common



OSC

iWARP - The new kid on the block

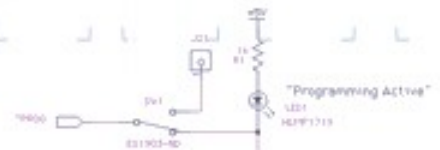
- iWARP = RDMA over Ethernet (TCP/IP)
  - Runs over existing network infrastructure
    - WAN Capable!
- IETF RFC specifications
  - RDMAP, DDP, MPA
- Downside
  - Switch cost for 10 Gigabit
  - New technology



OSC

Hardware History

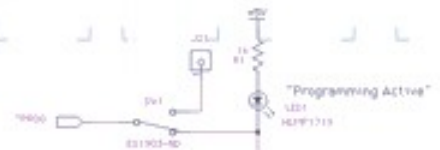
- Ammasso Inc
  - First commercially available
  - Only 1 Gigabit
  - Blazed the trail
  - Allowed researchers to experiment with iWARP
  - Ceased operations late 2005
- Allowed researchers to continue iWARP work
  - Everything learned is still applicable
- Ammasso presence still felt
  - OpenIB - now OpenFabrics driver



OSC

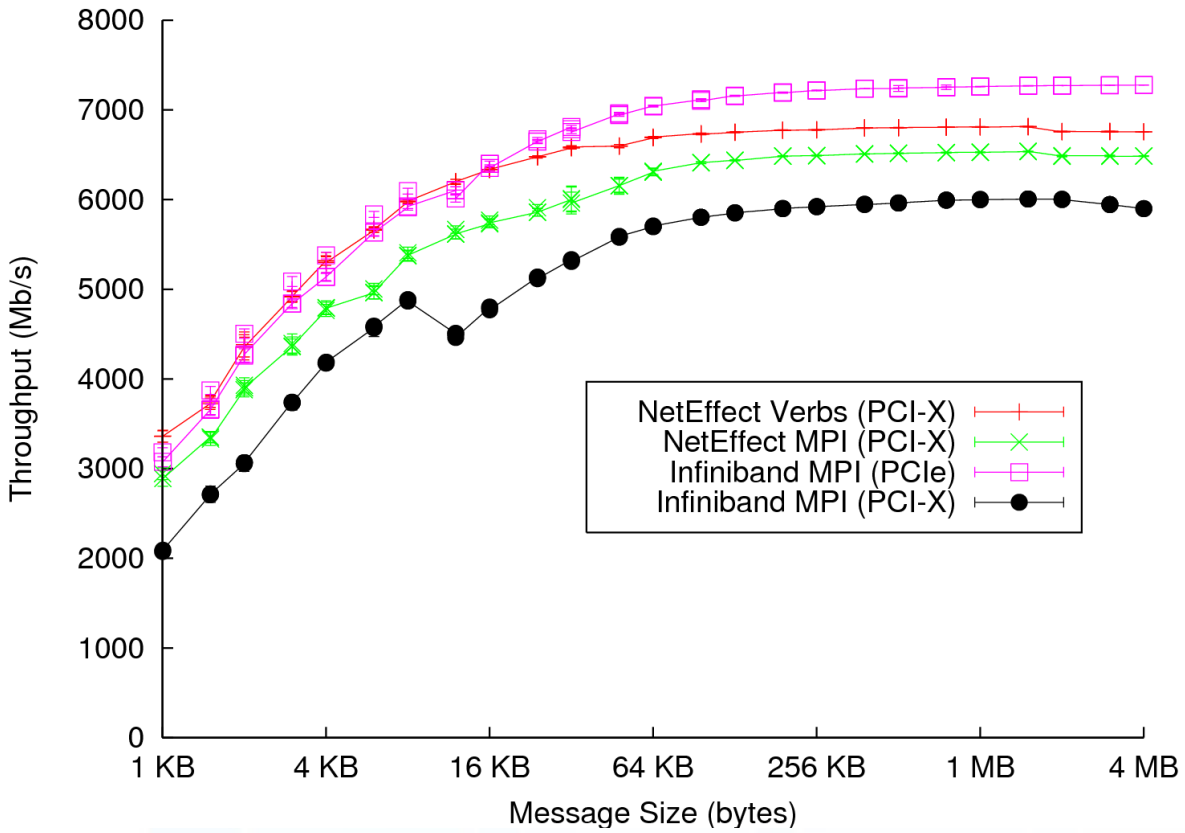
New players on the scene

- 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
  - ??????



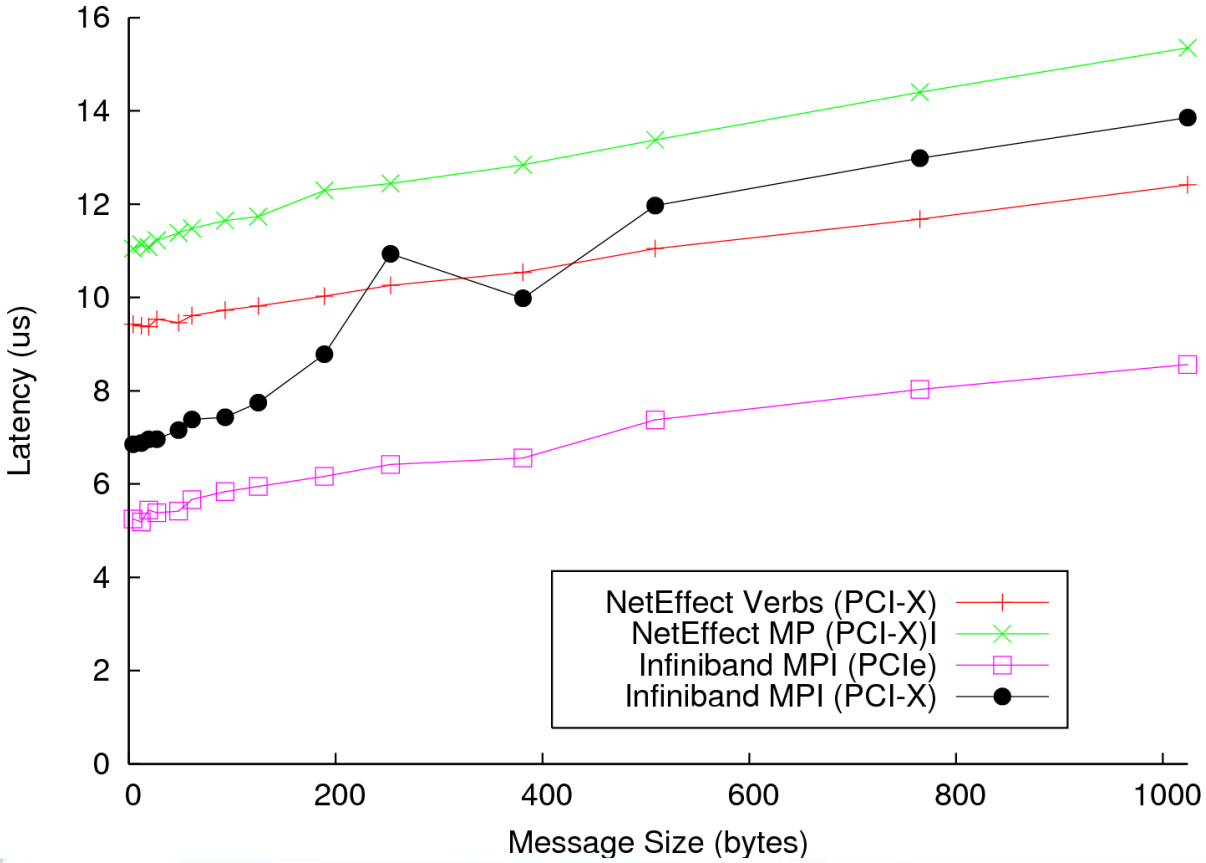
OSC

NetEffect performance



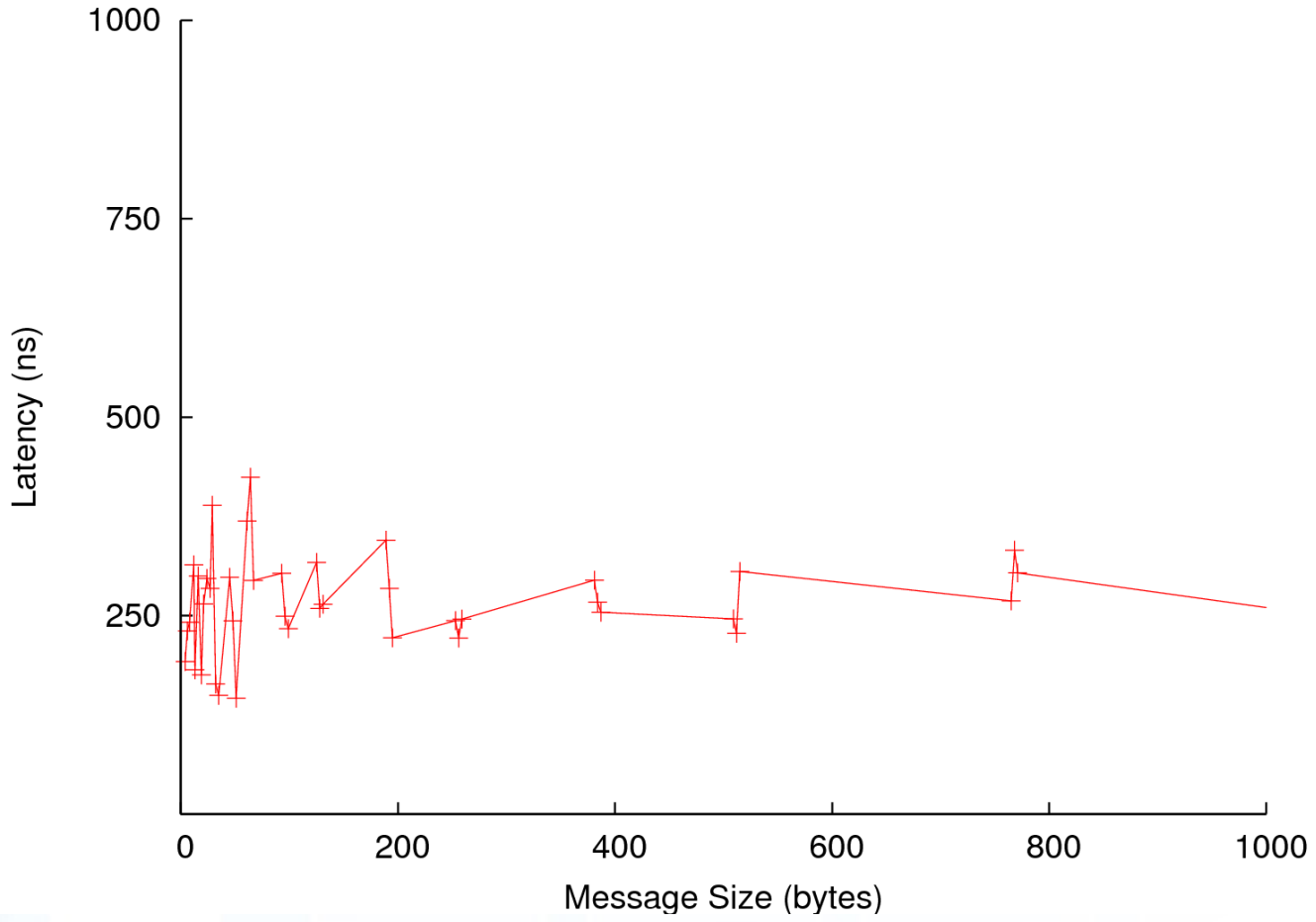
Throughput

NetEffect performance cont...



Latency

Switch overhead

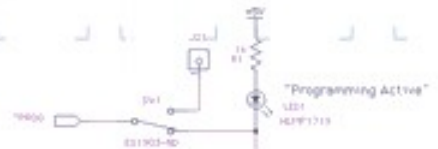


*Fulcrum Micro Test Platform

OSC

10 Gig iWARP

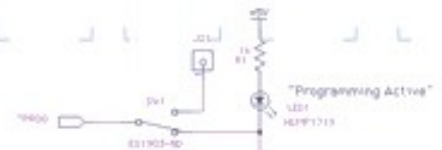
- Comparable (better?) in performance to IB
  - Higher throughput than standard 4X IB
  - Switch latency is comparable
  - A bit higher latency at small message sizes
  - Appropriate for cluster interconnect
  - Appropriate for high-end servers
  - Appropriate for storage (iSCSI)
- Just getting started with it
  - WAN tests
  - Interoperability with other iWARP HW



OSC

Current iWARP work

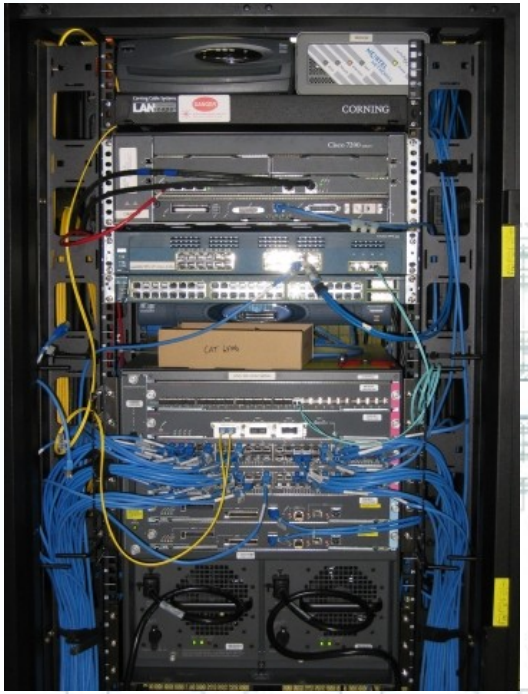
- iWARP in the WAN
  - Main point of this talk
- Interoperability of iWARP devices
  - Ammasso, NetEffect, Software iWARP
- RDMA enabled web server
  - Apache *mod_rdma* and proxy server
- RDMA enabled FTP client/server
- Real applications with NetEffect device



OSC

OSC iWARP resources

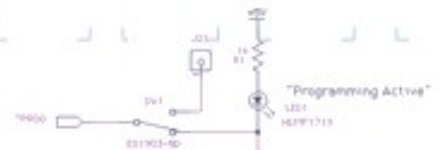
- NetRes Cluster
  - Up to 41 Ammasso
  - On TFN
- P4 Cluster
  - 17 Ammasso
  - On TFN
- NetEffect
  - 2 Servers
  - On TFN



OSC

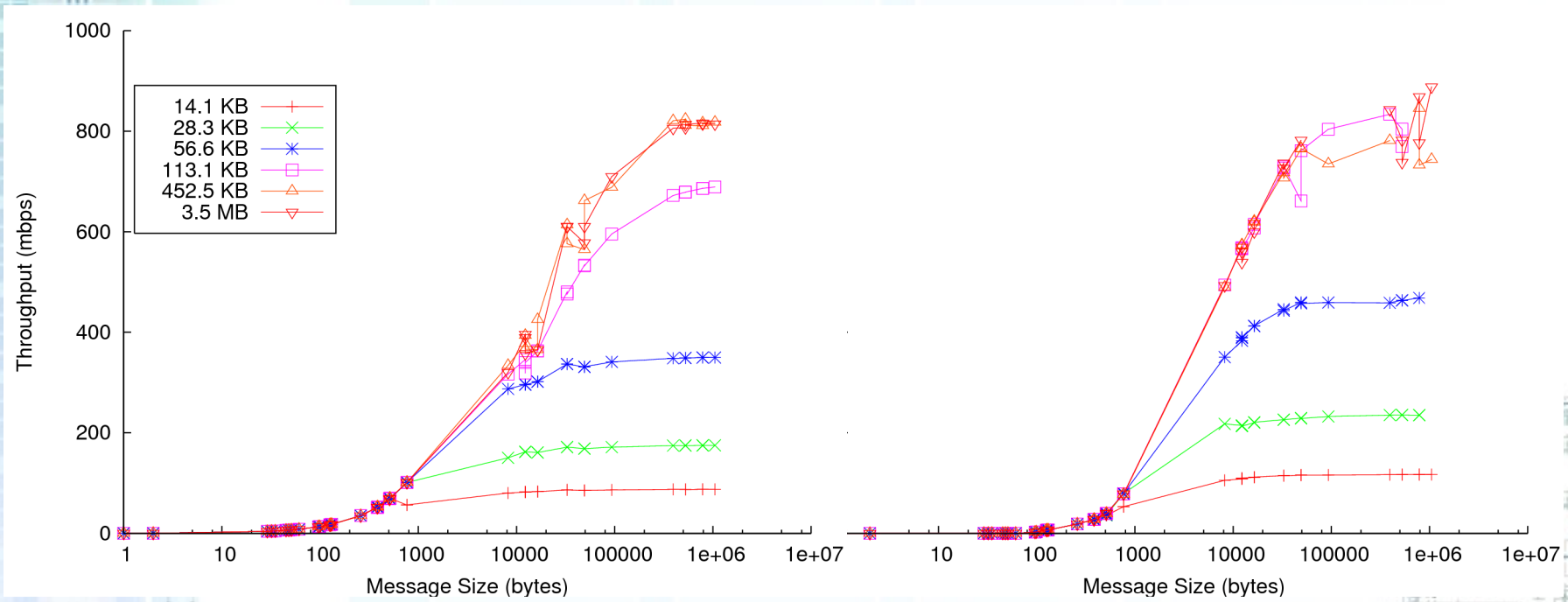
Basic performance

- 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?



OSC

Window size effect in WAN



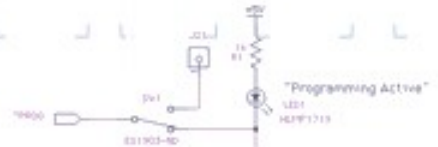
iWARP

TCP

*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

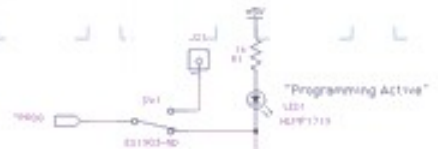


OSC

Basic FTP Performance

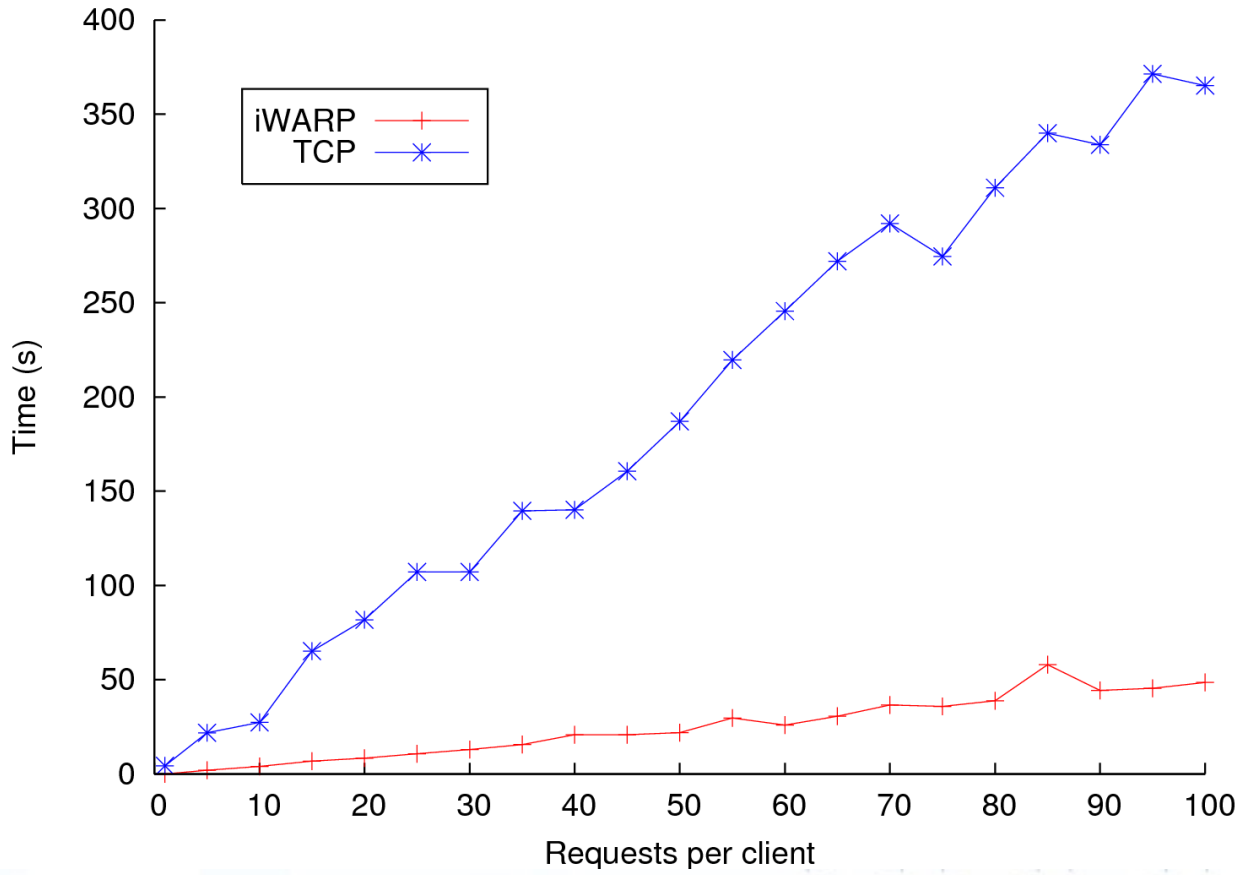
- Server: Springfield
- Client: Columbus
- Link: 10Gbps (TFN)
- About same perf

	iWARP	TCP/IP
200K	.010 s	.015 s
1M	.021 s	.031 s
10M	.117 s	.247 s
100M	1.05 s	2.59 s



OSC

The iWARP Benefit

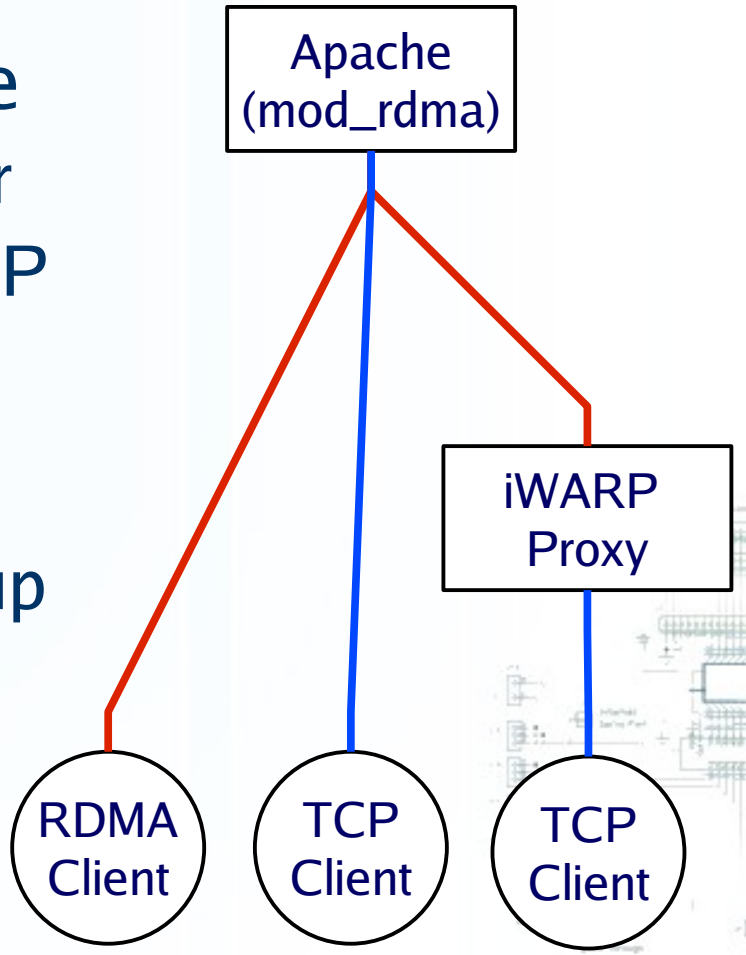


*16 clients each time

OSC

iWARP in the WWW

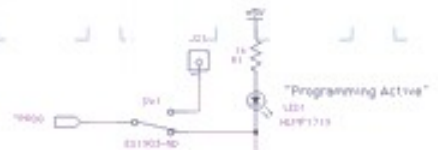
- RDMA module for Apache
  - Add RDMA info in header
  - Request/Response in TCP
  - Data transfer with RDMA
- Downside
  - RDMA connection build up
- Why not full iWARP port?
  - Extensive changes to Apache code base
  - New feature not new web server



OSC

RDMA enabled Apache web server

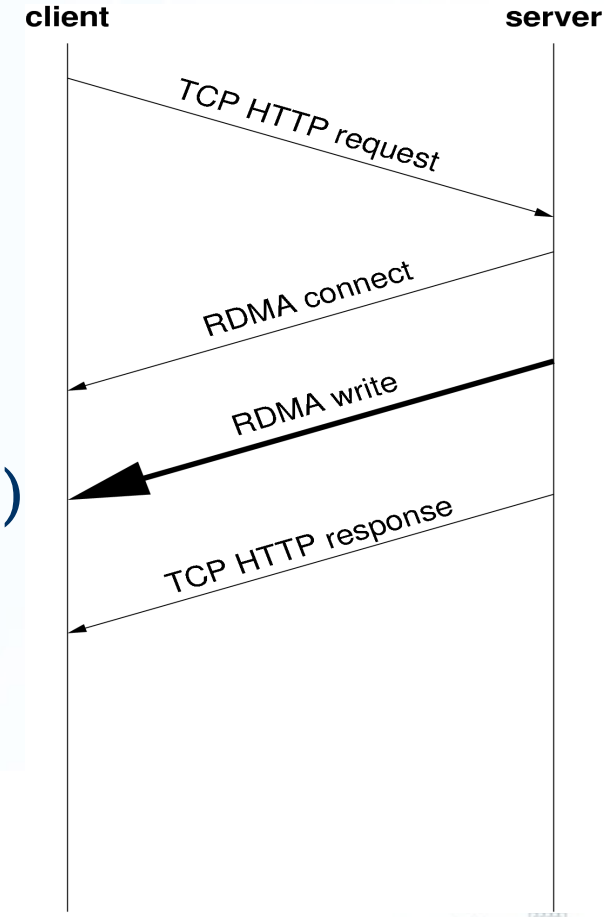
- mod_rdma
  - Apache module to handle RDMA transfers
  - “Grab” out going data and ship it with RDMA
  - Manipulate headers for minor changes
    - Simple changes, nothing fundamental
  - All the benefits of Apache
    - No rewrite of Apache code needed
    - Utilizes Apache *hooks*



OSC

mod_rdma cont.

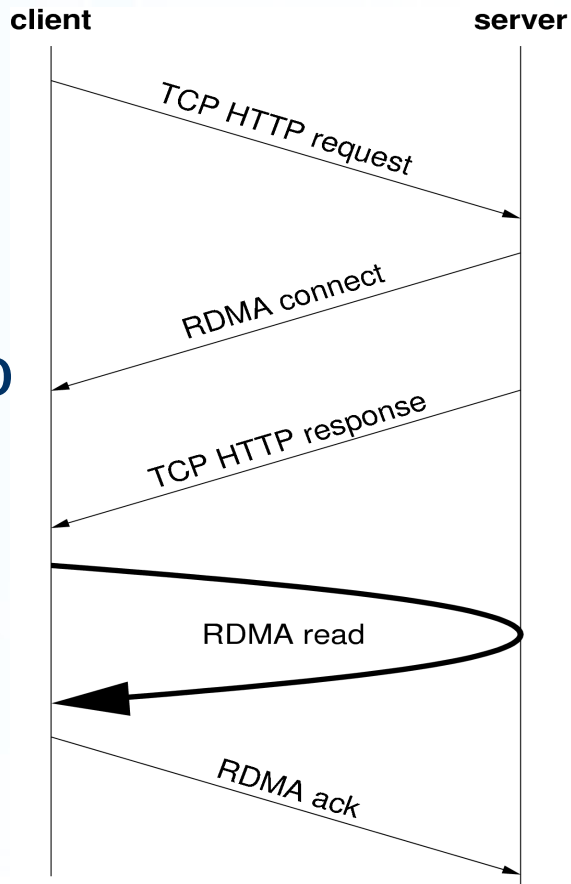
- Server Writes
  - Client has to guess size of file
  - RDMA connect takes time
    - Same as TCP connection (it is)




```
GET /index.html HTTP/1.1
Host: www.osc.edu
User-Agent: Mozilla/5.0
Connection: Keep-Alive
RDMA: server-writes, ip=10.0.0.15, port=3242, stag=642, to=0, maxlen=1048576
```

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

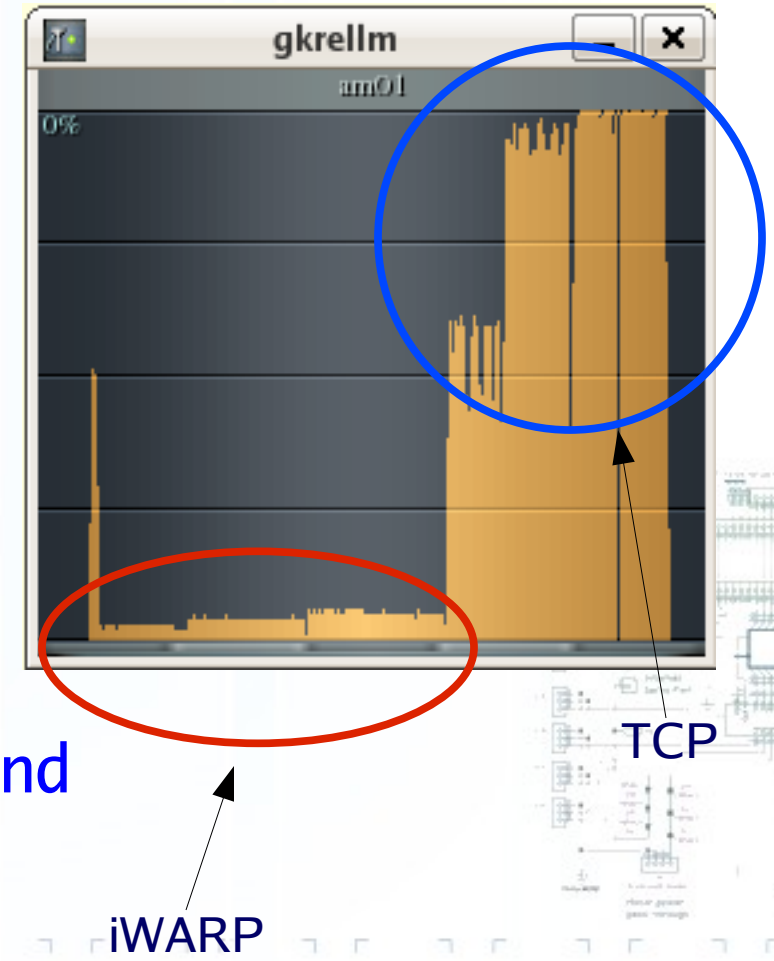


GET /index.html HTTP/1.1
Host: www.osc.edu
User-Agent: Mozilla/5.0
Connection: Keep-Alive
RDMA: clinet-reads, ip=10.0.0.14, port=3242 maxlen=1048576

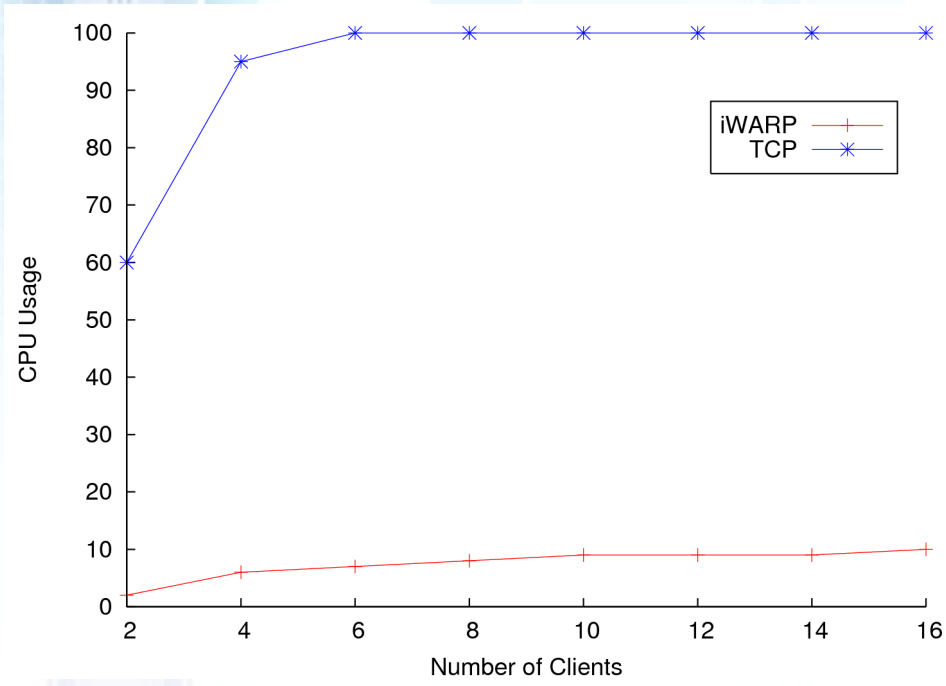
HTTP/1.1 200 OK
Host: www.osc.edu
Content-Length: 1327
Connection: Keep-Alive
RDMA: client-reads, stag=642, to=0

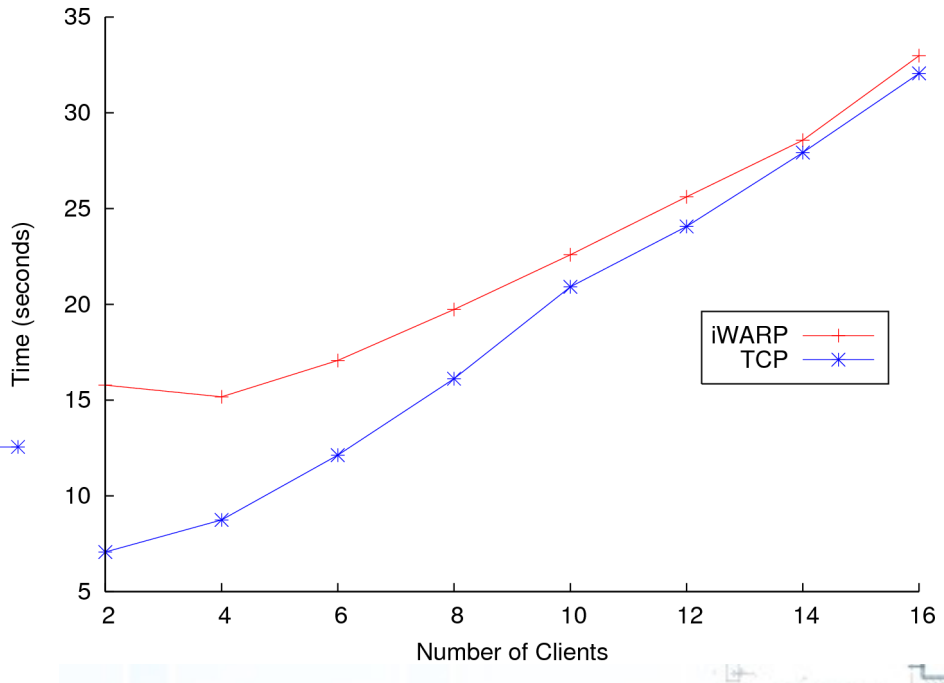
RDMA enabled Apache performance

- 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..





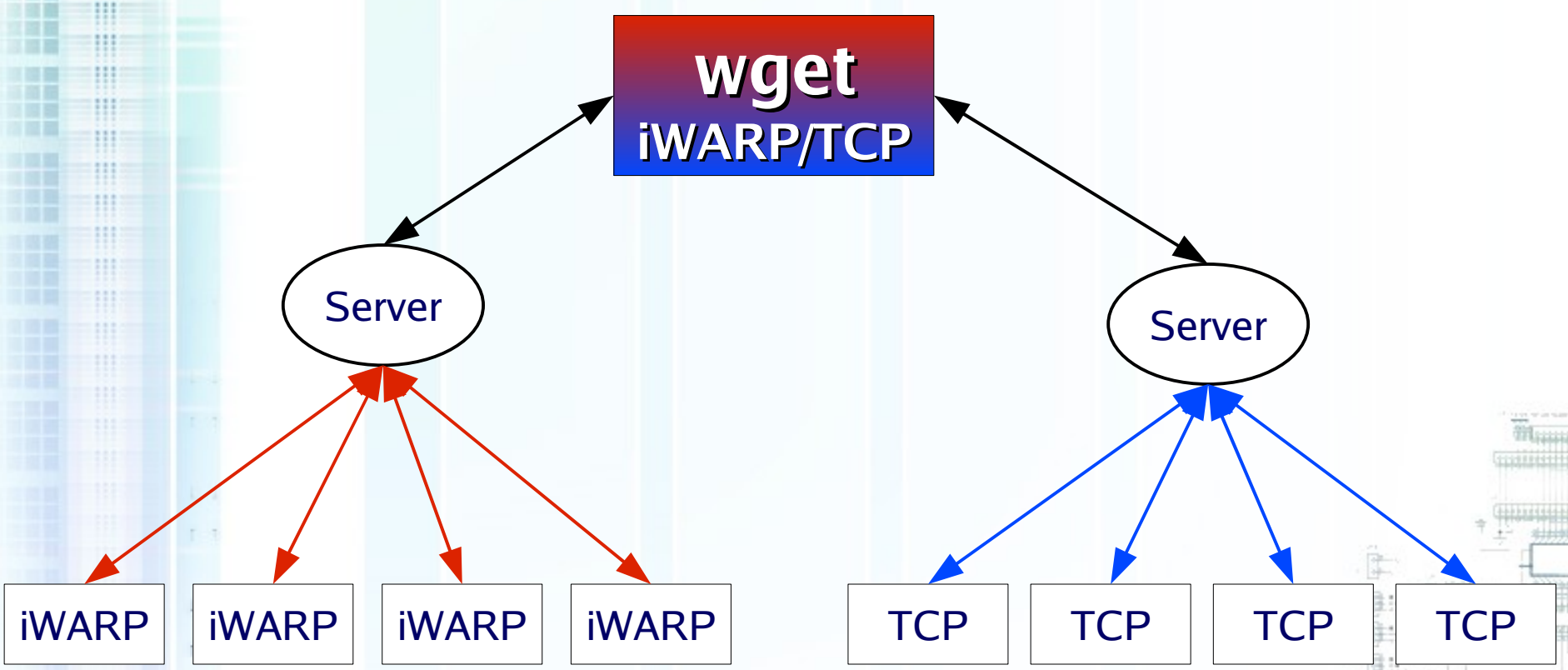
OSC

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

OSC

Two scenarios

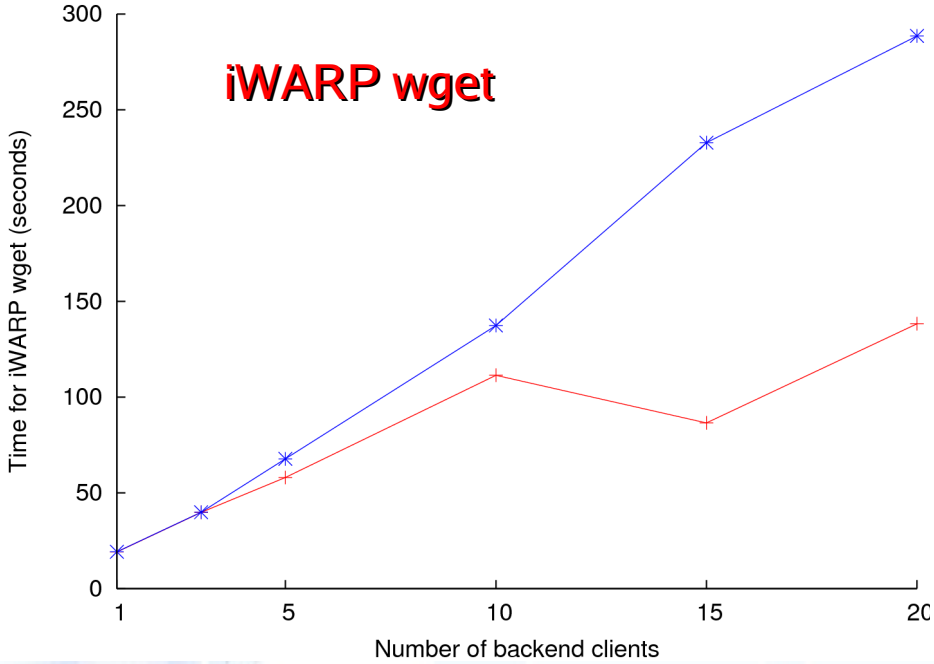


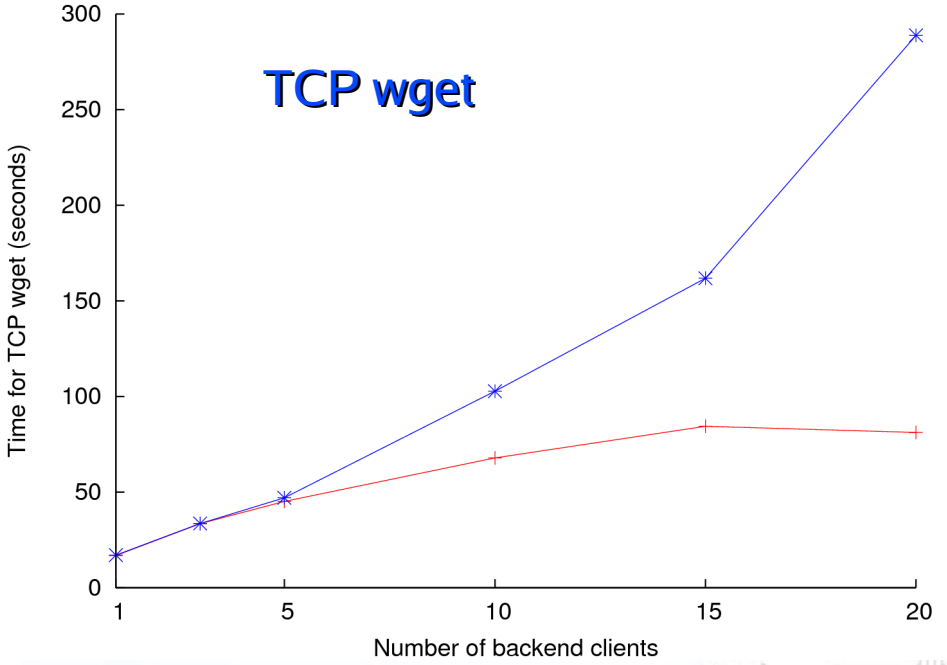
1. Back end RDMA clients

2. Back end TCP clients

OSC

Sample app performance



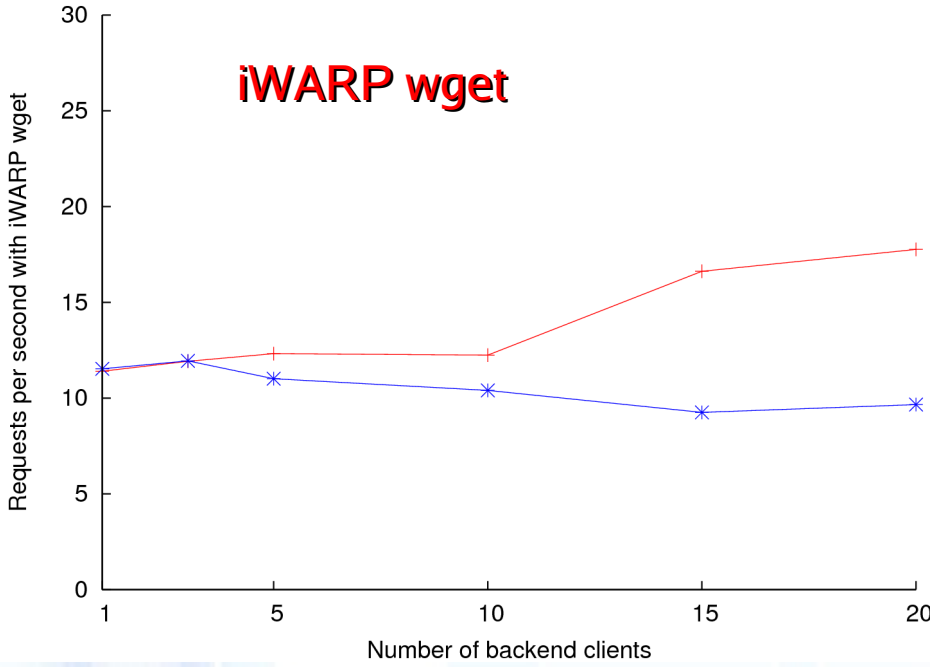


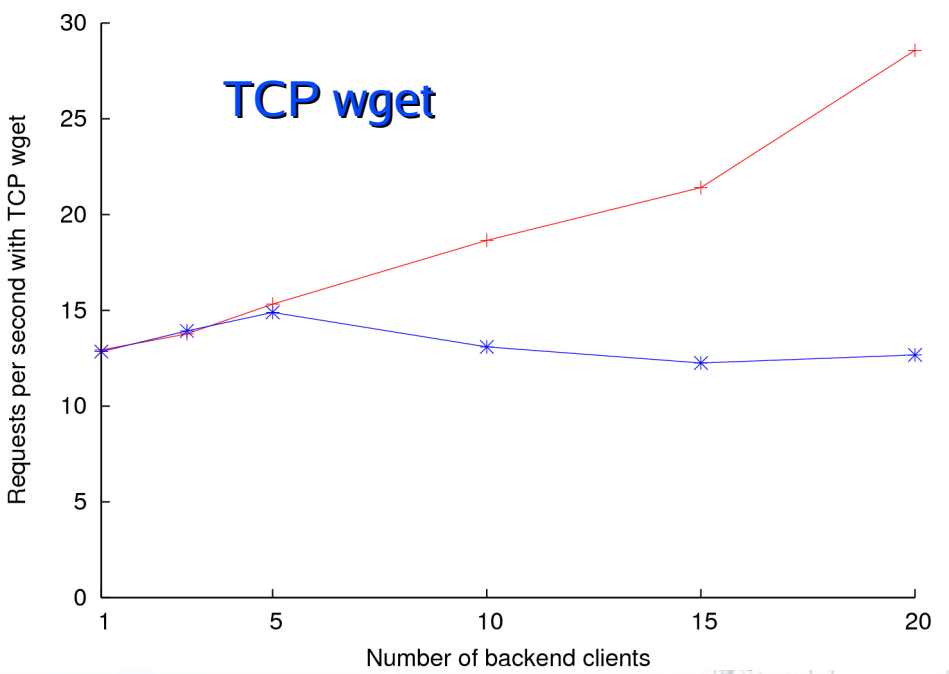
back end iWARP

back end TCP

OSC

Server performance





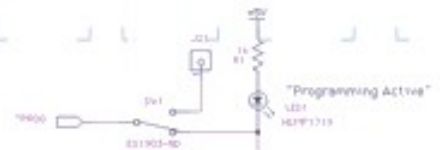
back end iWARP

back end TCP

OSC

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



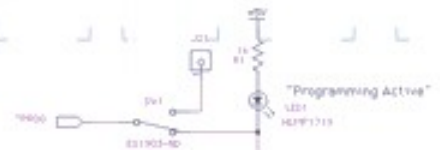
OSC

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



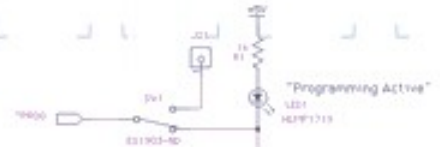
OSC

Questions?

Dennis Dalessandro

dennis@osc.edu

http://www.osc.edu/~dennis/iwarp



OSC