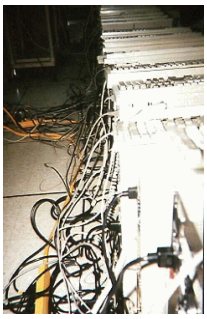
PRECISION™

Performance Networking

Commodity Ethernet Clusters - Maximizing Your Performance

Bob Felderman
CTO
Precision I/O

Ethernet Clusters: What do you think of?



PRECISION

Ethernet Clusters: What you'd like to see

High performance clusters using commodity Ethernet technology to accomplish HPTC tasks.

3

PRECISION

Ethernet Clusters: The Reality

- **Rack 'n' Stack with Gigabit Ethernet switches. Generally 1U or 2U systems.**
- **Often home-grown, especially at Universities**
- **Certainly available pre-configured from many vendors (Dual GigE is standard on many motherboards).**
- **Most clusters installed today *are* GigE.**
- **Blade-based system backplanes are now GigE and Infiniband too. Some PCIExpress soon.**

4

PRECISION

Why so few “big iron” Ethernet clusters?

- **Performance Performance Performance**
  - Bandwidth
    - Is 1 GigE sufficient?
  - Latency (time to get a message from here to there)
  - Overhead (work the CPU has to do per message)
  - Switch issues
    - Latency
    - Size (number of ports)
    - Flow-Control (xon/xoff messages versus hardware flow-ctrl)
- **What about Price/Performance? Does it matter?**
  - Acquisition cost and ongoing cost of operation
- **Ethernet clusters *are* employed successfully today for embarrassingly parallel problems or for small-scale parallel tasks.**

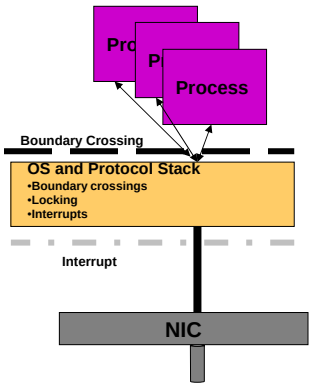
PRECISION

Bandwidth

- **Is 1GigE enough?**
  - Dominant cluster interconnect is Myrinet at 2Gig
- **10Gigabit Ethernet is available today**
  - Intel, Neterion, Chelsio and others
- **Various physical media**
  - Fiber is common (and expensive)
  - Copper will happen in some form
    - CX-4 used by IB and appearing in 10GigE
    - Twisted Pair - a must for economic deployment?
      - Fat copper versus thin fiber
- **Some standardization efforts at 2Gig and 5Gig**
  - Not likely to happen unless the PHY is the big issue

PRECISION

Latency/Overhead (*this* is the problem)



- **Ethernet Latency measured via the standard sockets kernel path.**
  - Traditionally this has been expensive
  - System calls, interrupts, kernel locks
  - May no longer be so bad
    - 3GHz Xeon running linux-2.4.20 20usec one-way latency via kernel path
  - But the 20usec is mostly time that the CPU is busy sending or receiving (overhead!)
  - Quoted Ethernet latency is highly CPU-speed dependent. IB, Myrinet, Quadrics mostly independent of host CPU.

PRECISION

Latency/Overhead Solutions

- **"TCP is heavyweight"**
  - TCP Offload Engines (TOEs)
    - Move protocol processing to the NIC
    - But, is protocol processing really the performance problem?
- **Data movement is expensive**
  - TOE + RDMA-NICs (RNICs)
    - Allows direct data placement like VIA/Infiniband
    - Also has OS bypass opportunities
    - Cache load is the expensive operation, so RDMA benefit may be less than expected for HPTC applications
      - data wants to meet processing cycles
- **Infiniband, Myrinet, Quadrics etc. have direct user-level access to the network**
  - Much easier to get low latency that way

PRECISION

What is the “Best” Host Interface/API?

- **Traditionally, HPC programmers have been willing to try anything to get performance.**
  - But, MPI is really the application programming paradigm of choice.
  - MPI over GM, over IB, over VIA etc.
- **MPI over Sockets (e.g. MPICH ch_p4) is the common API for Ethernet solutions.**
- **If you want to leverage mass market and commodity parts, then sockets/IP/Ethernet is the way to go!**
  - Assuming you can get the performance you want/need
  - You improve non HPC apps at the same time
  - CPU Overhead is the important metric
    - (UC Berkeley) R. Martin, A. Vahdat, D. Culler, T. Anderson. **Effects of Communication Latency, Overhead, and Bandwidth in a Cluster Architecture**. International Symposium on Computer Architecture (ISCA) , Denver, CO. June 1997

9

PRECISION

Ethernet: The Precision I/O Contribution

- **100% IP/Ethernet host-side solution**
  - Software-only at 1Gig – *in Beta test today*
  - Hardware and Software for 10GigE
- **Dramatically increases server application capacity**
  - Higher I/O throughput
  - Improved CPU utilization / Higher transaction capacity
  - Lower latency
- **General-purpose**
  - Supports a broad range of enterprise-class applications
  - Benefits both short transactions and streaming workloads
- **Non-disruptive**
  - Supports incremental deployment - one server at a time
  - Only required on one end of the connection - no client changes
  - No new fabric, protocols, packet formats, or infrastructure required
  - No application or operating system changes required

10

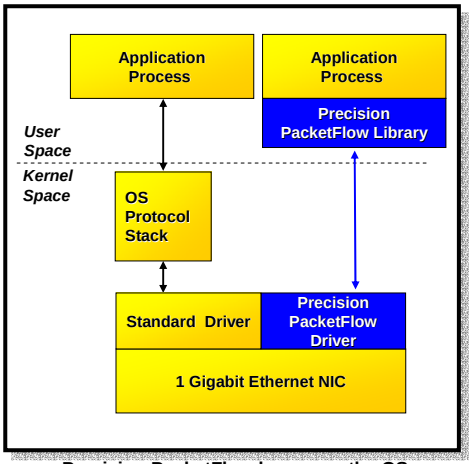
PRECISION

Precision I/O: Solution Overview

- **Speed Performance for Today's 1 Gigabit Ethernet Network**

- **PacketFlow MPI**
  - Software product that extends the life and investment in existing GigE networks
  - Increases MPI performance by 2x
  - Improves associated *application* performance 20-50%
  - Available for Beta Test Today!

Q3' 2005



Precision PacketFlow bypasses the OS, significantly speeding network I/O performance

11

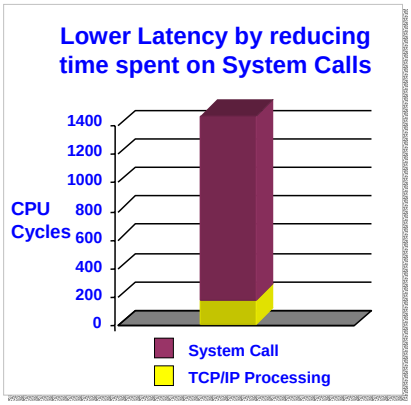
PRECISION

Precision I/O: Solution Overview

- **Speed Performance for Today's 1 Gigabit Ethernet Network**

- **PacketFlow MPI**
  - Software product that extends the life and investment in existing GigE networks
  - Increases MPI performance by 2x
  - Improves associated *application* performance 20-50%
  - Available for Beta Test Today!

Q3' 2005



12

PRECISION

Precision I/O: Solution Overview

- **Speed Performance for Today's 1 Gigabit Ethernet Network**
  - **PacketFlow MPI**
    - Software product that extends the life and investment in existing GigE networks
    - Increases MPI performance by 2x
    - Improves associated *application* performance 20-50%
    - Available for Beta Test Today!

Q3' 2005

- **Deliver Higher Bandwidth Standard Ethernet (10Gb) for Future Networks**
  - **PacketDirect 10Gigabit Ethernet**
    - Adds hardware in combination with Precision PacketFlow™ software to deliver 10-gigabit, wire rate performance.
    - Incremental migration from today's 10 GigE deployments
    - Standards based
    - Easy to support
    - Affordable

H1' 2006

13

PRECISION

Ecosystem Issues

- **What are the ecosystem concerns for Ethernet?**
  - Fortunately they are quite limited
  - Ethernet/IP is ubiquitous
  - Customers in other domains (and HPTC too) want Ethernet and IP everywhere
  - Key issue: Switch Performance and Scaling
    - Number of ports
    - Latency through switch

14

PRECISION

Current Ethernet Switch Landscape

- **Most current 1GigE switches have too few ports and too much latency to be interesting.**
  - Typically 5 to 50usec
    - when you have 100usec latency on the host, who cares about the switch latency?
    - Latency is often directly proportional to price! (*NOT* inversely)
  - Small number of ports means a deep tree of switches and more latency end-to-end.
- **10GigE switches don't look much better**

15

PRECISION

If you want it, you've got to ask

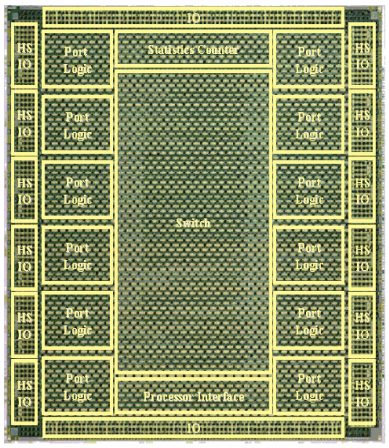
- **Switch vendors are not *yet* focused on HPC.**
  - These features don't help HPC clustering
    - Layer 4-7 "awareness"
    - Fire walling
    - Load balancing
  - Adding 10GigE as an uplink isn't sufficient
  - Go out and ask for a layer-2 switch with
    - high-port count
    - low-latency (1usec or less)

16

PRECISION

But at least one vendor is on board!

- **Fujitsu Labs of America, *"A Single Chip Shared Memory Switch with Twelve 10Gb Ethernet Ports"* Hot Chips 15, August 2003**
  - Focus on layer-2 switching
  - High-throughput, low-latency core
  - SerDes integration
  - Copper 700ns latency, Optical 1,200ns
- **Shipping today in switches (450ns port-to-port latency)**



Packing a dozen bidirectional 10-Gbit/s Ethernet ports and a memory-based switch fabric on a single chip, the MB7Q3050BYL dramatically reduces the cost, power, and size of high-performance Ethernet switches.

PRECISION

Summary: Fast Transparent Performance

- **Extends the life and investment of existing Gigabit Ethernet networks**.
  - Precision PacketFlow™ MPI significantly improves the performance of your existing MPI applications
  - Available for beta testing today!
- **Transparently increase application performance**
  - Precision I/O solutions deliver dramatic performance improvements yet require no changes to your application, OS or network
- **Provides a gradual migration path to future 10 GigE networks**
  - Start with Precision PacketFlow today to increase 1 GigE performance
  - Migrate to Precision PacketDirect 10GigE NIC gradually over time
    - Supports incremental deployment - one server at a time
    - Only required on one end of the connection - no client changes
    - No new fabric, protocols, packet formats, or infrastructure required
    - No application or operating system changes required

PRECISION

PRECISION™

Performance Networking

Precision I/O
4005 Miranda Ave., Suite 210
Palo Alto, CA 94304-1232
650-331-8000
http://www.precisionio.com
Bob.Felderman@precisionio.com

Company

Founded:	February 2003 (spun out of Packet Design LLC)
Location:	Palo Alto, CA
Employees:	45
Funded:	Q1 2004, Q3 2005: Foundation Capital, ATV, 3i
Market status:	1 Gigabit Ethernet performance boosting software in beta today

- **Derek Proudian,** CEO
  HP, Mohr Davidow, Zip2, Panasas
- **Bob Felderman,** CTO
  USC/ISI, Myricom
- **Ed Roseberry,** VP OEM & Biz Dev
  HP, 3Com, Alteon, Nortel, S2io/Neterion
- **Dan O'Farrell,** VP Marketing
  HP, N.E.T., interWAVE, Sun, Peribit
- **Judy Estrin**, Chairman
  Cisco, Precept, NCD, 3Com, Bridge
- **Van Jacobson**, Chief Scientist
  Cisco, Lawrence Berkeley Labs
- **Narendra Dhara**, VP Engineering
  LSI, NEC, Corona Networks, IntruGuard

PRECISION