HP Cluster Interconnects: The Next 5 Years

Michael Krause
mkrause@hp.com
September 8, 2003
• The Problem Cluster Interconnects Are Trying To Solve
• Proprietary vs. Industry Standard Interconnects
• Impact of Local I/O Technology
• Today’s Reality – Measured Performance
• What the Future Holds
The Problem: Taxes are Too High

- Like a “value-add tax”, OS + network stacks imposed taxes (overheads) at each stage of message and per network packet processing.
- As workloads becoming more distributed, growing percentage of solution cost goes to paying “taxes” rather than running applications.

- To provide customers with tax relief, the underlying solution infrastructure requires a new communication paradigm.

1. User / Kernel Context Switch
2. Copy to / from user buffer and network buffer
3. Packet protocol stack processing – per packet
4. DMA to / from network buffer
5. Device control including interrupt post processing for DMA read / write completions
RDMA delivers improved system efficiency

- Provides resource “tax” relief from the overheads imposed by operating system (OS) and network stack implementations.
- Provides message exchange “tax” relief from the overheads imposed by existing communication paradigms.

Benefits delivered through the:

- Elimination of intermediate network buffers – reduces system memory consumption associated with the network stack.
- Reduction / elimination of CPU required to access local or remote memory due to direct access and placement of data buffers – no copy operations or processing required.
- Reduction / elimination of CPU to perform message segmentation and reassembly via hardware-based acceleration. The CPU reductions on segmentation have been clearly demonstrated for the transmit path on various OS-based network stack implementations within the industry. RDMA enables these same savings to be accrued on the receive path for the system. A single receive completion is used to replace the per packet completion paradigm used in OS-based network stacks.
- Elimination of interrupts through well-defined completion semantics – no longer require DMA read / write completion processing. In addition, completion semantics enable implementations to reduce the number of process / thread context switches required to complete an application message exchange.
- Elimination of user / kernel context switch to send or receive data.
- Elimination of the network stack protocol processing within the system. This reduces CPU, memory, and I/O resource consumption as both the protocol processing costs as well as the exchange of non-application control traffic required by the network stack are off-loaded to the RNIC.
A. SDP (Sockets Direct Protocol) interposed between Sockets and RDMA NIC (RNIC). SDP enables SOCK_STREAM applications to transparently operate over RNIC. SDP interacts with the RNIC directly to process application and SDP “middleware” message exchanges. Enables OS Bypass.

B. Direct DMA to/from user buffer. No interrupts are required as completion processing is performed within SDP layer.

C. All protocol processing, memory access controls, etc. implemented in RNIC enabling complete off-load from the system.
Agenda

• The Problem Cluster Interconnects Are Trying To Solve
• Proprietary vs. Industry Standard Interconnects
• Impact of Local I/O Technology
• Today’s Reality – Measured Performance
• What the Future Holds
Data Center Infrastructure Evolution

**Today**
- Fibre Channel
- NAS (Storage over IP)
- 1 Gigabit Ethernet
- KVM over IP (Lights-out Management)
- Proprietary Solutions (ServerNet, Hyperfabric, Quadrics, etc.)

**Tomorrow**
- 10 Gigabit Ethernet
- 4 Gigabit Fibre Channel
- iSCSI (Storage over IP)
- IP acceleration (TCP/IP & IP Sec)
- IP Fabrics (RDMA/TCP)
- InfiniBand Fabrics

**Networking**
- 1 Gigabit Ethernet
- KVM over IP (Lights-out Management)

**Storage Fabric**
- 10 Gigabit Fibre Channel
- 4 Gigabit Fibre Channel
- iSCSI (Storage over IP)

**Data Center Fabric**
- 10 Gigabit Ethernet
- IP acceleration (TCP/IP & IP Sec)
- IP Fabrics (RDMA/TCP)

**Compute Fabric**
- InfiniBand Fabrics
InfiniBand Technology Overview

- **Point-to-Point, Switch-based Fabric**
  - Solid fault-containment, fault-isolation, management, hot-plug
  - Improved performance isolation
  - Scales up to tens of thousands of ports / nodes
  - Very low-latency switches available

- **Three bandwidth levels**
  - 0.5 GByte/s, 2.0 GByte/s, 6.0 GByte/S
    - New double / quad rate signaling under development
  - Three different link widths 1X, 4X, 12X

- **Message-based communication:**
  - Channel-based - Send / Receive
  - Memory-based - RDMA (Remote DMA) and Atomics
  - Asynchronous Completion Notification
  - Centralized fabric management

- **Multiple hardware topologies**
  - ASIC-to-ASIC, Board-to-Board, server-to-server
  - Copper and optical cabling
  - Primarily focused on cluster IPC market with some shared I/O
InfiniBand Benefits

• Ecosystem is nearing completion
  • IBTA (InfiniBand Trade Association) completed all requisite specifications
    – Continuing to evolve the technology where needed
  • IBTA members delivering hardware / software today
    – HCA, Switches, IP routers, Subnet Manager, etc.
    – Interoperable hardware is operational and demonstrated
      – Performance metrics are being gathered to feed into next generation designs / implementations
• Industry standard RDMA API Completed
  – ICSC (OpenGroup) completed IT API
  – Enables OS / RDMA fabric independent (portable) MPI, Sockets, Database, kernel subsystems, etc. to be delivered
• Multiple OS provide solid RDMA infrastructure today
  – Unix, Linux, etc.
• Customer-visible performance benefits of protocol off-load, OS bypass, and RDMA validated
  • Industry standard interconnects performance validated across wide range of design points and infrastructures
  • Clearly demonstrates that the concepts and associated technology are mature and moving to mainstream
iWARP (RDMA / TCP) Technology Overview

- Point-to-Point, Switch-based Fabric
  - Industry moving to develop low-latency switches
- Multiple bandwidth levels
  - Scales with Ethernet – 1Gbps to 10 Gbps (today)
  - Future will scale to 40 / 100 Gbps (2010)
- Re-uses existing IP / Ethernet ecosystem
  - Switch / router infrastructure, management, etc.
  - Solid, low-cost interoperability
- Message-based communication:
  - Channel-based - Send / Receive
  - Memory-based - RDMA (Remote DMA)
  - Asynchronous Completion Notification
- Multiple hardware topologies
  - ASIC-to-ASIC, Board-to-Board, system-to-system / storage
  - Copper and optical cabling
- Many opportunities for differentiation, e.g.
  - RNIC may expose all four interfaces, integrate IP Security, provide transparent fail-over between ports, port aggregation, QoS, etc.
iWARP Benefits

- Ecosystem under rapid development
  - RDMA Consortium (RDMAC) specifications completed
    - Wire protocols, Verbs, iSCSI Extensions (iSER), etc.
    - Nearing completion of Sockets Direct Protocol (SDP)
  - RDMAC provided drafts to IETF; working to align
  - Industry standard RDMA API Completed
    - ICSC (OpenGroup) completed IT API
      - Minimal extension needed to optimize for iWARP
      - Enables OS-independent (portable) MPI, Sockets, Database, kernel subsystems, etc. to be delivered
  - Multiple OS provide solid RDMA infrastructure
    - Unix, Linux, etc.

- Enables converged fabric for IPC, Storage, etc.
  - Re-uses existing data center / OS / Middleware management
  - Re-uses existing IP / Ethernet infrastructure
    - Lower cost to integrate into existing and new solutions
  - Reduces hardware costs

- Application across all design points
  - Can be integrated into chipsets, backplanes, adapters, etc.
Agenda

• The Problem Cluster Interconnects Are Trying To Solve
• Proprietary vs. Industry Standard Interconnects
• Impact of Local I/O Technology
• Today’s Reality – Measured Performance
• What the Future Holds
Why is Local I/O Technology Important

• Each I/O operation = multiple I/O transactions
  • Each transaction consumes local I/O bandwidth
    – Function of technology protocol efficiency as to how much bandwidth is available for I/O device
  • All discussion about speed matching, future-proofing is primarily all marketing hype
    – Problem is in the design / implementation of the I/O subsystem and I/O device (including its driver paradigm)

• I/O Latency to Memory Impact
  • Memory bandwidth is the gating factor in I/O performance
    – Memory bandwidth increases perhaps 10% per year
  • If limited concurrent transactions, I/O latency to memory will have negative impact on delivered performance
    – Potential 25-50% negative impact on device bandwidth

• Concurrency
  • Just as with processors, increased concurrency in the number of I/O transactions is a requirement
  • As I/O device performance increases, concurrency in the number of I/O operations is a requirement
    – Number of simultaneous application transaction rates will be gated by number of concurrent I/O operations
Local I/O Technology Recommendations

- Use point-to-point implementations
  - Provides strong fault isolation / containment, multiple management domains, predictable performance, improved / easy hot-plug, simplified configuration
  - HP pioneered and delivered point-to-point in late1990’s across entire range of design points illustrating these customer-visible benefits
    - These concepts and knowledge codified in PCI-X 2.0 and PCI Express specifications
- Require highly concurrent I/O infrastructure and devices
  - I/O latency to memory is the gating factor irrespective of local I/O technology used
- Use servers with flat local I/O topologies
  - Don’t add yet another switch domain between the application and the hardware
    - Only increases head-of-line blocking / congestion
      - Poor I/O operation throughput – (poor cost / benefit)
- Use balanced systems – Processors, memory, and I/O
  - Avoid the speeds-n-feeds trap
    - Examine delivered I/O bandwidth per slot as well as aggregate for system
    - Volume of server I/O requires less than 1 GB/s of local I/O B/W for many years to come
      - Multi-port 2 Gb FC, new 4 Gb FC, multi-port GbE, new 2.5-5.0 Gb Ethernet, etc.
- Keep in mind:
  - It is the solution design / implementation that matters most – technology != solution
The Growing Debate: PCI-X 2.0 and PCI Express

• PCI-X Ecosystem – strong, high-volume ecosystem shipping today
  • Strong interoperable, compatibility product offering enables fast adoption of new implementations in product environments with minimal customer validation
    – Broad OSV, IHV, ISV support with all major I/O device types shipping
    – Strong OSV, IHV, ISV support for PCI-X 2.0 for all major I/O device types

• PCI-X 266 will start ramp to volume in early 2004
  • Numerous designs and implementations completed, plug-fests starting, etc.
  • Strong customer need for high-speed I/O – 10 Gbps, multi-port, etc.
    – High-volume potential as well as strong customer investment / future proof protection

• PCI Express 2004/2005 will be spent on:
  • Fixing all errata since 1.0a PCI Express specification release
  • Incorporating key learning from initial client development experience (starts to ship in mid 2004)
    – Could lead to radical redesigns when combined with new volume process technology
  • Preparing for design requirements detailed by HP and by others within the industry
  • Evaluating new form factor (SIOM) impact on designs – function of cost / market segment need
  • Evaluating gen2 signaling and its value / impact on chipset, switch, bridge, and device designs
    – Given high volume potential in client space, this may have a major impact on many solutions provided
  • Developing the OS / Driver / PCI Express software infrastructure needed to meet customer requirements and provide value – expect delivery in 2006

• PCI-X 2.0 will be the dominant I/O device / slot for servers for many years
  • As PCI Express starts to ramp to volume in 2006 and with advent of integrated devices, expect to see a gradual decline in PCI-X 2.0 though continuing to ship through at least 2015-2020
Agenda

• The Problem Cluster Interconnects Are Trying To Solve
• Proprietary vs. Industry Standard Interconnects
• Impact of Local I/O Technology
• Today’s Reality – Measured Performance
• What the Future Holds
Where is the time spent

- **RDMA Infrastructure** spends majority of end-to-end packet exchange time in hardware
  - OS bypass / Protocol Off-load / RMDA enables system resources to be spent on application rather than networking
- **Prototype Measurements using 1st gen InfiniBand HCA** on volume OSV infrastructure with PCI-X 133
  - **Transmit endnode:**
    - S/W: Generate Work descriptor and ring doorbell 0.9 usec
    - H/W: Ring doorbell on PCI bus 0.3 usec
    - H/W: Process work descriptor: 1.3 usec
    - H/W: Read data and emit on IB port: 0.9 usec
    - IB Switch packet transmission latency 0.1 usec
  - **Receiving endnode:**
    - H/W: DMA Read receive work descriptor 1.0 usec
    - H/W: DMA Write data to host memory 0.7 usec
    - H/W: DMA Write completion queue 0.5 usec
    - H/W: Processor cache flush 0.3 usec
    - S/W: Poll for completion 0.6 usec
  - Some local I/O transactions to complete each I/O operation are pipelined with interconnect packet transmission thus do not impact end-to-end latency
• Observations:
  • Latency is relatively flat across relatively wide range of sizes where latency matters most
  • RDMA Write operations provide better performance and are more efficient
    – Use advertised target memory rather than post receive buffers
    – Fewer local I/O transactions as RDMA Writes do not require completion on receiver
  • RDMA Read latency (not shown) slightly higher than Send since two traversals of fabric and associated hardware interactions
• Observations:
  • Large amounts of idle CPU available for applications
    – Sender’s processor generally 98-99% idle irrespective of message size
    – Receiver’s processor generally 90-97% idle as a function of message size
    – Expect similar performance for iWARP
  • Load-balance across devices / ports will enable applications to reap benefits of RDMA infrastructure
    – Low-latency, High-bandwidth
  • Performance only gated by use of existing PCI-X 133 (1 GByte/sec of local I/O bandwidth)
    – Interconnect will linearly scale with improved local I/O
    – PCI-X 266 will provide 2GByte/sec of raw bandwidth – expect ~2x interconnect bandwidth perf
Overall Performance Observations

- Open, industry standard RDMA technology is real!
  - Experience from proprietary interconnects successfully transferred to standards
- Clear performance benefits of OS Bypass, Protocol Off-load, RDMA
  - Today,
    - Demonstrated with proprietary interconnects: HP Hyperfabric, HP ServerNet, etc.
    - Demonstrated with open, industry standards: InfiniBand
  - Starting in 2004
    - Soon to be demonstrated with open, industry standards: iWARP
      - Start with 1Gb Ethernet and progress through multi- and 10 Gb Ethernet
- RDMA Infrastructure is coming together
  - Hardware and software ready to evaluate and start deployment
  - Demonstrated performance will increase customer confidence in deploying in production environments beyond just the technical compute space
    - Distributed Database, ERP, Tier 1/2/3 data centers, etc.
  - Increased confidence will increase acceptance across design points and allow customers to choose which open standard interconnect best meets their needs
Agenda

• The Problem Cluster Interconnects Are Trying To Solve
• Proprietary vs. Industry Standard Interconnects
• Impact of Local I/O Technology
• Today’s Reality – Measured Performance
• What the Future Holds
Cluster Interconnect: Future Themes

- Economic “Darwinism” is reaping havoc with OSV / IHV / ISV
  - Technology consolidation occurring
  - Focused on fundamental interoperability at each solution “layer”
- Open, industry-standard infrastructure
  - Hardware standards
    - InfiniBand and iWARP will become the dominant interconnect technology
    - InfiniBand available today – demonstrated performance values and cost structure
    - “Ethernet Everywhere” will make iWARP high-volume / commodity solution in future
      - Combined with iSCSI / iSER to deliver converged fabric for higher volume
  - Software standards
    - IT API, Sockets API Extensions (OpenGroup), etc.
      - Enables application portability across OSV, platforms, etc.
    - SNMP, CIM, XML, etc. management infrastructure with plug-ins enables faster, transparent deployment of new technology and services
    - Adaptations of Sockets and MPI over industry standard RMDA
- Utility Computing
  - Efficiency gains from use of RDMA technology provide customer-visible value
    - Higher, more efficient utilization of hardware; improved endnode / fabric responsiveness
  - Interoperable interconnect enables dynamic, multi-tier / Grid services to transparently reap benefits
Cluster Interconnect: 
**Software Differentiation**

- Transparent single node, multi-device / port load-balancing
  - Multi-port devices enable port aggregation
    - Provide software controlled service segregation
  - Port device aggregation
    - 16 million (IB) / 4 billion (iWARP) endpoints per OS instance – transparently spread across multiple devices
  - Dynamically rebalance service to match:
    - Workload performance requirements
    - Hardware hot-plug / failure events

- Transparent multi-node load-balancing
  - Utilize standard infrastructure / wire protocol to redirect to “best fit” node
    - Multiple polices available: Available capacity, service resource locality (data), etc.

- Virtualization
  - Transparent port fail-over – enables recovery from external cable / switch failure
    - Core functionality specified in InfiniBand
    - Value-add functionality for vendors to implement in iWARP
      - Leverage existing port aggregation and fail-over infrastructure
  - Transparent device fail-over
    - Value-add functionality for vendors to implement over either interconnect type
Cluster Interconnect: Hardware Differentiation

- IB relatively consolidated
  - HP supplied industry “consolidated” solution requirements in May 2001
    - Industry executed to meet these requirements – have demonstrated interoperability
- RNIC designs have large opportunity for vendor differentiation
  - HP helping industry understand solution requirements though more variability expected
- Multi-port and port aggregation
  - Transparent fail-over across a set of ports
- Access to all protocol off-load layers:
  - TCP Off-load (TOE)
  - IP Security Off-load
  - IP Routing Off-load
  - Ethernet Off-load
    - Checksum Off-load (IPv4 and IPv6)
    - Large TCP Send
    - QoS Arbitration + Multi-queue + MSI-X
    - 802.1p Priority / 802.1q VLAN
- Ethernet Virtualization
  - Multiple MAC Address support
- Connection Caching (impacts of thrashing on ASIC and local I/O)
  - Side memory to provide high-speed device local cache
RDMA Infrastructure: 
**Solution Components**

- **Sockets API** (Sync/Async)
- **Ethernet**
- **IP (v4, v6) Offload**
- **TCP Offload**
- **IP (v4, v6) Offload**
- **Ethernet**
- **RDMA**
- **TCP/UDP/etc**
- **IP**
- **Host Stack**
- **RNIC**
- **RDMA Services**
- **SDP**
- **RNIC HAL**
- **IB HAL**
- **RDMA Sockets Application**
- **RDMA/Cluster Middleware**
- **IT API**
- **InfiniBand Transport**
- **InfiniBand Link**
### Product Availability

*Estimates for the Industry as a whole, product offerings from multiple vendors*

<table>
<thead>
<tr>
<th>Product</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>10GbE switch infrastructure</td>
<td>Now</td>
</tr>
<tr>
<td>10GbE NIC</td>
<td>2003</td>
</tr>
<tr>
<td>iSCSI to FC bridging</td>
<td>Now</td>
</tr>
<tr>
<td>iSCSI HBAs</td>
<td>2003</td>
</tr>
<tr>
<td>iSCSI HBAs with integrated IPSec</td>
<td>2004</td>
</tr>
<tr>
<td>iSCSI storage targets</td>
<td>2004</td>
</tr>
<tr>
<td>iSER storage targets</td>
<td>2005</td>
</tr>
<tr>
<td>InfiniBand HCAs, switches</td>
<td>Today</td>
</tr>
<tr>
<td>RDMA-based NAS</td>
<td>Today (IB), 2004/5 (iWARP)</td>
</tr>
<tr>
<td>RNICs (1GbE, 10GbE)</td>
<td>2004-2005</td>
</tr>
<tr>
<td>Low-latency Ethernet switches</td>
<td>2004-2005</td>
</tr>
<tr>
<td>IT API-based middleware</td>
<td>2004</td>
</tr>
<tr>
<td>RDMA-enable Async Sockets applications</td>
<td>2004-2005</td>
</tr>
</tbody>
</table>

*Does not indicate specific product plans from HP*
• HP is the technology invention engine for the industry
  • PCI, hot-plug, PCI-X, PCI-X 2.0, PCI Express, InfiniBand, iWARP, iSCSI, SAS, etc.

• HP drives technology invention in the industry
  • Founding member of the PCI SIG, RDMA Consortium, ICSC, IBTA, etc.
  • Lead developers / authors / co-chairs of numerous industry workgroups:
    – Electrical and Protocol for PCI, PCI-X, PCI-X 2.0, SHPC
    – Protocol, Electrical, Graphics, Mechanical, Software, etc. for PCI Express
    – RDMA, SDP, iSER for RDMA Consortium as well as iWARP within the IETF
    – iSCSI protocol, SNS, etc. for complete storage over IP solutions, SAS, T10/T11, etc.
    – Interconnect Software Consortium – APIs for new Sockets and RDMA services

• HP sets the industry direction by focusing on customers:
  The **right** solution using the **right** technology, at the **right** time
RDMA – Just Better Networking

Fast and secure communications
- remote direct memory access (RDMA) provides efficient memory to memory transfers between systems
  - much less CPU intervention needed
  - true “zero copy” between systems, data placed directly in final destination
  - makes CPU available for other tasks
  - dramatically reduces latency
  - maintains current, robust memory protection semantics

RDMA enables:
- Increased efficiency for networking apps
- Increased scaling for distributed database, technical applications
- Increased scaling for distributed and cluster file systems
- New application models:
  - Fine-grained checkpointing
    - Remote application memory as a diskless, persistent backing store
  - Distributed gang scheduling