What is RDMA?

An Introduction to Networking Acceleration Technologies

Fred Worley

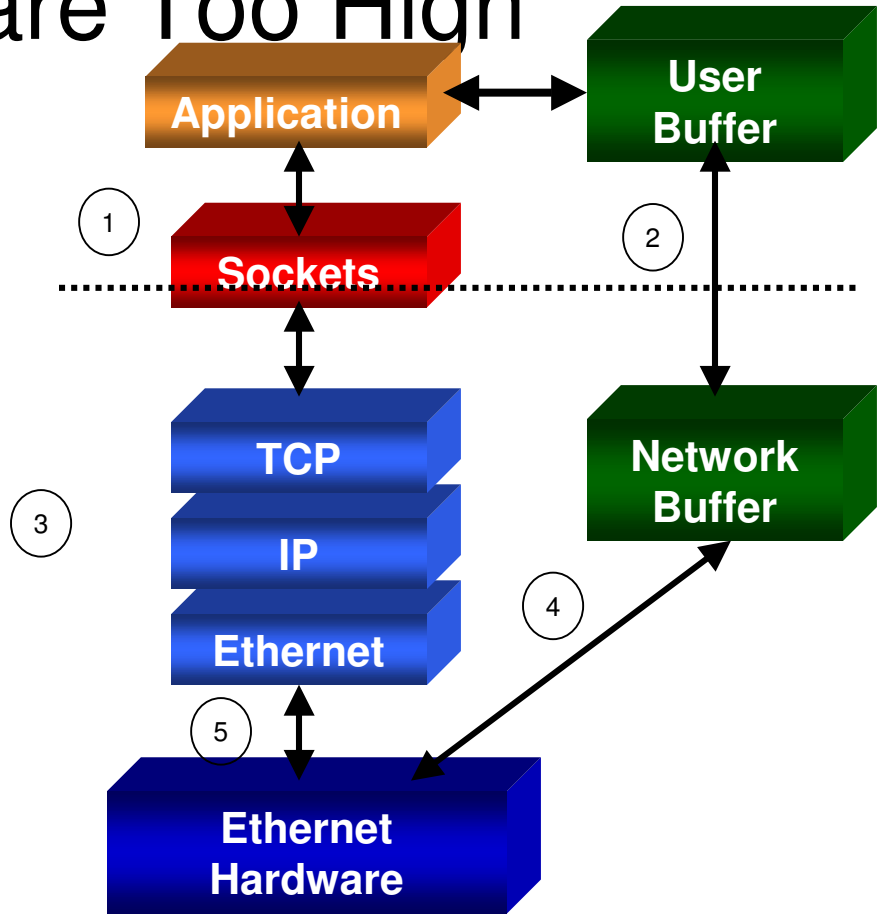
Software Architect
Hewlett-Packard

Outline

- The Networking Tax
- Why we need relief
- Tax Refund Technologies
  - RDMA
  - Local Protocol Acceleration
  - Performance Implications
- Applying your Tax Refund
  - Storage Implications
  - Solution Implications
- Summary

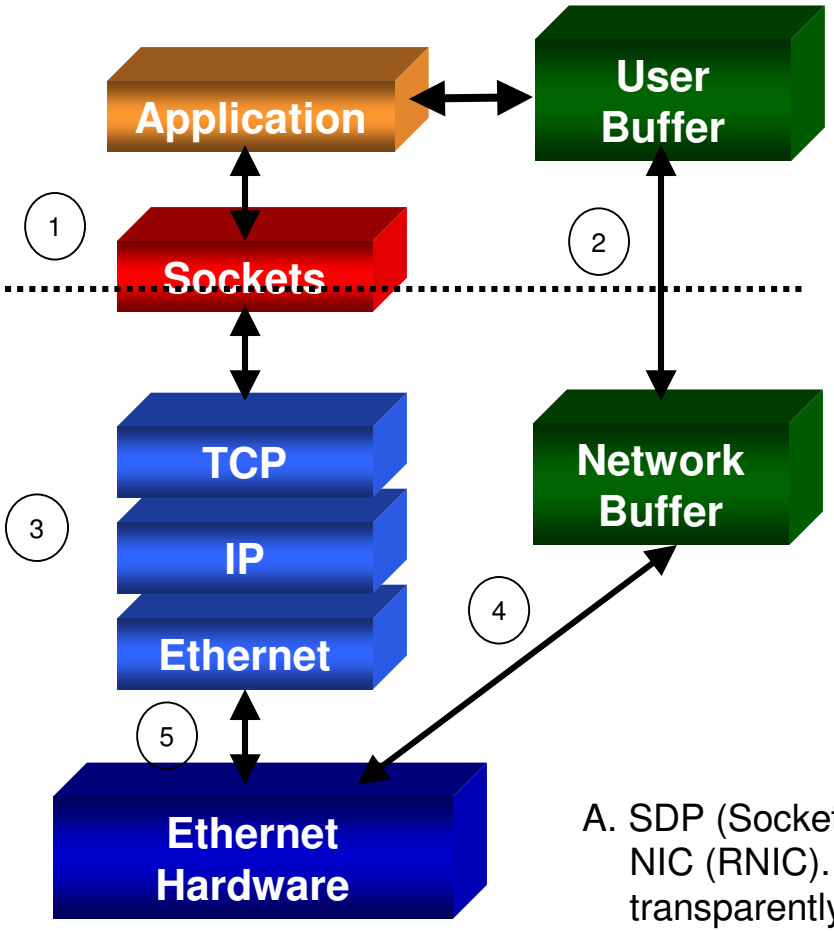
The Problem: Taxes are Too High

- Like a “value-add tax”, OS + network stacks impose taxes (overheads) at each stage of message processing
- As workloads become more distributed, a 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 / infrastructure

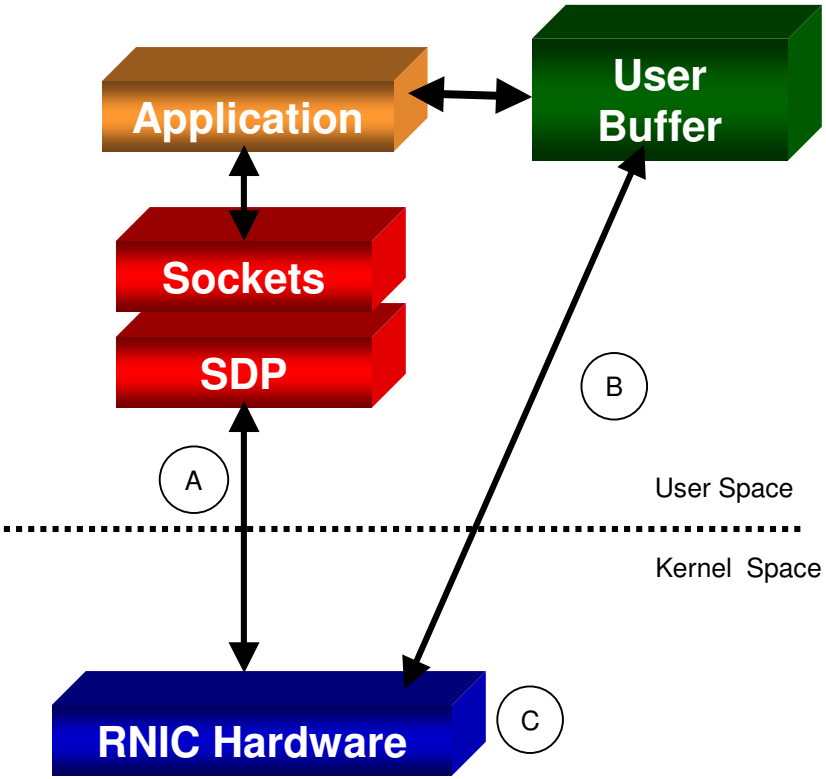


1. User / Kernel Context Switch
2. Copy to / from user buffer and network buffer
3. Packet protocol stack processing – per packet
4. DMA to / from network buffer
5. Device control including interrupt post processing for DMA read / write completions

Existing Architecture



RDMA Architecture



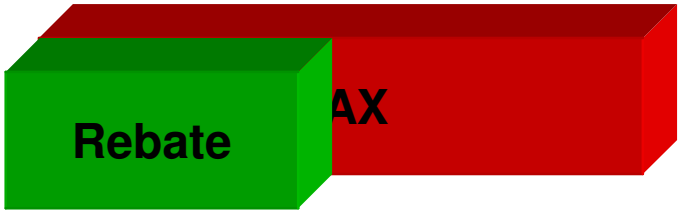
A. SDP (Sockets Direct Protocol) interposed between Sockets and RDMA NIC (RNIC). SDP enables SOCK_STREAM applications to transparently operate over RNIC. SDP interacts with the RNIC directly to process application and SDP “middleware" message exchanges. Enables OS Bypass.

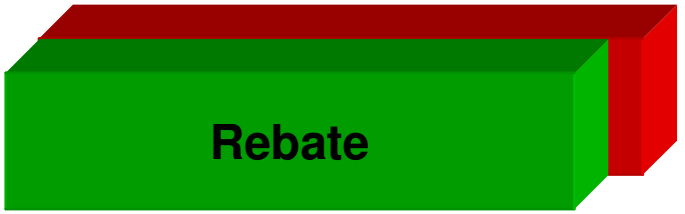
B. Direct DMA to / from user buffer. No interrupts are required as completion processing is performed within SDP layer.

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

Alternatives: Progressive Tax Relief

- Protocol Assist technologies
  - Provides “simple” protocol acceleration with very low cost
  - CKO, TSO, Jumbo Frames, RSS
- Network Stack Acceleration
  - Accelerates higher levels of protocol stack
  - ETA, TOE, iSCSI
- RDMA
  - Acceleration + Zero Copy + OS Bypass
  - Eliminates most host overhead
  - InfiniBand, iWARP, iSER





Why we need relief

Industry Trends drive need for Tax Relief

- Faster link speeds
- Increase in distributed workloads
  - Grid computing
  - Clustered computing
- Decrease in “unused” compute resources
  - Server Consolidation
  - Virtual Partitions (VPARs)
- Effects are felt across a wide class of workloads
  - Database, HPTC, Block and File storage, Backup, Web Server, Video Server, etc.

Critical Technology Rules of Thumb

- CPU performance increases 60% per year (Moore's Law)
- Optical bandwidth performance increases 160% per year
- Large systems double the number of CPUs every 4 years
- Software algorithms improve 5-10% per year
- Memory performance increases 10% per year
- Conclusion:
  - Memory performance remains the critical bottleneck for many years to come
    - New technology / algorithms must evolve to conserve memory bandwidth
    - Applications must be able to adapt to take advantage of these technologies / algorithms

Implications of high taxes

Application	I/O Requirements
DB/Application servers	Up to 40% of System resources to I/O
Client/Server applications	Sensitive to server overhead and latency
Scientific Computing	Large, distributed working sets, latency sensitive
Block Storage over Ethernet	IP data copy limits performance, ASIC solution risky with immature protocols
File Storage over Ethernet	Sensitive to server overhead and latency
Backup over Ethernet	IP data copy limits performance, 24x7 operation requires backup on live system

- Application use of I/O fabrics increasing over time
- Increased communications load limits scaling of single system, multi-computer and tiered solutions
- **More efficient communication** can compensate

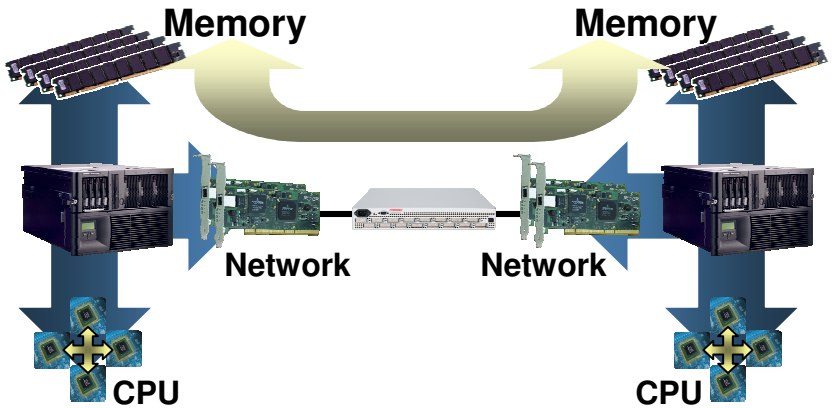
Tax Refund Technologies

RDMA

RDMA – Just Better Networking

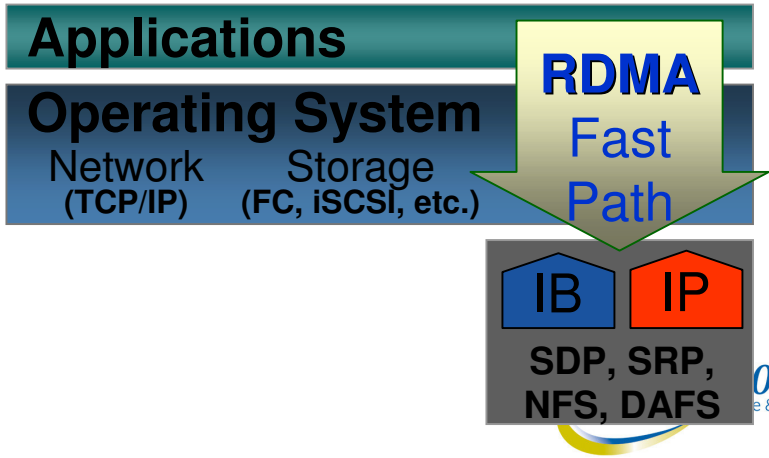
Fast and secure communications

- **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

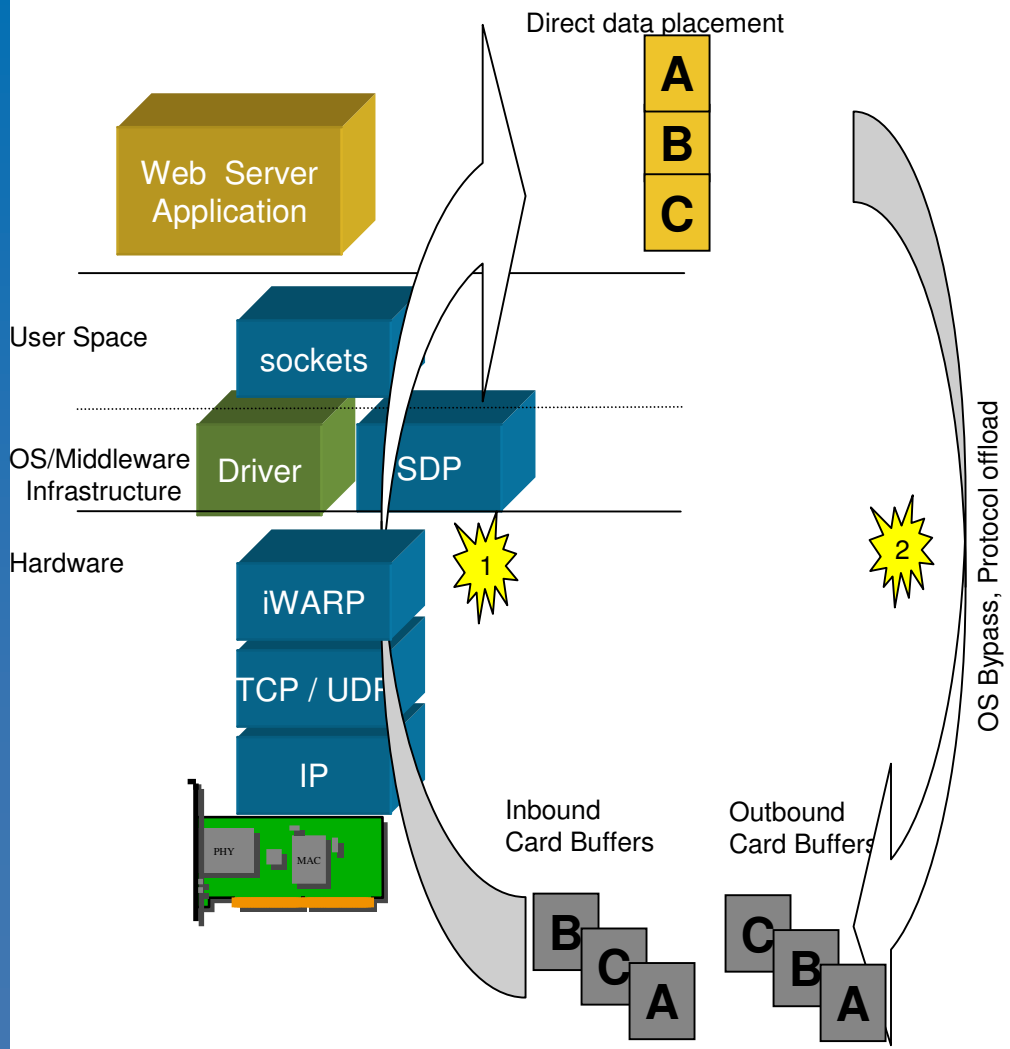


RDMA enables:

- 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



RDMA Example: Dynamic Web Server



Dynamic Content Web Server

- Direct Memory Placement
  - Eliminate processor copy
- Protocol offload
  - Eliminate protocol CPU, Memory overhead
  - TCP, iSCSI, SSL, IPSec, iWARP
- OS Bypass
  - Eliminate context switch, reduce latency
- RDMA
  - Combines the above in an industry standard, link independent manner

RDMA Summary

- RDMA
  - Combines Direct Memory Placement, Protocol Offload and OS Bypass
  - ***Eliminates*** copy overhead, protocol overhead, cache pollution effects, buffer management overhead, reduces NIC cost for high speed links
  - Enabled by sockets-level enhancements (Async sockets), OS enhancements (memory management, iWARP support, SDP, OS bypass library), intelligent interface cards (IB, iWARP), new applications
  - Enables new applications, greater server efficiency, higher performance
- Application level performance benefit (transactions per unit time):
  - Existing applications benefit 10-35%
  - Modified applications benefit 20-100%
  - Benefit is combination of increased bandwidth and reduced CPU util.
  - Benefit dependent on workload

Tax Refund Technologies

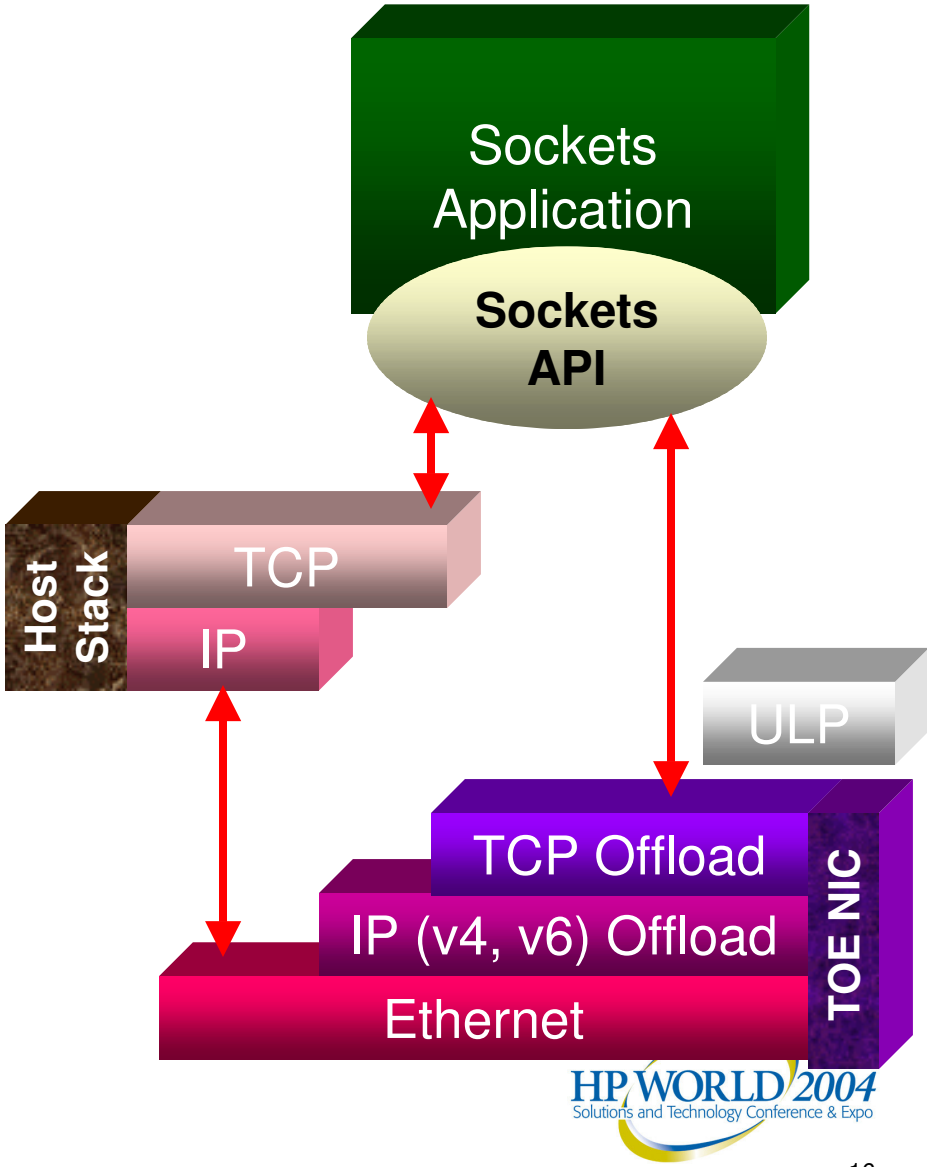
Local Acceleration

Protocol Assist Technologies

- TCP/IP Checksum Offload
- Jumbo Frames
  - Uses larger frames (typically 9K bytes)
  - Reduces TCP/IP segmentation/reassembly and interrupts
- TCP Segmentation Offload ("Large Send Offload")
  - Works for transmits only
  - Similar benefits to jumbo frames
- Receive Interrupt Distribution (a.k.a. "Receive Side Scaling")
  - Distributes networking workload across multiple CPUs
  - Reduces CPU bottleneck, esp. for 1500B frames/10GbE

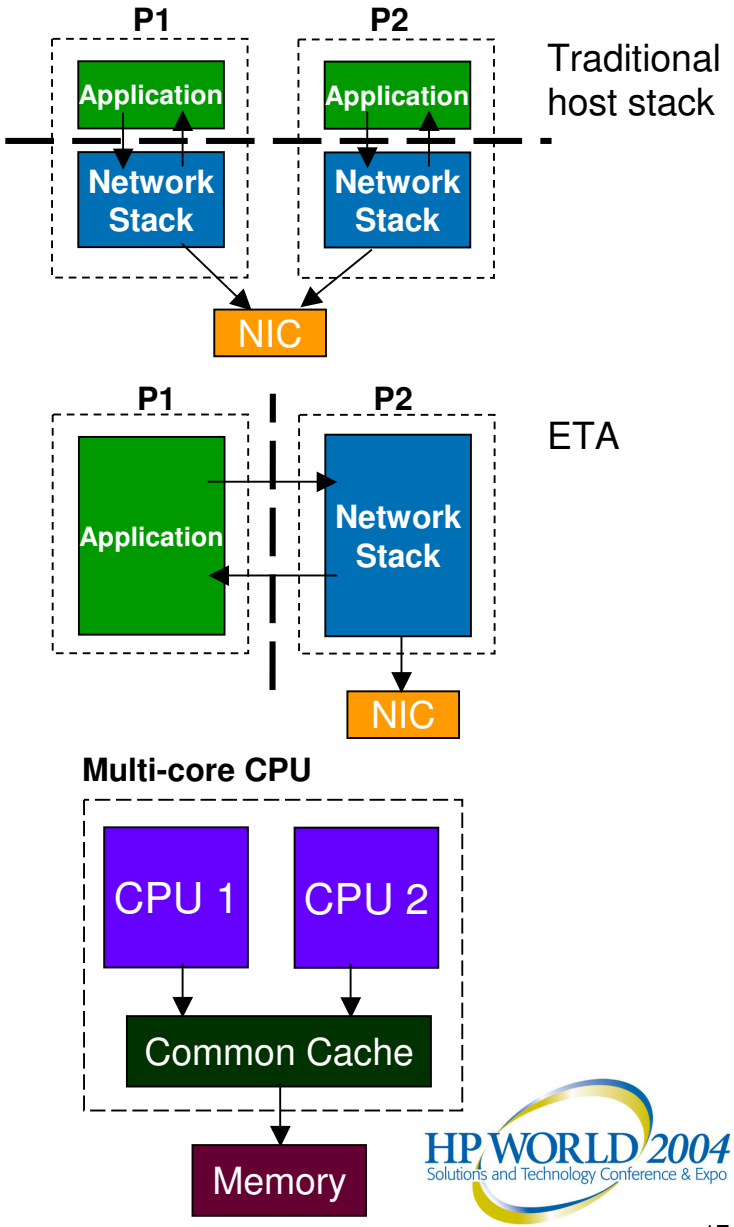
TCP Offload Engine (TOE)

- Implements the TCP/IP protocol stack on the card
  - Implementations will vary for IHV differentiation
    - Main path to entire stack
- Simultaneous support for host-based protocol stack over Ethernet PHY
- Enables other protocol offload:
  - Upper Layer Protocols:
    - RDMA, iSCSI, NFS, CIFS
  - Security
    - IPSec, SSL
  - IP Routing



Embedded Transport Acceleration (ETA)

- Dedicate a processor to network protocol processing
- Advantages:
  - Stack acceleration with a “standard” NIC
  - Makes use of “unused” CPU cycles
    - Industry trend towards multi-core, multi-threaded CPUs
- Issues:
  - Processors & the cache-coherent interface are a precious, expensive commodity.
  - Power, cooling, cache pollution, memory bus utilization
  - Opportunity cost of not running applications on that processor
  - Industry trend towards consolidation, virtual machines
- Research effort by Intel Labs
  - http://ieeexplore.ieee.org/xpl/abs_free.jsp?arNumber=1268989



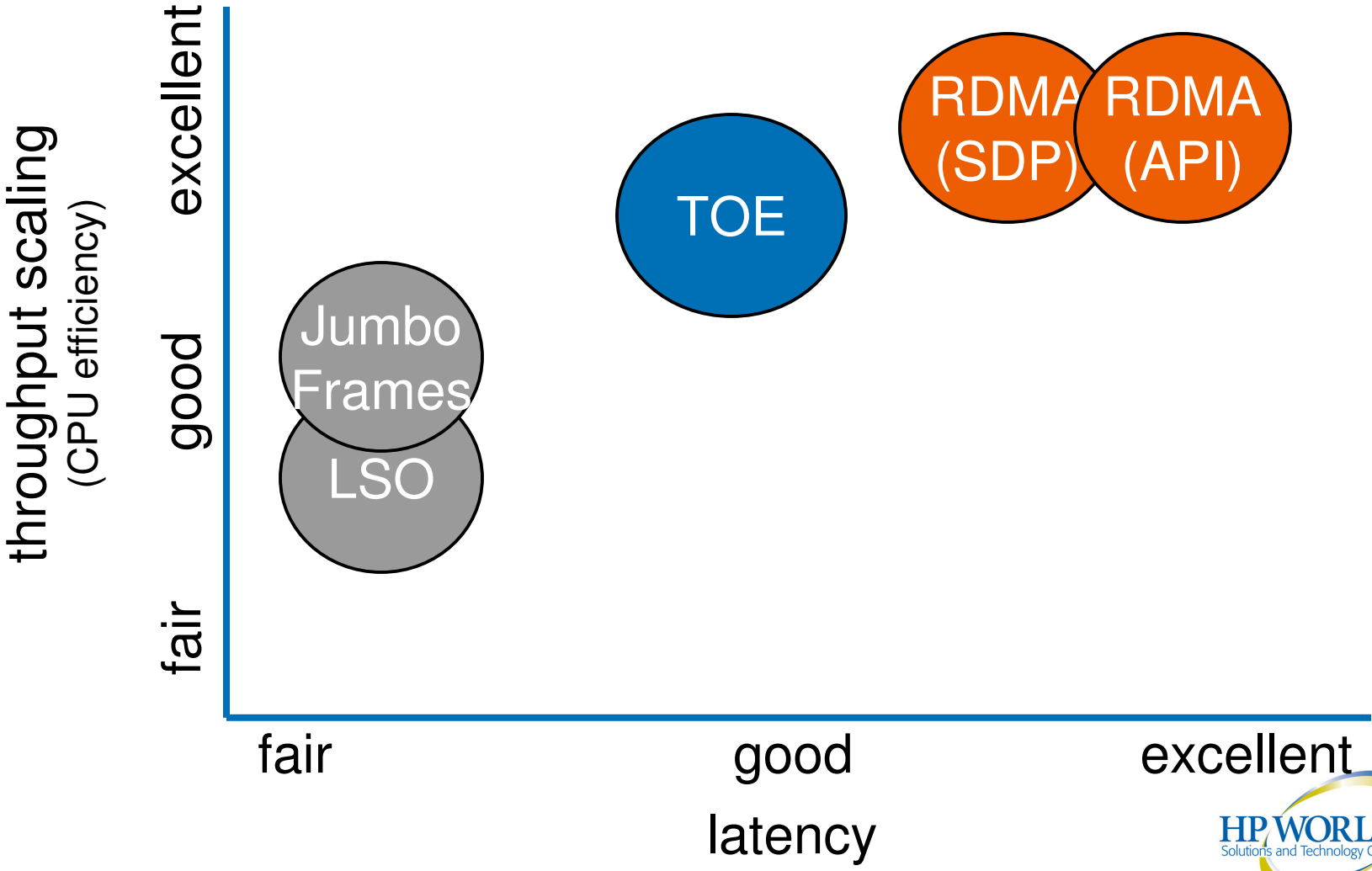
Technology Summary

	RDMA	TOE	ETA	TSO	Jumbo
Reduces overhead from:					
Data copy	Yes	No	No	No	No
Context Switch	Yes	No	Yes	No	No
Network stack processing	Yes	Yes	No	Some	Some
Interrupt processing	Yes	Some	Some	No	Some
Impact to environment:					
Application transparency	w / SDP	Yes	Possible	Yes	Yes
Better perf for modified apps	Yes	No	Yes	No	No
Most benefit to async apps	Yes	No	Yes	No	No
Local host change only	No	Yes	Yes	Yes	No

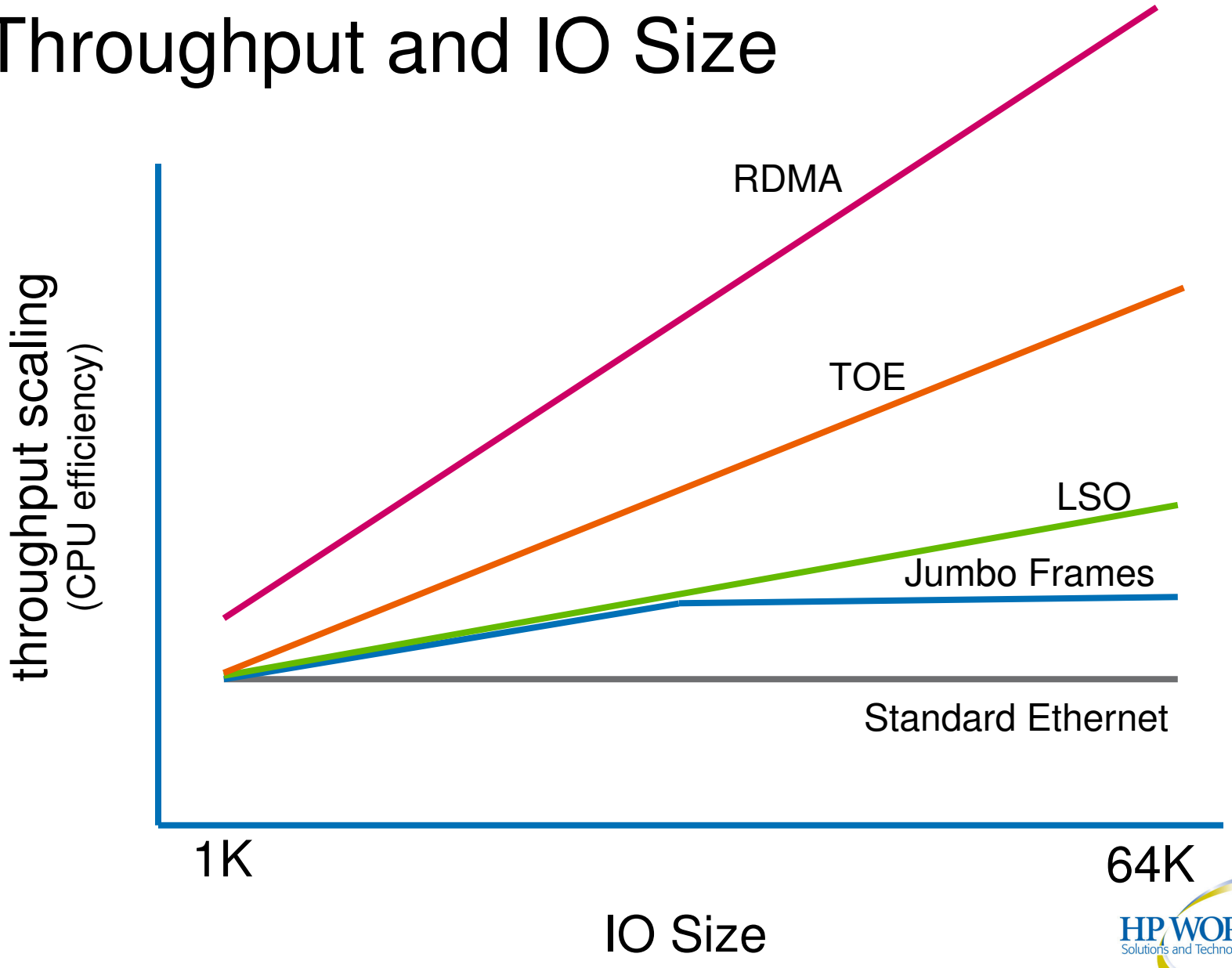
Tax Refund Technologies

Performance

Throughput and latency



Throughput and IO Size



Performance Benefits Summary

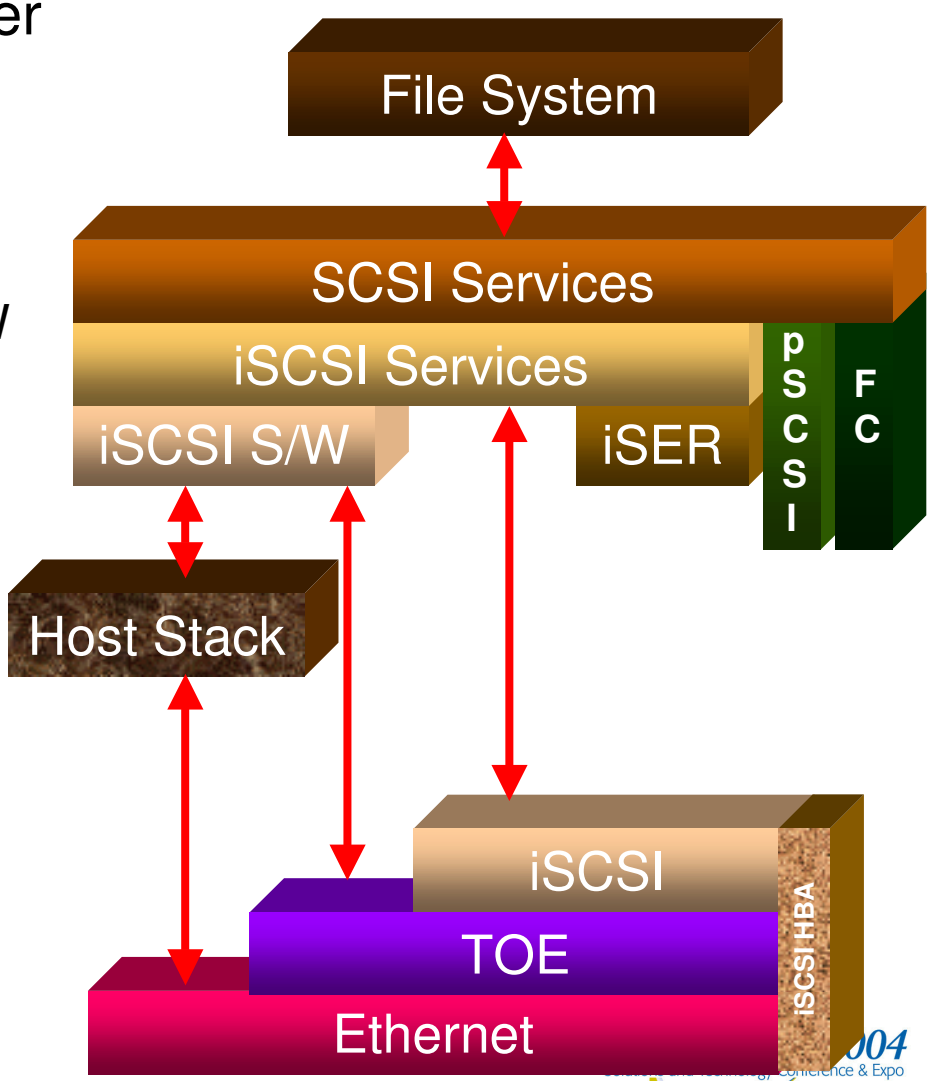
Technique	Benefit	Where?
Jumbo Frames	Reduces CPU utilization (segmentation and reassembly) and interrupts for large transfers.	Requires equipment that supports jumbo frames all through the network.
Large Send Offload	Reduces CPU utilization (segmentation and reassembly) and interrupts for large transmits.	Only helps transmits.
Receive Side Scaling	Distributes connections and receive processing across CPUs. Improves scaling, but not efficiency.	Works well for short-lived connections where other techniques will not work well; helps standard packets on fast links.
TOE host-based connections	Reduces CPU utilization and interrupts for large transfers. Zero copy possible on transmits (receives with pre-posted buffers).	Needs long-lived connections.
TOE card-based connections	Same as above.	Better handles short-lived connections, but possible security issues.
RDMA via Sockets	TOE benefits plus zero copy on both transmit and receive. Reduced latency.	Long-lived connections. Passing information from point to point.
RDMA via APIs	Benefits as above PLUS ability to post once for many reads. Best low-latency messaging.	Long-lived connections Multiple nodes (single posted buffer can be read by many).

Applying your Tax Refund

Storage Implications

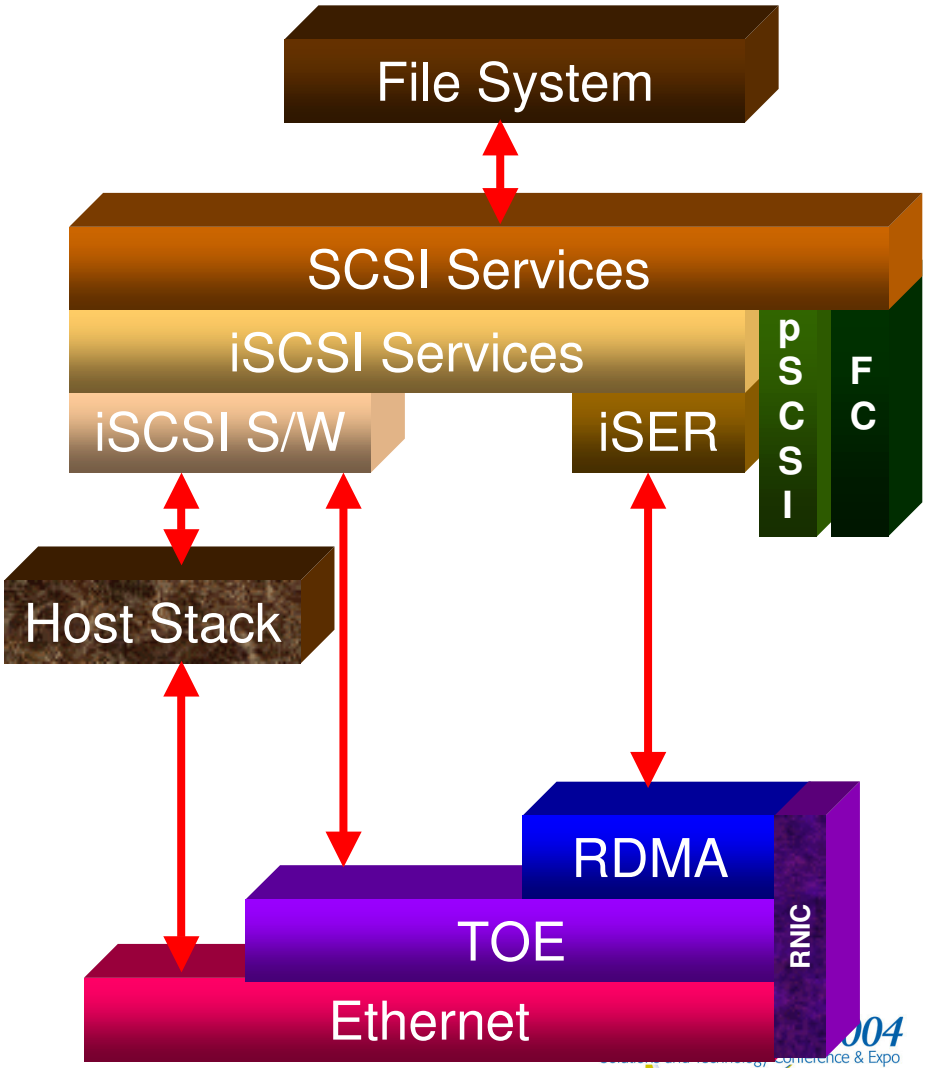
Storage over IP: iSCSI

- iSCSI is block storage access over TCP/IP
  - Provides alternate transport to SCSI over TCP
- Software and hardware solutions
  - TOE vendors are adding special HW to accelerate iSCSI (iSCSI HBA)
  - OS provides iSCSI support in S/W
    - Can use host stack or TOE
  - Products available today:
    - iSCSI HBAs from multiple vendors
    - iSCSI software OS support on HP-UX, MS, others
      - HP-UX iSCSI software on 1GbE: 110 MB/s using 50% of a CPU
    - iSCSI to FC bridge from HP, others

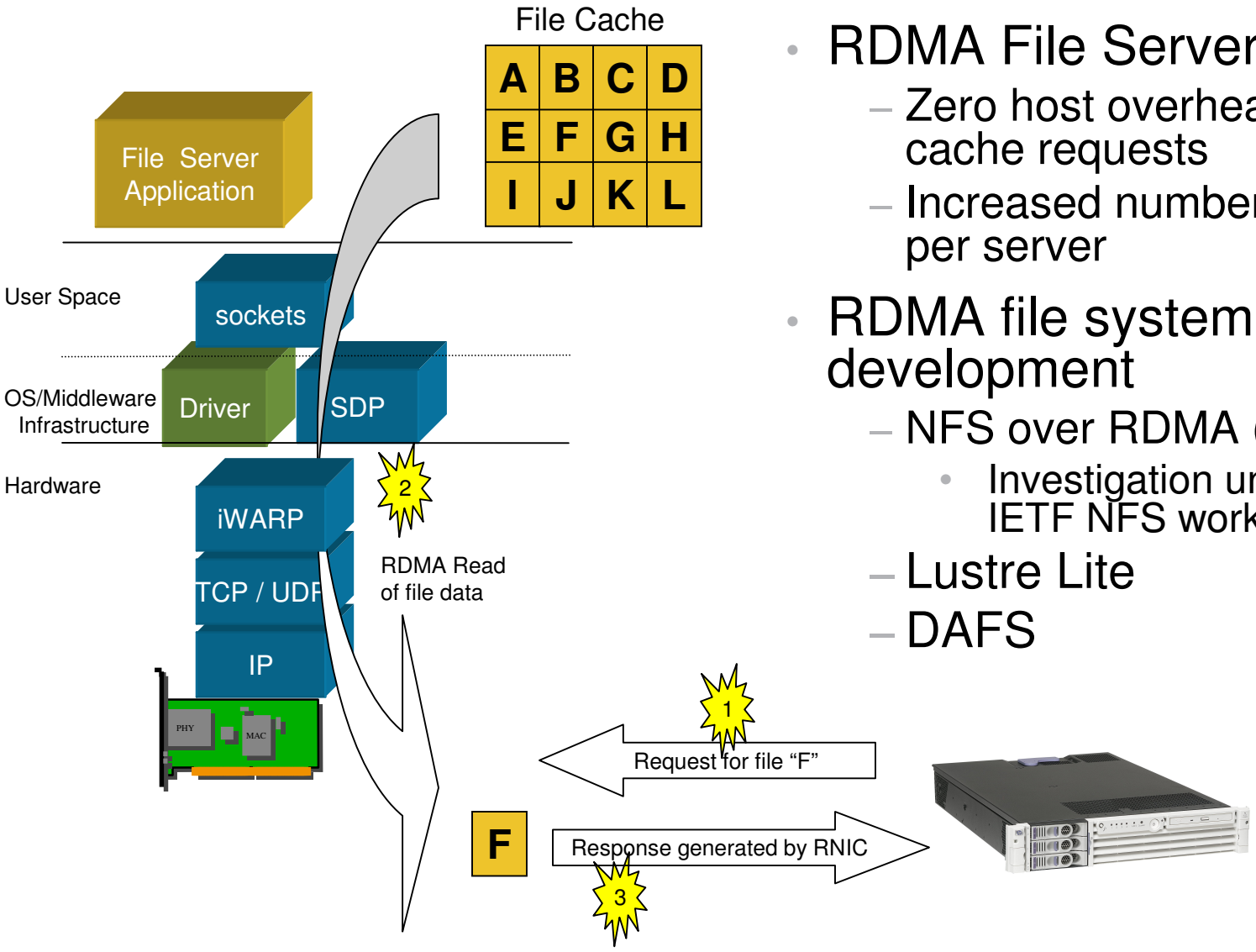


Storage over IP: iSER

- iSER: iSCSI Extensions for RDMA
  - An adaptation layer for iSCSI so it can us the accelerations provided by the RNIC
  - Doesn't reinvent the wheel, leverages all the iSCSI work
  - Separates storage management (host software) from data movement (RNIC RDMA interface)
- Allows the general purpose RNIC to support high performance storage
  - Standard iSCSI (I.e. w/o iSER) requires an iSCSI HBA for optimal performance



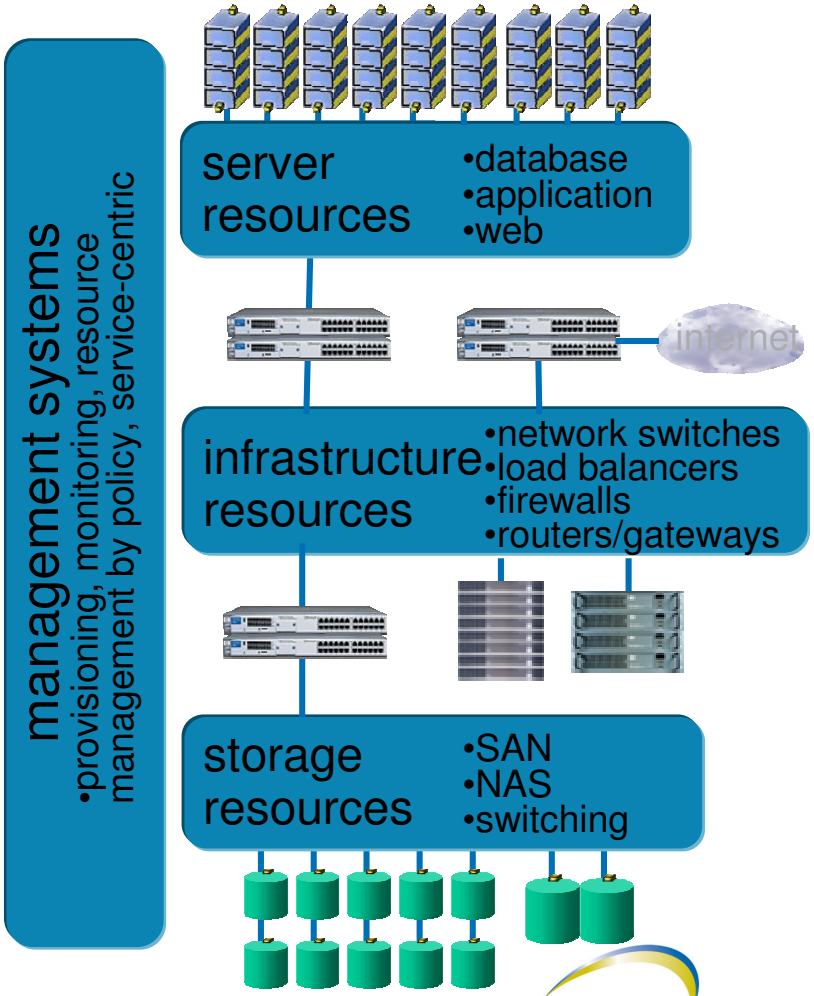
RDMA NAS



- RDMA File Server enables
  - Zero host overhead for in-cache requests
  - Increased number of clients per server
- RDMA file systems under development
  - NFS over RDMA (NFS v4)
    - Investigation under way in IETF NFS work group
  - Lustre Lite
  - DAFS

Fabric Consolidation

- Maximize customer value with low TCO
- Simplify Fabric Structure
  - Reduced infrastructure complexity with improved availability
    - Fewer parts to purchase / manage / points of failure
    - Consolidated workload management and policy driven controls
    - Flat switched fabric topologies for improved QoS
  - Reduced infrastructure cost structure
    - Fewer administrators
    - Reduced recurring costs leverage common knowledge base
    - Flatten switch structure
    - Commodity pricing for components
    - Simplified stocking of replacement components
- Simplify Fabric Structure, cont
  - Simplified server configuration
    - Adaptability without physical reconfiguration
  - Instant Access to Any Content, Any Time
- Enabled by new technologies
  - Remote Direct Memory Access (RDMA)
    - Enables Ethernet as SAN, CI technology
  - 10 Gigabit Ethernet
    - Bandwidth for consolidation, flattens switch structure
  - Flexibility of Ethernet
    - Continual industry development, interoperability, Security, QoS, Management Suite, etc



Applying your Tax Refund

Cutting through the Red Tape

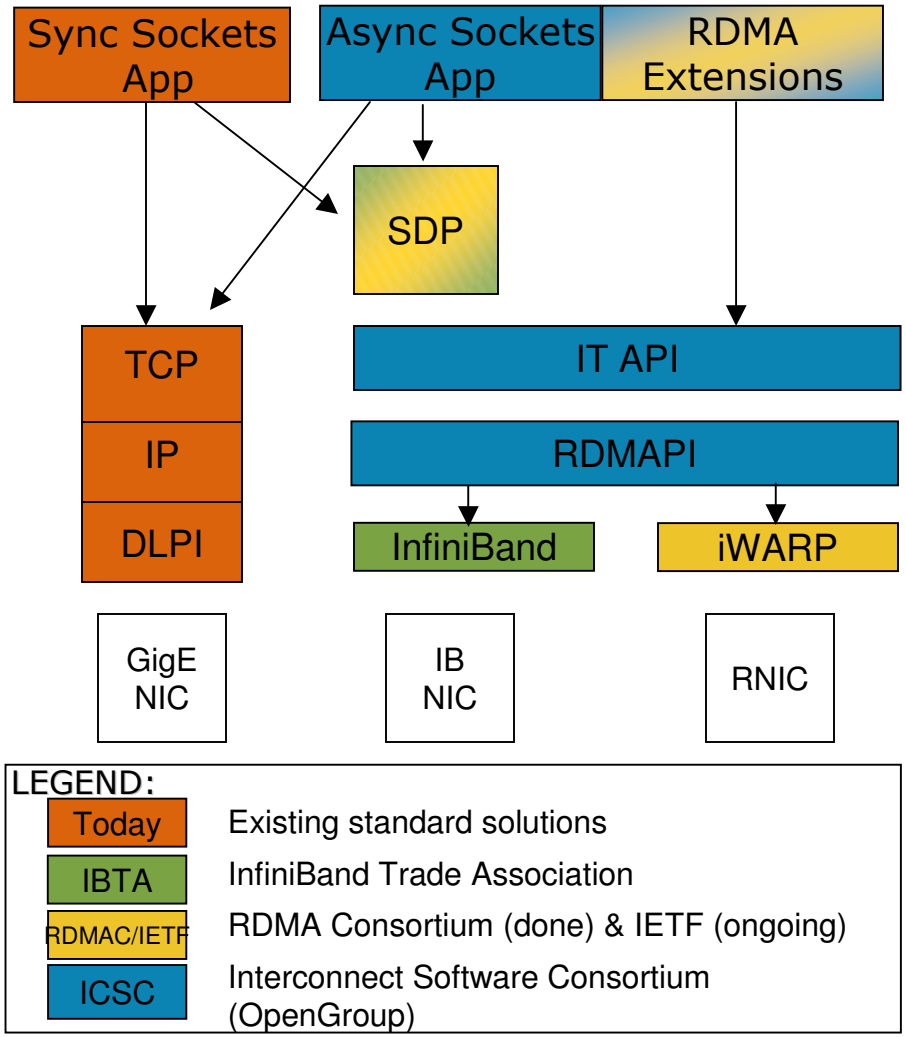
Cutting the check requires: Delivering a full solution

- Infrastructure
  - Standards
  - Industry deployment
- Application behavior
  - Application transparency
  - Asynchronous benefits
  - The benefits of change…
- Solution differentiation
  - QOS
  - HA / Failover
  - Cluster management
  - Continued evolution

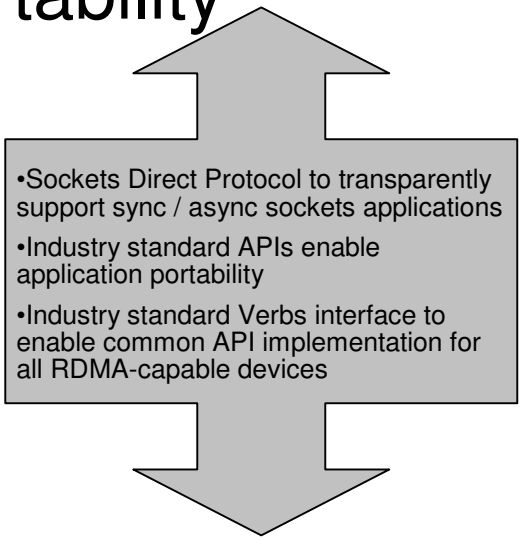
RDMA Software Infrastructure

- SDP: Sockets Direct Protocol
  - Provides transparent support of existing Sockets applications over RDMA
    - Accelerates Sockets; fewer touches of data, particularly for large transfers
    - Gets most of the benefit of RDMA protocols without any change to the app
- IT API: Interconnect Transport API
  - Industry Standard transport APIs for RDMA-capable fabrics
  - InfiniBand, VI, iWARP (IT API v1.2)
- Async Sockets Extensions for RDMA
  - Industry Standard Sockets API for RDMA-capable fabrics

RDMA Software Infrastructure, cont.

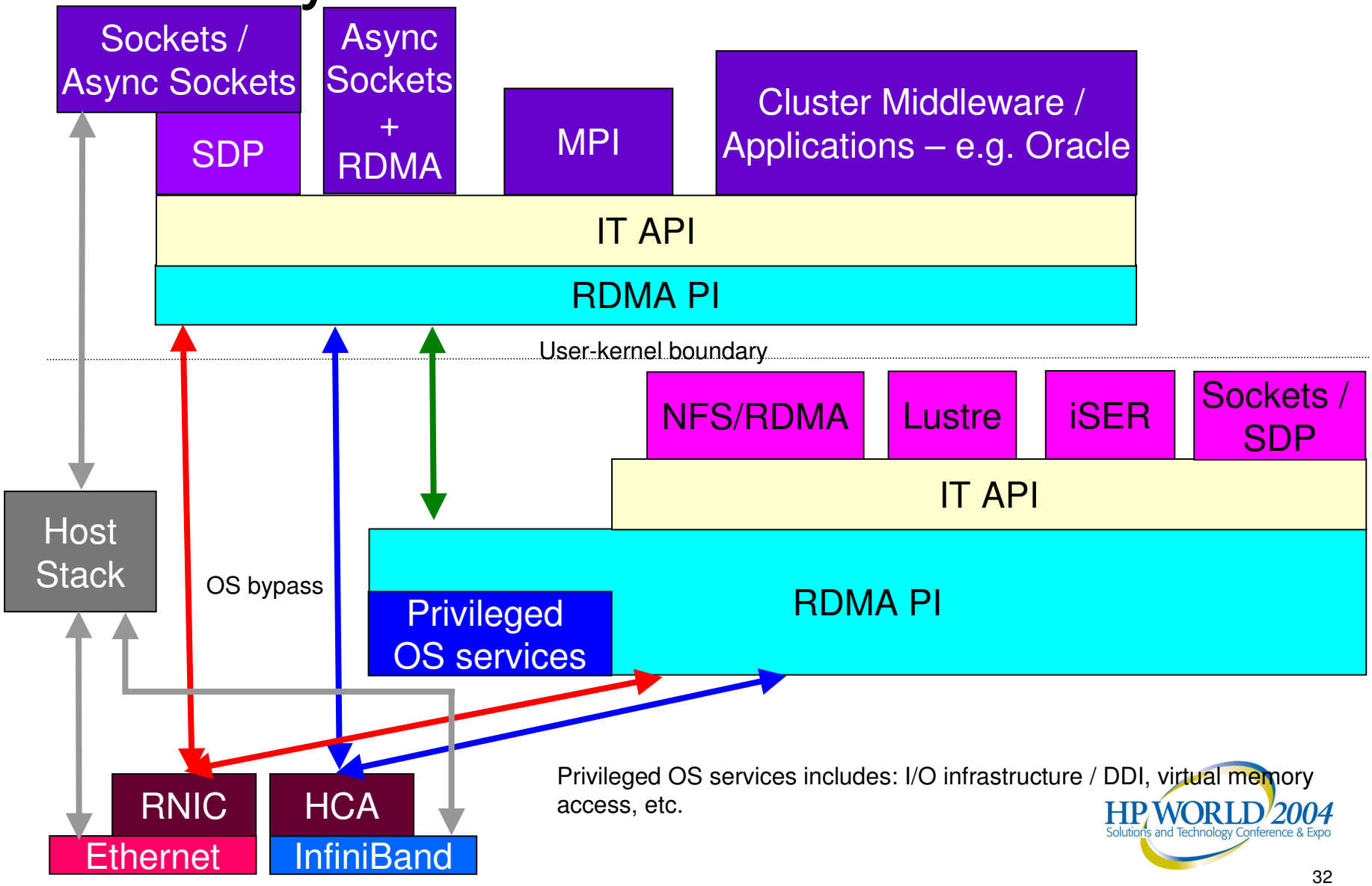


- Maximize benefit to existing applications
- Maximize application portability



- Maximize hardware technology independence

RDMA Software Infrastructure Summary



Privileged OS services includes: I/O infrastructure / DDI, virtual memory access, etc.

Industry Momentum for RDMA

- **Establishing industry-wide single interconnect independent RDMA paradigm**
  - **InfiniBand Trade Association**
    - Developed significant momentum for industry standard RDMA
    - Founding Members: **HP**, IBM, Intel, Microsoft, Sun Microsystems, Dell (joined 2002)
  - **Interconnect Software Consortium (OpenGroup)**
    - Creating Unix APIs for RDMA (application and HDM), Async Sockets
    - Founding Members: Fujitsu, **HP**, IBM, Network Appliance, Sun, Vieo [Principal member: Intel]
  - **RDMA Consortium**
    - Defined standard for RDMA over TCP/IP transports (includes framing, DDP, RDMA, verbs definitions)
    - Founding Members: Adaptec, Broadcom, **HP**, IBM, Intel, Microsoft, Network Appliance
      - Now includes EMC, Dell
      - RDMAC covers 90+% of Server, Storage, OS and NIC/HBA vendors
  - **IETF RDMA over IP workgroup**
    - Creating DDP, RDMA as transport independent solution
  - **Storage over IP**
    - iSCSI 1.0 specification complete, wide support in industry
    - RDMAC specification for iSCSI over RDMA (iSER) complete
    - IETF working to adopt iSER definition

RDMA 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

RDMA 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
    - Side memory to provide high-speed device local cache

RDMA 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 (OpenGroup) , 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

iWARP or InfiniBand?

Fabric	Strengths	Weaknesses	Outlook
iWARP (RDMA over Ethernet)	Ubiquitous; Standard Affordable adapters and switches. Minimal fabric training costs. Extends beyond the datacenter. Common infrastructure with IB.	Switch Latency. 10GbE cost.	Enables fabric consolidation for the data center. Potential for volume ramp in 2006. It's Ethernet!
InfiniBand	Available now. Lowest latency. Most affordable 10Gb link (today). Mature, standard protocol. Common infrastructure with iWARP.	Unique fabric and fabric management. Bridges needed to go to FibreChannel and Ethernet fabrics.	Best solution for clustered applications where optimal performance is critical. Provides mature RDMA solution through iWARP ramp.

InfiniBand on HP-UX

- 7us Latency, 767MB/s at 6% CPU utilization (32k msgs)
- HPTC release for HPUX 11i V2 -- Q2/2004
  - OpenGroup ICSC IT API version 1.0
  - Full O/S bypass and RDMA support
  - InfiniBand RC and UD transports
  - IP support on InfiniBand (IPoIB)
  - PCI OLAR support (H204)
  - Bundled with HPTC 4X InfiniBand HCA
  - **HP MPI 2.0.1**
  - Supported on all HP Integrity servers
- HA & DB release for HPUX 11i V2 – H2/2004
  - Added features:
    - **+** HA capabilities
      - Link and session failover
      - HCA virtualization
      - Auto path migration (APM)
      - MC/SG support
      - IPoIB virtualization
    - **+** Oracle 10gRAC
    - **+** Bundled with HA & DB 4x InfiniBand HCA
  - Supported on all HP Integrity servers

Industry-Wide Product Availability

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

10GbE switch infrastructure	2003
10GbE NICs	2003
iSCSI to FC bridging	2004
iSCSI HBAs	2003
iSCSI HBAs with integrated IPSec	2003
iSCSI storage targets	2004
iSER storage targets	2005
InfiniBand HCAs, switches	2003
RDMA-based NAS	2003 (IB), 2004-2005 (iWARP)
RNICs (1GbE, 10GbE)	2004-2005
Low-latency Ethernet switches	2004-2005
IT API-based middleware	2004
RDMA-enabled Async Sockets applications	2005-2006

Does not indicate specific product plans from HP

Summary

New I/O and IPC Technology : invent

- 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

Summary

- Networking Taxes are too high
  - System overheads decrease server efficiency, add cost to the data center
- Industry trends demand tax relief
  - Faster link speeds
  - Reduction in “unused” CPU cycles
  - Increase in distributed / Grid workloads
  - Improved memory latencies lag improved CPU and wire speeds
- Increased efficiency provides relief
  - Protocol acceleration and assist technologies increase efficiency with minimal impact to infrastructure
  - RDMA technologies *eliminate* copy, protocol, OS overheads
  - Industry standards for protocols AND infrastructure enables broad deployment, maximum ROI for customers and hardware, software and OS vendors

Summary , cont.

- "Tax Refund" technologies enable new solutions, simplified data center infrastructure with lower TCO
  - New networking paradigms are being created to allow server scaling to keep pace with advances in network and processor technology
  - Increased server efficiency through protocol acceleration
  - Ethernet with RDMA efficiently supports LAN, SAN, CI workloads
  - RDMA Software Infrastructure enables application portability across platforms, hardware technology independence for software
    - Enables clean migration between CI technologies
    - Preserves investment in software
  - Value add features create complete solution
    - Integrated tools for cluster management, transparent failover, QoS, dynamic configuration combine higher performance with ease of use
- What to expect from HP
  - Technological leadership
  - Enterprise-class RDMA solutions
    - Hardware, software, management, tools
  - Total solutions for the data center built on a consistent infrastructure
    - Integrated product families for support of existing and future technologies
    - Transparent integration of new technologies into Adaptive Enterprise environment


hp
invent

HP WORLD 2004

Solutions and Technology Conference & Expo

Co-produced by:

interex

shared knowledge • shared power

encompass

AN HP USER GROUP

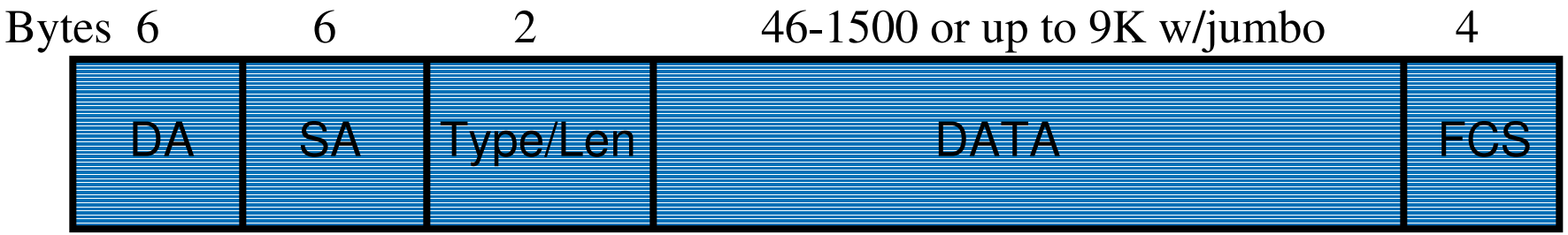
RECOMMENDED TRAINING VENUE FOR THE

HP Certified Professional

Additional Material

Jumbo Frames

- Jumbo Frames increases Ethernet MTU (1500B to 9000B)



Pros

- Reduces CPU overhead
  - Process 1/6th as many packets
- Increases NIC throughput
  - Substantial performance increase when CPU limited, e.g. 10GbE

Cons

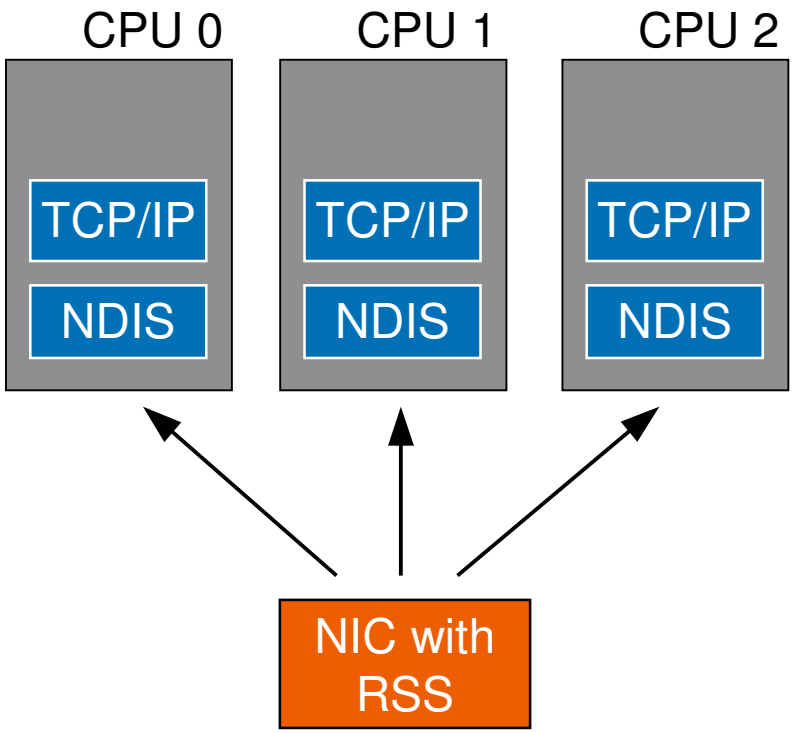
- All devices in network need to support Jumbo Frames
- NOT an IEEE standard

TCP Segmentation Offload (a.k.a. “Large Send”)

- Provides performance improvement for Send path (no change to Receive path)
- Supports segmentation of the TCP byte stream into multiple IP packets in hardware
- Reduces the amount of CPU processing per outbound message by offloading some of the work to the NIC
- Reduces I/O subsystem overhead by transferring data to the card in larger chunks
- Now commonly supported on GigE NICs
- Feature of NIC hardware and operating system
- Independent of the IP MTU size

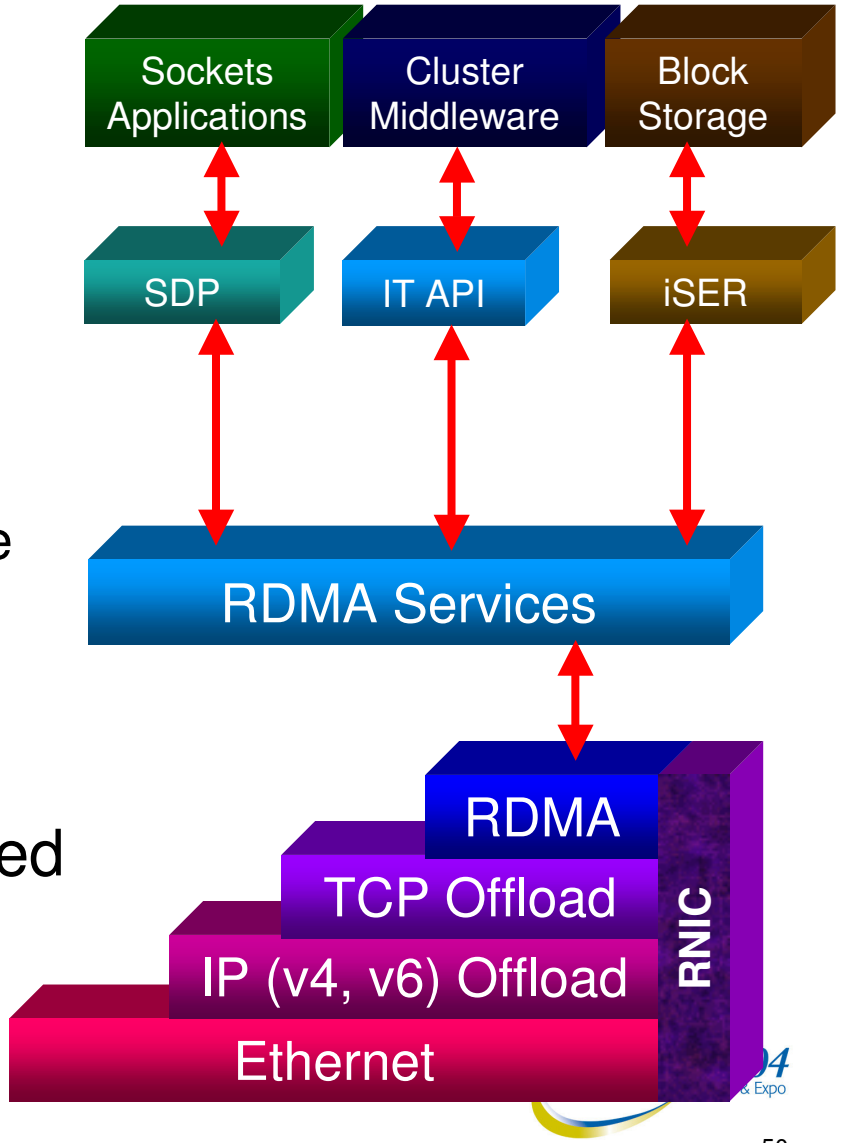
Receive Side Scaling (RSS)

- Spreads incoming connections across the CPUs within a server.
- Overcomes the single CPU bottleneck.
- Works well in applications with lots of short-lived connections (where TOE doesn't work well).
- Supported on Windows 2003 with Scalable Networking Pack (Beta in 2H2004).
- Similar technology supported in HP-UX 11i and Linux 2.6



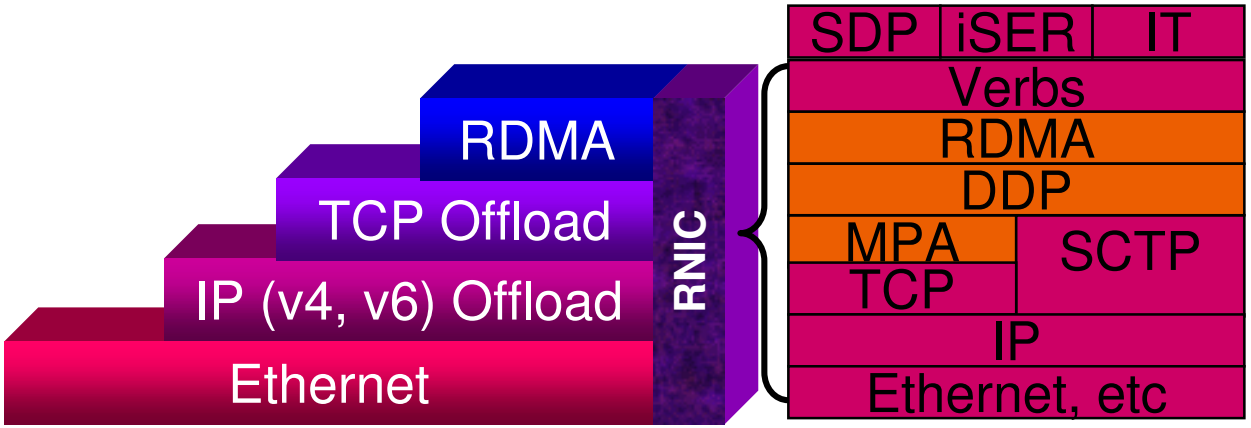
RNICs

- RDMA Network Interface Card
- Supports Upper Layer Protocols:
  - *Networking:* SDP: Sockets Direct Protocol
  - *Storage:* iSER: iSCSI Extensions for RDMA
- Supports standard APIs
  - *Clustering:* IT API: Interconnect Transport API (Interconnect Software Consortium's Unix API)
  - *Networking:* Sockets RDMA Extensions
- Simultaneous support for RDMA offload, TCP offload and host-based protocol stack

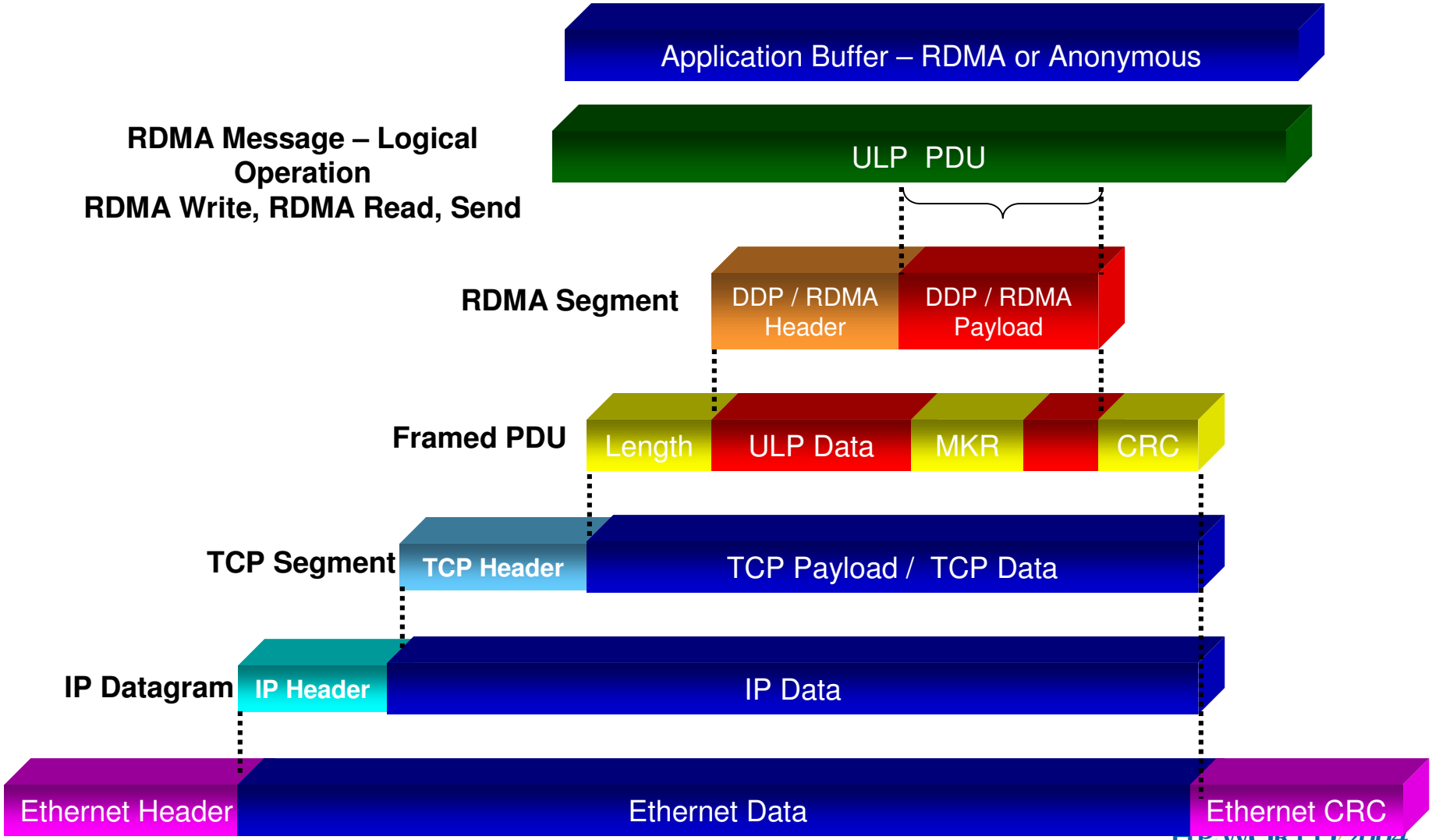


RNICs, cont.

- Implements “iWARP” (RDMA over TCP) and TCP/IP protocol stacks on the card
  - RDMA wire protocol (iWARP)
  - DDP: Direct Data Placement
  - MPA: Marker based PDU Alignment
- Provides RDMA “Verbs” interface
  - Verbs provides standard functional semantics to hardware

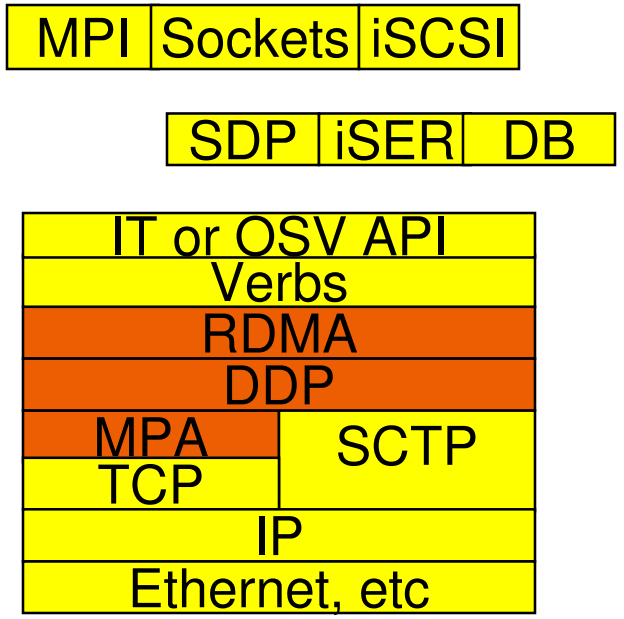


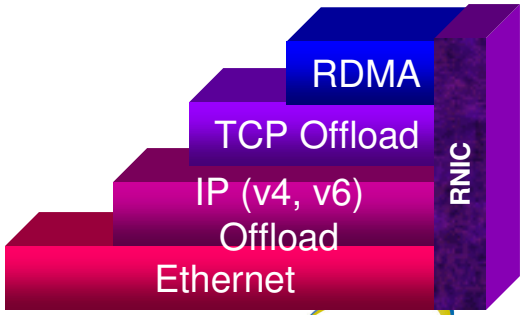
RDMA Headers



RDMA/TCP Benefits Summary

- Ecosystem under rapid development
  - RDMA Consortium (RDMAC) specifications completed
    - Wire protocols, Verbs, iSCSI Extensions (iSER), 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 RDMA / TCP
    - 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.

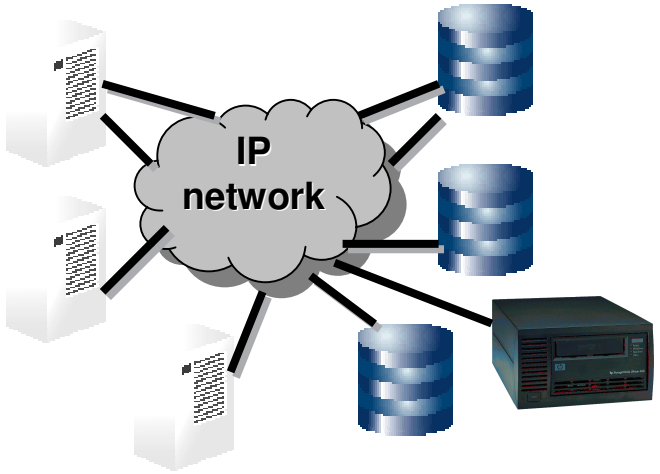




Emerging RDMA Standards

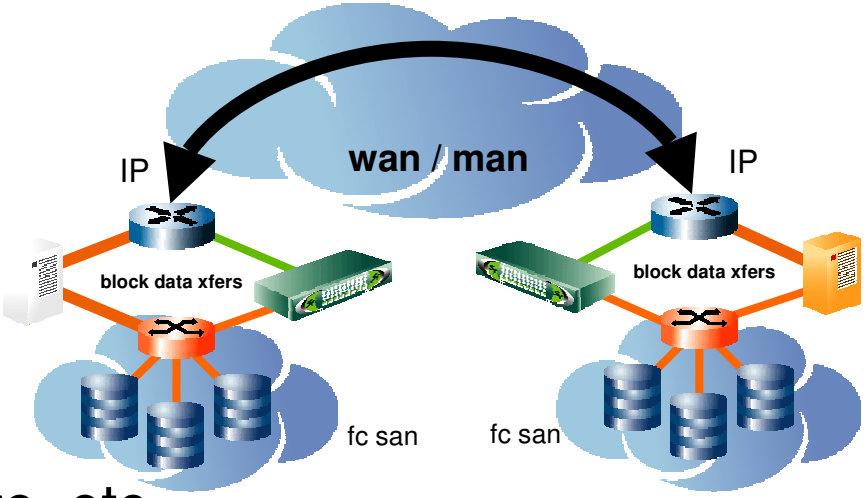
- Roots in Virtual Interface (VI) and InfiniBand
  - HPL Hamlyn and other research led to the creation of VI
  - VI defined the basic RDMA / protocol off-load model
  - InfiniBand completed user-mode Verbs model
- RDMA Consortium formed
  - HP, Microsoft, Adaptec, Broadcom, Cisco, Dell, EMC, IBM, Intel, NetApp
  - Developed v1.0 protocols for MPA/DDP/RDMA
  - Developed iSER/DA Upper Layer Protocols
  - Evolved IB Verbs to include Kernel / Storage
  - Developed SDP Upper Layer Protocol
  - http://www.rdmaconsortium.org
- Interconnect Software Consortium (OpenGroup ICSC)
  - Developed RDMA API (IT API) – V1.0 completed in August, 2003
    - Being enhanced now to include RDMA/TCP and InfiniBand verbs extensions
  - Developing RNIC Provider Interface to simplify OS RNIC integration
  - http://www.opengroup.org/icsc/
- RDMA Consortium submitted specs over to IETF as drafts
  - Moving to last call now
  - http://www.ietf.org/html.charters/rddp-charter.html

Storage over IP deployment



- Common fabric for both block and file storage access
  - Lower cost structure:
    - Host connectivity (single interconnect type), deployment, operations, commodity parts, etc.
  - Leverage IP / Ethernet infrastructure and expertise
  - Integrates into HP Utility Data Center unified fabric utility
  - Integrates into existing storage and network management infrastructures
    - Easily extensible while maintaining interoperability and storage / network management infrastructures

Storage over IP deployment, cont.



- Example solutions:
  - Native IP Storage
    - iSCSI / iSER block storage devices, NAS
  - Distributed file systems
    - CIFS, CFS, NFS, Lustre Lite, etc.
    - Movement to RDMA based solutions:
      - NFS / RDMA, Lustre Lite, etc.
  - IP Infrastructure to bridge FC networks
    - Broader access to isolated data islands
    - Builds upon existing FC deployments
    - Disaster recovery and replication solutions

Cluster technology deployment

- Immediate solution
  - Maintain separate fabric for Cluster Interconnect
  - Deploy InfiniBand (available today)
    - 10Gb link rates
    - Middleware available (Oracle, MPI, OpenIB)
  - Take advantage of common, industry standard infrastructure for RDMA technologies
    - Future-proof software investment
    - Solutions deployed today on InfiniBand will migrate to iWARP
- Choose RDMA software/middleware using Industry standard APIs
- Near-future technologies offer:
  - RDMA over Ethernet (iWARP)
  - Low-latency Ethernet switches
  - QOS support to share large capacity pipes between multiple traffic types (LAN, SAN, CI)

Availability estimates for the Industry as a whole, product offerings from multiple vendors
Does not indicate specific product plans from HP


invent