Enabling the Adaptive Enterprise: I/O Fabric Technology Evolution

Fred Worley

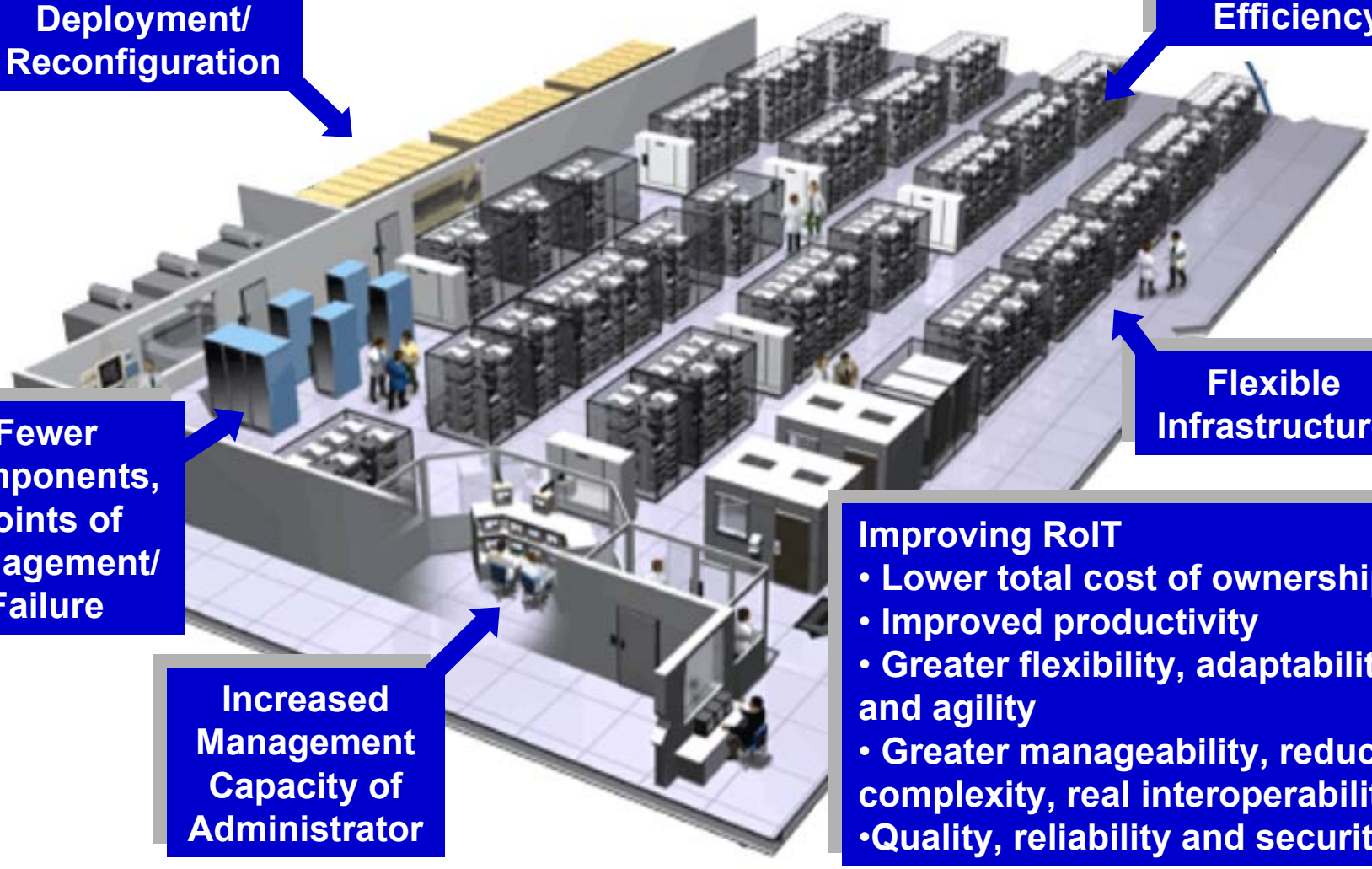
Software Architect
Hewlett Packard
System Interconnect Solutions Lab

HP WORLD 2003
Solutions and Technology Conference & Expo

Agenda

- Adaptive Enterprise Overview
- The Data Center Toolbox
- Data Center opportunities
- New Tools for the Data Center
- Deployment and Migration
- Summary

Adaptive Enterprise



Rapid Service Deployment/ Reconfiguration

Increased Server Efficiency

Flexible Infrastructure

Fewer Components, Points of Management/ Failure

Increased Management Capacity of Administrator

Improving RoIT

- Lower total cost of ownership
- Improved productivity
- Greater flexibility, adaptability and agility
- Greater manageability, reduced complexity, real interoperability
- Quality, reliability and security

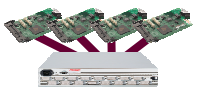
Data Center Infrastructure Evolution

Expanding the Data Center Toolbox

HP WORLD 2003
Solutions and Technology Conference & Expo

storage

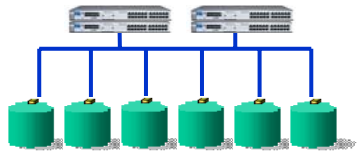
networking



clustering

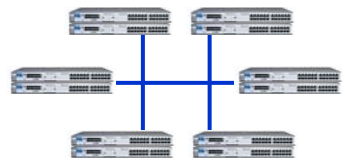
Today	Tomorrow
• Fibre Channel	• 4 Gigabit Fibre Channel • 10 Gigabit Fibre Channel
• NAS (Storage over IP)	• iSCSI (Storage over IP)
• 1 Gigabit Ethernet	• 10 Gigabit Ethernet
• KVM over IP (Lights-out Management)	• IP acceleration (TCP/IP & IP Sec)
• Proprietary Solutions (ServerNet, Myrinet, etc.)	• IP Fabrics (RDMA/TCP) • InfiniBand Fabrics

storage fabric



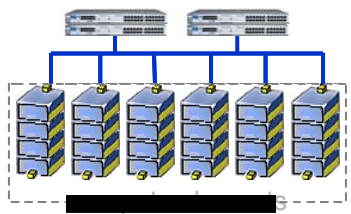
storage elements

data center fabric



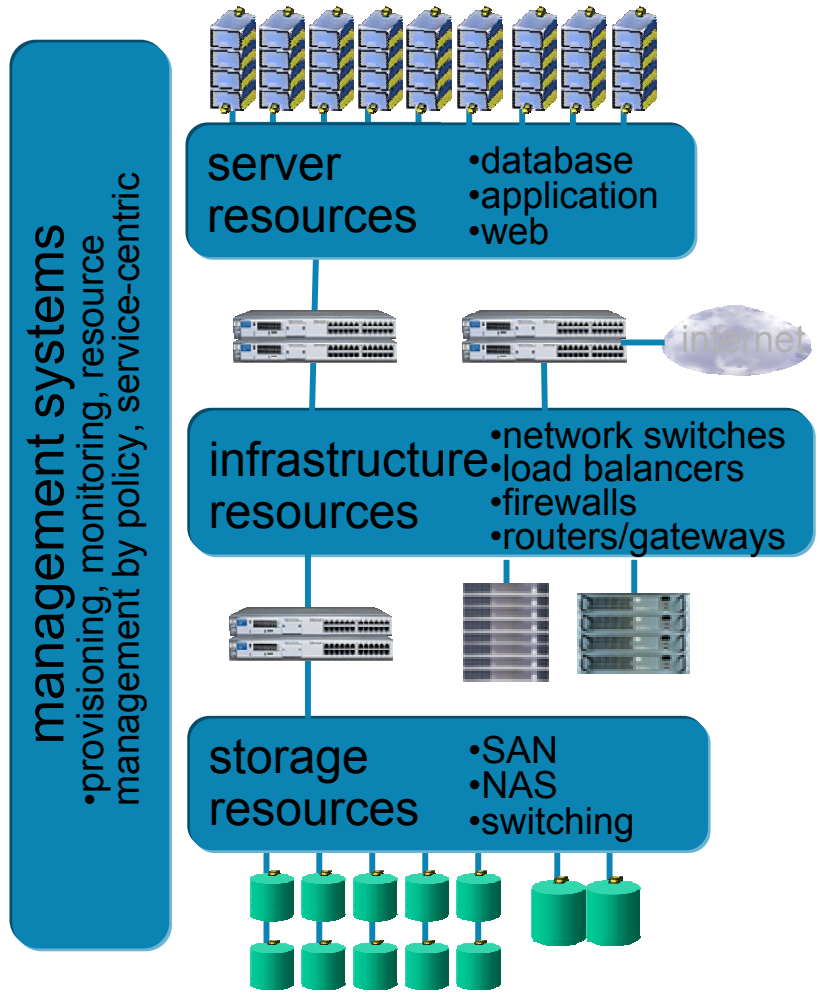
fabric switches

compute fabric



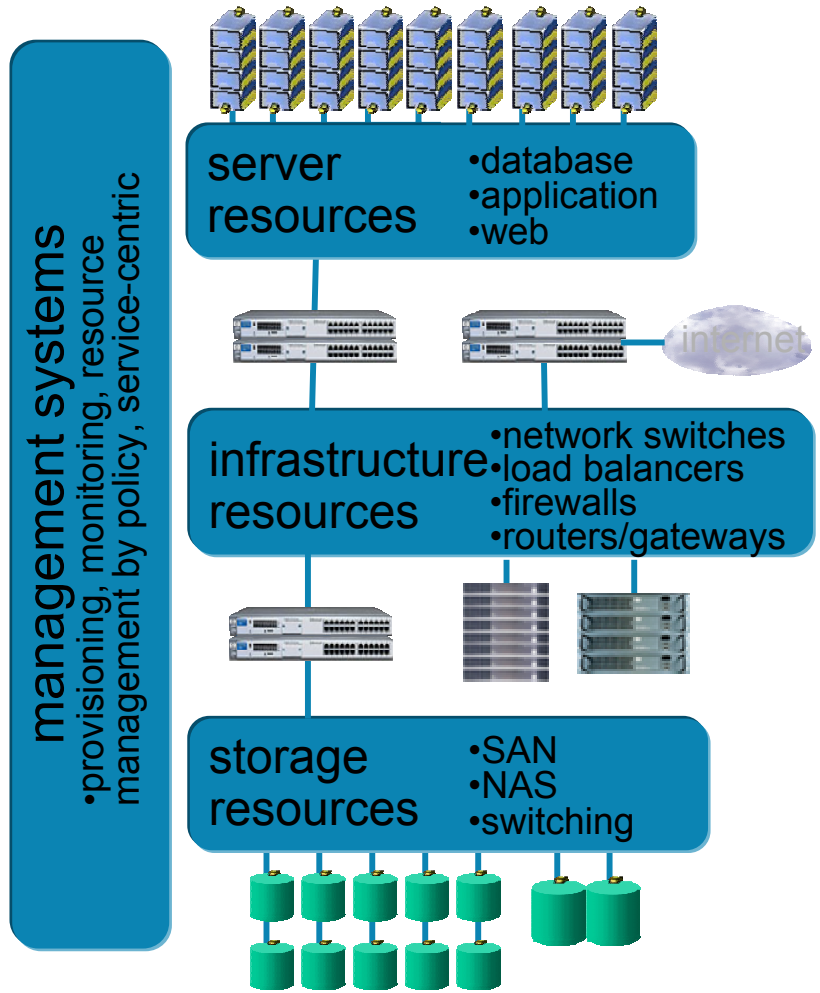
Reduce cost: Fabric Consolidation

- 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



Reduce cost: Fabric Consolidation, cont.

- 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 Cluster Interconnect technology
    - Enables high-performance Storage over IP
  - 10 Gigabit Ethernet
    - Provides bandwidth necessary for consolidation
    - Flattens switch structure, reduces component count
  - Flexibility of Ethernet
    - Continual industry development , guaranteed interoperability, Security, QoS, Management Suite, etc



Reduce cost: Increase server efficiency

- Server Scalability limited by communications

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

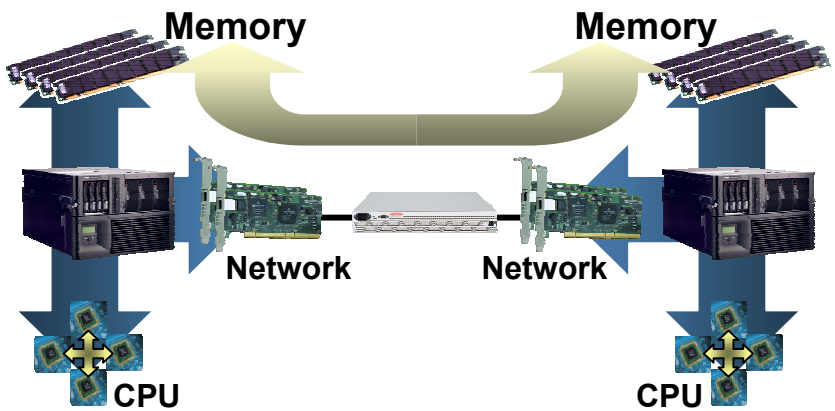
Critical Technology Rules of Thumb

- 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

RDMA – Just Better Networking

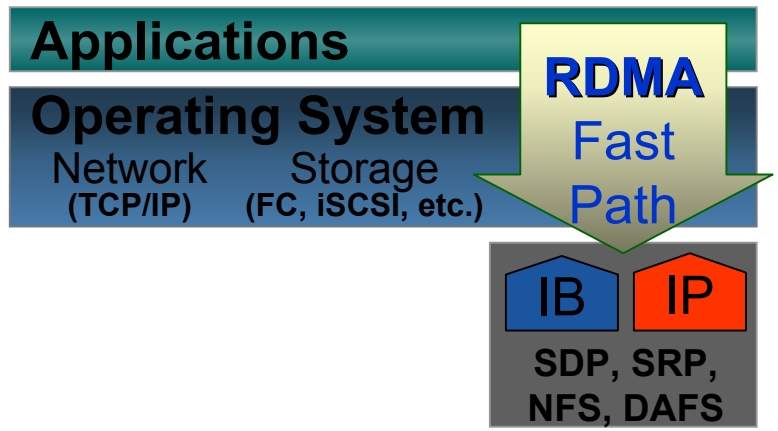
Fast and secure communications

- **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

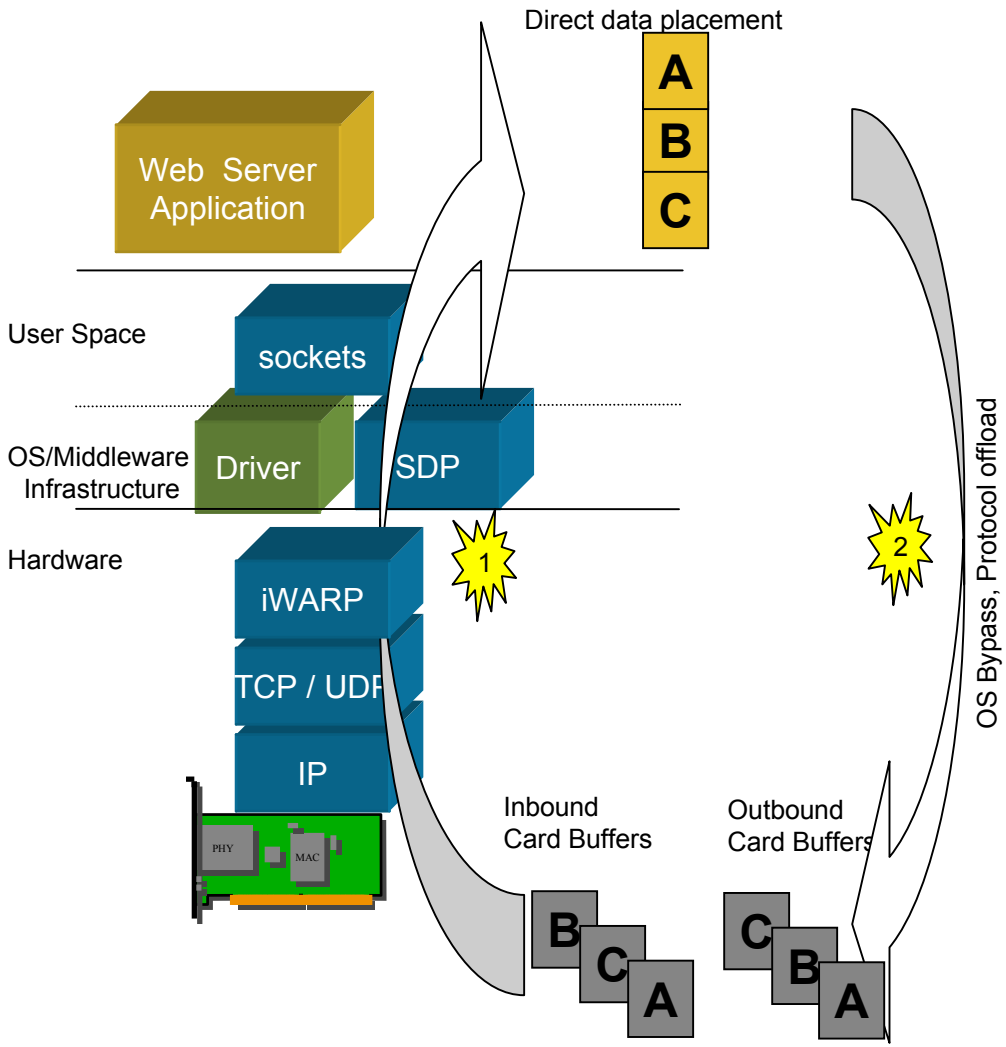


RDMA enables:

- 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



- 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

Dynamic Content Web Server

RDMA Summary

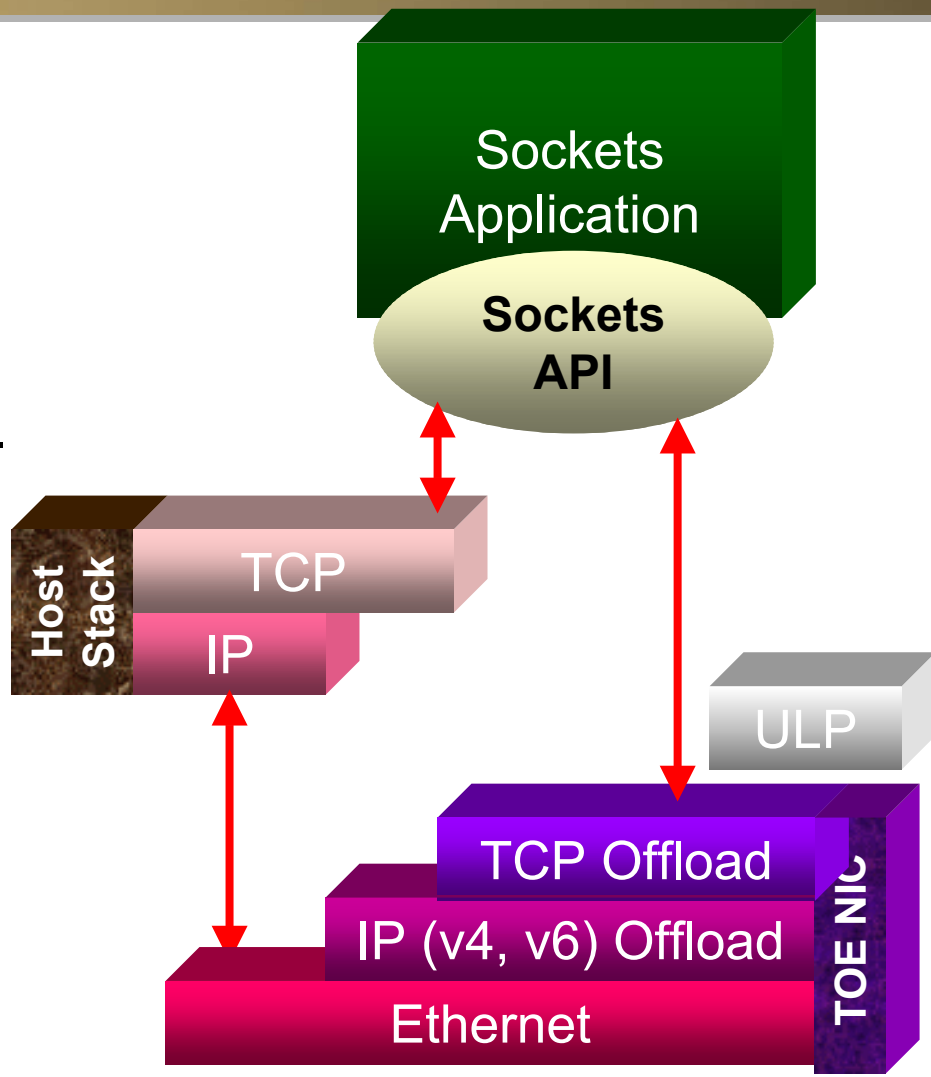
- RDMA
  - Combines Direct Memory Placement, Protocol Offload and OS Bypass
  - ***Eliminates*** copy overhead, protocol overhead, cache pollution effects, buffer management overhead, reduces NIC cost for high speed links
  - Enabled by sockets-level enhancements (Async sockets), OS enhancements (memory 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

Agenda

- Adaptive Enterprise Overview
- The Data Center Toolbox
- Data Center opportunities
- New Tools for the Data Center
- Deployment and Migration
- Summary

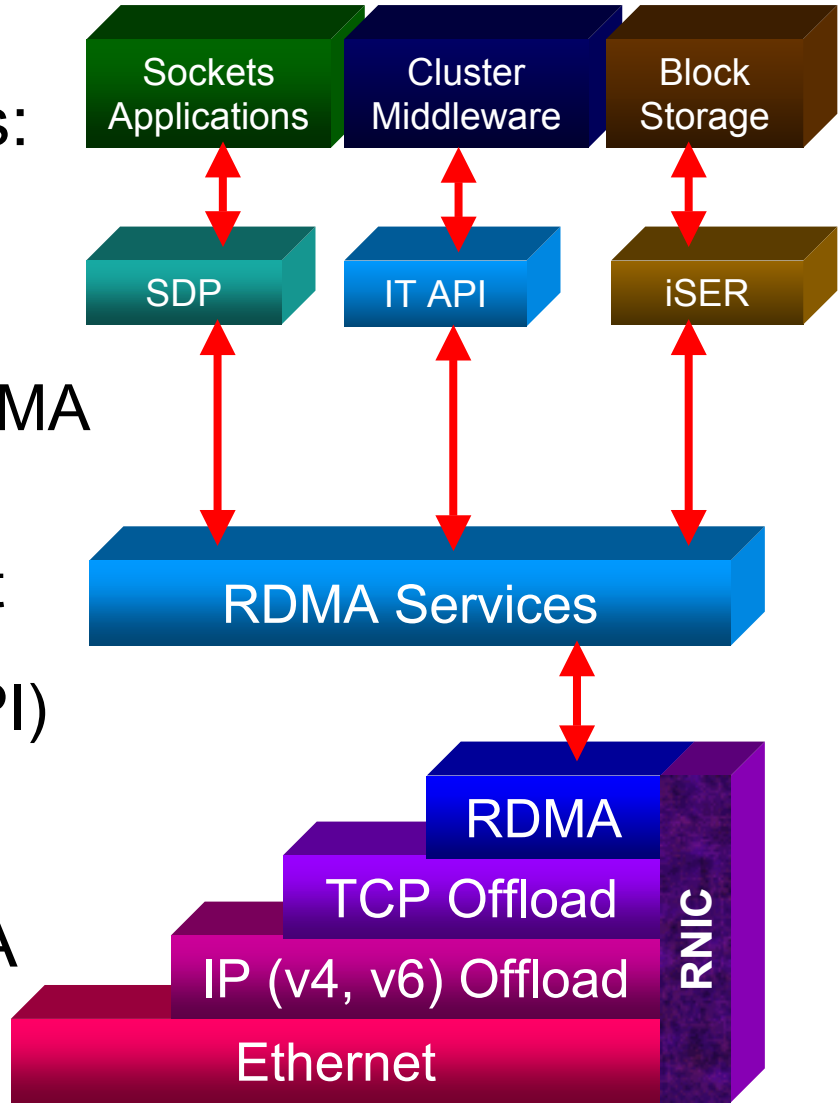
New tools for the Data Center: TCP Offload Engine (TOE)

- 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



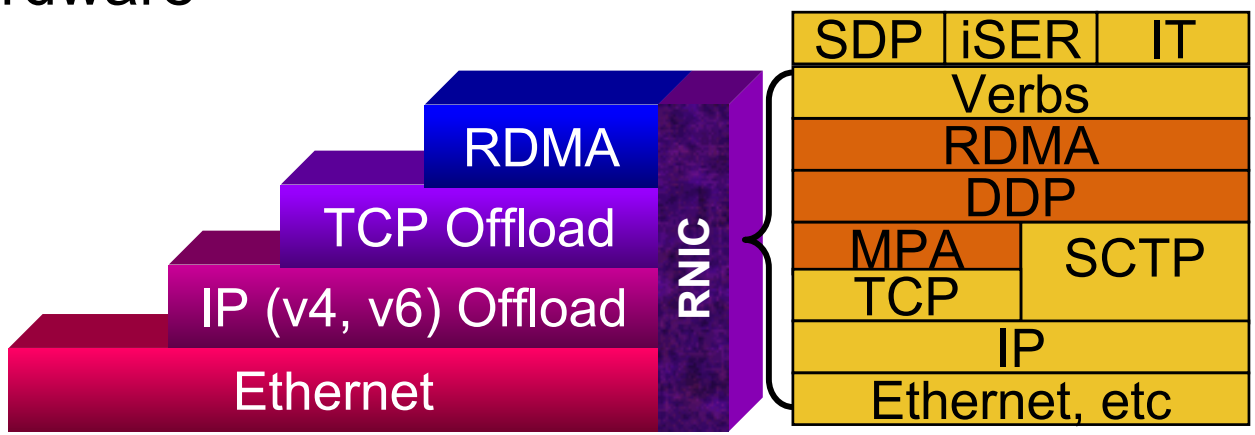
New tools for the Data Center: RNICs

- 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



New tools for the Data Center: RNICs, cont.

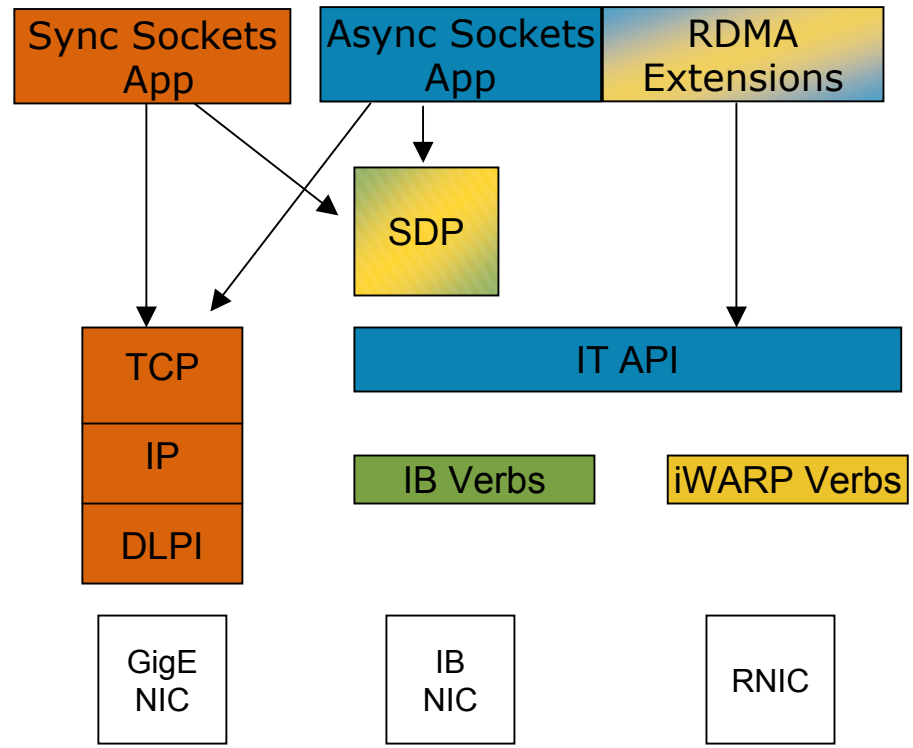
- Implements “iWARP” (RDMA over TCP) and TCP/IP protocol stacks on the card
  - RDMA wire protocol (iWARP)
  - DDP: Direct Data Placement
  - MPA: Marker based PDU Alignment
- Provides RDMA “Verbs” interface
  - Verbs provides standard functional semantics to hardware



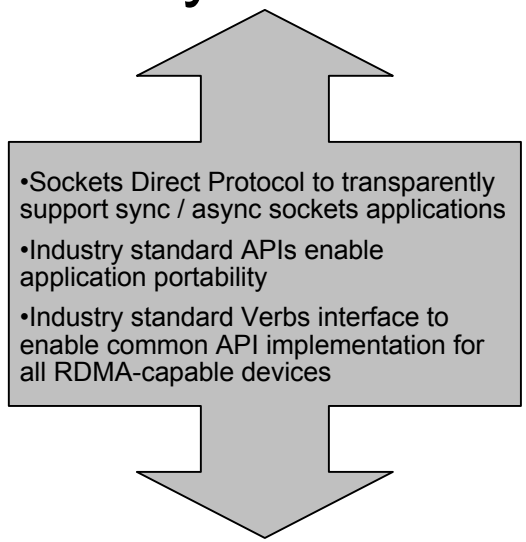
New tools for the Data Center: RDMA Software Infrastructure

- 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

New tools for the Data Center: RDMA Software Infrastructure, cont.

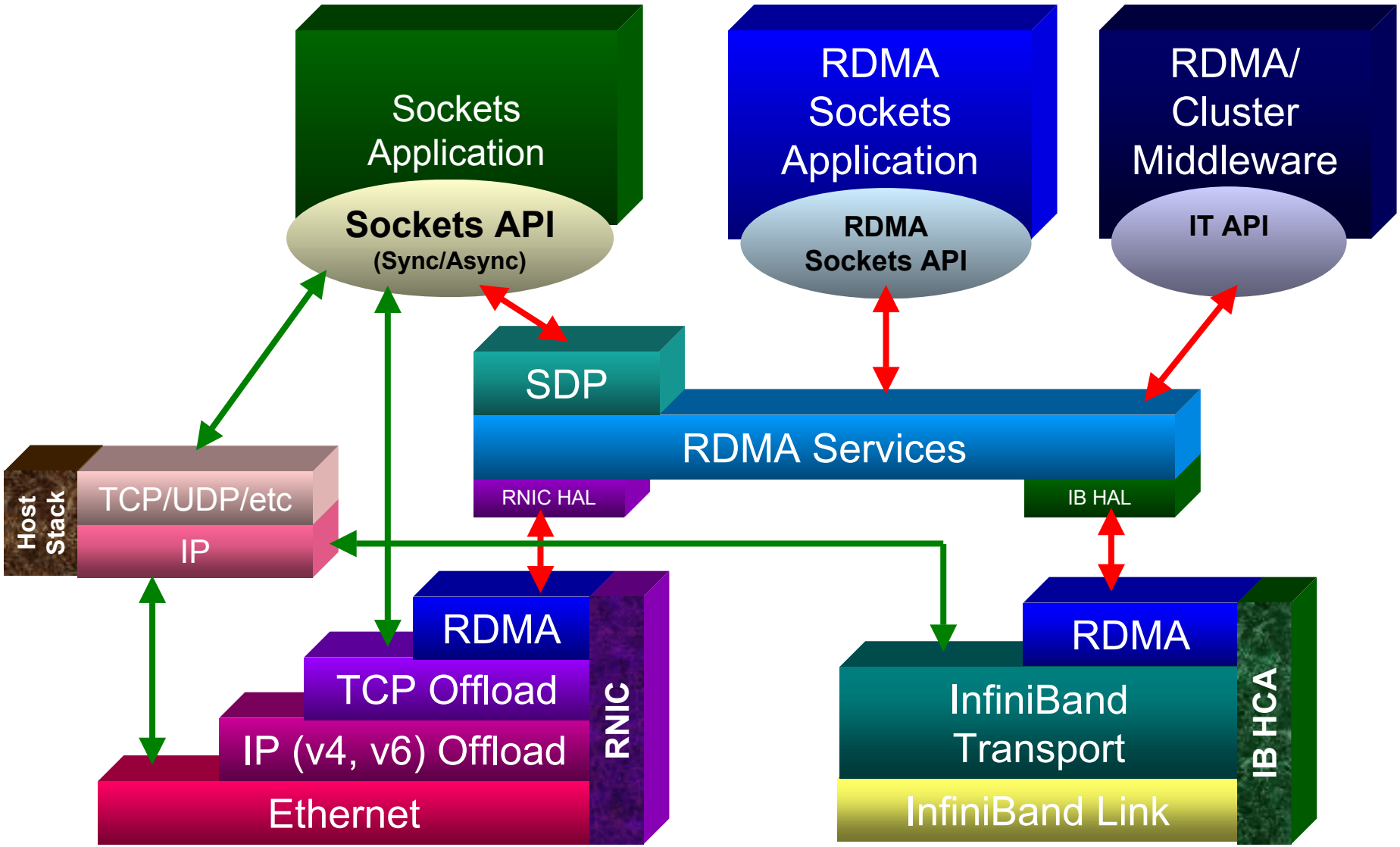


- Maximize benefit to existing applications
- Maximize application portability



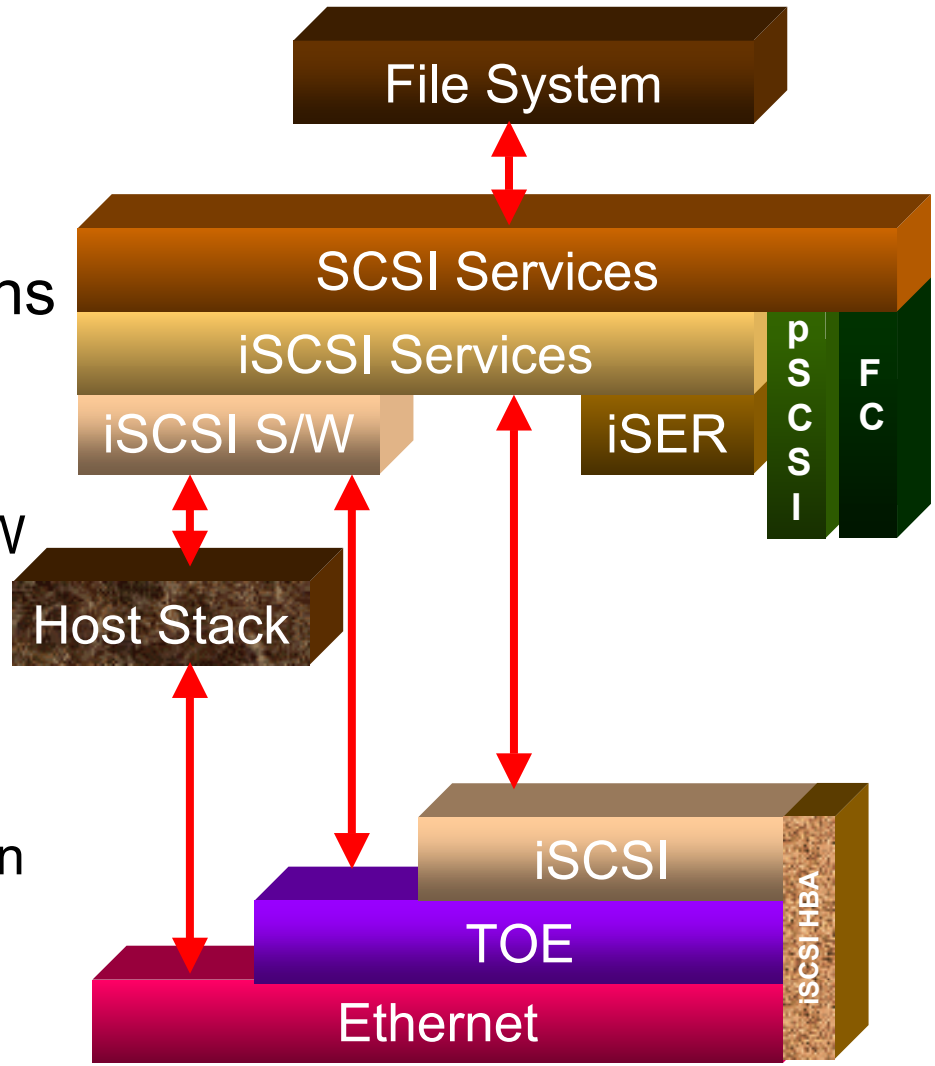
- Maximize hardware technology independence

New Tools for the Data Center : RDMA Networking Summary



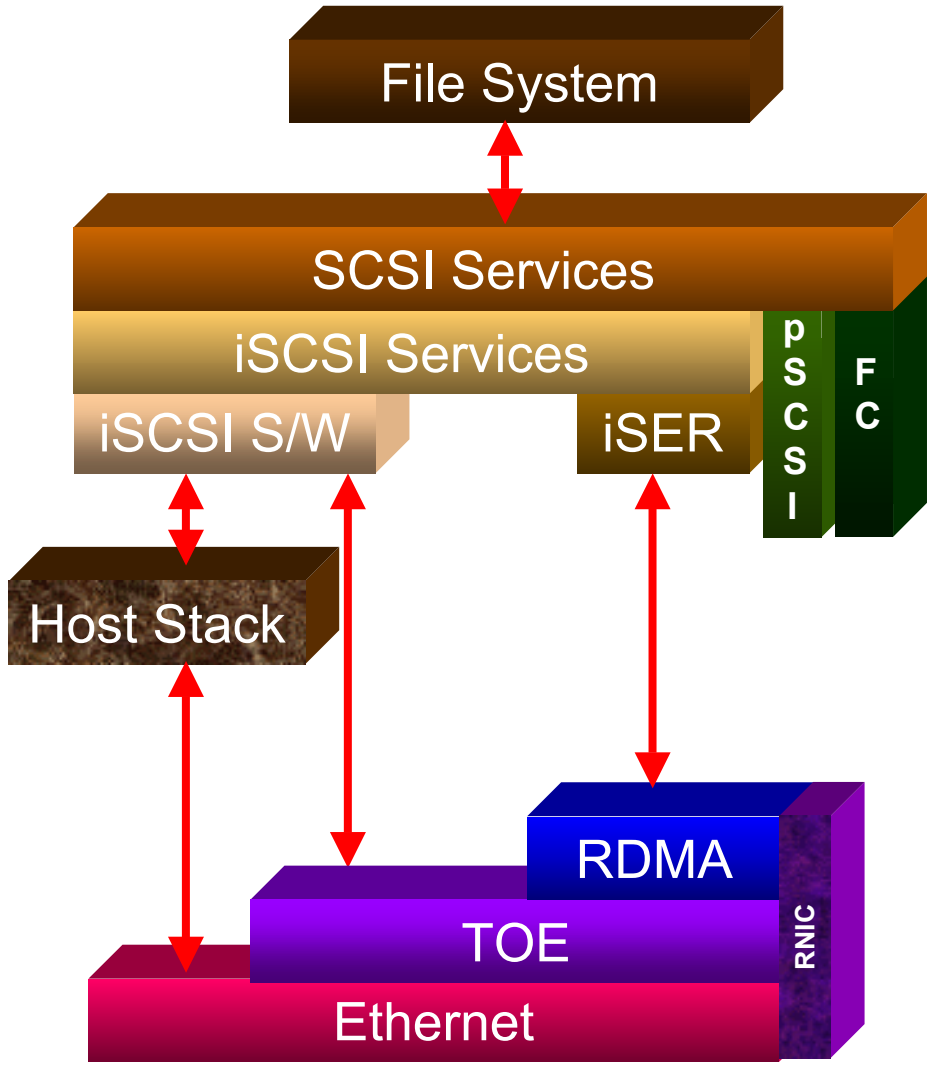
New tools for the Data Center: Storage over IP: iSCSI

- 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
    - iSCSI to FC bridge from HP, others

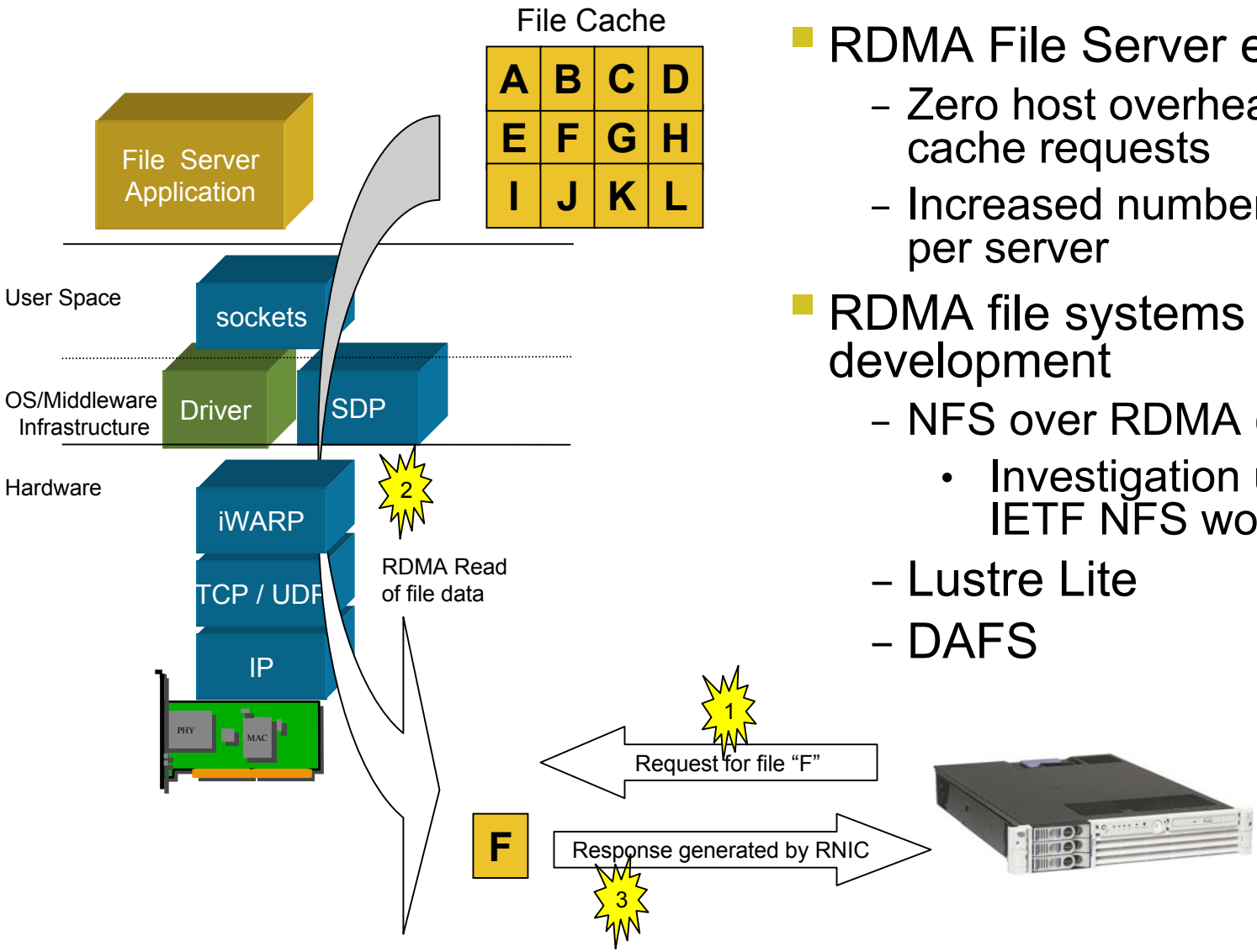


New tools for the Data Center: Storage over IP: iSER

- 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



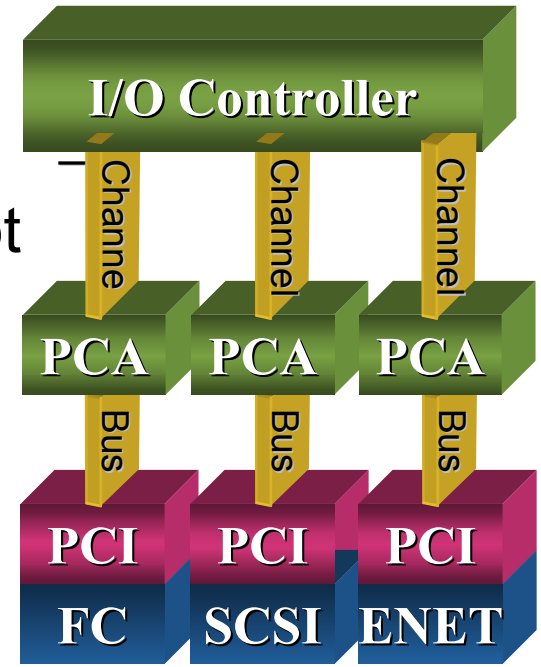
New tools for the Data Center: RDMA NAS



- 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

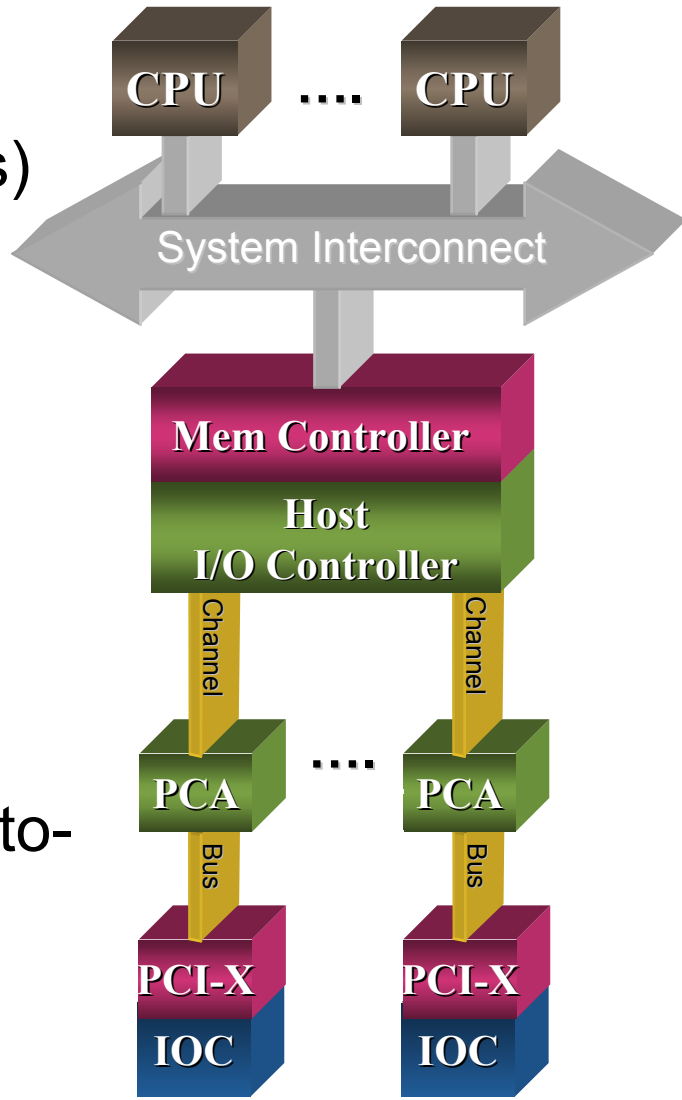
New tools for the Data Center: Dedicated PCI bus per I/O slot

- I/O Controller scales with host processor speeds
  - Point-to-point serial interconnect
- Predictable Performance
  - No bus arbitration contention
  - Full PCI / PCI-X bandwidth for each slot
- Superior RAS
  - Fault containment
  - Fault isolation
  - Simple, independent hot-plug
- Independent management domains
- No configuration restrictions
- Shipping with PCI-X 133 **today** (1GByte/second per slot)



New tools for the Data Center: PCI-X 2.0

- PCI-X 2.0: Higher speed with 100% customer investment protection
  - PCI-X 266 and 533 (2.1 - 4.3 GBytes/s)
  - Complete forward and backward compatibility with PCI-X
    - ***Any card, any slot, any time***
      Same card form factor and connector
    - No software (OS, driver) changes required
  - New protocol highlights:
    - ECC Data Integrity protection
    - Device ID Messaging (DIM) – Peer-to-peer messaging
    - Updated hot-plug signaling driver
    - Optional, improved error reporting

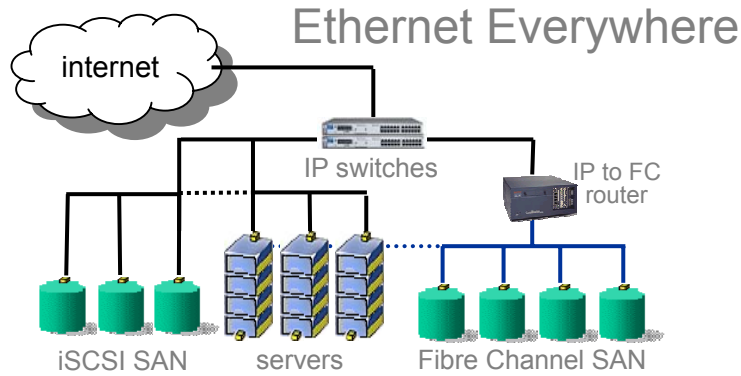


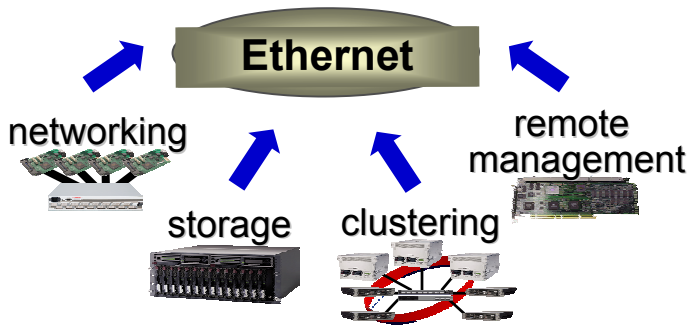
Agenda

- Adaptive Infrastructure Overview
- The Data Center Toolbox
- Data Center opportunities
- New Tools for the Data Center
- Deployment and Migration
- Summary

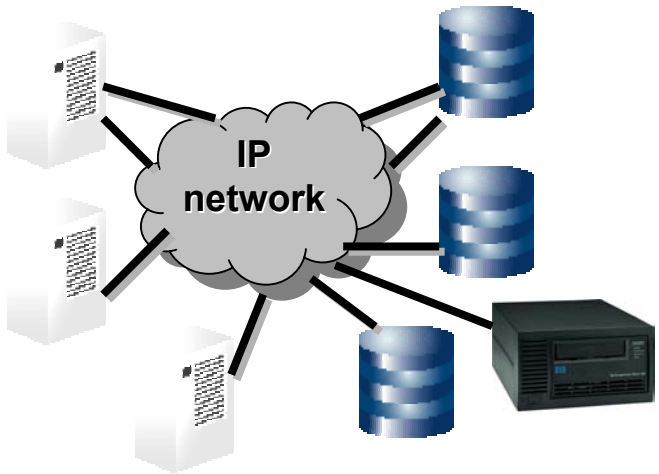
Creating a Simpler, Unified Infrastructure

- Consolidate ports
  - Leverage Ethernet pervasiveness, knowledge, cost leadership and volume
  - Headless console servers
    - Consolidate KVM over IP and reduce switch port costs (up to $1000 per port)
- Converge functions
  - Multiple functions can be consolidated over a common type of network
  - Blade server storage connectivity (low cost)
  - Packaged “end-to-end” Ethernet solutions
- Broad connectivity - Ethernet Everywhere
  - Bridge “islands” - ex: iSCSI to FC bridge
  - Extend geographic reach - Remote sites
  - Centralize management



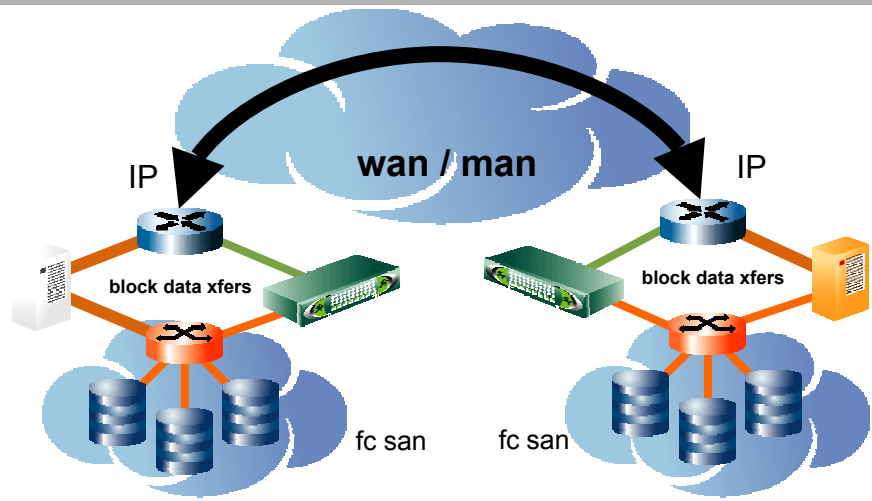


Storage over IP deployment



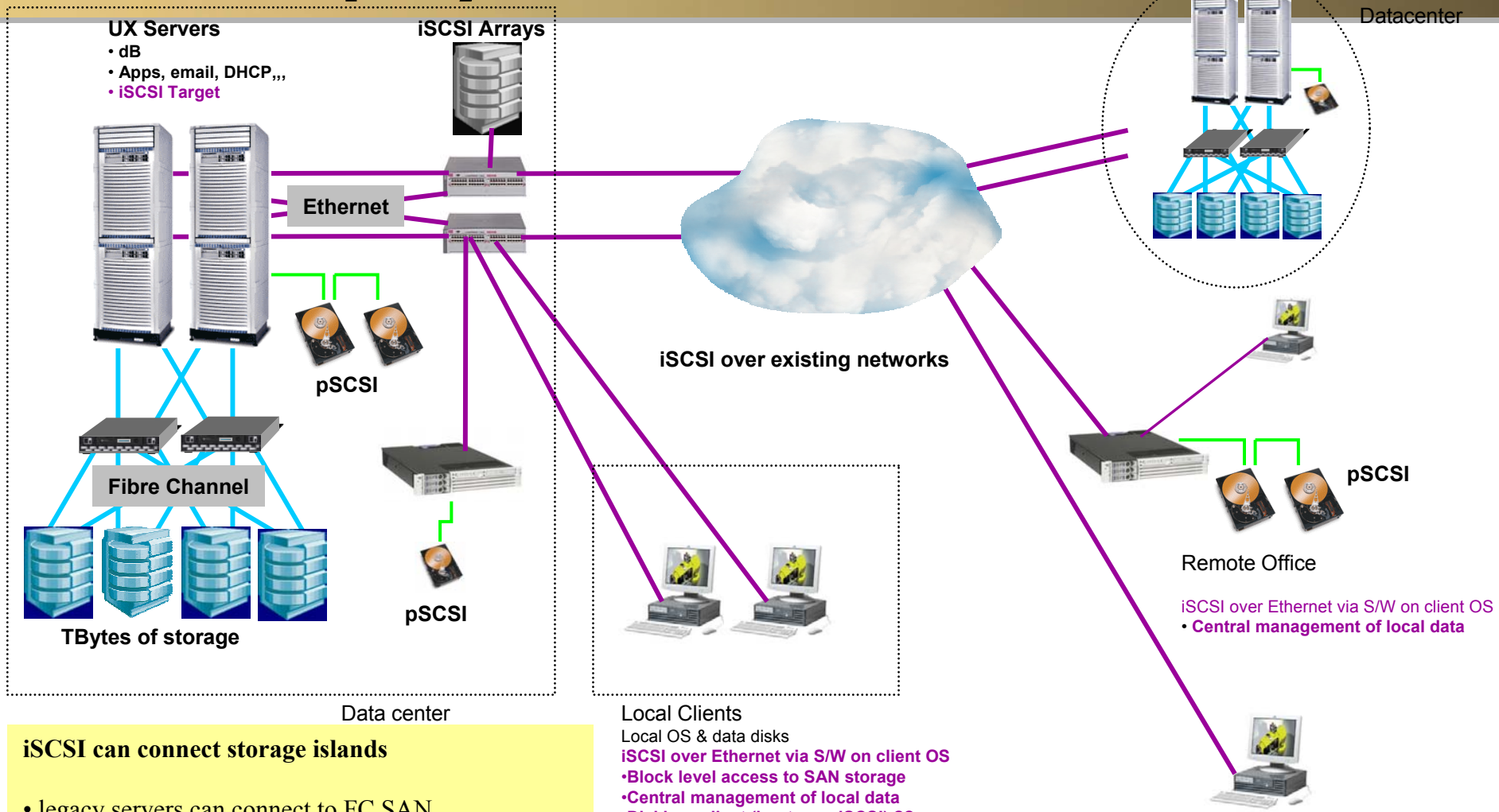
- Common fabric for both block and file storage access
  - Lower cost structure:
    - Host connectivity (single interconnect type), deployment, operations, commodity parts, etc.
  - Leverage IP / Ethernet infrastructure and expertise
  - Integrates into HP Utility Data Center unified fabric utility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

iSCSI deployment



iSCSI can connect storage islands

- legacy servers can connect to FC SAN
- clients can access SAN storage at block level
- allows central management of local client data
- effective protocol to connect remote data centers

Cluster technology deployment

- Immediate solution
  - Maintain separate fabric for Cluster Interconnect
  - Deploy InfiniBand (available in 2003)
    - 10Gb link rates
    - PCI-X 133 point of attach today, PCI-X 266 in 2004
  - 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

Product Availability

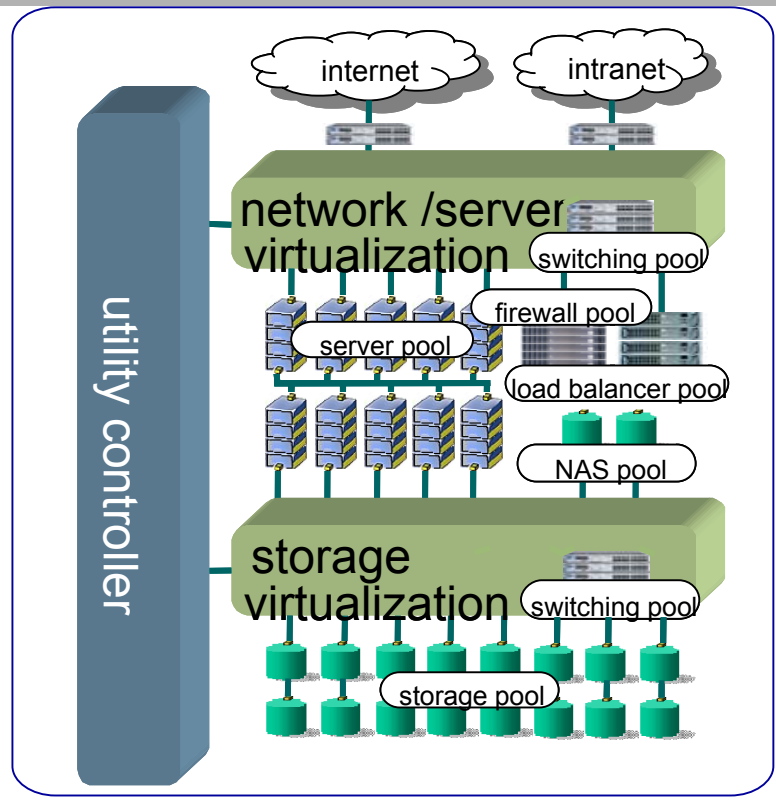
Estimates for the Industry as a whole, product offerings from multiple vendors

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

Does not indicate specific product plans from HP

HP Utility Data Center

- Wire Once
  - Support virtual allocation of resources for the entire system
- Resource Virtualization
  - Allocate networking, storage and server components without rewiring
- Utility Controller
  - Simple user interface to architect and activate new systems using available resources



UDC is a fully integrated software and hardware solution that enables virtual provisioning of application environments to optimize asset utilization and reduce administrative staff

Agenda

- Adaptive Infrastructure Overview
- The Data Center Toolbox
- Data Center opportunities
- New Tools for the Data Center
- Migration Strategy
- Summary

New I/O and IPC Technology =

- 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

- New technologies enable simplified data center infrastructure, lower TCO
  - Fabric consolidation, increased server efficiency through protocol offload
  - Enabled by RDMA hardware and software infrastructure
  - New networking paradigms are needed to allow server scaling to keep pace with advances in network and processor technology
- New technologies create complete solution
  - Ethernet feature set extended to encompass LAN, SAN, CI
  - iSCSI to FC bridging enables deployment of each technology when and where it meets the needs of your data center
  - RDMA Software Infrastructure enables application portability across platforms, hardware technology independence for software
    - Enables clean migration between CI technologies
    - Preserves investment in software
  - VLAN, QOS enables dynamic deployment, simplified management, shared use of higher capacity networks

Summary, cont.

- New technologies are here, more on the way
  - Industry standard RDMA solutions, 10GbE, iSCSI ready to deploy
  - iWARP, low latency Ethernet switching, greater server I/O capacity coming to market over next 2 years
  - Near term investment in high-speed Ethernet fabric technology and careful investment in RDMA-capable software will ease future migration and lower long-term TCO
- 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 UDC environment

Interex, Encompass and HP bring you a powerful new HP World.

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, Async Sockets, Fabric Management
    - Founding Members: Fujitsu, **HP**, IBM, Network Appliance, Sun, Vieo [Principal member: Intel]
  - **RDMA Consortium**
    - Developing 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/IETF working to define iSCSI over RDMA (iSER)