IP Fabrics – The Future of Networking!

Chuck Hudson

Manager, Network System Engineering
Industry Standard Servers, HP

HP WORLD 2003
Solutions and Technology Conference & Expo

Agenda

- Introduction
- Challenges for IP/Ethernet
- Improving Throughput Scaling
- Remote DMA
- Networked Storage
- Conclusions

The server environment today

A collection of special-purpose interconnects…

Network

- Gigabit Ethernet
- Limited processing offload (checksums, LSO)
- Virtual Local Area Networks (VLANs)

LAN

Storage

- Fiber Channel Storage Area Network (SAN)
- Network Attached Storage (NAS) using CIFS, NFS, etc.
- SCSI direct attached storage (DAS)

SAN, NAS

System

- Cluster Interconnect
- High-speed (>1Gb/s)
- Low-latency (<50 µS)
- IB and Proprietary Solutions (ServerNet, Myrinet, etc.)

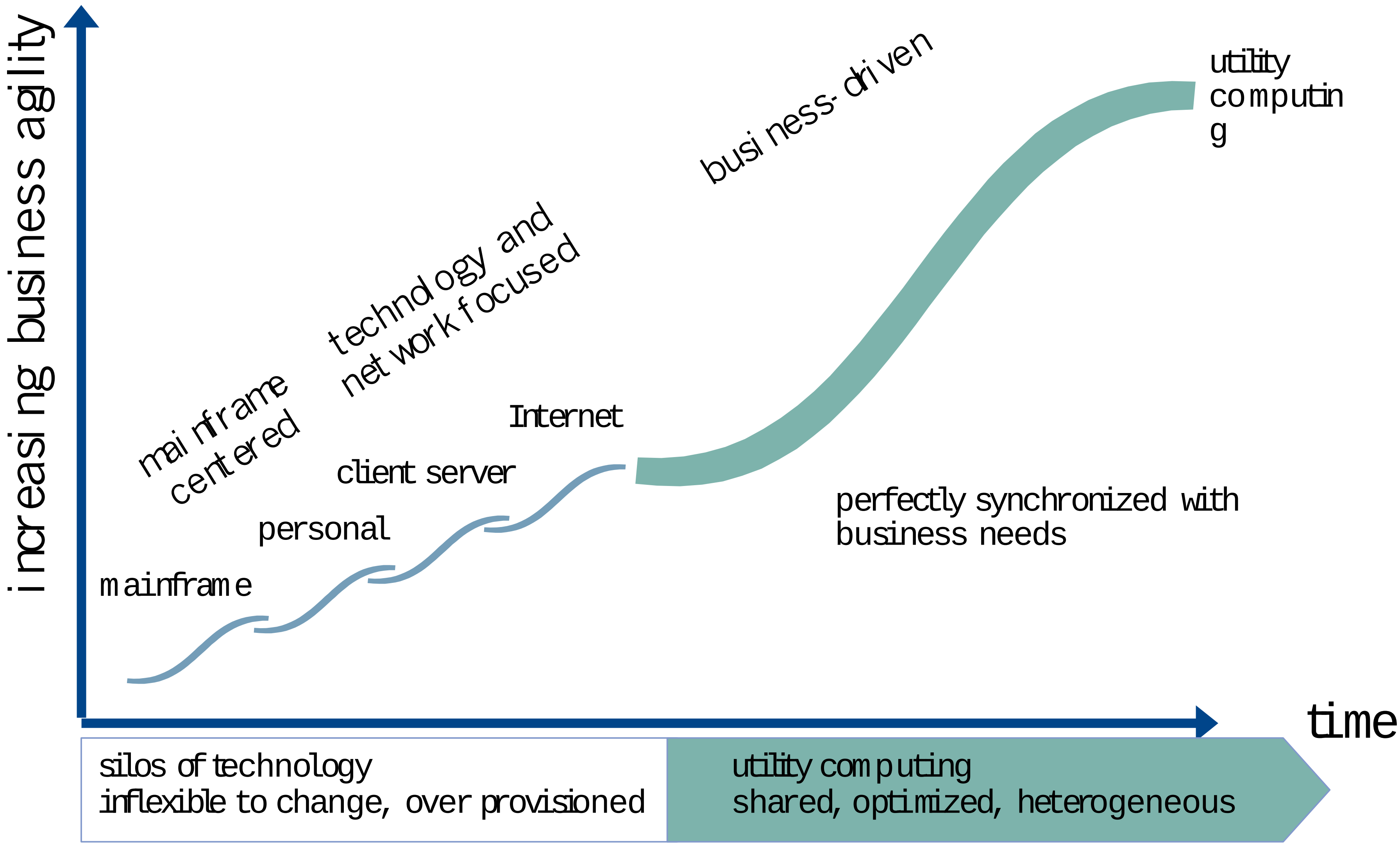
IP
C

Management

- KVM
- HP OpenView
- HP Insignt Manager
- Lights Out Management
- ProLiant Essentials RDP

KVM → iLO

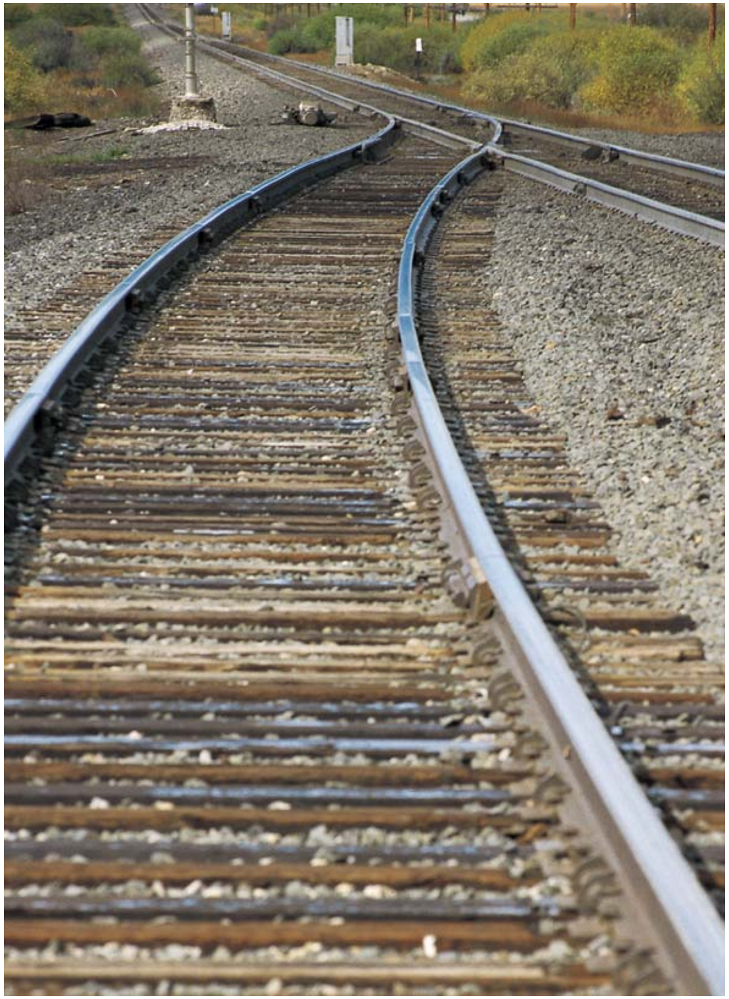
Utility computing needs an adaptive interconnect fabric



Wanted… a single interconnect fabric

- A single media that
  - provides a simpler, unified infrastructure
  - improves performance
  - increases flexibility
  - supports utility computing
- A single media to handle
  - networking
  - block and file storage
  - management
  - cluster interconnect

Requirements for our single fabric…



- Single medium
- Standards-based
- Scalable throughput
- Low-cost
- Reliable
- Low-latency
- Flexible
- Secure
- Familiar

Is IP/Ethernet our single fabric?

- Strengths
  - Ubiquitous; standard
  - Extends beyond the data center
  - Minimal training costs
  - Understood management model
  - Affordable adapter, cabling, switches
  - Mature foundation
- Weaknesses
  - Scalability
    - CPU consumption
    - Memory bandwidth consumption
  - Latency

Agenda

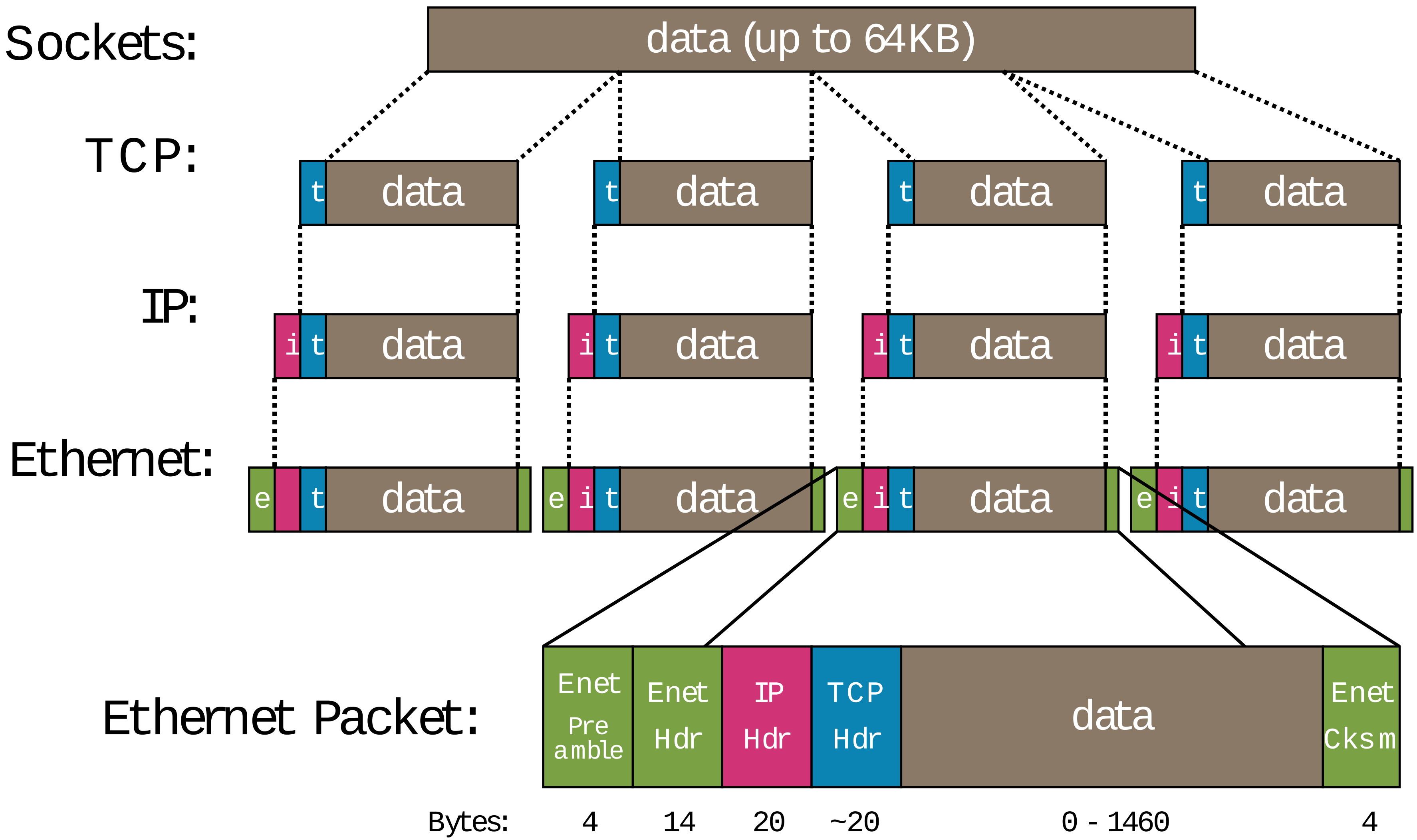
- Motivation
- **Challenges for IP/Ethernet**
- Improving Throughput Scaling
- Remote DMA
- Networked Storage
- Conclusions

Challenges for IP/Ethernet:
TCP/IP CPU utilization

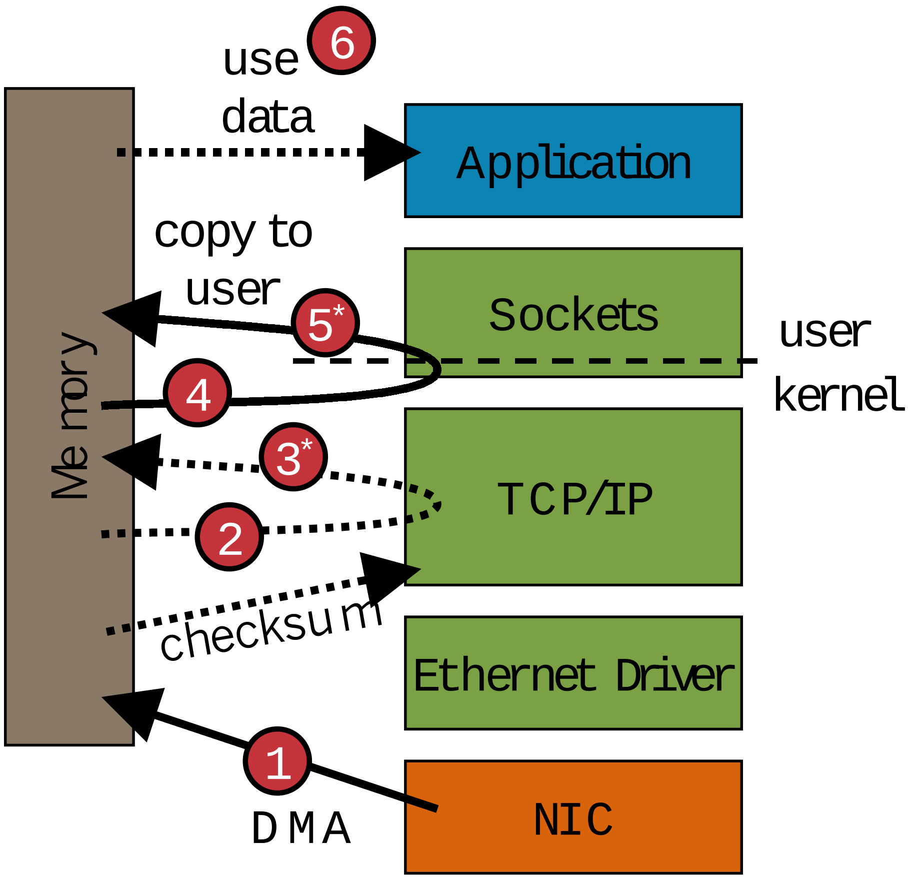
'CPU utilization'

- CPU computation
  - segmentation & re-assembly
  - checksum calculation
  - memory management
  - sync. data structures
- Context switches
  - caused by
    - user/kernel transitions
    - interrupts
  - may result in pre-emptions
- Buffer copies
  - between user and kernel memory
  - between kernel memory and network interface card

Challenges for IP/Ethernet:
TCP segmentation & reassembly



Challenges for IP/Ethernet:
Buffer copies (RX)



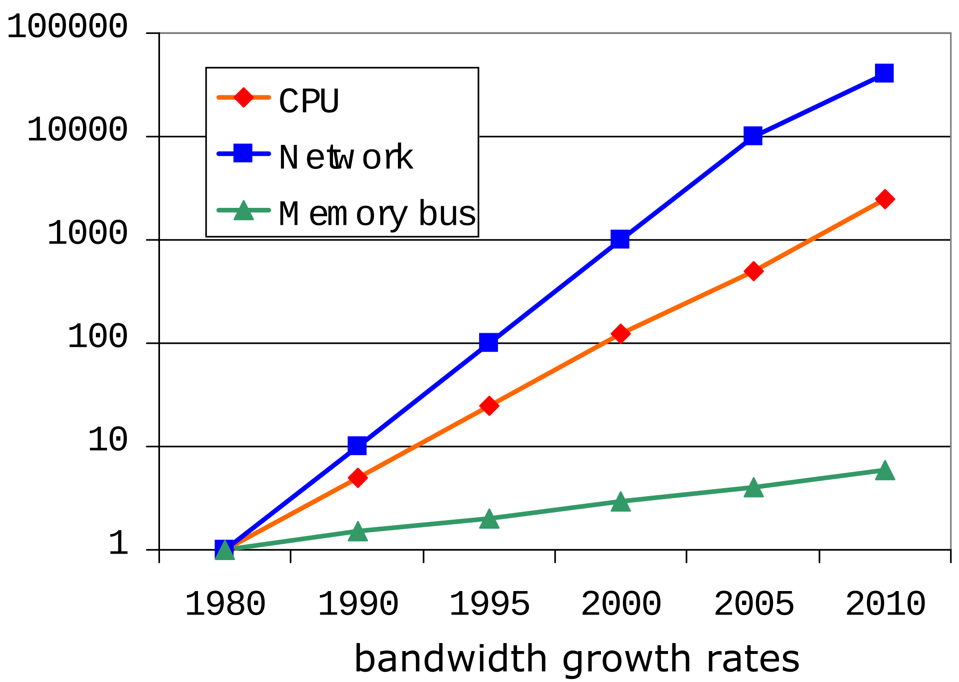
- transfer from NIC to server memory via DMA ①
- checksum calculation (may be offloaded)
- data may be copied to free up the NIC receive buffer ② ③
- data is copied from kernel TCP buffer to user buffer ④ ⑤
- application may copy data into other data structures ⑥

* Note: writes consume 2x memory bandwidth of read due to cache line reads

Challenges for IP/Ethernet:
Memory bandwidth limitations

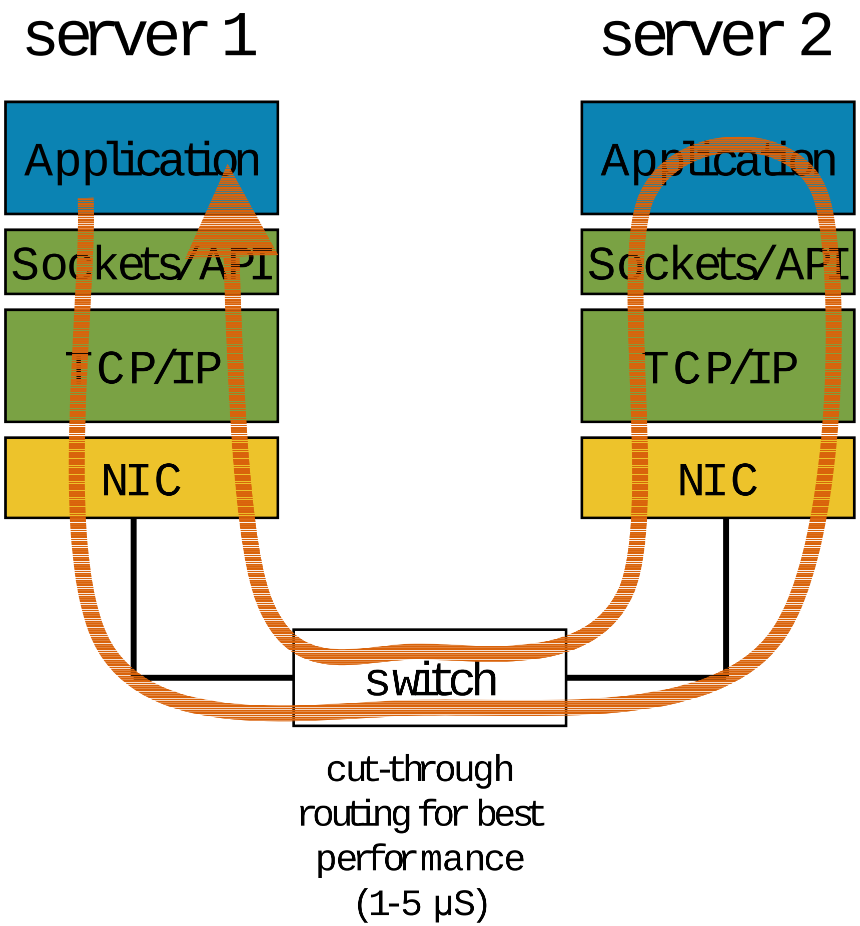
- Host-based TCP/IP consumes memory bandwidth equal to 4x to 7x the raw data rate.
  - 1-2 buffer copies + DMA
  - each buffer copy = 3x memory touches
- Memory Controller Bandwidth is not keeping up with CPU and network bandwidth.
- Current memory controller bandwidth is ~ 3-6 GB/sec

Ethernet	Raw Data Rate	Required Memory Bandwidth (RX)
1 GbE	125 MB/sec	500-875 MB/sec
10 GbE	1250 MB/sec	5000-8750 MB/sec



Challenges for IP/Ethernet:
Round-trip message latency

- server 1 protocol stack (µS)
- network latency (µS)
  - NIC1 latency
  - Switch latency
  - NIC 2 latency
- server 2
  - protocol stack (µS)
  - **wake-up application for response (mS)**
  - server 2 protocol stack (µS)
- network latency (µS)
- server 1
  - protocol stack (µS)
  - **wake-up receiving application (mS)**



Agenda

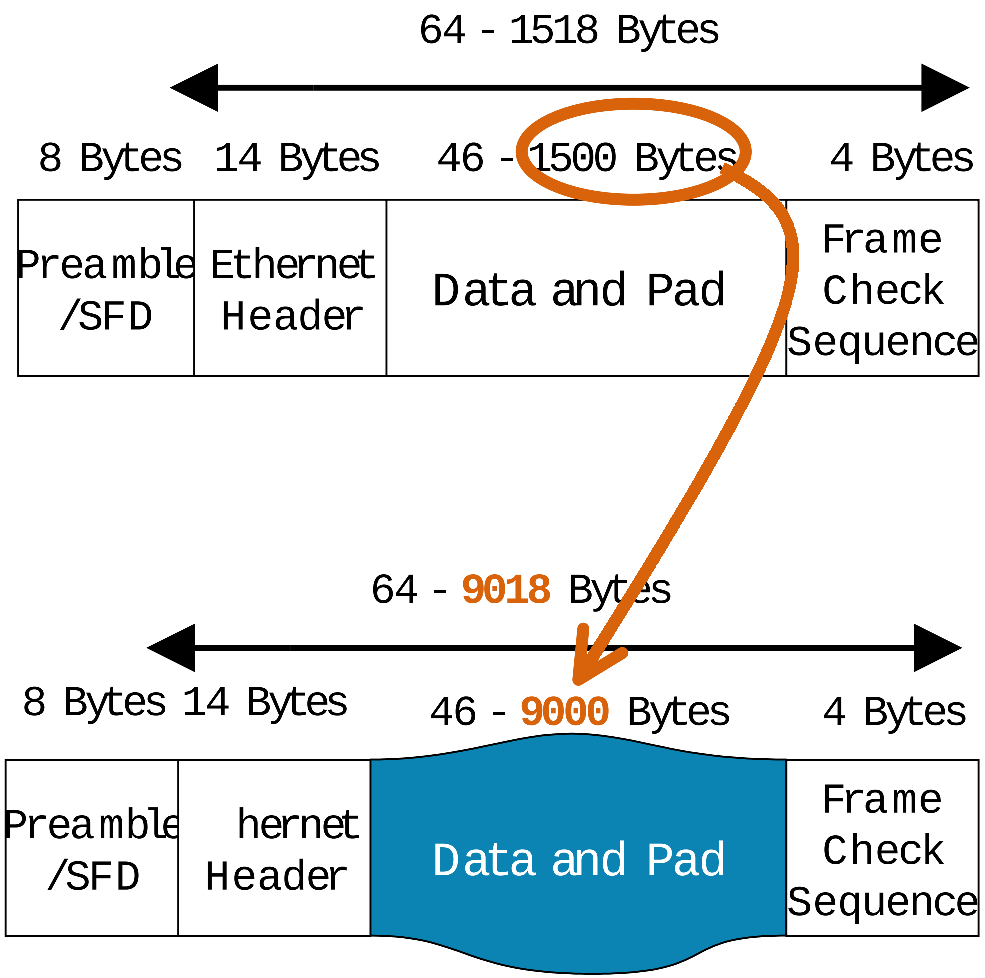
- Motivation
- Challenges for IP/Ethernet
- **Improving Throughput Scaling**
- Remote DMA
- Networked Storage
- Conclusions

Improving throughput scaling

- Jumbo frames
- Asynchronous IO
- Large Send Segmentation Offload
- Receive Side Scaling
- TCP/IP Offload Engines (TOE)
- Remote Direct Memory Access (RDMA)

Jumbo Frames

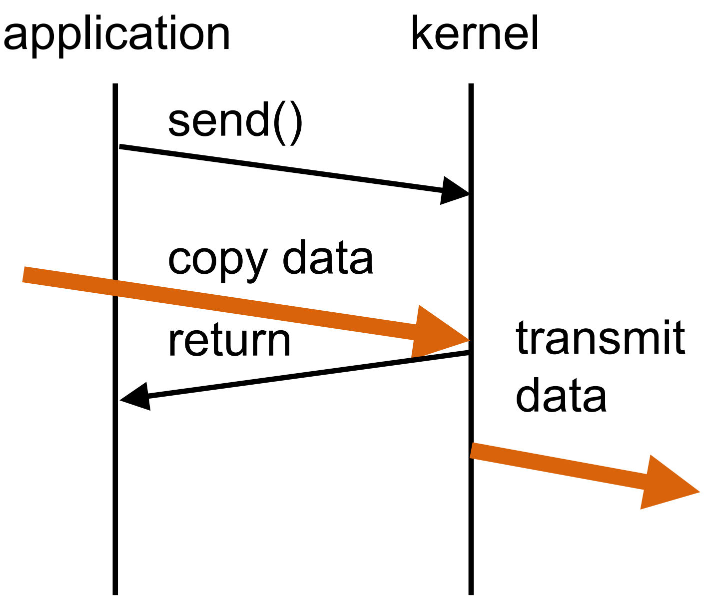
- Ethernet frame size is increased.
- Reduces the amount of segmentation and reassembly overhead.
- Requires all points on the network to support jumbo frames (limits deployment).
- Informal standard.



Improving Throughput Scaling

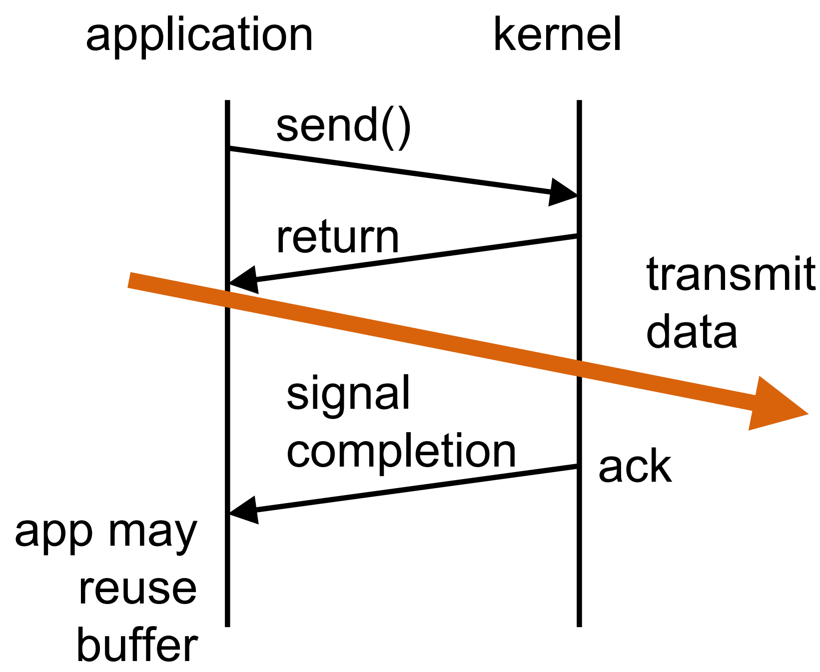
Asynchronous I/O

synchronous send



kernel copy needed

asynchronous send



No copy required!

Improving Throughput Scaling

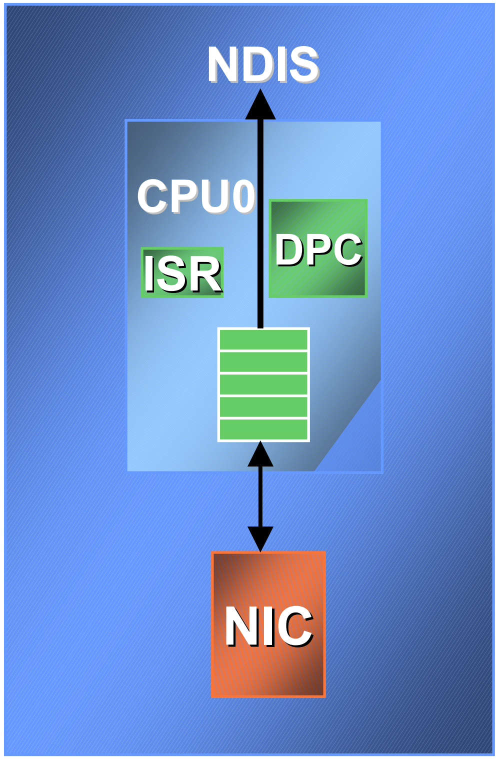
Large Send Offload (LSO)

- Pushes segmentation of out-going data to NIC
- TCP passes large blocks (up to 64KB) to NIC hardware
- NIC partitions into Ethernet frames (1.5KB)
- Only works for sends
- Reduces segmentation CPU utilization
- No special infrastructure support required
- **Available in Microsoft Windows Server 2003**

Improving Throughput Scaling

Receive Side Scaling (RSS)

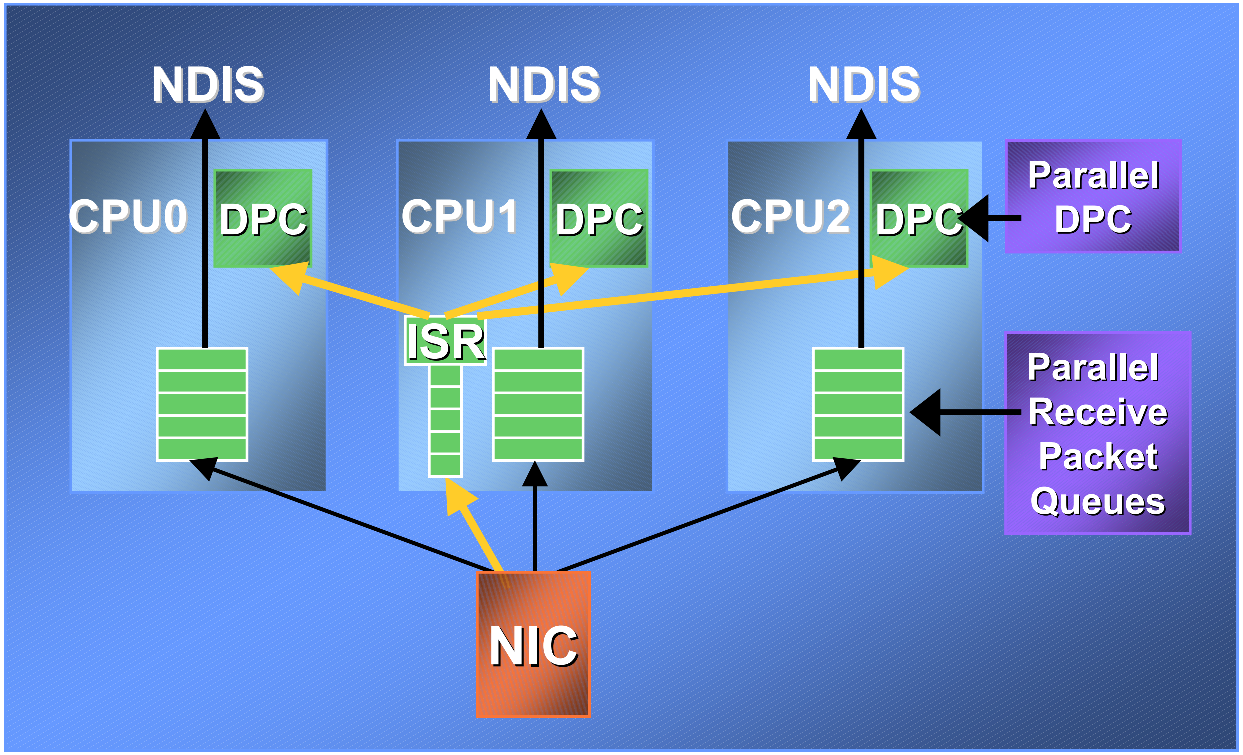
Today



One processor
per NIC

Figure courtesy of Microsoft,

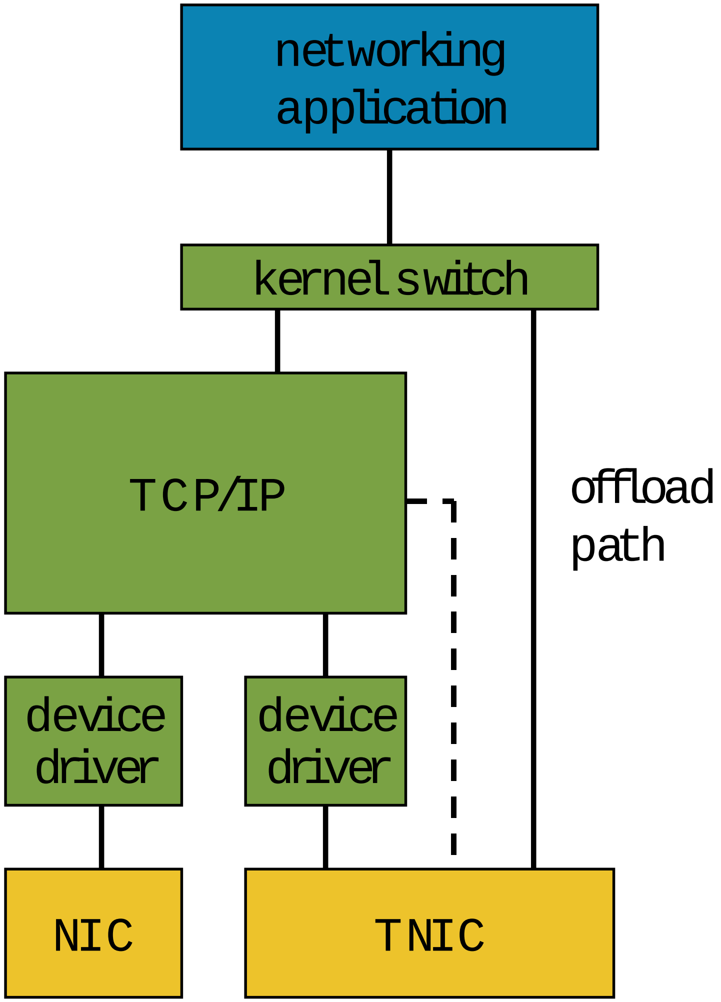
NIC with RSS



- One RSS Implementation
  - Single Interrupt Service Routine (ISR), ISR queue tells which hardware packet queue received a packet
  - Second-level lookup to find which CPU to run DPC
  - DPC processes receive packet queue

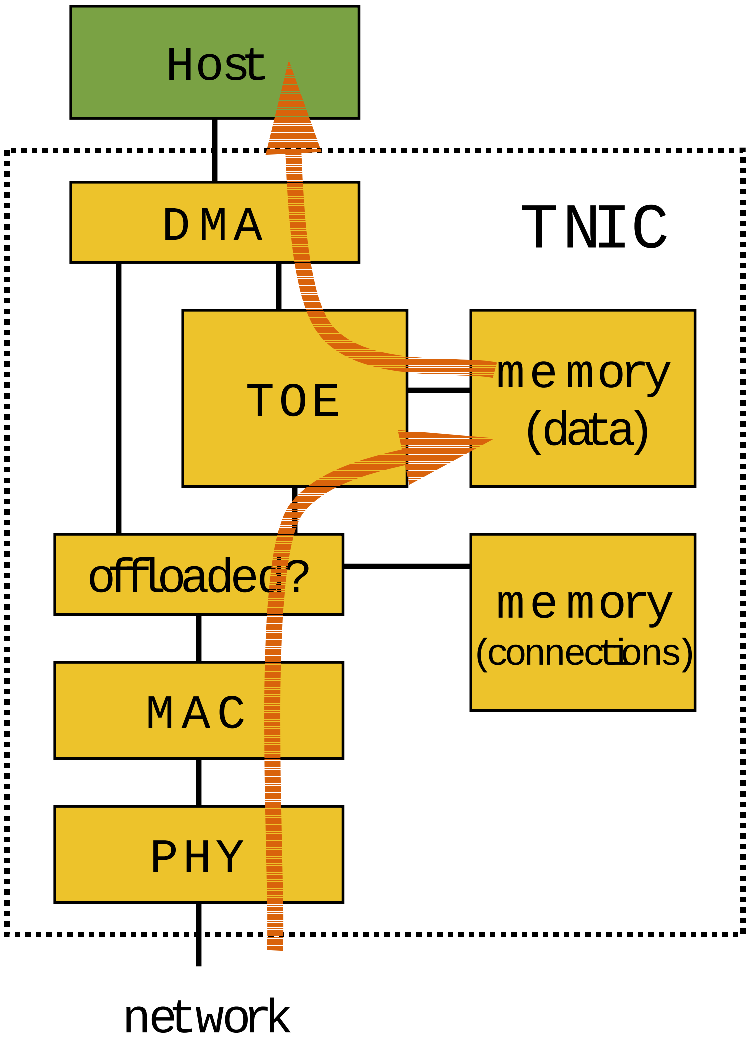
Improving Throughput Scaling
TCP/IP Offload Engines (TOE)

- TCP/IP processing moved from the host CPU to TOE NIC (TNIC)
- TCP connections may be established in TNIC or by host
- Reduces CPU utilization for segmentation and reassembly
- Reduces interrupts and context switches
- Allows for zero-copy receives to kernel memory buffers
- Works best with async IO



TOE NIC operation

- TCP connection state retained on NIC
- Incoming packet sequence
  - headers inspected to see if associated with offloaded connection
  - if so, TCP/IP processed on-chip
    - packet re-ordering may be required (data memory)
    - data transferred to host
  - if not, packet sent for host processing

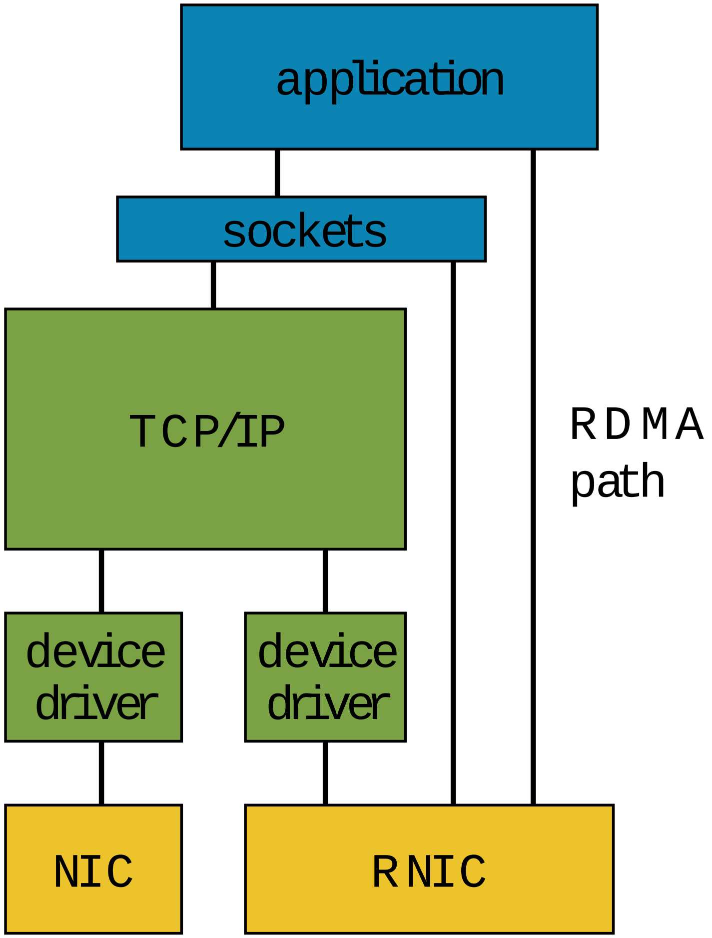


Agenda

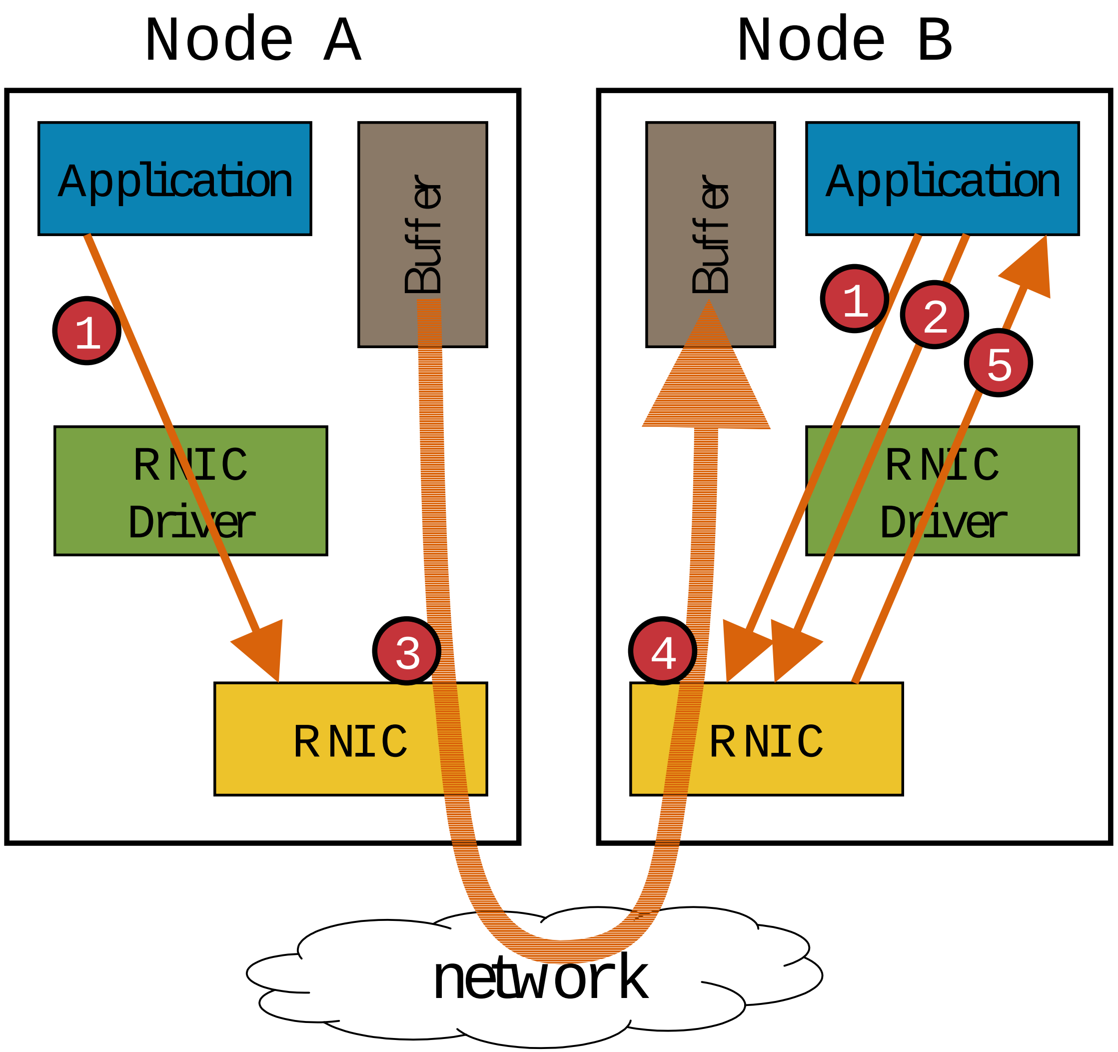
- Motivation
- Challenges for IP/Ethernet
- Improving Throughput Scaling
- **Remote DMA**
- Networked Storage
- Conclusions

Remote DMA NIC (RNIC)

- Provides direct communication between application buffers in separate servers.
- Bypasses the OS kernel
  - avoids protocol processing
  - avoids context switches
  - avoids interrupt processing
  - yet, preserves kernel protections
- Improves both
  - throughput scaling
  - message latency
- Provides the performance needed by networking, IPC, and storage

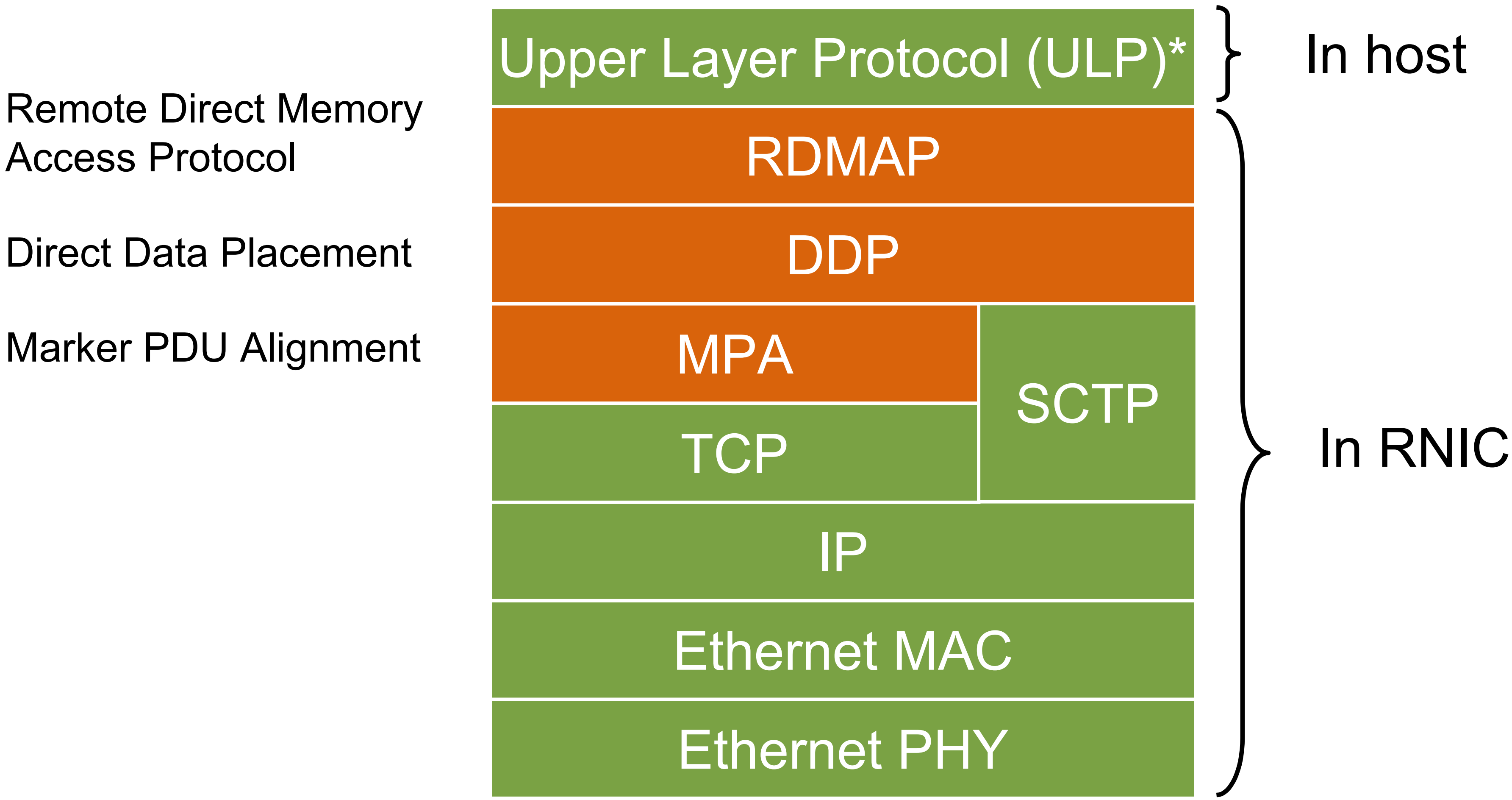


RDMA read operation



1. Both nodes have suitable memory regions registered
2. Node B initiates RDMA Read
3. RNIC in Node A sends data
4. RNIC in Node B places data in final buffer destination
5. RNIC in Node B completes read (w/o kernel intervention)

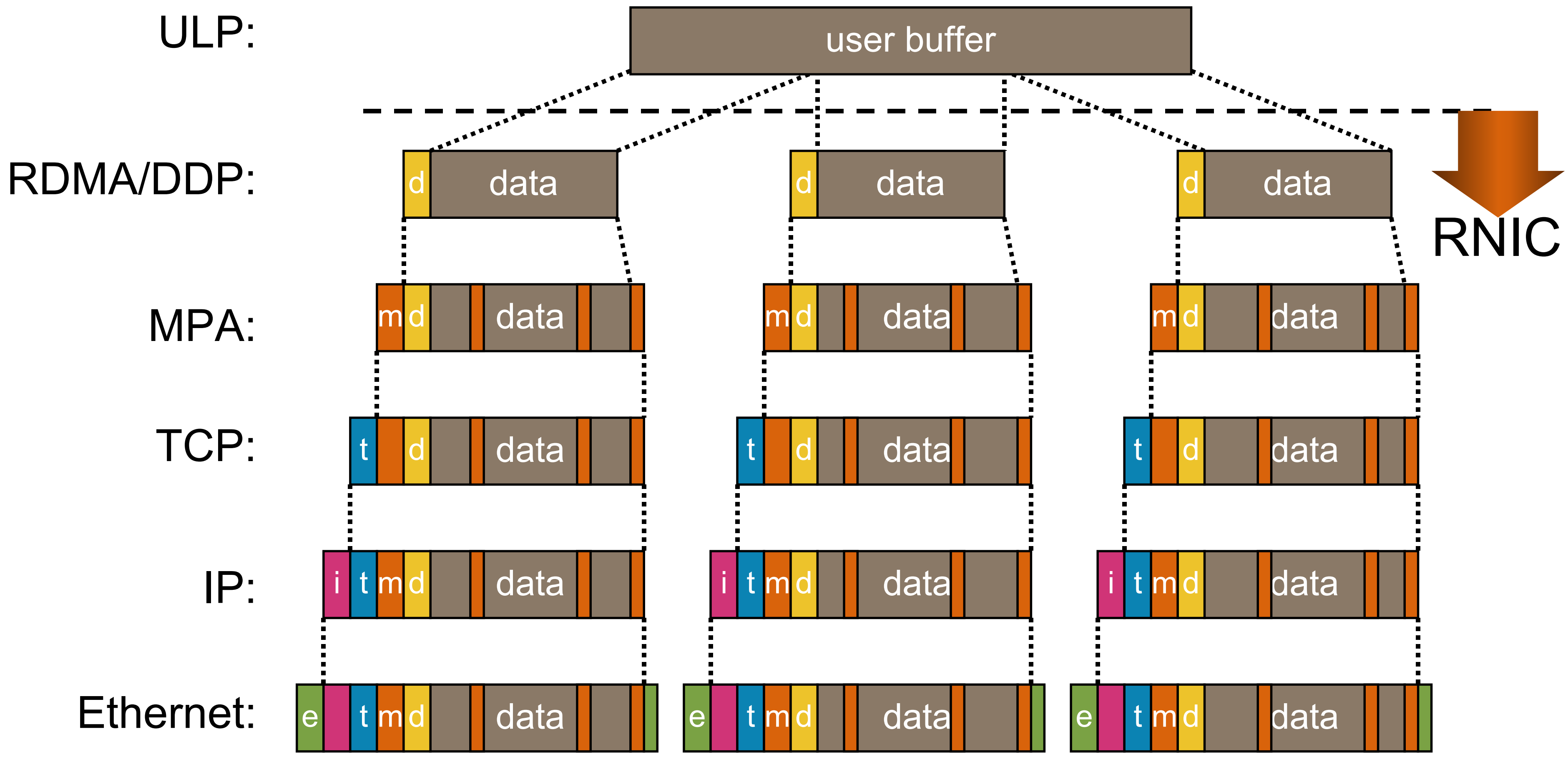
RDMA protocol stack



*Application or, e. g., iSCSI, SDP, WSD, NFS.

RDMA

Protocol wire formats



Emerging RDMA Standards

- Roots in Virtual Interface (VI) and InfiniBand
  - VI established the offload 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
  - Evolved IB Verbs to include Kernel/Storage
  - Developing SDP & iSER/DA Upper Layer Protocols
  - http://www.rdmaconsortium.org
- RDMA Consortium turning specs over to IETF
  - http://www.ietf.org/html.charters/rddp-charter.html

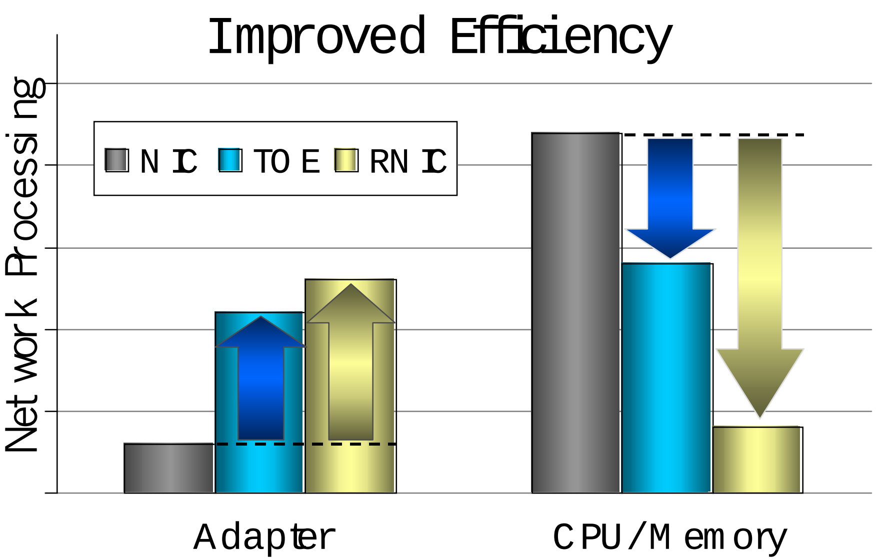
Interfacing applications to RDMA

- Sockets (existing applications)
  - Microsoft Windows WinSock Direct (WSD)
  - IETF Sockets Direct Protocol (SDP)
- RDMA-specific APIs
  - Linux/Unix:
    - The Open Group's Interconnect Software Consortium (ICSC) APIs
  - Microsoft Windows:
    - 'Named Buffer' API
    - Future OS release; described briefly at WinHEC 2003

"Big wins" for RDMA

- Accelerate sockets applications
  - User space sockets → WSD/SDP → RDMA
  - Universal 25% - 35% performance gain in Tier 2-3 application communication overhead (long lived connections)
- Parallel commercial database
  - <100us latency needed to scale real world apps
  - Requires user space messaging and RDMA
- IP based storage
  - Decades old block storage access model (iSCSI, SRP)
    - Command/RDMA Transfer/Completion
  - Convergence of NAS and SAN storage (DAFS, NFS, CIFS)

RNICs – Just Better Networking



Networking Benchmarks	BW Mbps	CPU Util %	Perf. Index
1Gb/s Enet	1000	60%	17
TOE	1000	40%	25
1Gb/s RDMA	1250	15%	74
10Gb/s RDMA	8500	15%	567

4X

30X

Note: Based on internal HP projections

RDMA enabled NICs (RNICs)

- More efficient network communications
- TOE moves TCP/IP work from the CPU
- RDMA reduces the communication work

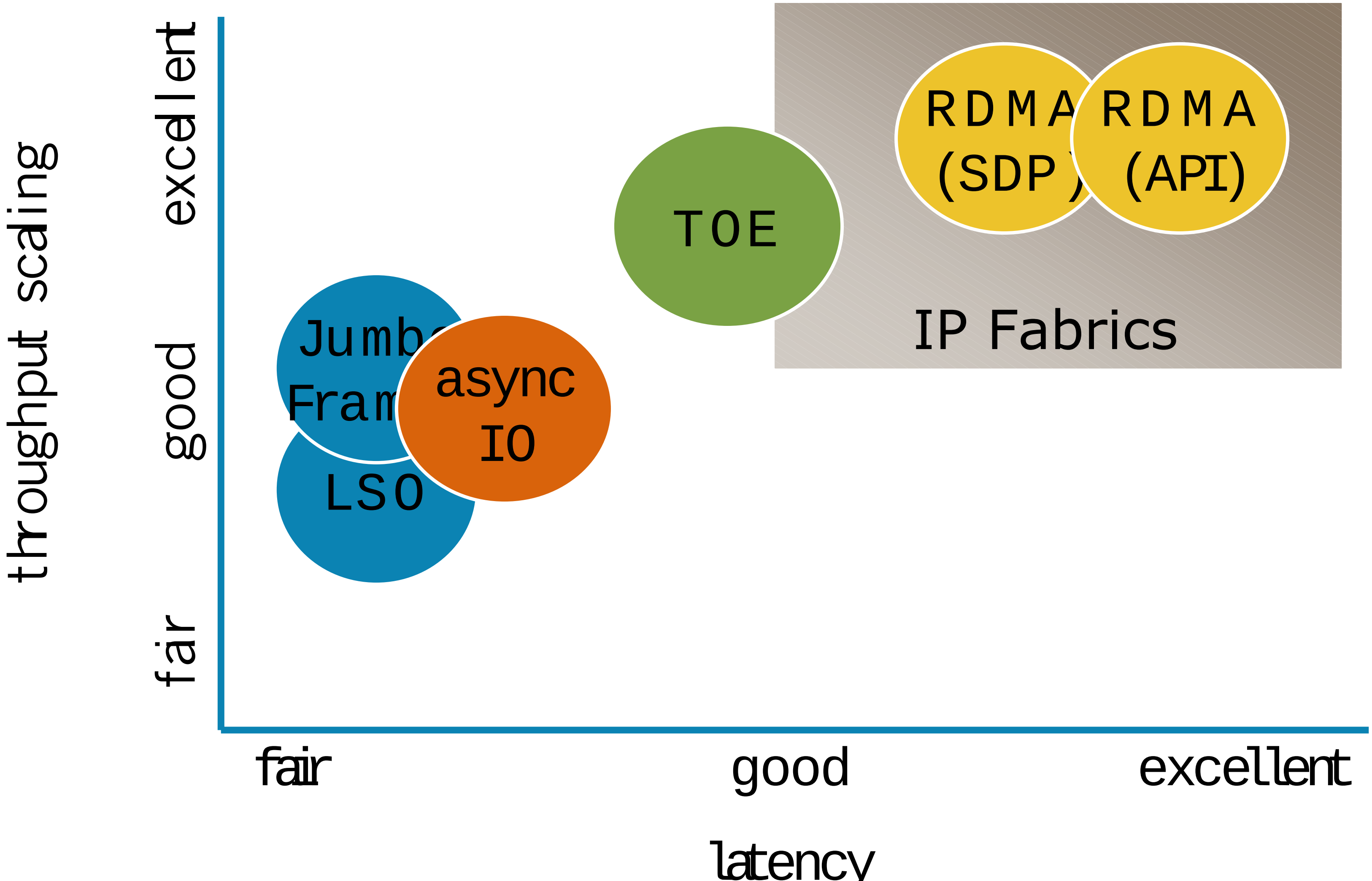
CPU/memory freed up for applications

- Zero-copy RDMA protocol conserves valuable memory bandwidth
- Much lower CPU utilization
- Per message communication overhead

Improved application performance

- Opportunity for increased application throughput or server consolidation
- Improved scalability for streaming applications or large data files

Comparing TCP/IP Networking Performance Features



Agenda

- Motivation
- Challenges for IP/Ethernet
- Improving Throughput Scaling
- Remote DMA
- **Networked Storage**
- Conclusions

Storage Fabric Directions

Fibre Channel

Proven storage fabric choice for data centers and backbones

- Faster speeds and resource virtualization under development
- New SAN disaster recovery over WAN option with bridged FC over IP

Serial Attached SCSI (SAS) / Serial ATA (SATA)

Drive interface technology migration - parallel to a common serial interconnect

- Chassis and controllers can accommodate both types of drives
- Drives remain differentiated by performance, reliability, cost per gigabyte
  - SAS (SCSI) remains highest performance, reliability
  - SATA (ATA) great bulk storage for online archival

iSCSI (storage over IP)

Unified network and storage infrastructure possible

- Geographic flexibility - Broader access to FC SAN via iSCSI proxy
- Consolidate file & block storage access with one Ethernet wire

iSCSI – Block storage

- Proposed Standard published Jan 2003.
- iSCSI initiators (Host)
  - Software-based iSCSI initiators provide connectivity at lowest host cost
  - Windows, HP-UX and Linux support
  - Multi-purpose NICs will integrate iSCSI functionality with other host IP functions (TOE)
- iSCSI targets (Storage)
  - Variety of SW/HW implementations possible
  - iSCSI to FC bridges available today
  - Native iSCSI targets will emerge as TOE technology matures

NAS – File Storage

- NAS just means File Oriented IO Services (instead of block)
- There are many standard wire protocols:
  - CIFS (SMB), NFS, NCP, Appletalk, HTTP, FTP.
  - Just wire protocols, so they operate over Ethernet as well.

- HP has a full product line:

  From the NAS b2000,

  To b3000, Up to

  NAS e7000 & 8000

Conclusion:

- For Storage, there are no technical barriers preventing the development & deployment of IP-based block and file oriented storage.
- iSCSI is the emerging block storage standard

Agenda

- Motivation
- Challenges for IP/Ethernet
- Improving Throughput Scaling
- Remote DMA
- Networked storage
- **Conclusions**

Is IP/Ethernet our single fabric?

- Strengths
  - Ubiquitous; standard
  - Extends beyond the data center
  - Minimal training costs
  - Understood management model
  - Affordable adapter, cabling, switches
  - Mature foundation
- Weaknesses
  - Scalability

    solution: **Jumbo frames, Async IO, TOE, and RDMA**
  - Latency

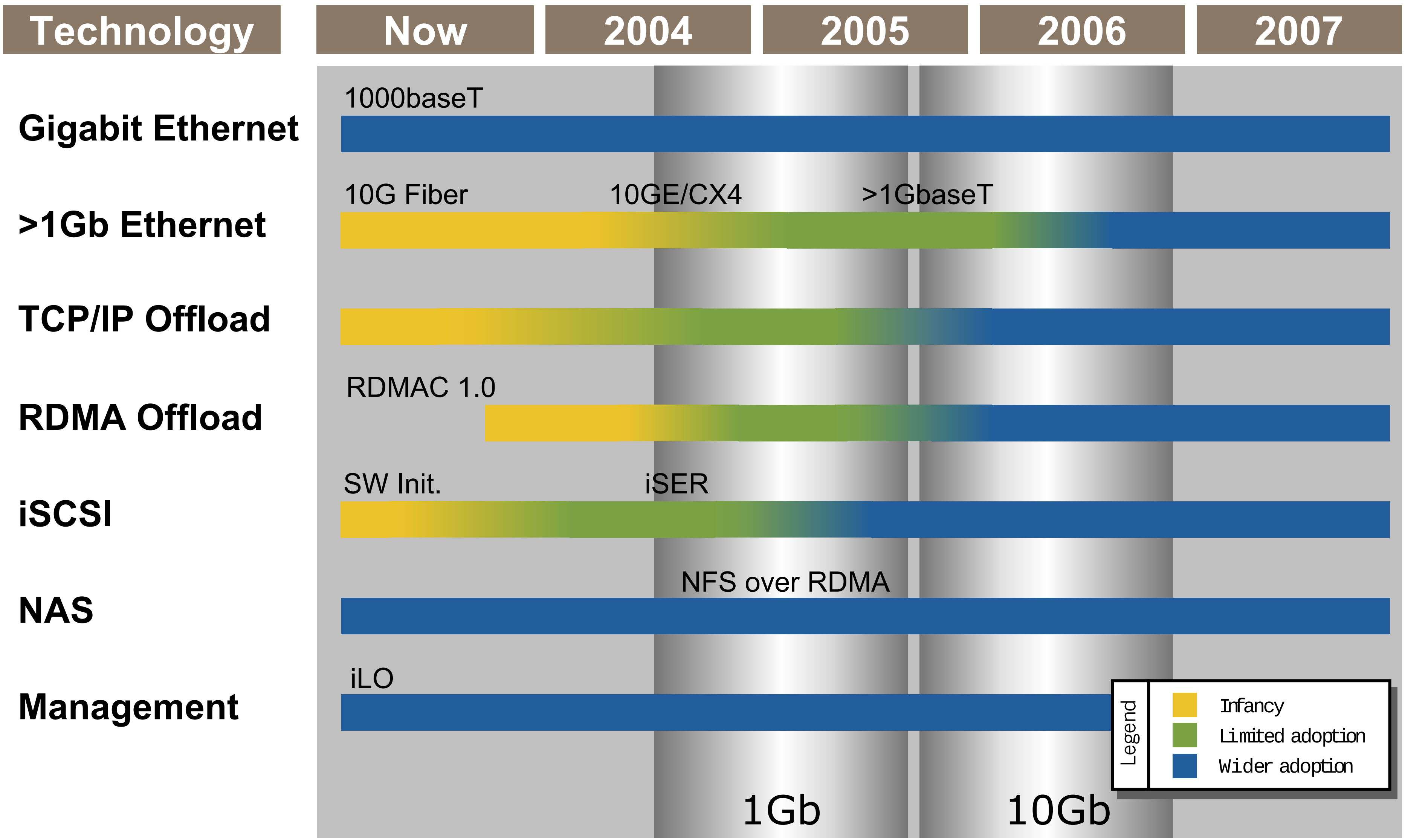
    solution: **TOE and RDMA**

When will IP fabrics emerge?

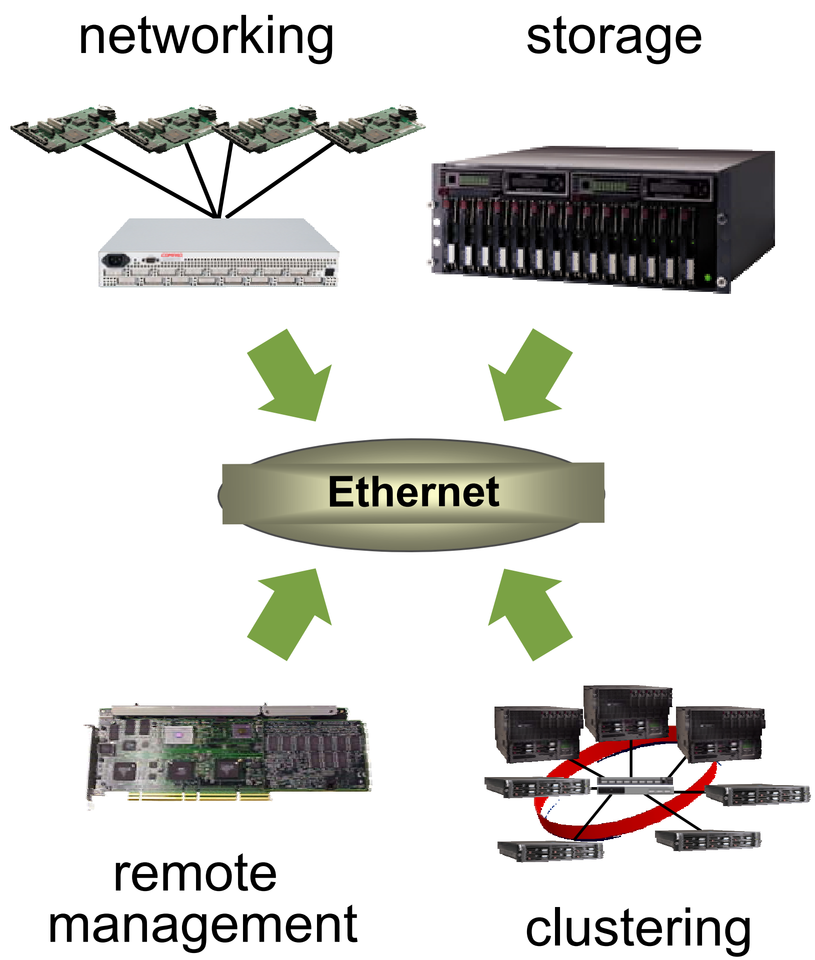
- Legend:

(yellow)	Infancy: limited suppliers, premium pricing
(green)	Limited adoption: >1 supplier, moderate pricing
(blue)	Wider adoption: affordable, integrated

When will IP fabrics emerge?



IP Fabrics:
a simpler, unified infrastructure



Converges functions

- Multiple functions (SAN, LAN, IPC, Mgmt.) can be consolidated to a single fabric type.
- Blade server storage connectivity (low cost)
- Packaged “end-to-end” Ethernet solutions

Consolidates ports

- Leverage Ethernet pervasiveness, knowledge, cost leadership and volume
- Consolidate KVM over IP and reduce switch port costs

Ethernet Everywhere

- Bridge storage & network “islands”
- Extend geographic reach globally
- Centralized management

HP WORLD 2003

Solutions and Technology Conference & Expo

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

interex
shared knowledge • shared power

encompass
AN HP USER GROUP

hp
invent