Aspects of the InfiniBand™ Architecture

Gregory Pfister
IBM Server Technology & Architecture, Austin, TX

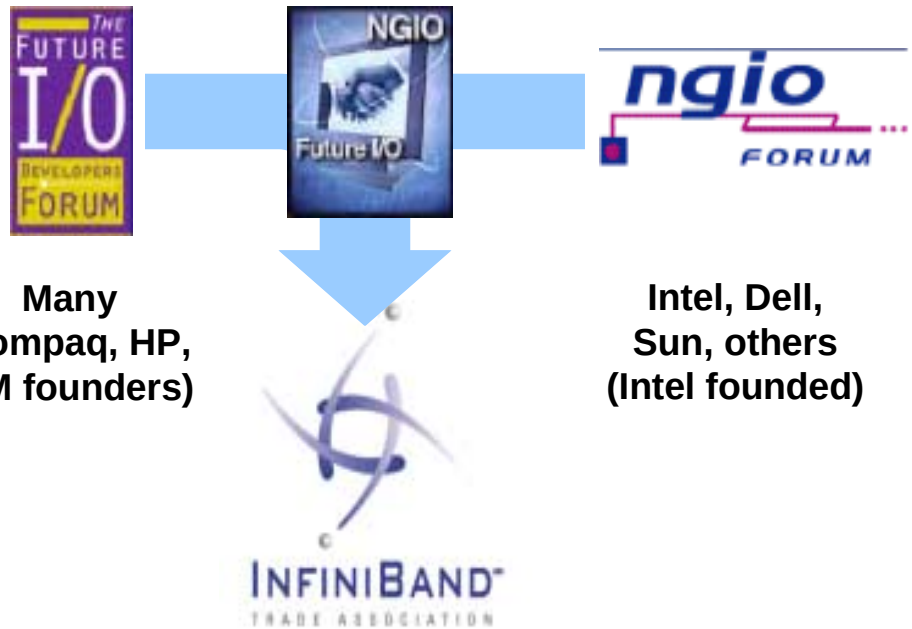
Legalities...

- InfiniBand is a trademark and service mark of the InfiniBand Trade Association.
- All other product names mentioned herein may be trademarks or registered trademarks of other manufacturers. We respectfully acknowledge any such that have not been included above.
- **None of the opinions expressed here necessarily reflect the position of the IBM Corporation.**

Agenda

- Where did it come form?
- What is it?
  - Overview
  - Selected sub-topics
- When?
- Conclusions

InfiniBand Trade Association℠: A Merger, 9/99



Many (Compaq, HP, IBM founders)

Intel, Dell, Sun, others (Intel founded)

Steering Committee

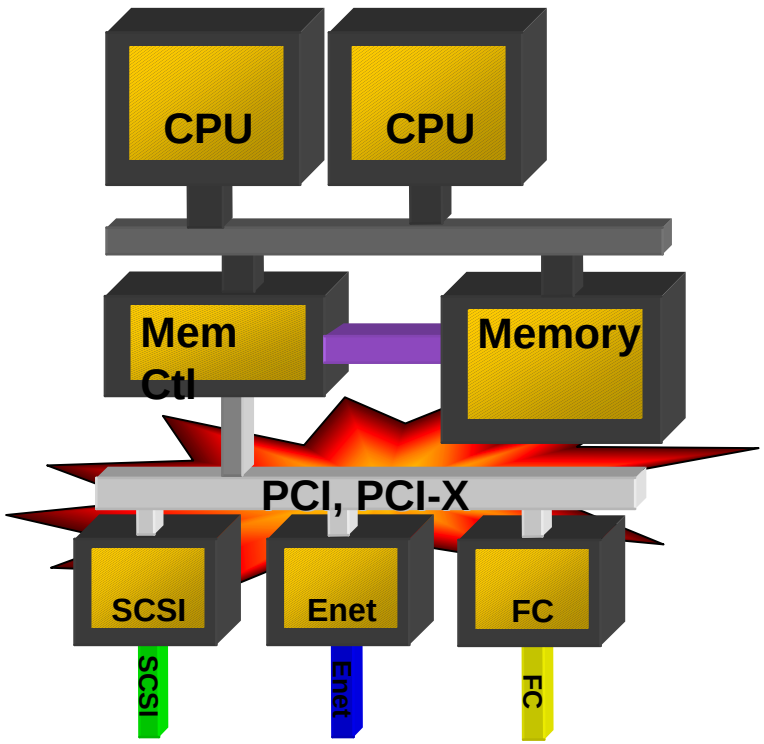
COMPAQ · HEWLETT PACKARD · DELL · IBM · intel · Microsoft · Sun microsystems

Sponsor Member Companies

3Com · adaptec · Cisco Systems · Fujitsu Siemens · HITACHI · NEC · Lucent Technologies · Nortel Networks

- 220+ companies
- Right Ts&Cs for wide adoption: “fair & non-discriminatory” licensing
- Like PCI SIG:
  - Anyone can join with member or associate status.
- Spec 1.0 published 11/23/00; 1.0.a errata in 6/00. See http://www.infinibandta.org - free download

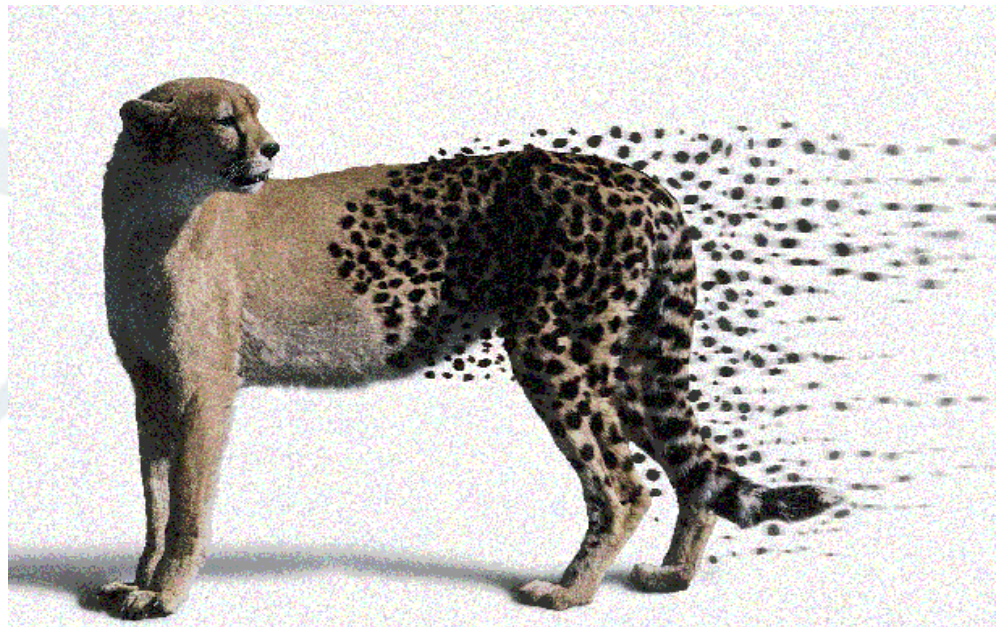
Why IBTA? I/O Busses Simple, Useful, but Running Out



- Widespread realization:
  - Standard I/O busses can't keep up with processors & communication.
- Bus frequency just 2X / 3 years
- Arbitration limits real bandwidth
- Load/store model limits throughput
  - e.g., loads can't pass stores in most cases
  - particularly hurts at distances required for scaling
- Single fault domain

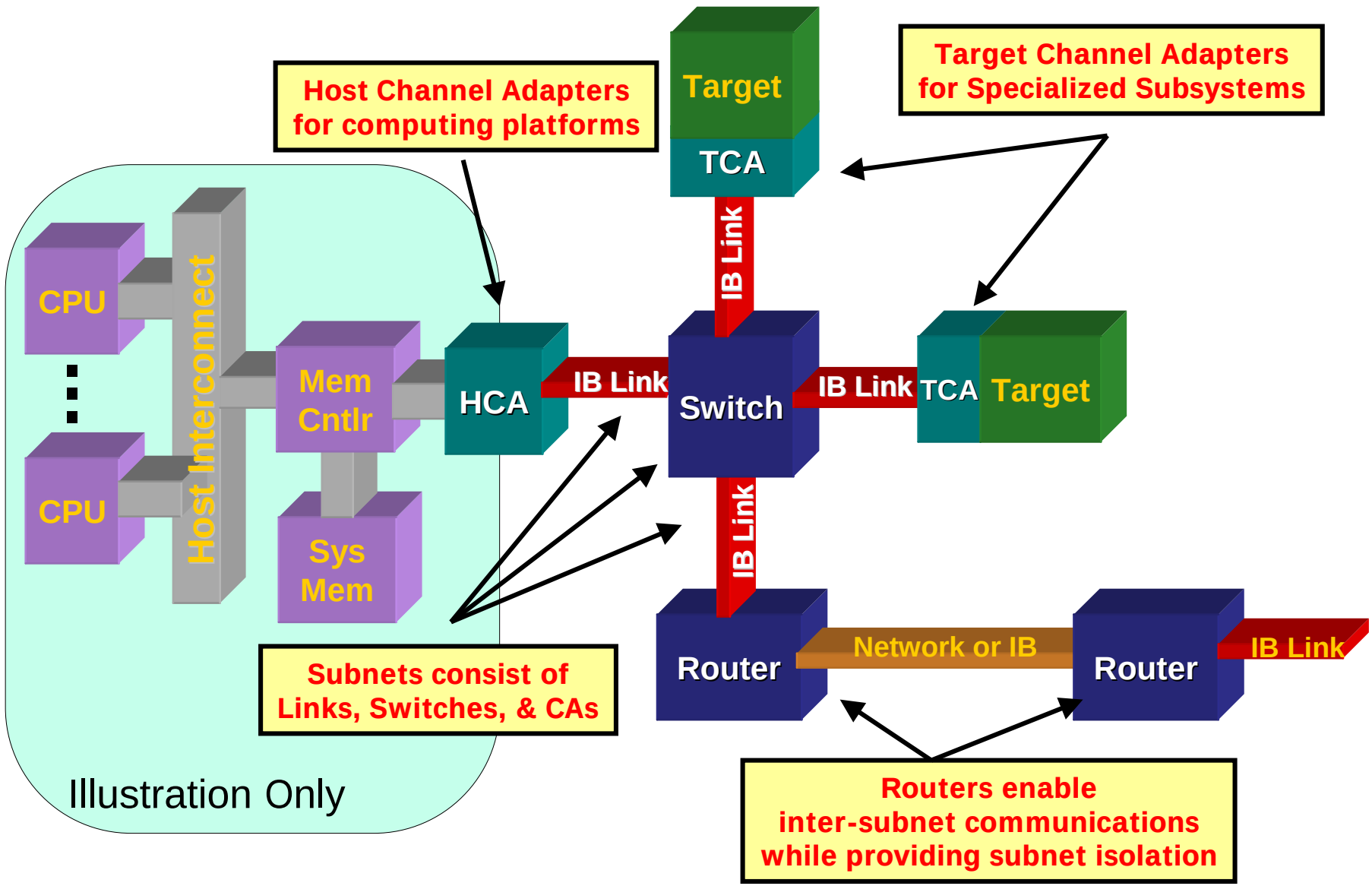
Agenda

- Where did it come form?
- What is it?
  - **Overview**
  - Selected sub-topics
- When?
- Conclusions

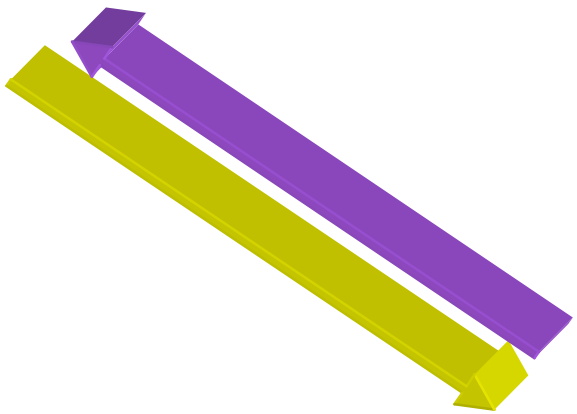


Random gratuitous clipart

IBA Elements



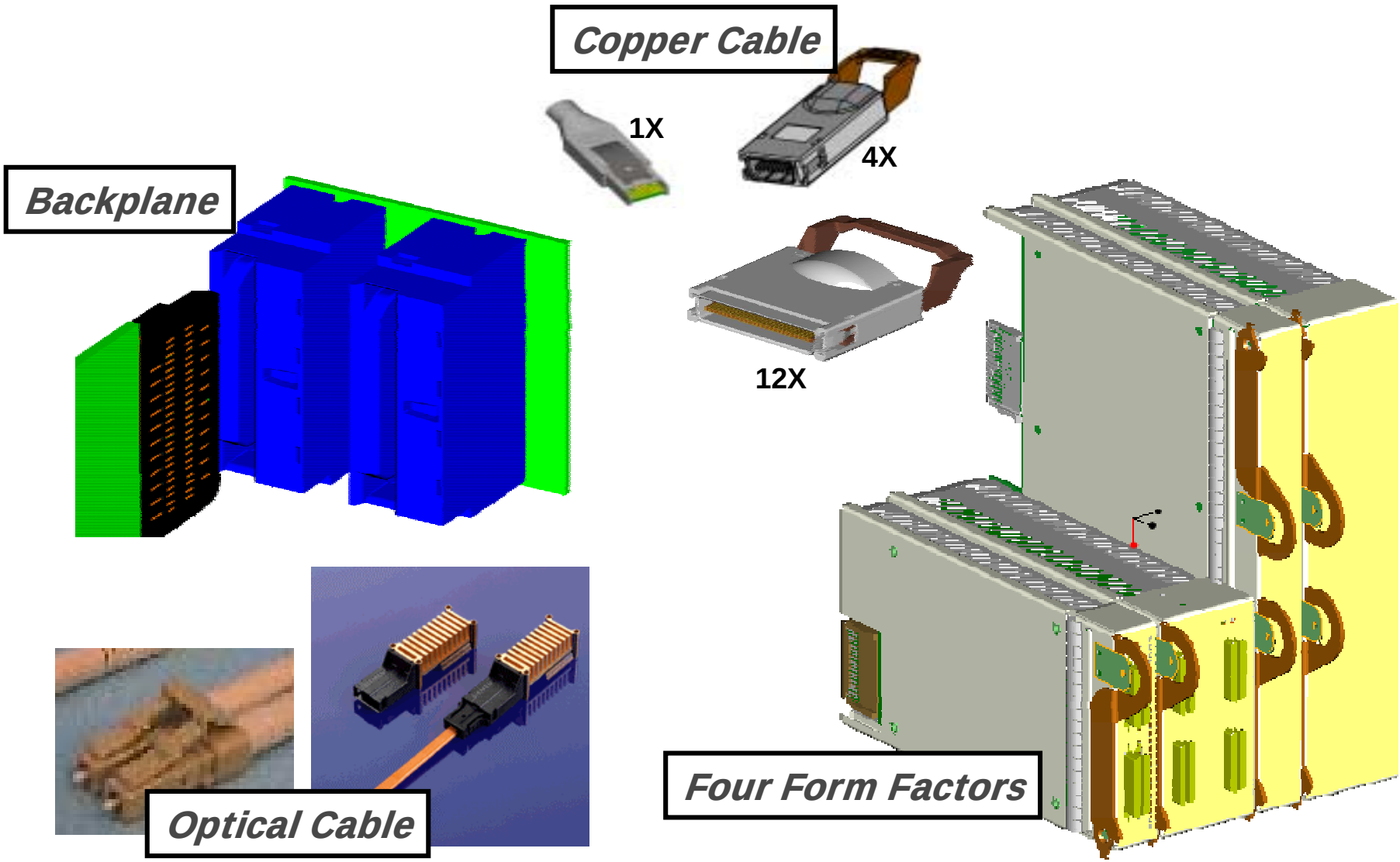
The Link



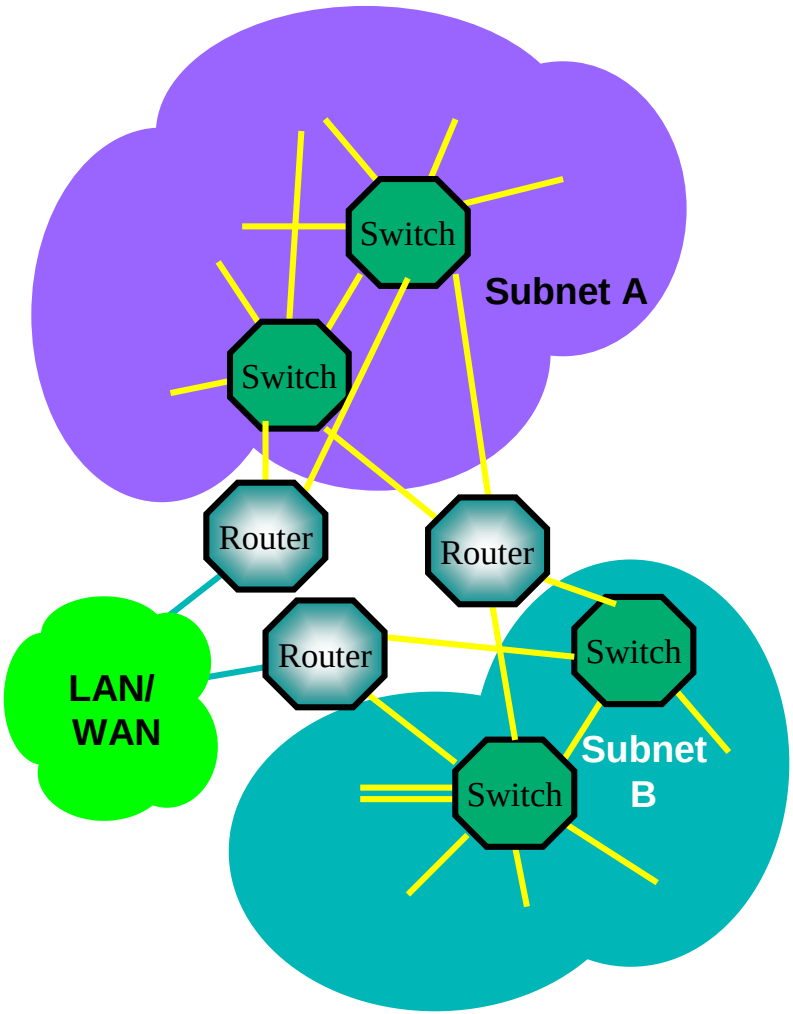
Width	Bi-directional Bandwidth
1	500 MB/s
4	2 GB/s
12	6 GB/s

- Bidirectional, 4 wires (copper)
  - Parallel links for 4X, 12X widths
- 2.5 Gbaud signal rate
- No length spec
  - attenuation budget: 15dB
- Multimode and single mode fibre
  - single only 1X, but goes 10Km
- Hot plug, of course
- Training sequence and credit exchange when connected.
- MTU 256Bytes to 4KBytes

Electromechanical

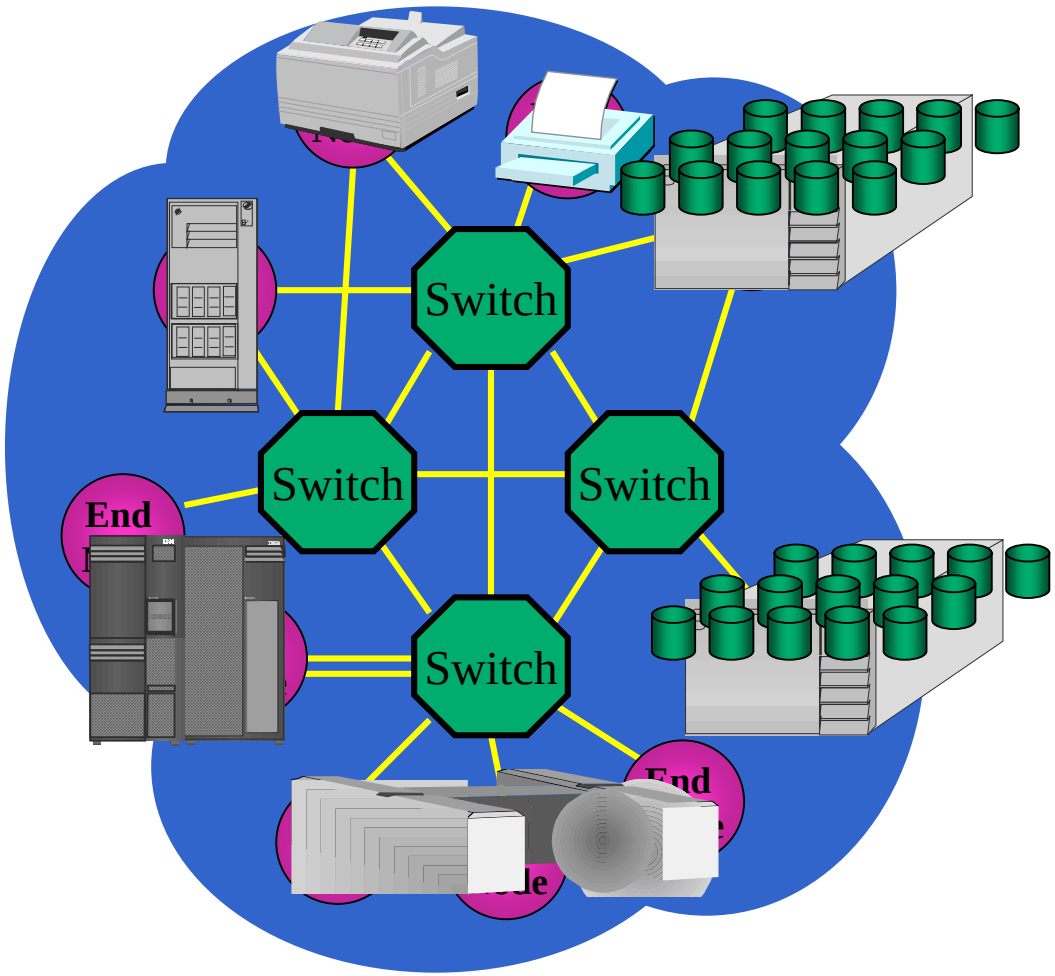


Switches and Routers



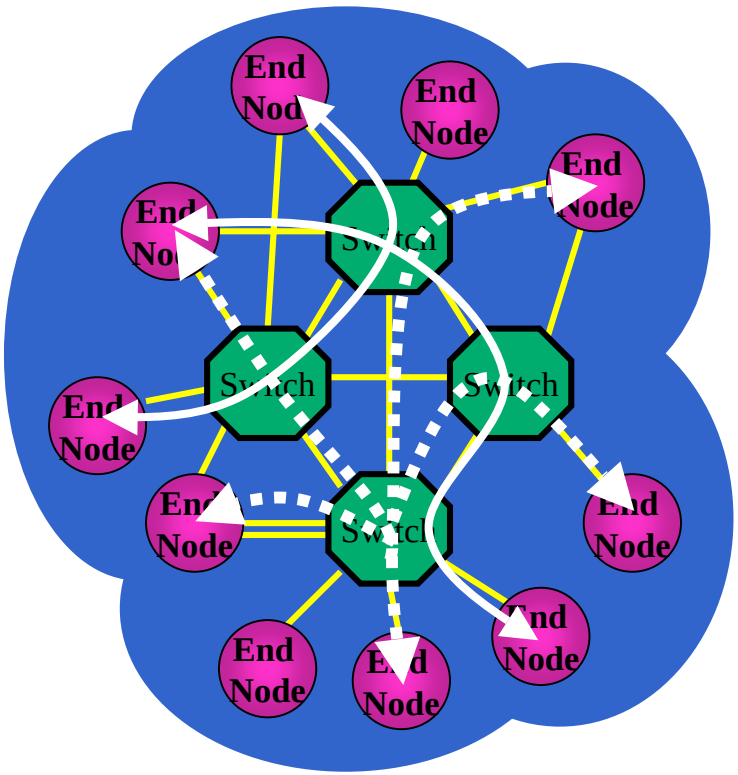
- Switch: routes packets within subnet.
  - Destination routed, based on LID
    - Direct routing for initialization
  - Up to 48K unicast LIDs per subnet.
  - SLs provide service differentiation.
  - Multicast (optional)
  - Switch size, network topology are vendor-specific
- Router: routes packets between subnets
  - Based on GID (128 bit IPv6 Address)
  - Can transfer through disparate fabrics

Endnodes



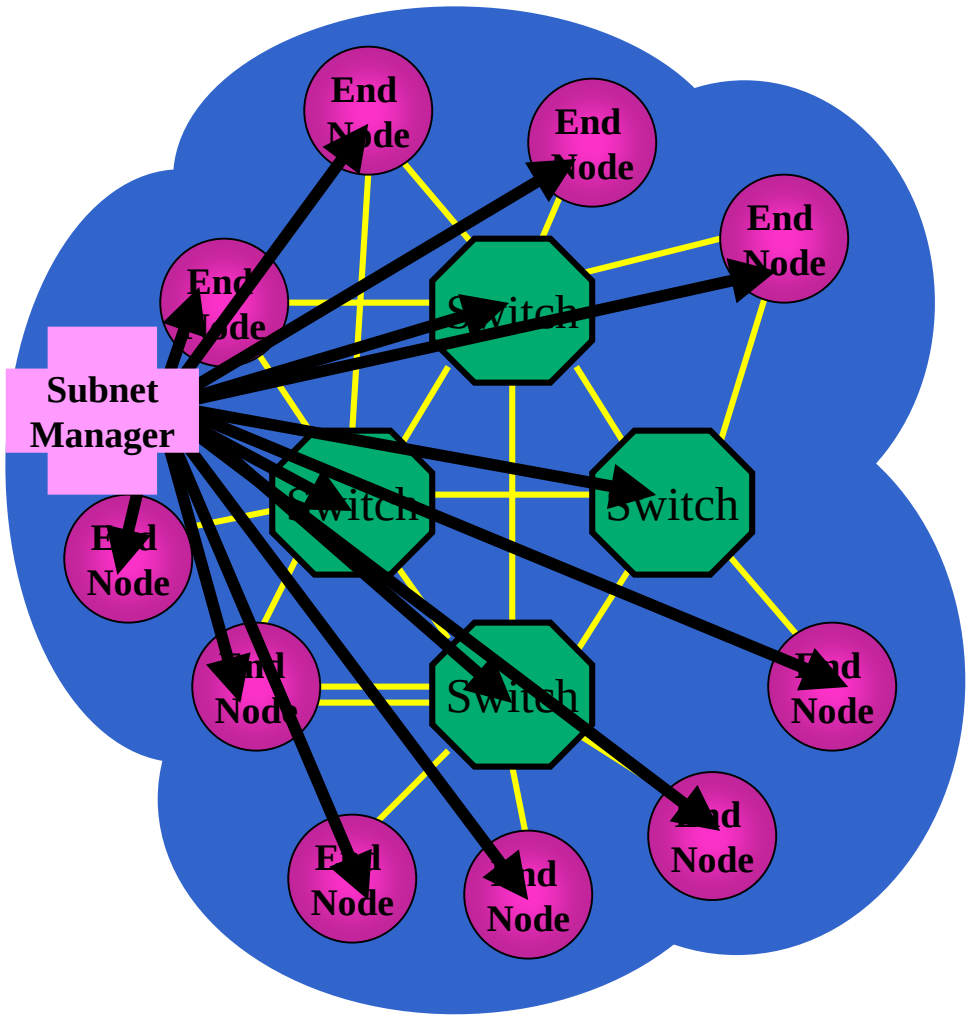
- Hosts
  - processors, memory
- Devices
  - Storage, network adapters, etc.
- Bridges
  - to “legacy” I/O busses: PCI, etc.; vendor unique; not part of spec
- Channel Adapters attach endnodes to links
  - Only HCA/TCA difference: TCA has no defined software interface.

Channel Adapters



- Attach nodes to links: data engines
- Service types:
  - Reliable Connection, (Unreliable) Datagram, Unreliable Connection, Reliable Datagram (optional)
- Very low software overhead
  - reliable = in-order, correct, receipt acknowledged - ***provided by hardware***
  - ***zero-copy*** data transfer operations
  - ***in user mode***; no switch to OS
- Low-overhead byte-gran mem protection
- Remote DMA on reliable services
  - user-mode virtual addresses; memory windows
- Optional: atomic operations (inter-node); (Unreliable) Multicast

Subnet Management



- Each subnet has a master subnet manager
  - resides on endnode or switch
- Discovers & initializes network
  - assigns LIDs, determines MTUs, loads switch routing tables
- Provides path information
  - what devices can I access?
  - what path(s) to a device?
- Scans/traps for hot plug/unplug
- Multiple SMs for HA failover
- Other managers: Baseboard, Performance, Device, etc.

Topics Not Covered

InfiniBand spec is over 1500 pages long.

Some topics not covered here:

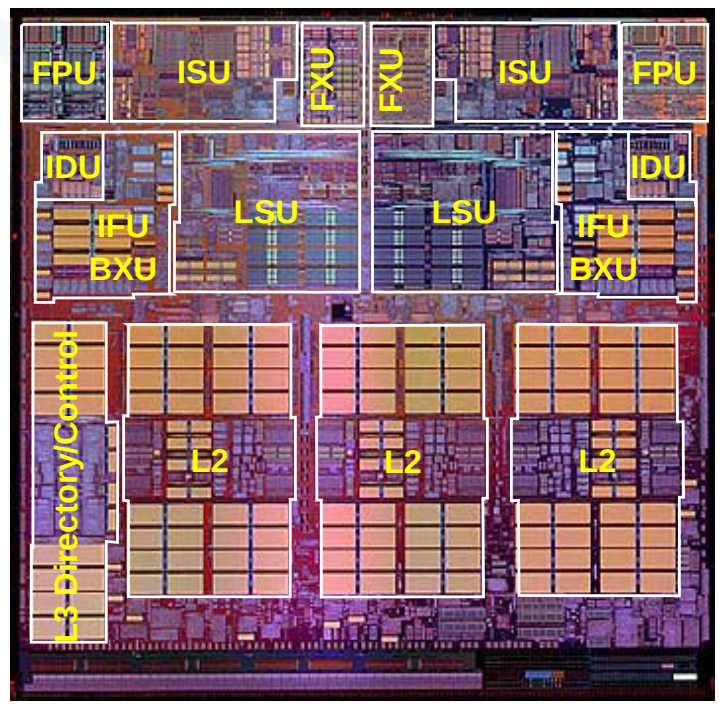
- Compliance and interoperability
- Automatic Path Migration
- Verbs (no API)
- Subnet Management
- Initialization
- Performance monitoring
- Packet formats
- Addressing – relation to EUI64 and IPv6
- Electronic/Mechanical issues
- Operation of virtual lanes

Agenda

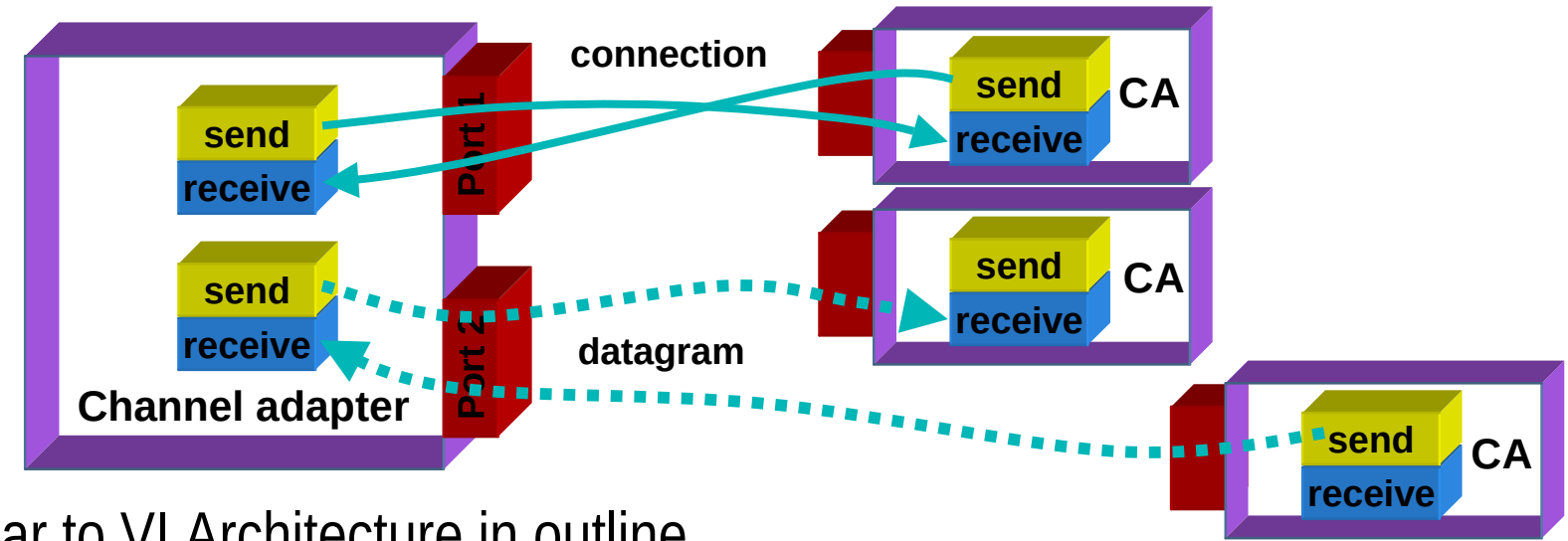
- Where did it come form?
- What is it?
  - Overview
  - **Selected sub-topics**
    - **Queues**
    - **Partitioning**
    - **Reliable datagram**
    - **Sockets over IB**
- When?
- Conclusions

More random gratuitous clipart

POWER4



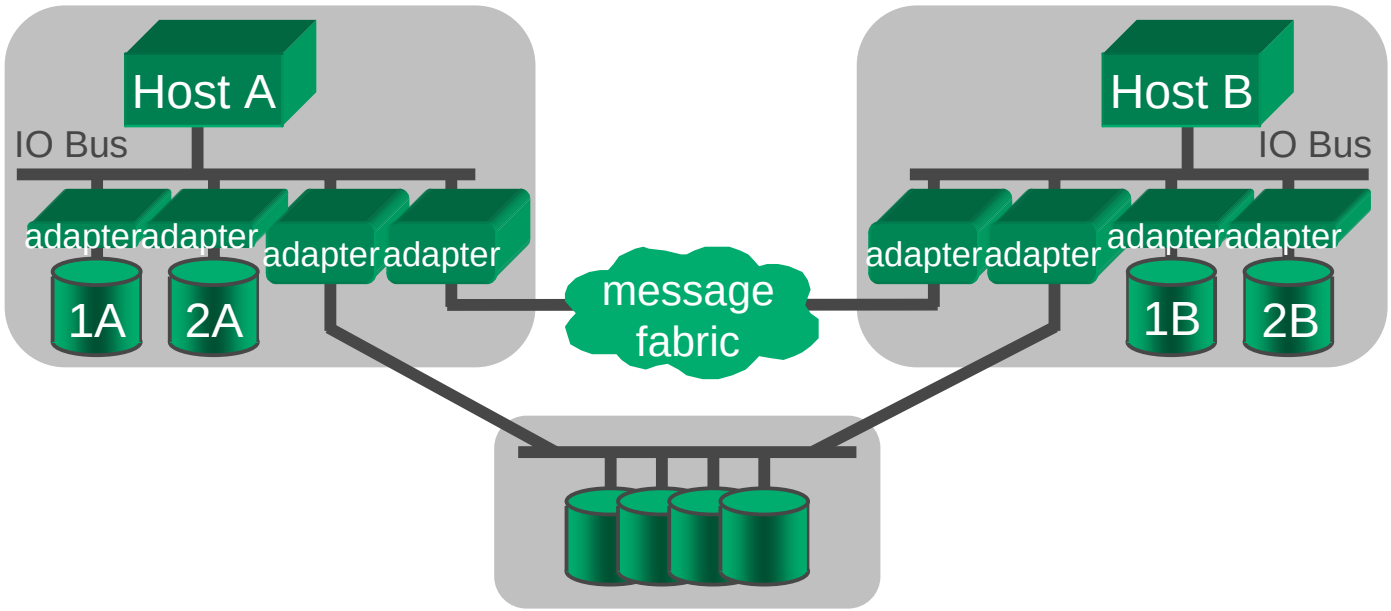
Queue Pairs



- Similar to VI Architecture in outline.
- Queue Pairs (QPs) are the means by which messages are sent and received. They specify, among other things:
  - service type (reliable connection, unreliable datagram, etc.)
  - CA port (network connection)
  - max queue depth
- All communication ultimately targets a QP in a CA.
- QPs are created as needed by consumer using verbs: 2**24 QPs max/CA

Partitioning

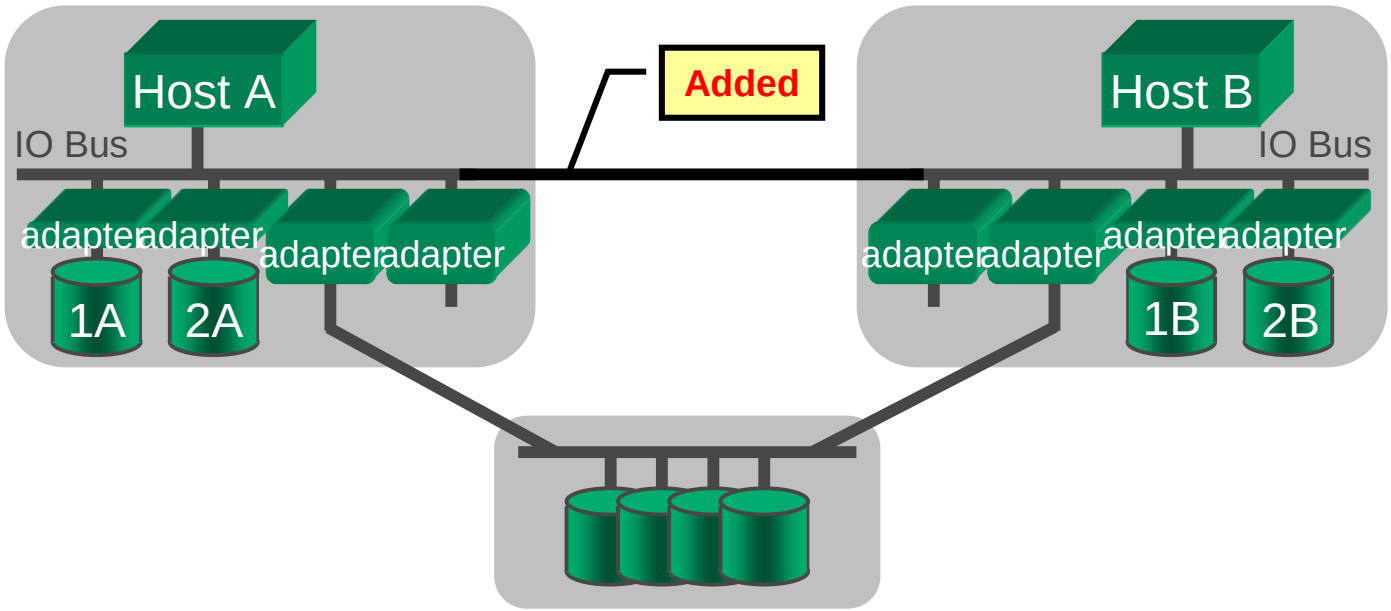
Clusters Today



- Each node has its own local devices, and adapters specifically dedicated to shared devices
- Shared devices have known special semantics: Inter-OS locking, etc.

Partitioning

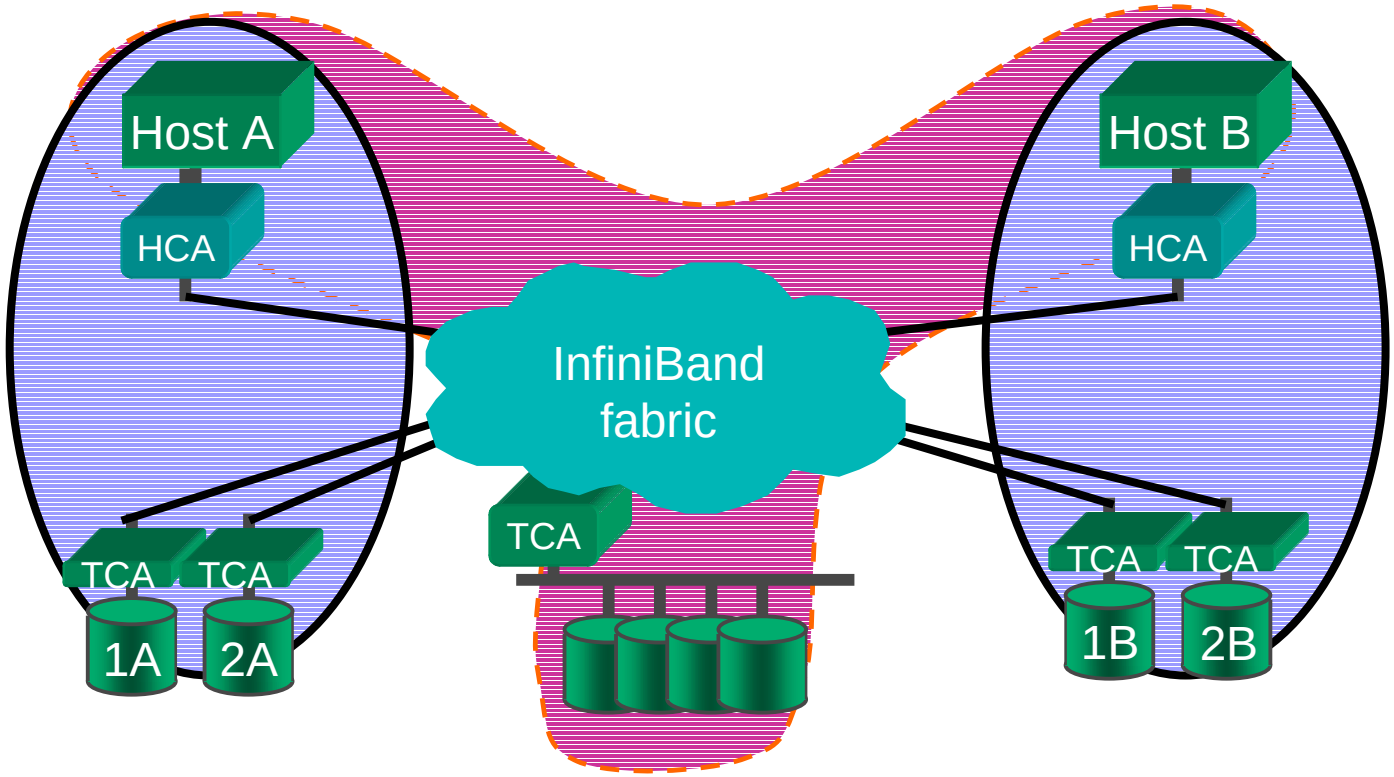
What If...



- On boot, each host scans entire bus to find all devices, make them their own:
  - OS of A writes on 1A, 2A, 1B, 2B
  - OS of B writes on 1A, 2A, 1B, 2B
- Chaos. Probably will not even boot.

Partitioning

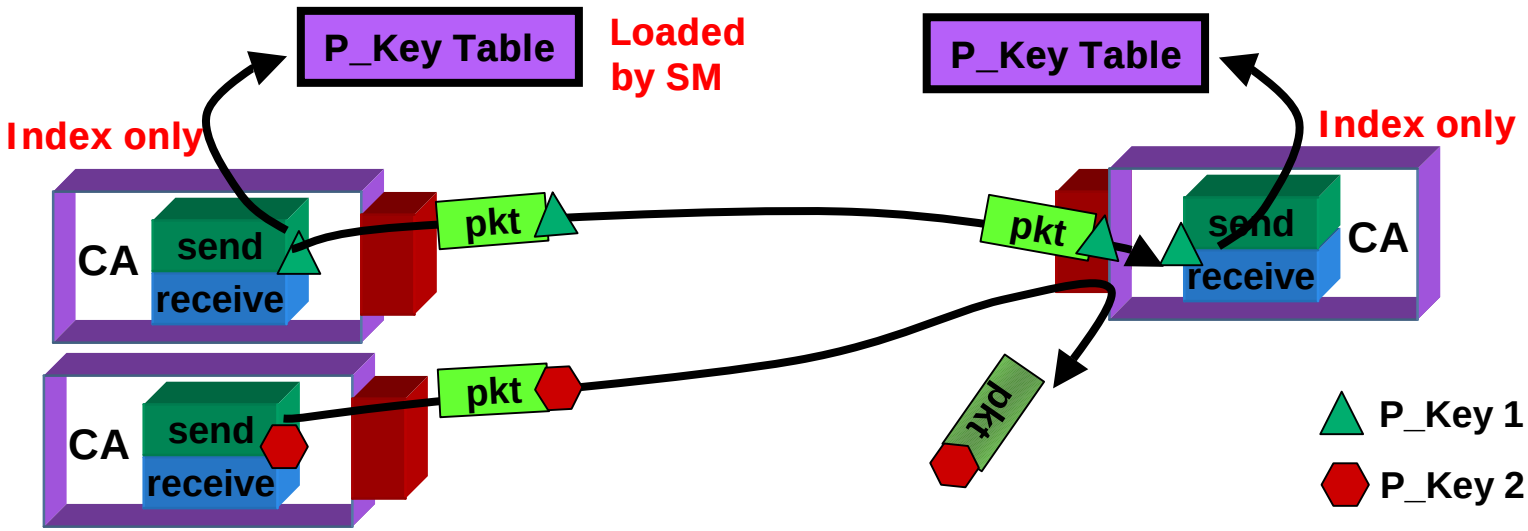
That's what IBA Is.



- Must separate private devices and provide controlled sharing.
  - That's the purpose of partitioning. It's not optional.
- Full/Partial not shown: Allows many to reach "server" without being aware of each other

Partitioning

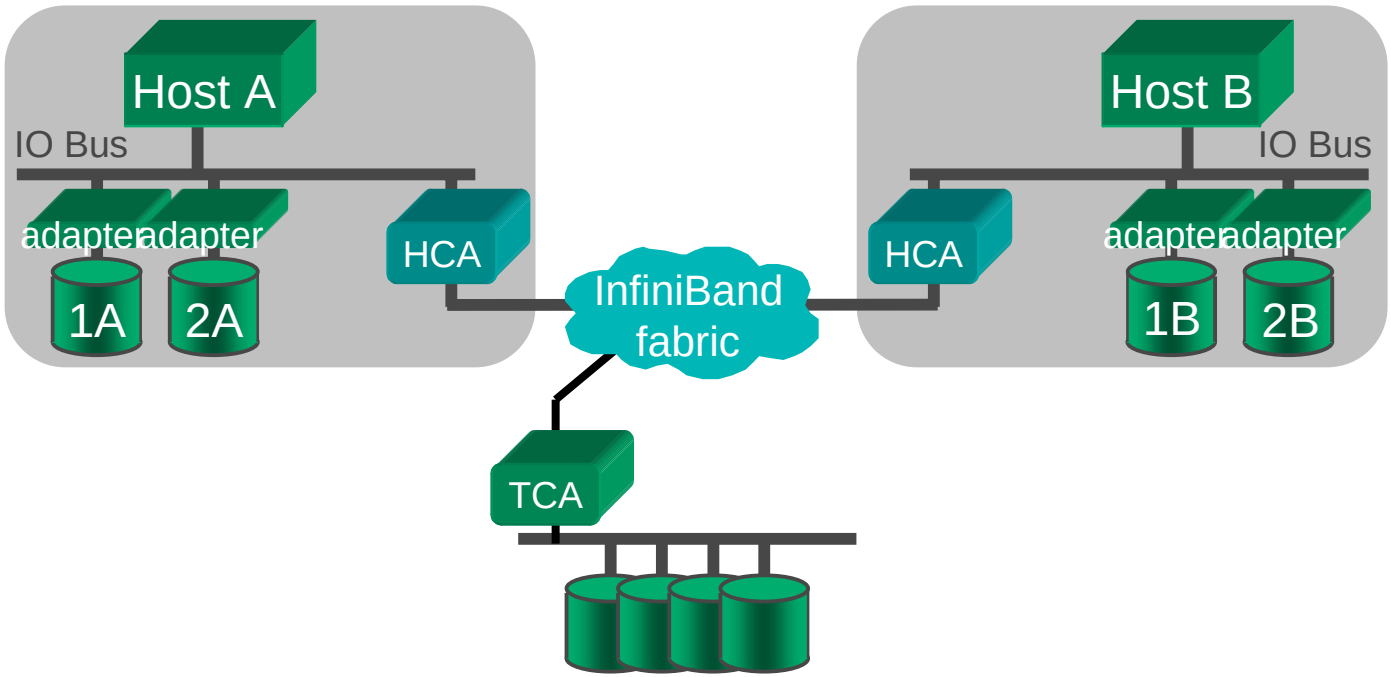
Implementation



- Packets have a Partition Key (P_Key) attached by Channel Adapter (CA)
- Matched on receive: if no match, **silently dropped** - no NAK returned
  - same semantics as attempt to contact nonexistent endnode: don't know it exists
  - notice given to subnet manager
  - can also be enforced in switches (optional) on inbound or outbound sides
- Verbs can only specify index into P_Key Table in each HCA; table content set only by Subnet Manager
  - 64-bit M_Key used to authenticate message from SM; P_Key only 16 bits

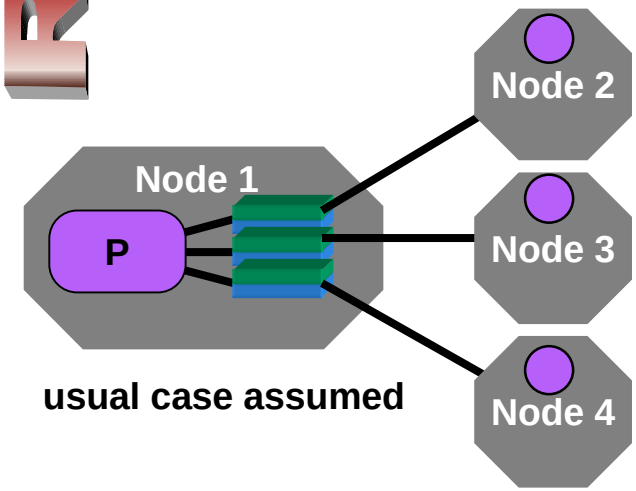
Partitioning

Possible Confusion in Initial IBA Implementations

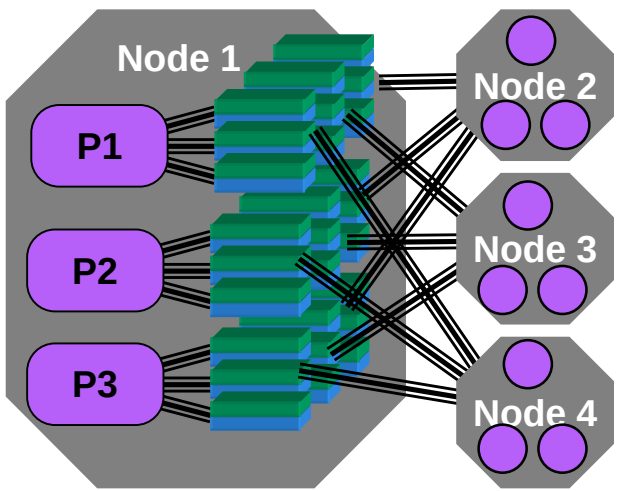


- Provides for bringup, software development, perhaps deployment
- Effectively uses the physical partitioning of prior cluster systems

Reliable Datagram: The Problem



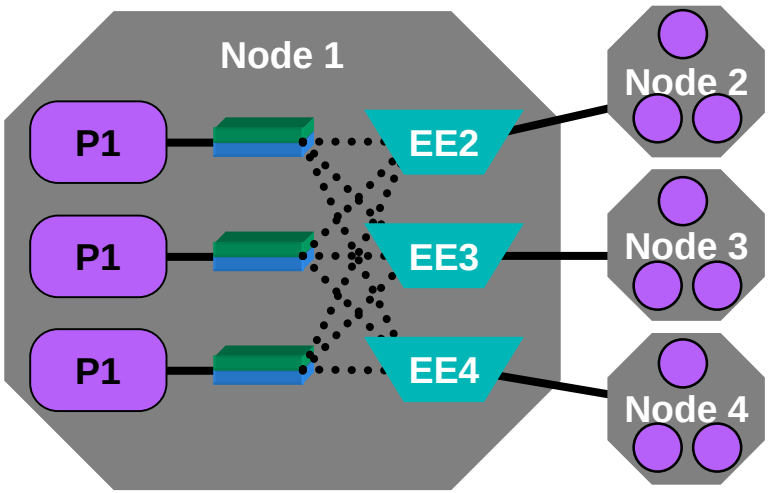
usual case assumed



SMP endnode Reality: Does not scale.

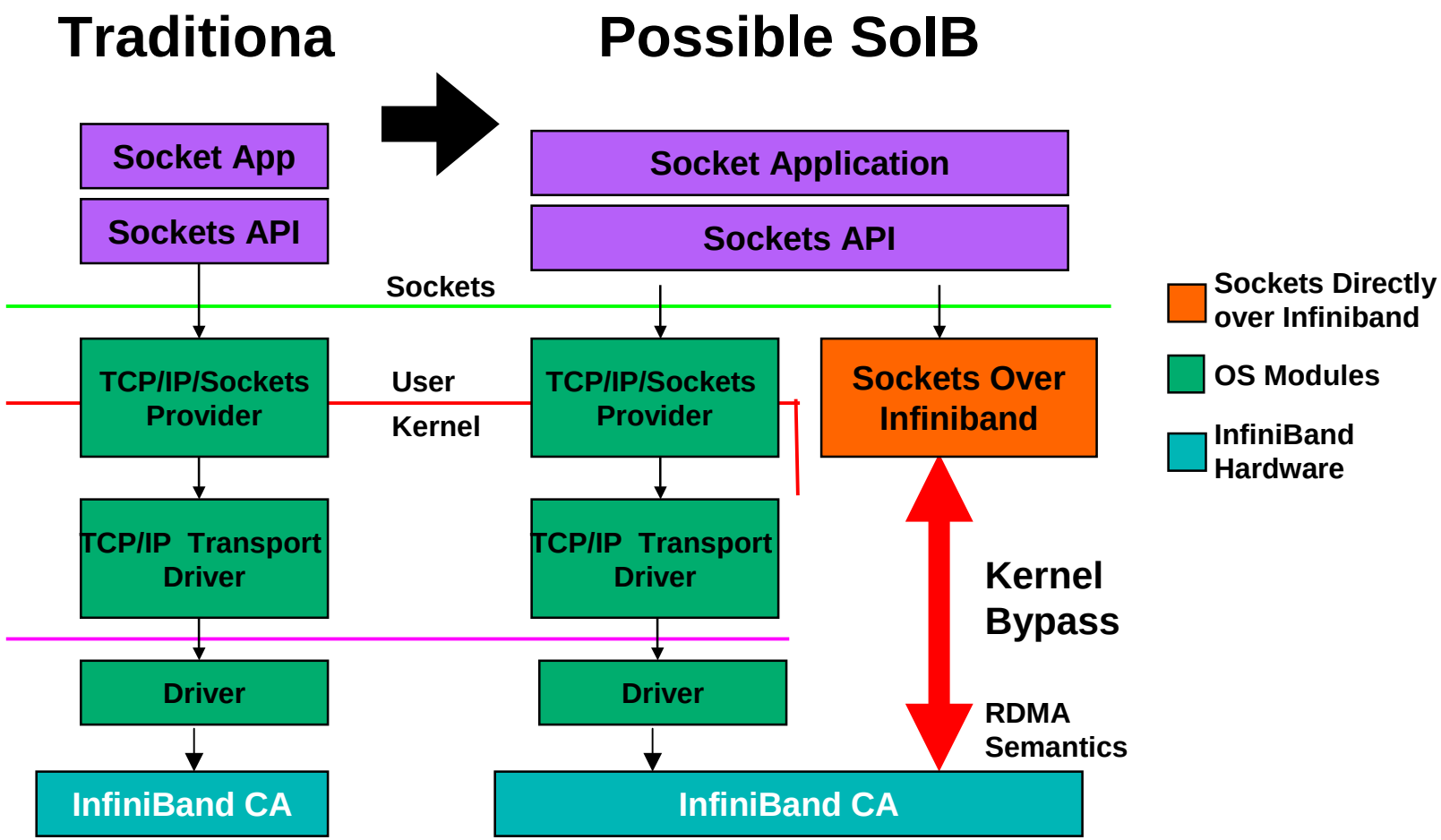
- Assume: parallel program, needs reliable commo to each of N nodes
  - e.g., distributed database, parallel techical.
  - N-1 QPs on each node, each RC
- But w/SMPs, really need commo among all processes on all nodes - P processes/node
  - $\Rightarrow$ (N-1)×P QPs per process
  - $\Rightarrow$ (N-1)×P^2 QPs per node
  - database: 16 processors/node, P = 1000s
  - 4 nodes, P=1000: 4,000,000 QPs
- Alternative1: mux workers w/commo process
  - software overhead sending and receiving
- Alternative 2: Use Unreliable Datagram
  - more software overhead to attain reliability

Solution



- Real problem: QP holds reliability context.
  - sequence #, retry count, etc
- Separate the reliability context from queue pair as end-to-end context.
  - 1 EE context / endnode (typical)
  - One on each side of communication
  - Set up prior to use
- RD work requests specify
  - EE context to use (hence target node)
  - target QP on that node
- Reliable, and Scales
  - P QPs + (N-1) EEs on each node
- Note – EE setup is a kind of per-node connection setup; might better be called “multiconnected” instead of RD.

Sockets over IB (SoIB): The Intent

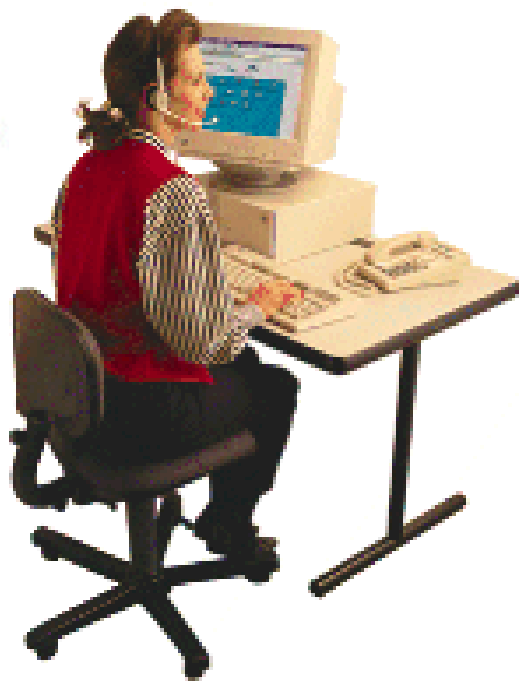


Sockets Over IB (SoIB)

- Full specification not yet published; early 1Q02 (draft form now)
- “On the wire” packet format and protocol only
  - Interoperability, but no implementation specification
- Characteristics:
  - Complete SOCK_STREAM semantics with TCP error semantics
    - Including, e.g.: graceful & abortive close, out of band data, socket duplication, socket options
  - Full protocol offload, including reliability
  - Allow no/minimal data copying
    - Can switch between RDMA and SEND based on length of transfer
  - Kernel bypass, interrupt avoidance
  - Implementable using just IBA 1.0 required features, but can take advantage of options and later optional additions.

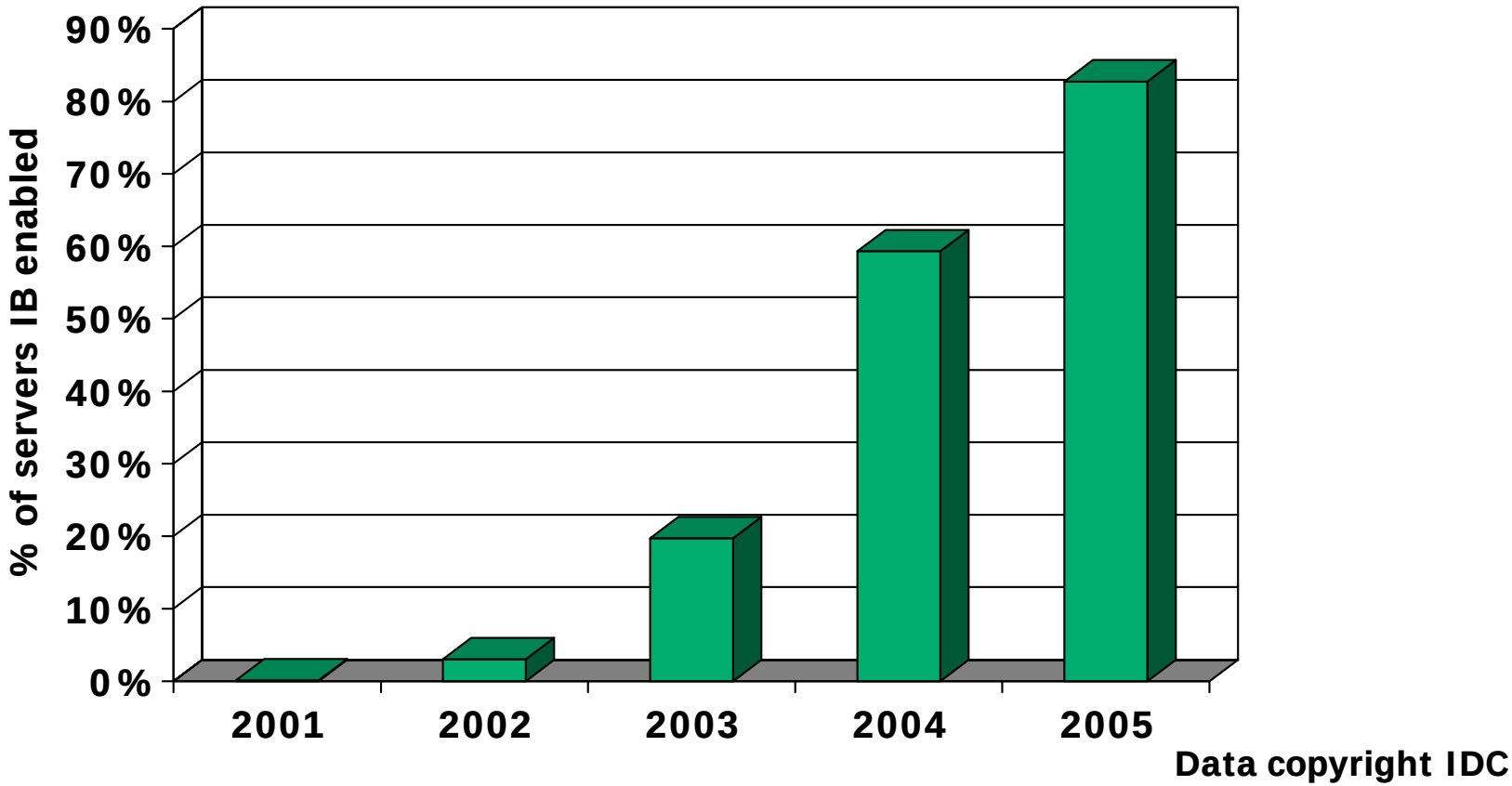
Agenda

- Where did it come form?
- What is it?
  - Overview
  - Selected sub-topics
- **When?**
- Conclusions



Totally random gratuitous clipart

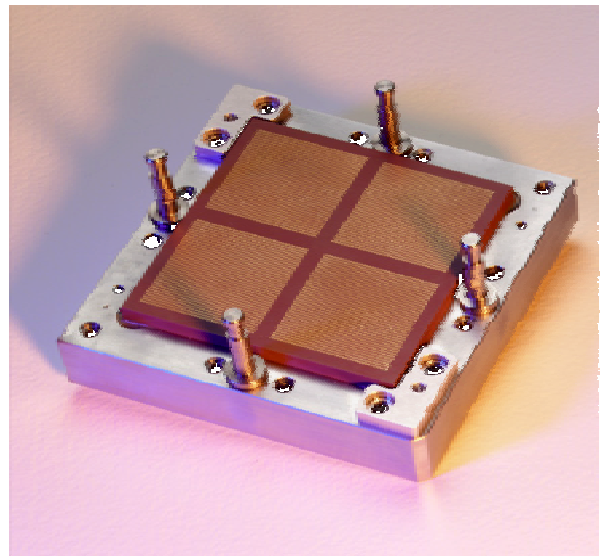
IDC Forecast, May 01



- Server support first as PCI cards, then IO hub chips, then integrated with memory controller.
- In particular, IB and server blade architectures are a natural fit.

Activity

- 6/01: First interoperability plugfest
  - Over 200 developers participated
  - Will be held 3-4 times a year
- 6/01: IBTA Developers' Conference
  - 70 node heterogeneous fabric initialized, managed
  - IBM DB2 EEE parallel database demo across 4 nodes
- 8/01: Intel Developers' Forum
  - 100 node heterogeneous fabric initialized, managed
  - Three-tier application demonstrated: SAP applications driving IBM DB2 EEE parallel database
- Over 50 vendors have announced over 100 IB-related products
- VC money is still there – going into IB startups.
- All this prior to shiping any products!
  - Expect initial products late '01 or early '02
  - Integration into memory subsystems will take longer.



Indeed, yet more random gratuitous clipart

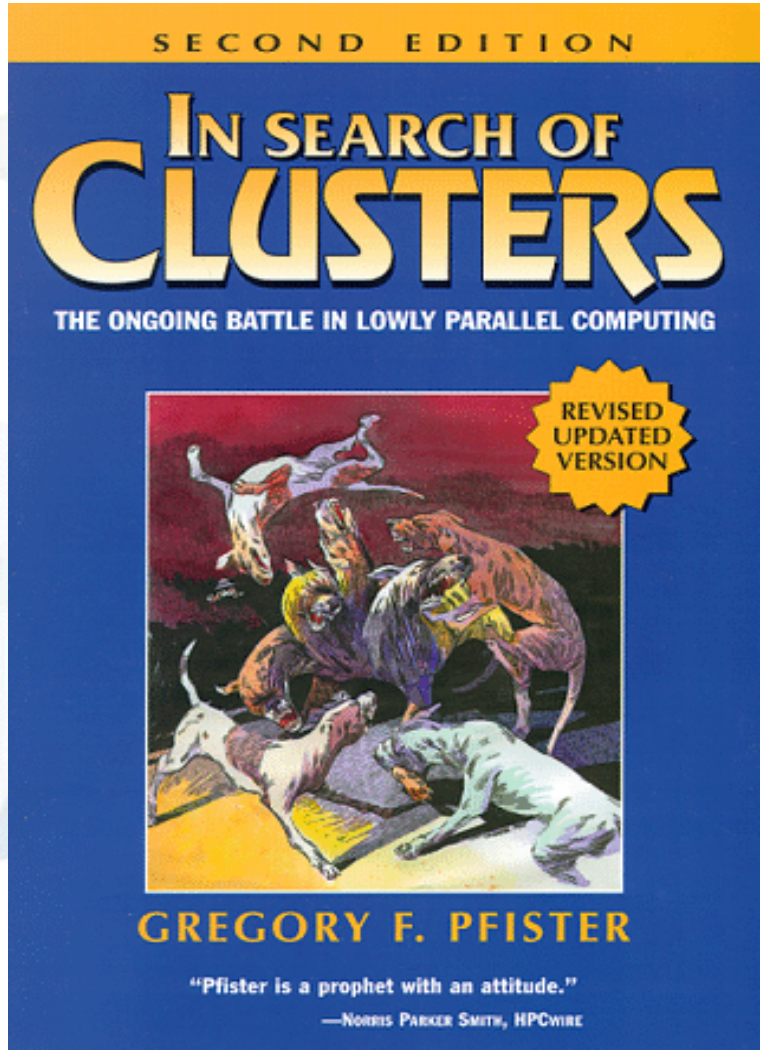
InfiniBand is a &Big_Deal

- Standard, high-volume enterprise-class server fabric:
  - RAS; management; performance; scalability
- Non-proprietary, low-overhead inter-host communication
  - enables open function now only on proprietary systems
  - will result in new cluster multi-tier server solutions/markets that have been impossible
- Scalable sharing of devices and host-I/O separation
  - Perhaps data sharing deserves another look?

Any of those, alone, would be very significant.

Together: foreshadow widespread new hardware/software system structures: A Golden Era for Clusters.

- Thank you for listening.
- Any (more) Questions?



Extremely nonrandom clipart

Just in case any of you were wondering...

(No, I can't give a presentation without plugging my book.)