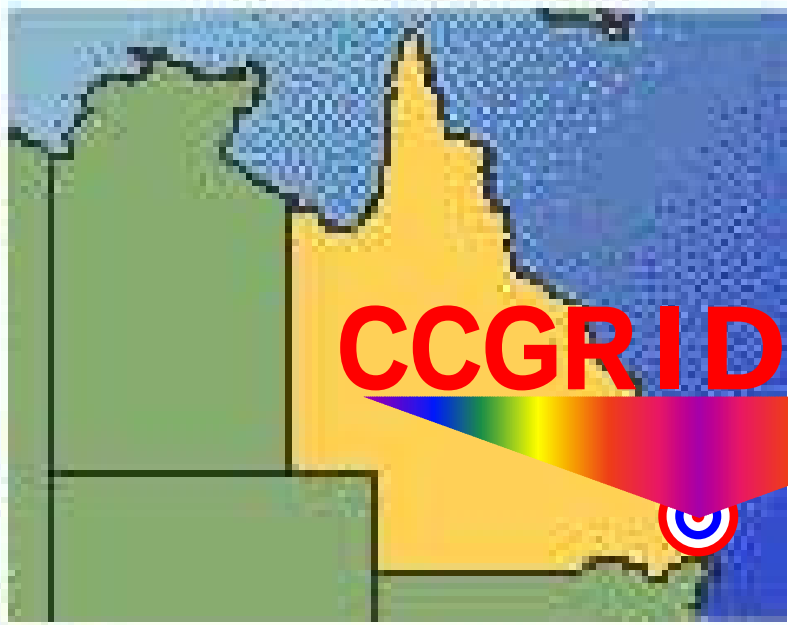
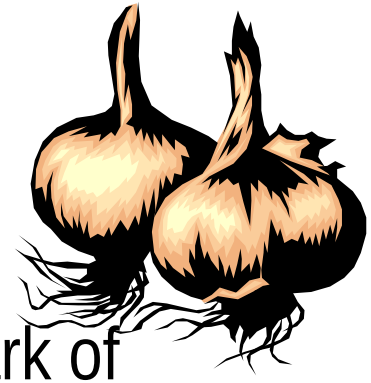
The Promise of the InfiniBand™ Architecture for Cluster Computing



CCGRID2001

Gregory Pfister
IBM Server Technology & Architecture, Austin, TX

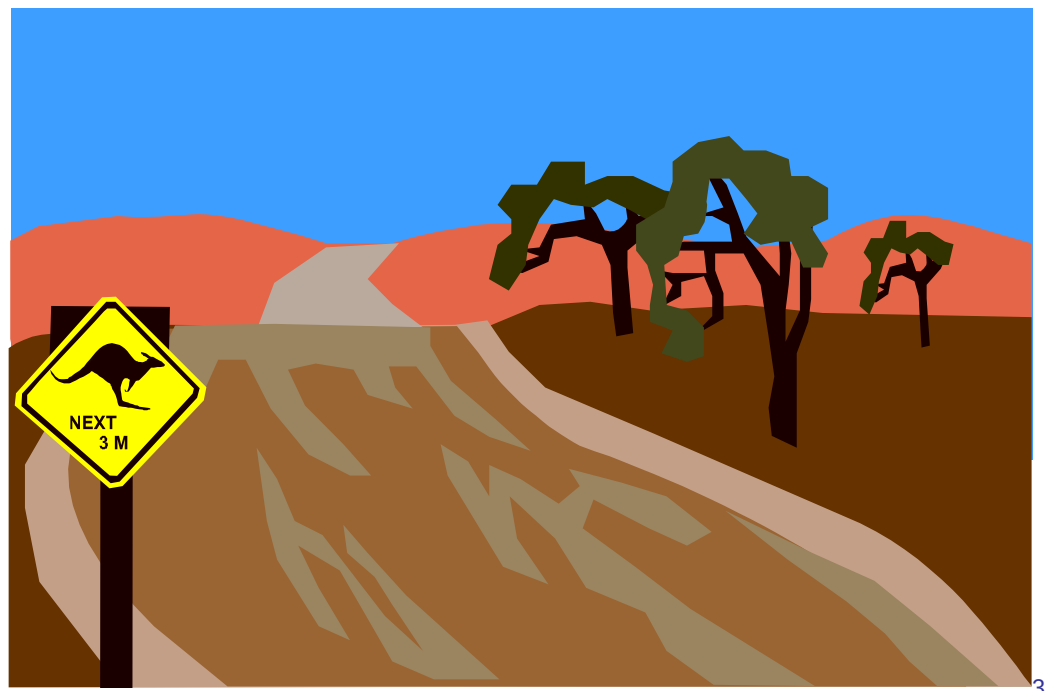
Legalities...



- InfiniBand is a trademark and service mark of the InfiniBand Trade Association. RapidIO is a trademark of the RapidIO Trade Association. Lightning Data Transport and Hyper Transport are trademarks of Advanced Micro Devices, Inc.
- 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.**

Before I forget:

- Ask questions!
- Or make comments.
- Please.



Random gratuitous clipart

Agenda

- Something Strange is happening
- What is the problem?
- InfiniBand and the InfiniBand℠ Trade Association
- The InfiniBand Architecture
- Industry Implications and Conclusions

More random gratuitous clipart

A Very Strange Thing is Happening

- Flamboyant words are in the industry press:

- They have nothing to do with microprocessors.
- They are being used for, of all things,

I/O Systems

Eh What?

- Isn't I/O that dull stuff called names like PCI, ISA, Fibre Channel, scuzzy, … ?
- Not any more.
- Moving bits into and out of computers is **hot**.
- Definitely includes moving bits *between* computers:
  - High bandwidth, low overhead
  - Broadly implemented – industry standards
- **Key hardware barriers to cluster exploitation are collapsing.**

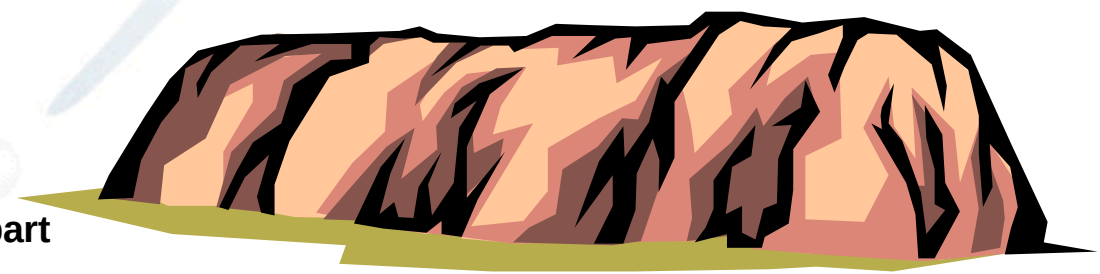
Even more random gratuitous clipart

What Happened?

- PC market saturation
  - So everyone has decided they better do servers
  - Which actually need good I/O (other than graphics)
- “Good enough” inexpensive microprocessors
  - Hence web server farms as an industry-wide paradigm
- Genuine technical problems with busses,
  + general march of technology
  - networks can replace busses in high volume products.

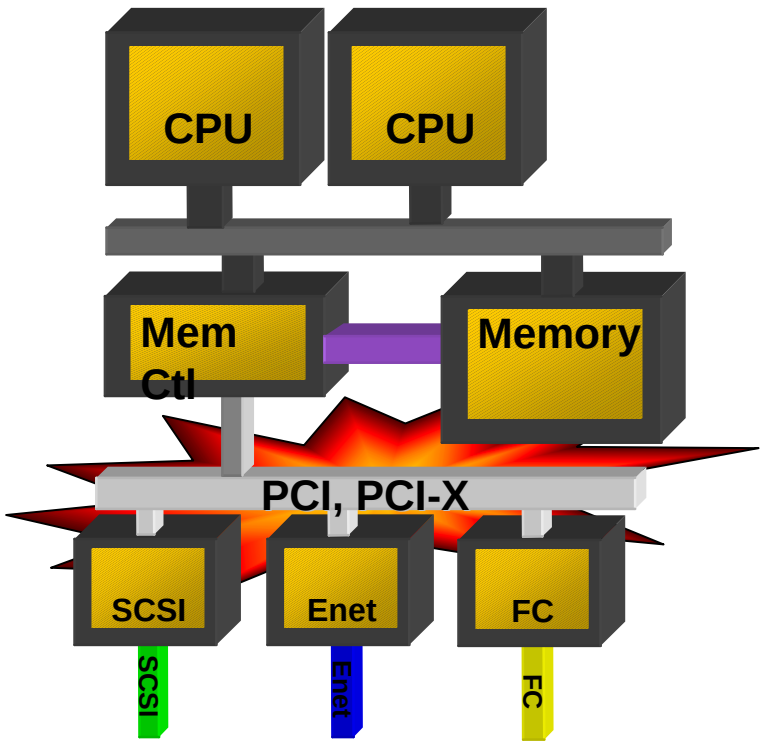
Agenda

- Something Strange is happening
- **What is the problem?**
- InfiniBand and the InfiniBand℠ Trade Association
- The InfiniBand Architecture
- Industry Implications and Conclusions



Yet more random gratuitous clipart

The Problem With Busses: Simple, Useful, but Running Out



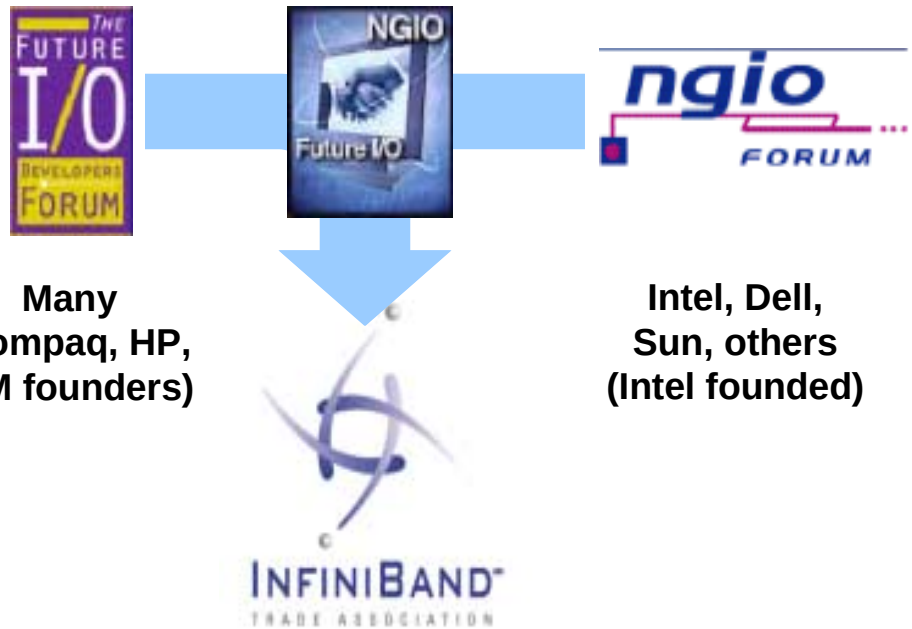
- Widespread realization: standard I/O busses (PCI familiy) can't keep up with
  - processors
  - LAN, etc.
- Bus frequency just 2X / 3 years
- Arbitration limits real bandwidth
- Load/store memory model
- Single fault domain for all I/O

Note: I/O includes sys-sys comm.

What Is InfiniBand?

- Replaces bus with industry-standard network
  - Connecting to devices and other systems
- Standard across the industry: 220+ companies
  - backed by all the major players
- Aim: an architecture able to track future technology and server requirements:
  - scalable bandwidth & fanout, up and down
  - high reliability, availability
  - low overhead
- Spec 1.0 published 11/23/2000
  - available for download: http://www.infinibandta.org

IBTA: A Merger, 9/99



Steering Committee

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

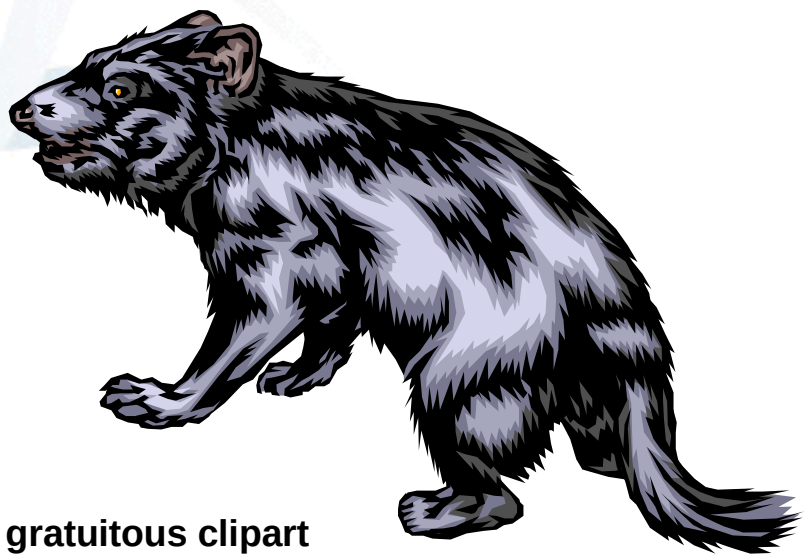
Sponsor Member Companies

3Com · adaptec · Cisco Systems · Fujitsu Siemens · HITACHI · NEC · Lucent Technologies · NORTEL NETWORKS

- Right Ts&Cs for wide adoption: "fair & non-discriminatory" licensing
- Not an open standards group (time to market)
  - Anyone can join with member or associate status.
- Managed like a SIG

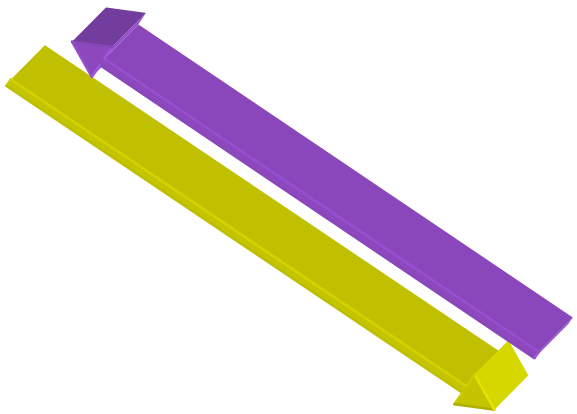
Agenda

- Something Strange is happening
- What is the problem?
- InfiniBand and the InfiniBand℠ Trade Association
- **The InfiniBand Architecture**
- Industry Implications and Conclusions



Still more random gratuitous clipart

The Link

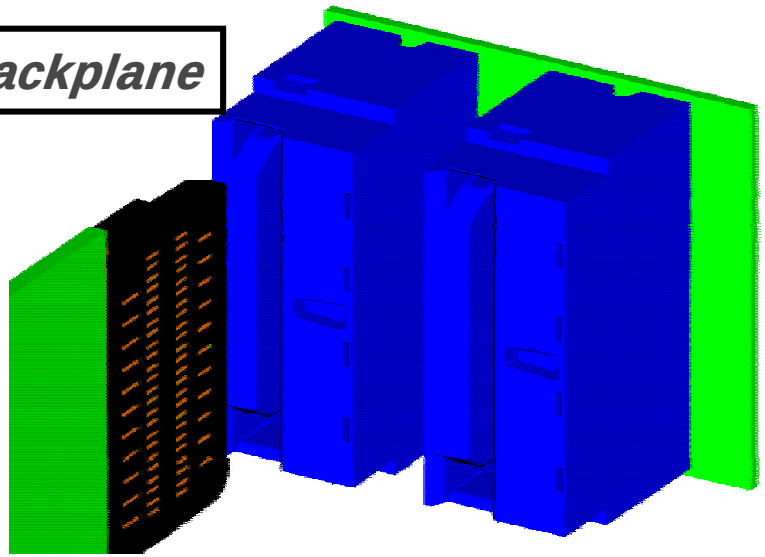


Width	Bi-directional Bandwidth
1	500 MB/s
4	2 GB/s
12	6 GB/s

- 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 256B to 4KB

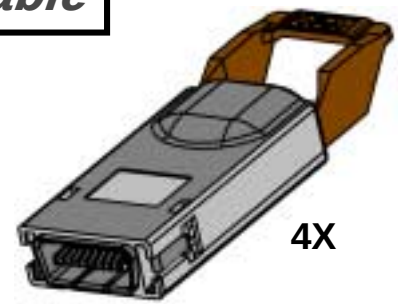
Connectors

Backplane

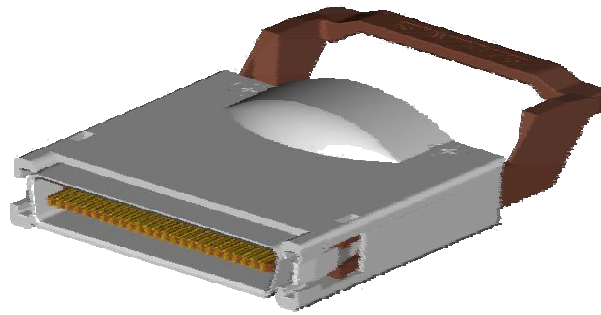


Copper Cable

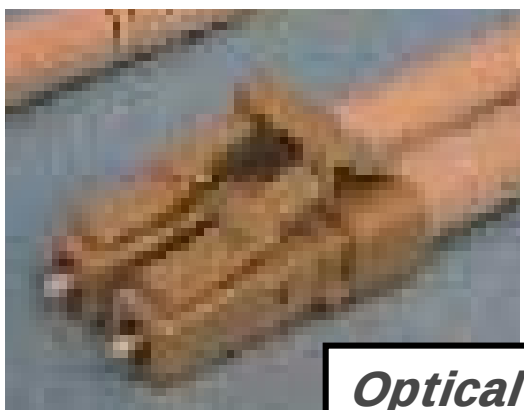
1X

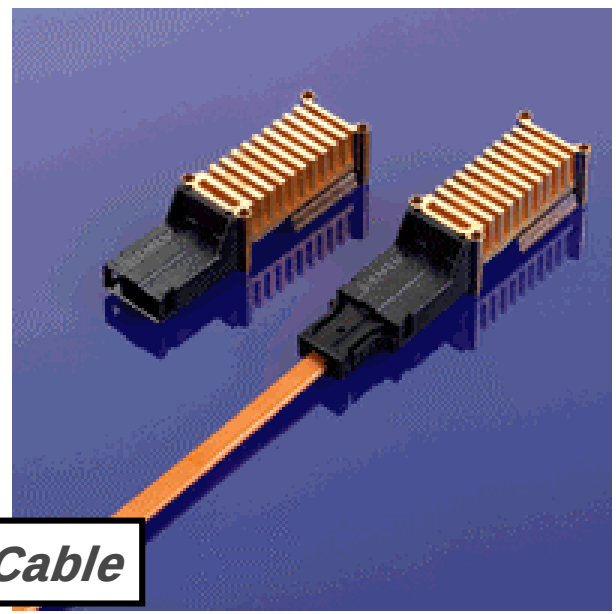


4X



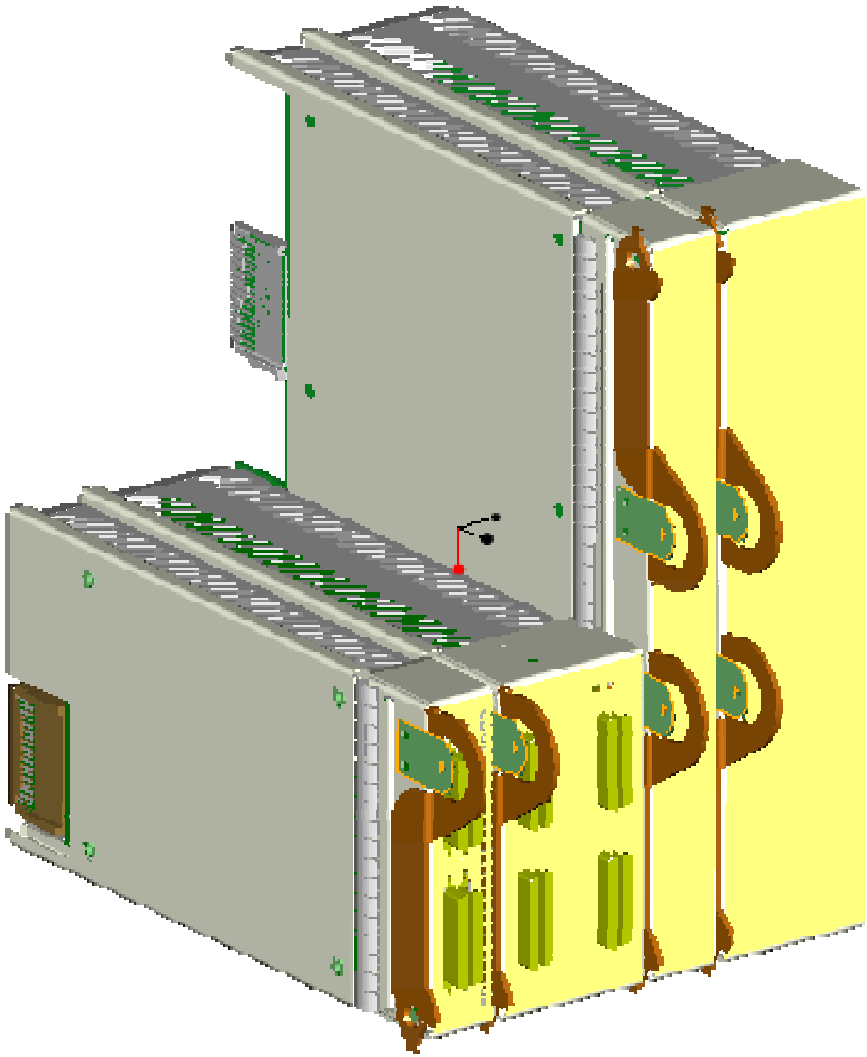
12X





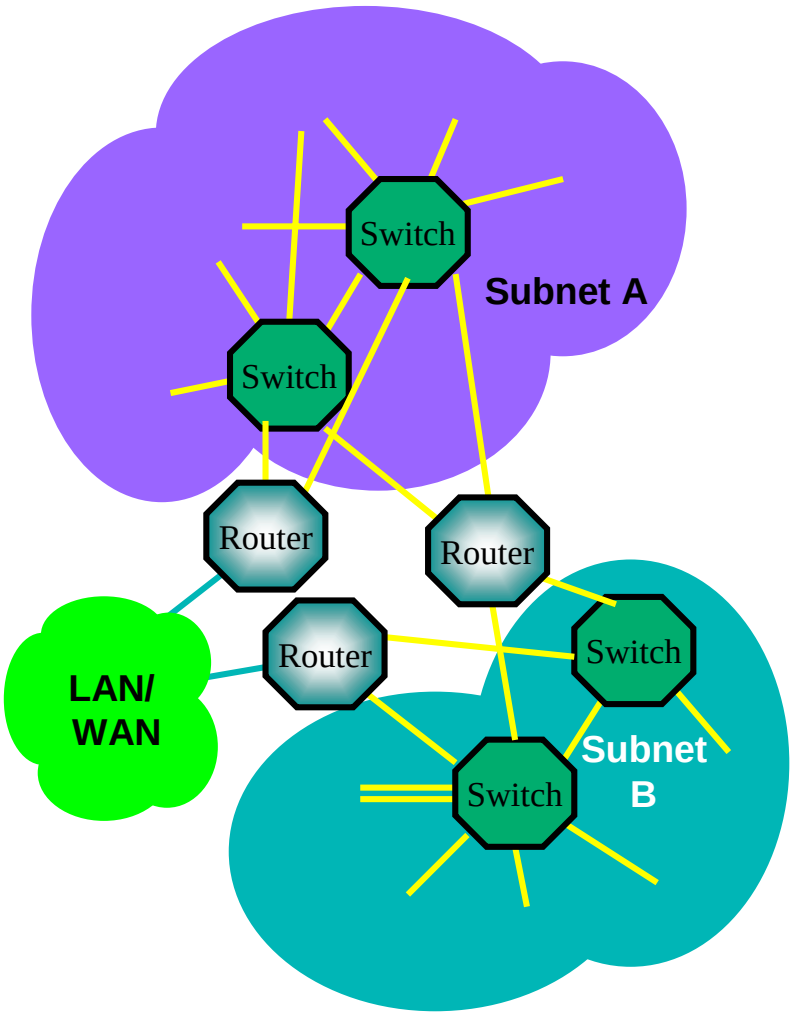
Optical Cable

Electromechanical



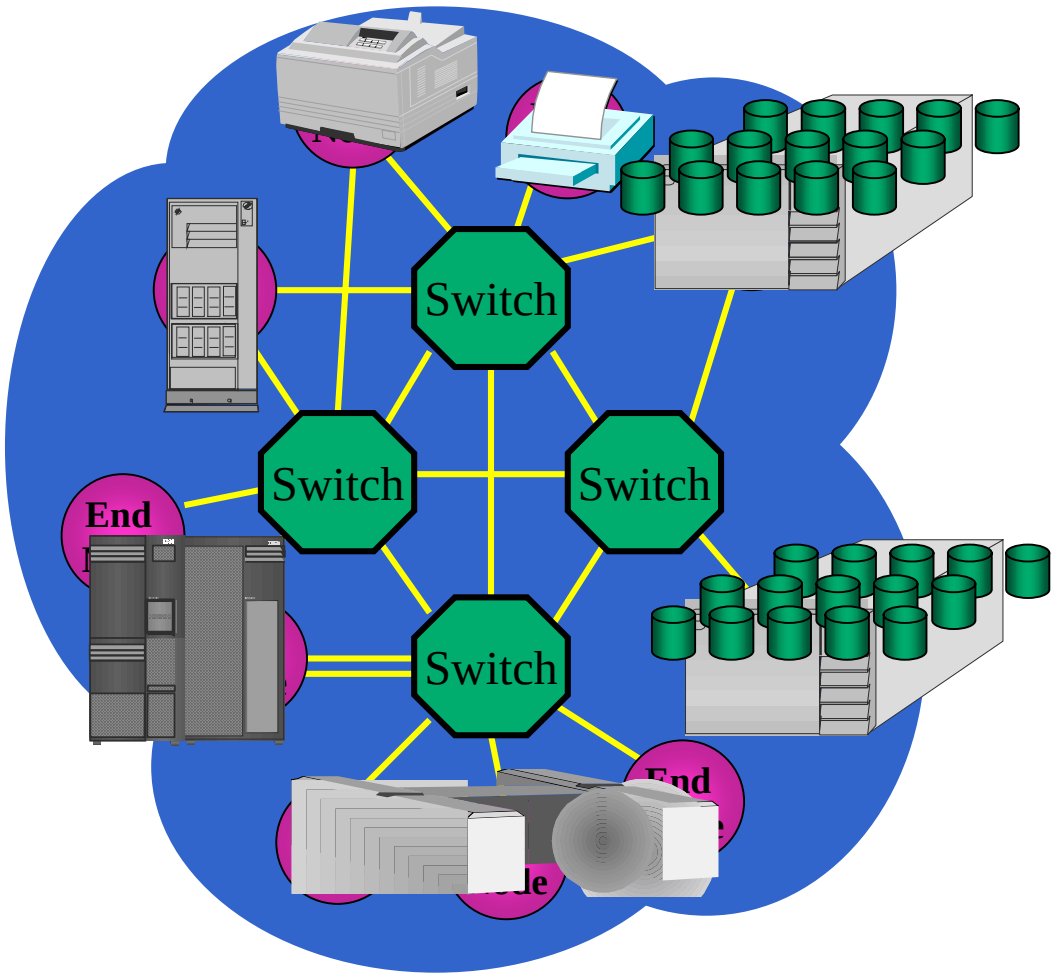
- Four adapter form factors:
  - standard, standard wide, tall, tall wide
  - standard approx. 20/100/220 mm wide/high/deep
  - face plate can support quad SCSI / HSSDC connectors
- Standardized baseboard mgmt, thermal, EMC, hot swap, E/M interfaces, modules, slots, LEDs.

Switches and Routers



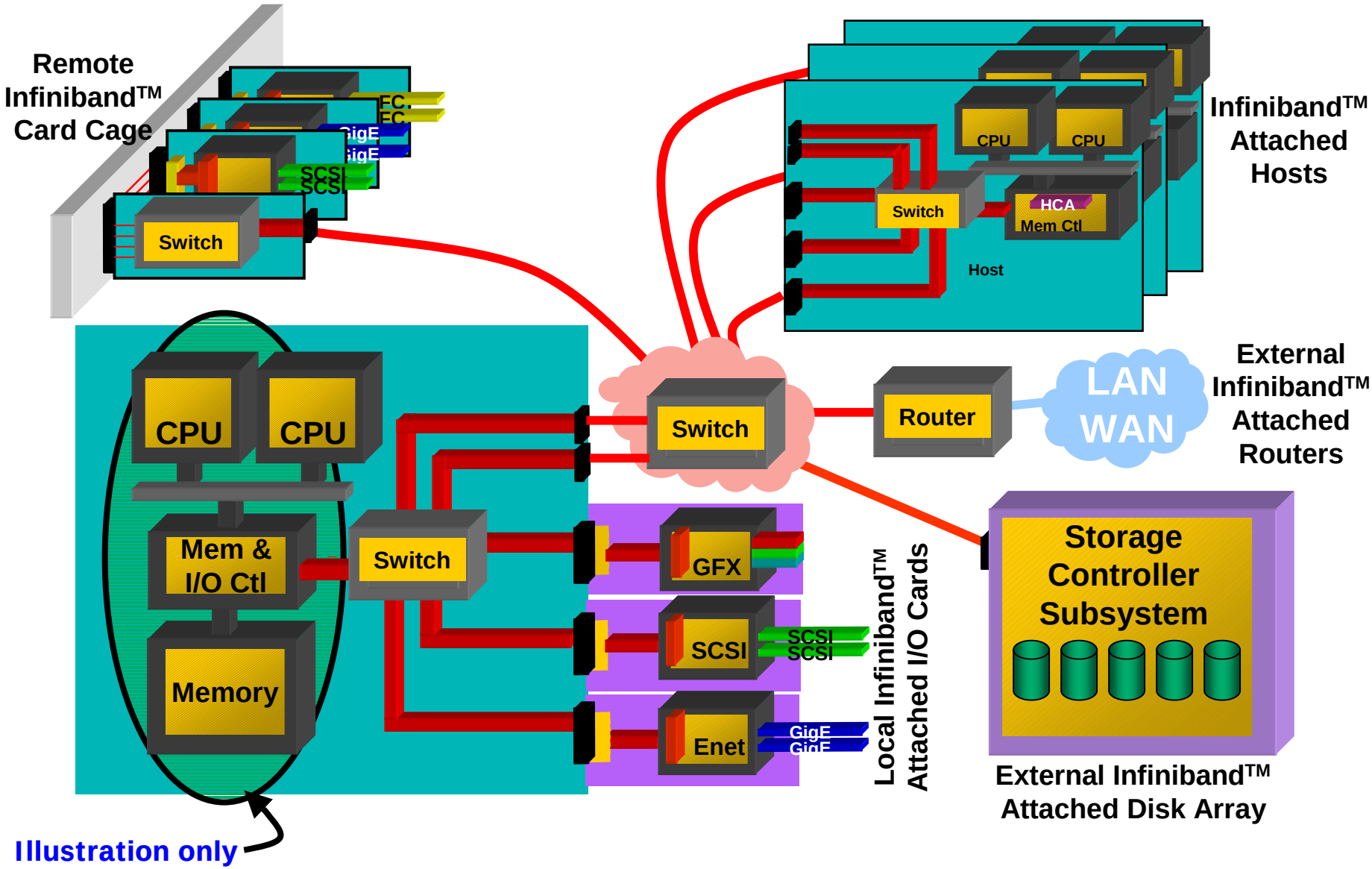
- Switch: routes packets within subnet.
  - Destination routed, based on LID
    - Special direct route for initialization
  - Up to 48K unicast LIDs per subnet.
  - SLs provide service differentiation.
  - Multicast (optional)
  - Switch size, network topology are vendor-specific
- Router: routes packets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

Endnodes

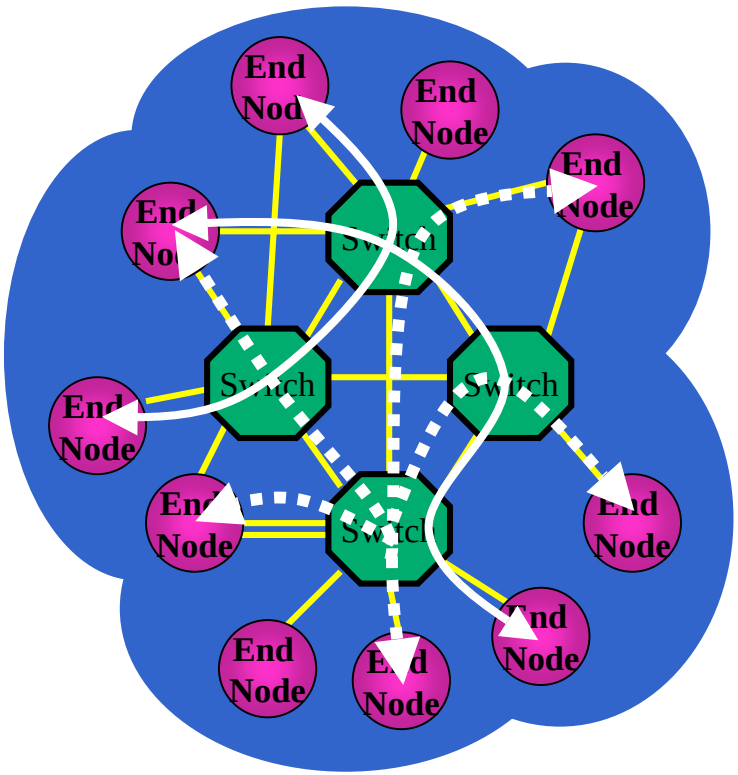


- Hosts
  - processors, memory
- Devices
  - Storage, network adapters, etc.
- Bridges
  - to “legacy” I/O busses: PCI, etc.; vendor unique; not part of spec
- Channel Adapters attach endnodes to links
  - Host (HCA) vs. Target (TCA)
  - Only difference: TCA has no defined software interface.

A Less Cloudy View

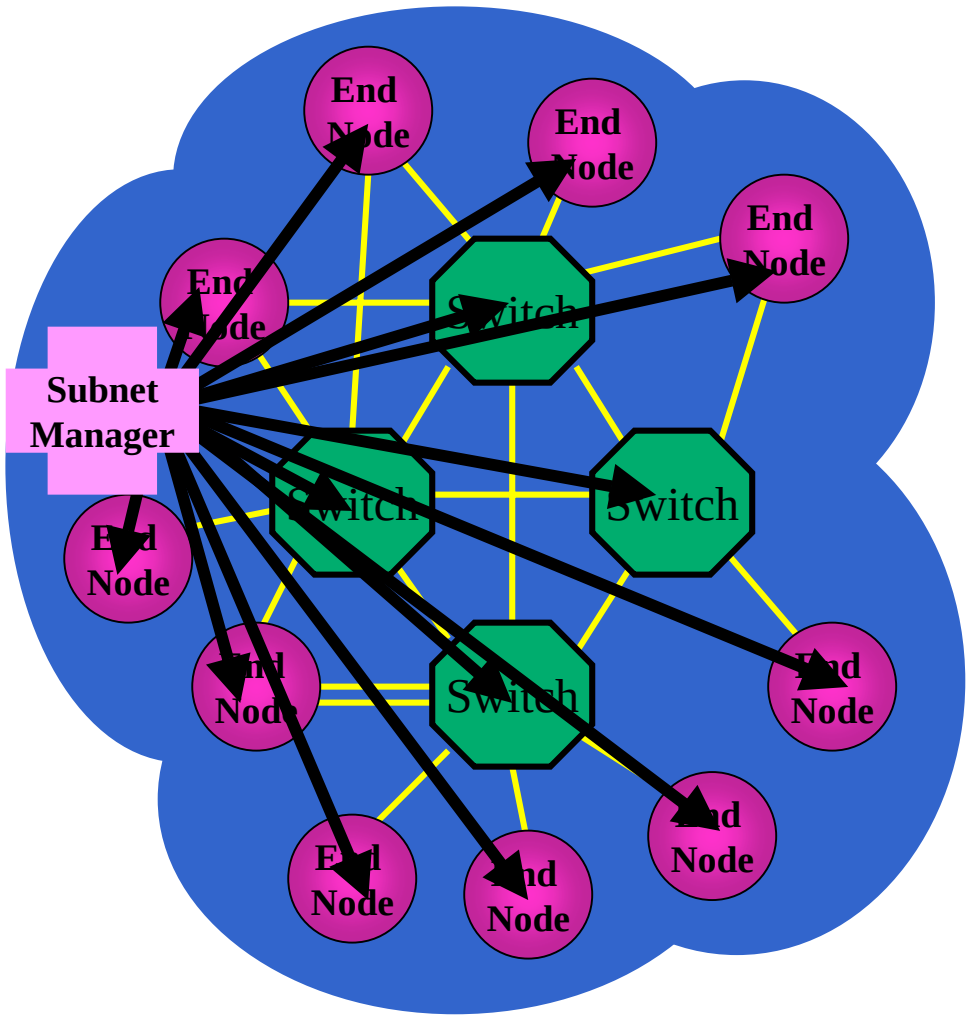


Channel Adapters



- 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.

Other Topics Not Covered

- InfiniBand spec is over 1500 pages long.
- Some other topics that could be covered:

- Compliance and interoperability
- Partitioning
- How Reliable Datagram works – and why it's there
- Queue Pairs
- Automatic Path Migration
- Various management functions:
  - Subnet administration, performance, device, configuration, boot, etc.
- Verbs
- Link vs. transport layers
- Electronic/Mechanical issues

Agenda

- What is the problem?
- InfiniBand and the InfiniBand℠ Trade Association
- The InfiniBand Architecture
- **Industry Implications and Conclusions**

Yet still more random gratuitous clipart

What About 10 Gb Ethernet & IP?

Semi-gratuitous clipart (featuring really stupid armor)

- Enet+IP: widespread, incumbent, familiar.
  - NAS and iSCSI compete directly with IB
  - Why don't they just win in the market?
- Why this may not happen (but it might):
  - IP software overhead:
    - serious server I/O requires major IP offload
      - hard! full offload never commercially successful.
      - must re-invent IB-like zero-copy/user-mode
    - IB ***is*** I/O: direct IB communication must be more efficient than anything going through IB to an adapter
  - Volumes & Presence:
    - If IB already comes out of every SHV system, and IB switches = cost/port of 10 GigE switches -- it's already there!
    - Rapid adoption rate predicted.

InfiniBand is a &Big_Deal.

(All the terms are overused. Use the one in your context.)

- 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
- Host-I/O separation enable higher density and data-centric system organizations

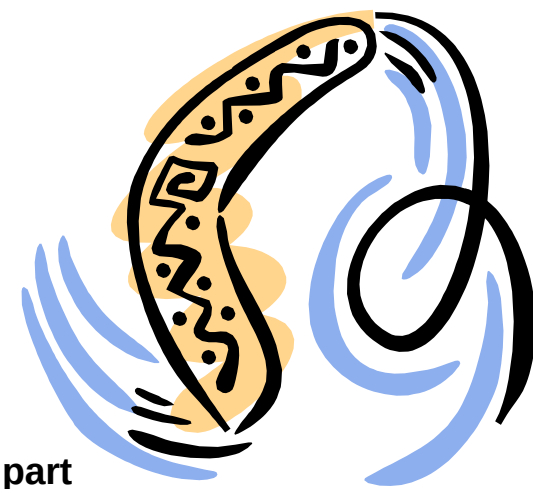
Separately, any of those would be very significant.

Together: foreshadow widespread new hardware/software system structures.

Indeed there is more random gratuitous clipart

But It's Still the Software, Stupid!

- … even though hardware vendors would much prefer otherwise.
- Still must deal with unfinished business:
  - Programming models; sharing vs. shared-nothing; security & authorization; accounting & chargeback; scheduling; process and data co-location; resource discovery and/or recruiting; global naming of several sorts; QoS support; heterogeneous interoperability
  - Many of these now being visited (or re-visited) in work on Grids.
- However, a new context:
  No longer hamstrung by hardware that's slow, inefficient, or nonstandard.



Another fine example of semi-gratuitous clipart

Which Means

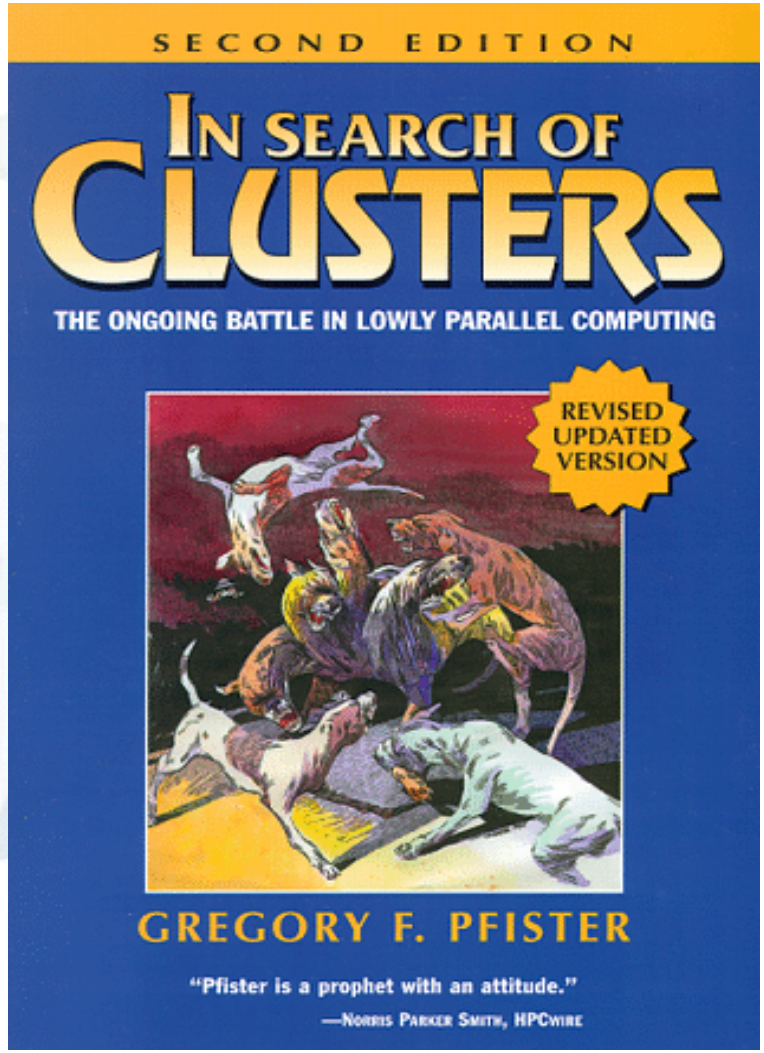
- **We may be at the start of a new cluster era.**
- **Those industry trumpets are sounding for you**

Cluster Computing is

Hot!

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

Just in case any of you were wondering...



Extremely nonrandom clipart

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