
HPC高速网络

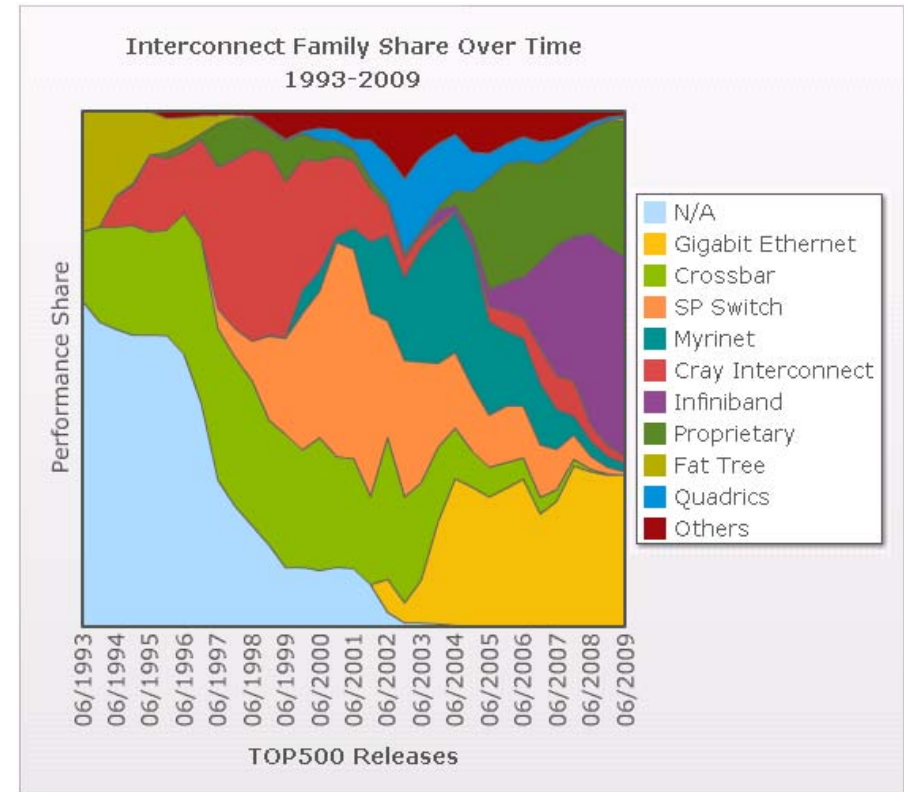
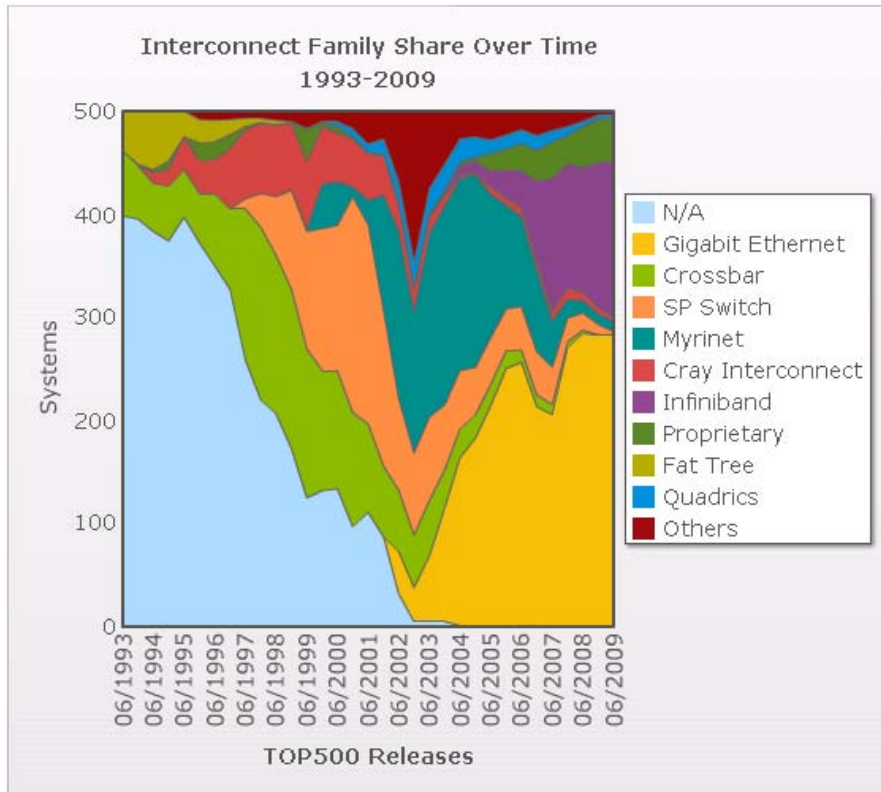
霍志刚

国家智能计算机研究开发中心
中国科学院计算技术研究所

2009年10月28日



Top500 Statistics



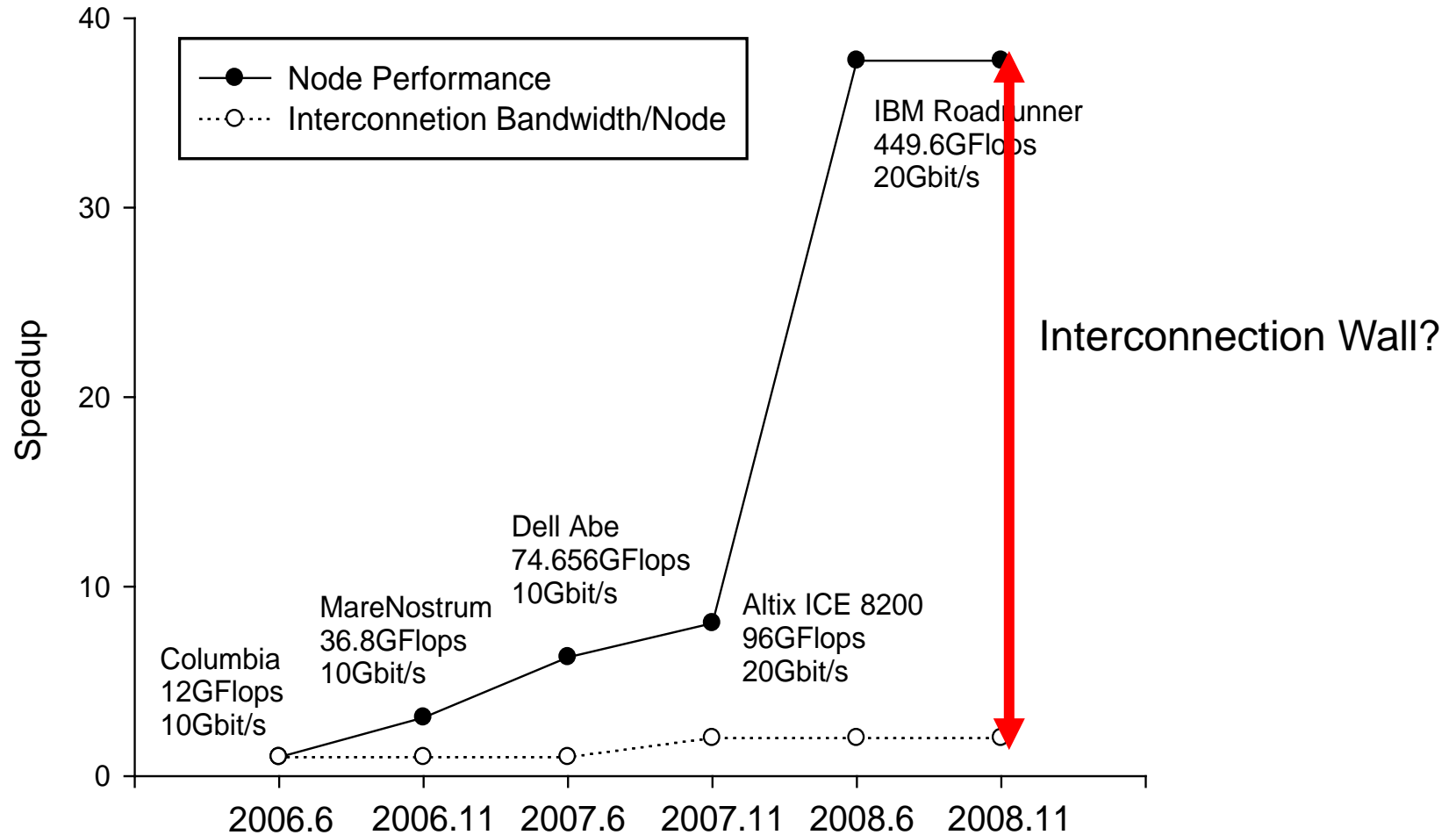
- What lead to the demise of earlier technologies?



Considerations in choice of HSN

- Cost
- Performance
 - Latency matters most
- Scalability
- Usability
 - Long long time ago, “Only the guys from Israel can save you”
- Support
- Anti-monopoly

Demands for HSN (1)



Demands for HSN (2)

- High-density nodes put more pressure on interconnects.
 - Multi-core/many-core
 - FPGA
 - GPU
 - Hybrid design
- Outdated rule of thumb
 - 1GHz ~ ?GB ~ ?Gbps



- What a mess!

HSN: The sky is the limit?

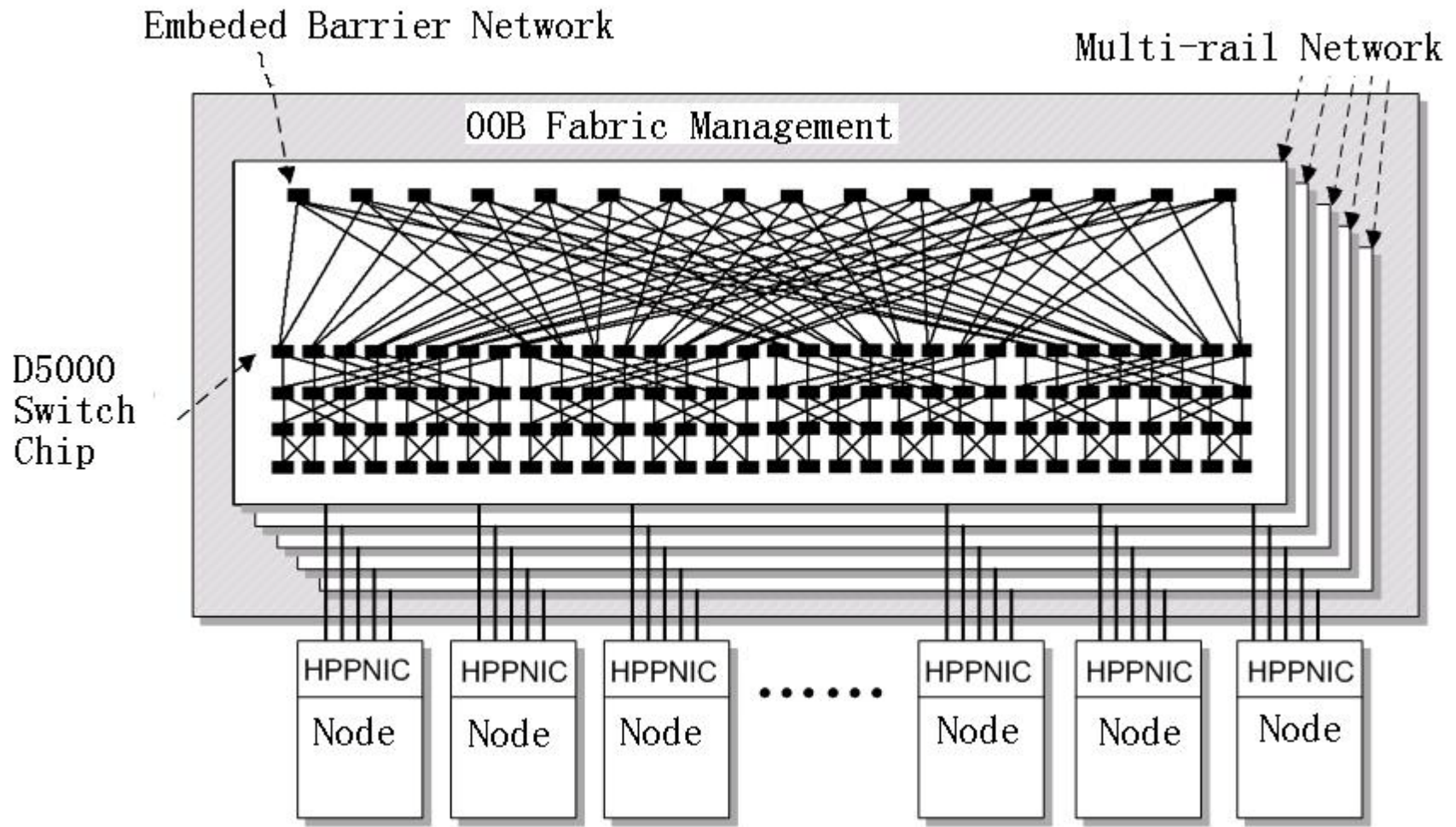
- Fiber Optical Technology
 - 30Tbps (KDDI/NICT, Aug.2009)
 - ~100 fibers in a single cable
- Latency
 - Ruled by physical law (Light in fiber: 5ns/m)
 - SerDes

Barrier Optimization

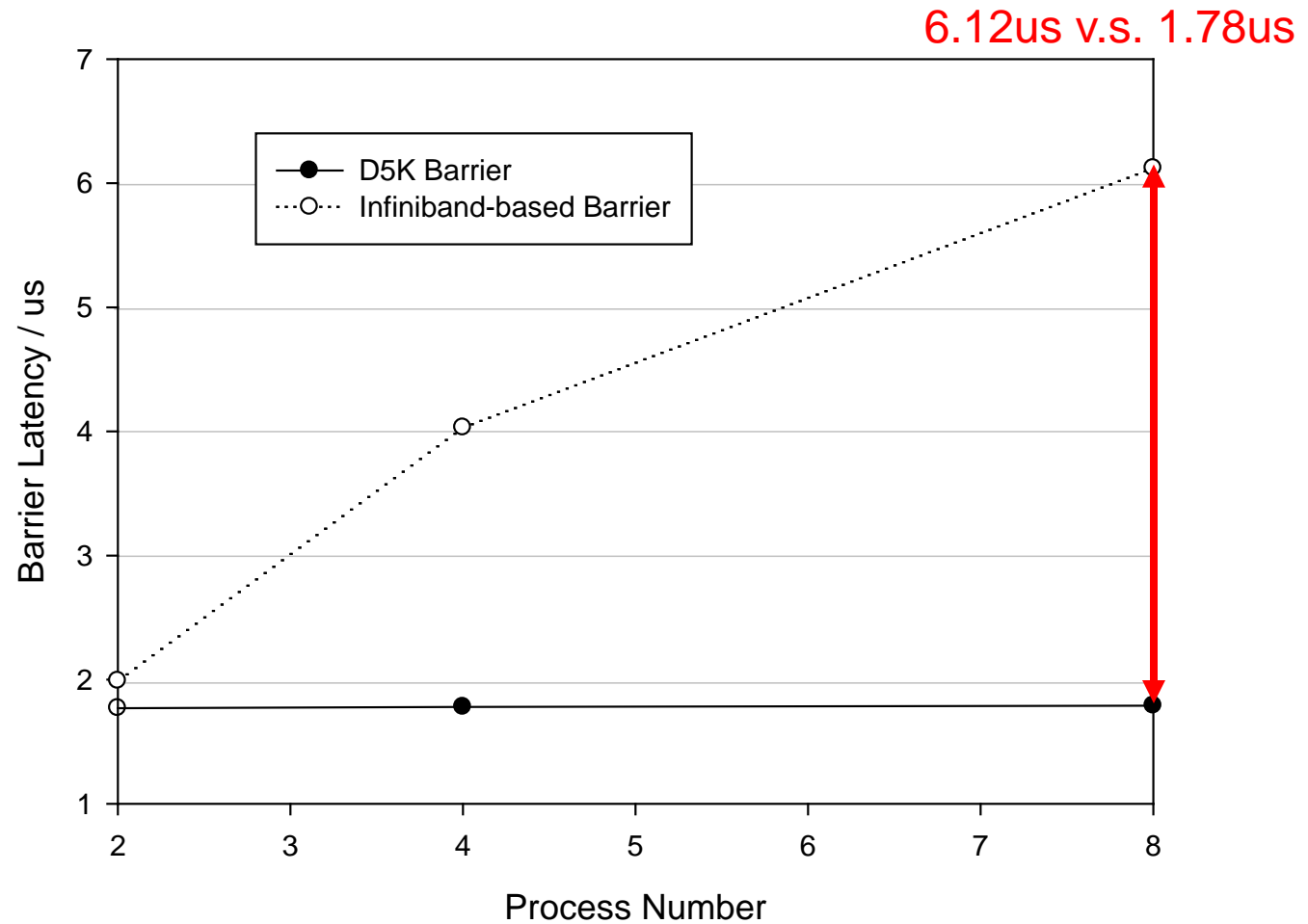
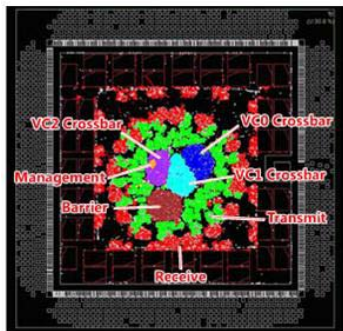
System	Year	Topo.	Scale	Lat.
IBM CM5	1993	Simple Barrier Tree	256	5ms
Cray T3D	1994	Degree 4 Hard-wired Barrier Tree	256	35us
FujitsuAP1000	1996	Hard-wired Barrier Tree	256	5.2us
NEC ES	2002	Single State Barrier Counter	640	3.4us
IBM BG/L	2006	Hard-wired Binomial	16k	1.3us
Cray T3E	1995	3D Torus/32 Barrier Tree	512	5.0us
Quadrics	-	Fat Tree/32 Barrier Tree	64	6.0us

- A long history.

D5000 prototype



Barrier opti. in D5000 prototype



Logistics of High Speed Networking

- Fabric Management
- Routing
- OS support
 - High Speed Networking: A SW View
- Fault-tolerance



Corner latencies



- Phy-virt addr translation
 - Headache for decades
 - Eliminate it?
- System noise
 - Process mgt.
 - Cache/TLB miss



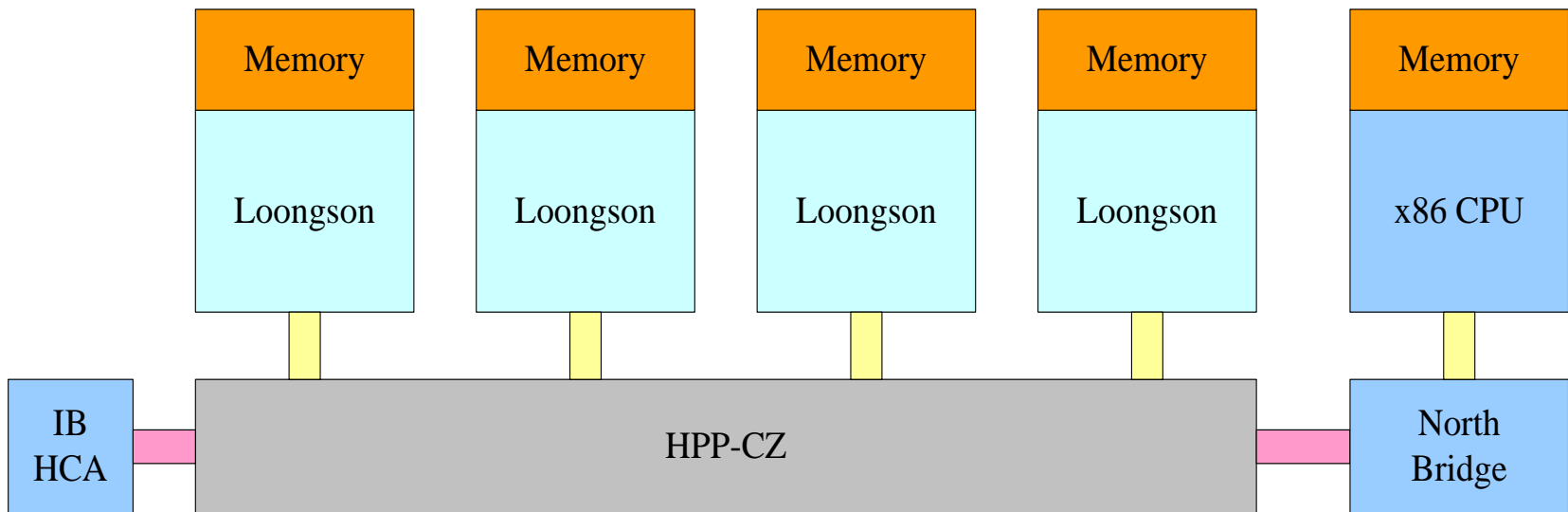
Resource consumption

- Lightweight endpoints
 - A simple calculation
 $128 * 32 * 128 = 524,288$ (Jack, 2007)
 - Learn from BSD Sockets
- Addressing

What we are doing NOW?

- IB support for Loongson
- IB virtualization in D6000
- Coll-op optimization
 - Offloading
 - Specialized network
 - Multi-core-Aware intra-node comm

Node structure of D6000



Any comments?

