INFINIBAND FAST INTERCONNECT

Yuan (Rick) Liu
Institute of Information and Mathematical Sciences Massey University
May 2009

Overview

History

- What is IB
- Where to use IB
- Why IB
- Interconnect Now & Future

History

- Why need IBA
- Future I/O (FIO)
  - aims to increase server I/O throughput
  - IBM, Compaq and HP
- Next Generation I/O (NGIO),
  - enabling relatively rapid PCI replacement in volume segments
  - Intel Microsoft and Sun

History

- FIO + NGIO = System I/O (SIO) (2000)
- SIO → InfiniBand

Where to use InfiniBand

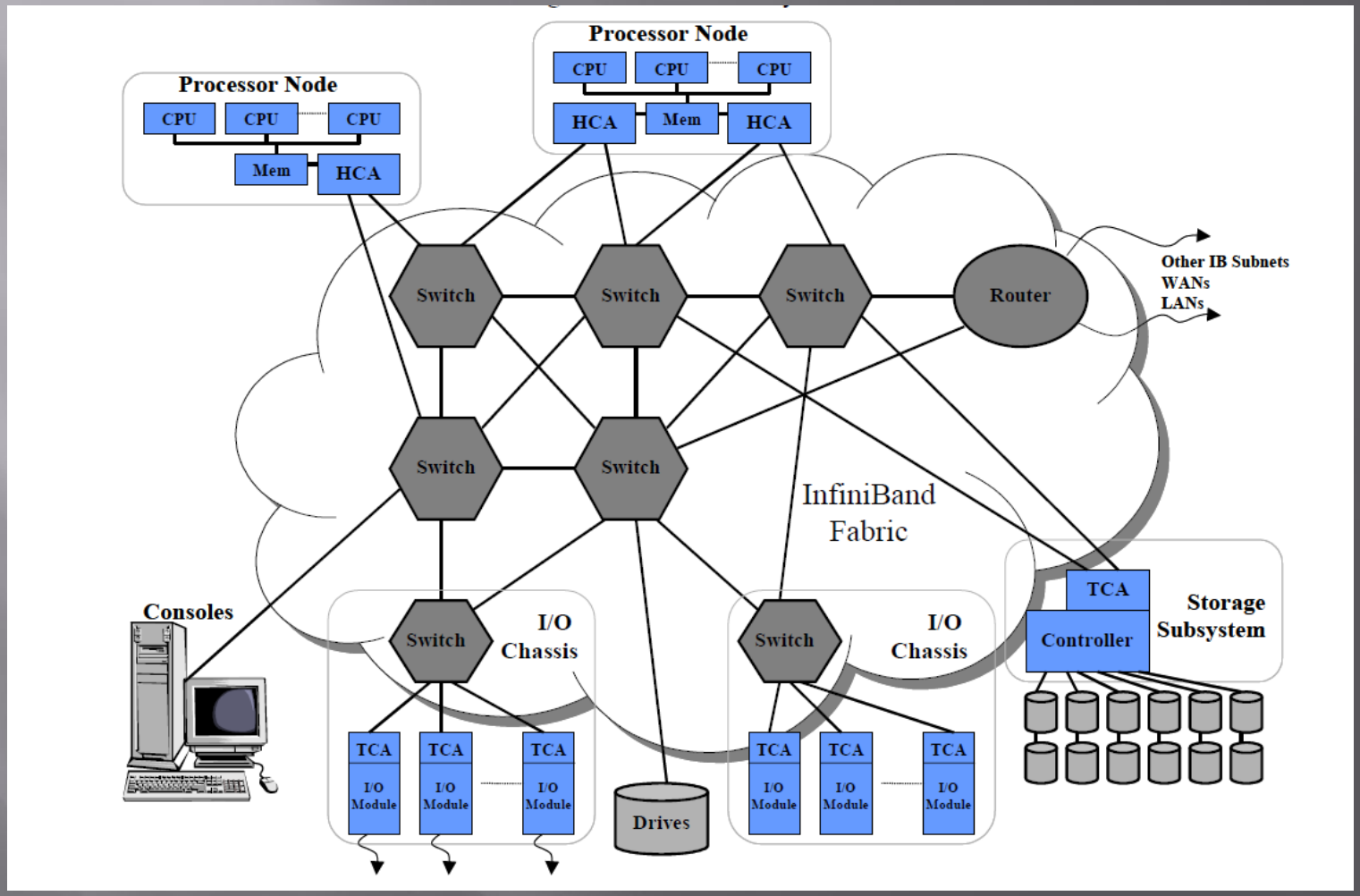
- Bandwidth “inside the box”
- replace bus based structure

- Bandwdth “out of the box”
- Data center, clusters, HPC interconnect

What is InfiniBand

- Best of the technologies from each side
- Industry standard switch-based serial I/O interconnect architecture
- High speed
- Low latency
- Cost efficient
- Enhanced reliability
- Interconnect

What is IB


Processor Node
CPU
CPU
CPU
HCA
Mem
HCA
Processor Node
CPU
CPU
CPU
Mem
HCA
Switch
Switch
Switch
Router
Other IB Subnets
WANs
LANs
Switch
Switch
InfiniBand
Fabric
Consoles
Switch
I/O
Chassis
Switch
I/O
Chassis
TCA
Controller
Storage
Subsystem
TCA
I/O
Module
TCA
I/O
Module
TCA
I/O
Module
Drives
TCA
I/O
Module
TCA
I/O
Module
TCA
I/O
Module

What is IB

- Switches
  - Do not generate or consume packets
  - Support 1X,4X,12X mode, each mode can support SDR/DDR/QDR

	Single (SDR)	Double (DDR)	Quad (QDR)
1X	2 Gbit/s	4 Gbit/s	8 Gbit/s
4X	8 Gbit/s	16 Gbit/s	32 Gbit/s
12X	24 Gbit/s	48 Gbit/s	96 Gbit/s

What is IB

- Host Channel Adapters (HCA)
  - Very intelligent
  - Capable of handling large numbers of concurrent connections
  - Large number of send/receive buffers

  - Connect from processor node to switches
  - Can have one or more ports
  - Can use local system bus interface

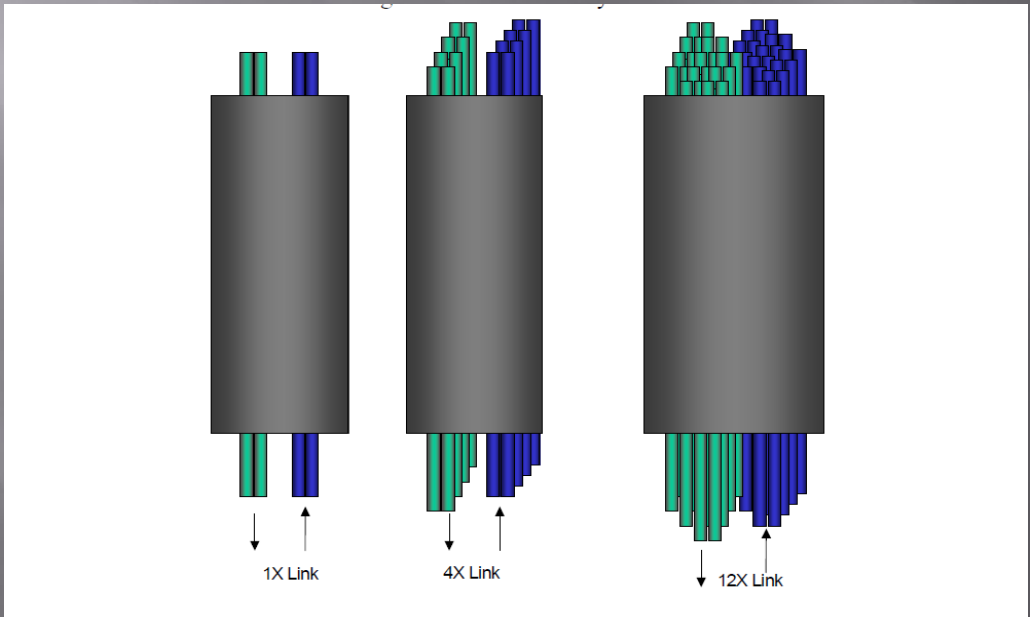
What is IB

- Target Channel Adapters(TCA)
  - Less intelligent
  - Less numbers of concurrent connections
  - Less number of send/receive buffers

  - Connect from I/O node to switches
  - One I/O node = one or more storage devices

What is IB

- Wirings
  - twisted pair copper wires / Fiber cables
  - 30m /10KM



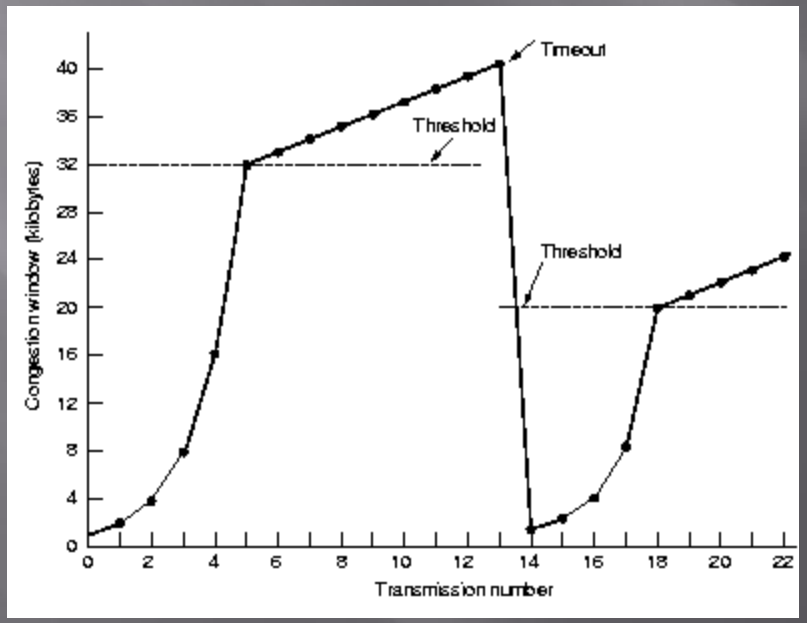
Why IB?

- IB V.S. Ethernet
- IB protocol V.S. TCP

- high bandwidth
- low CPU utilization
- low latency
- scalability
- RAS

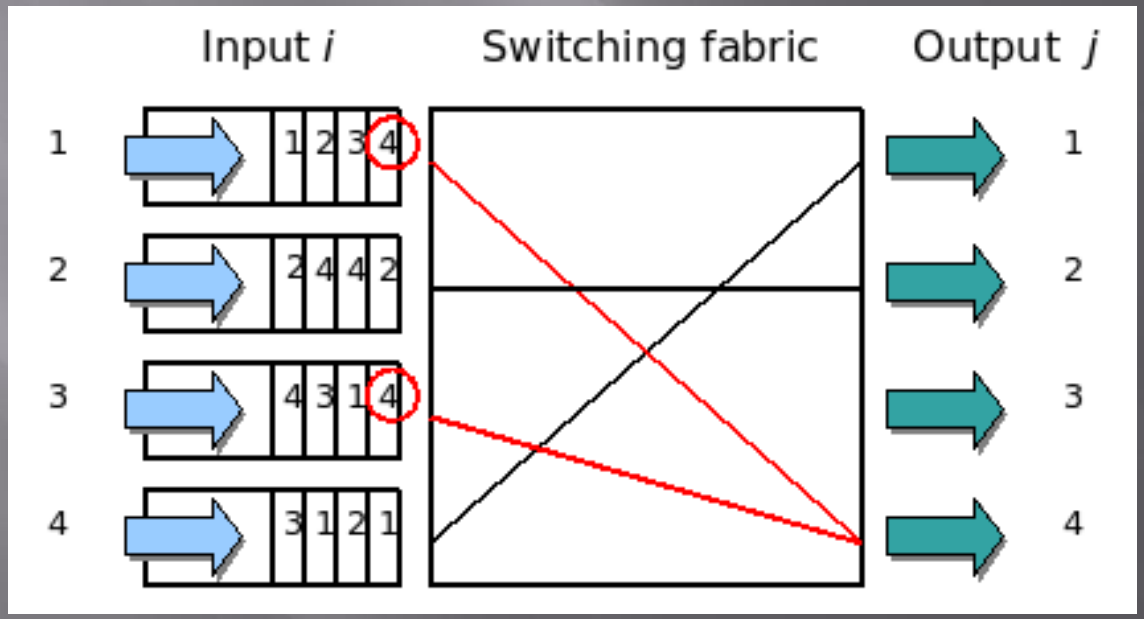
Why IB?

- The TCP bottleneck



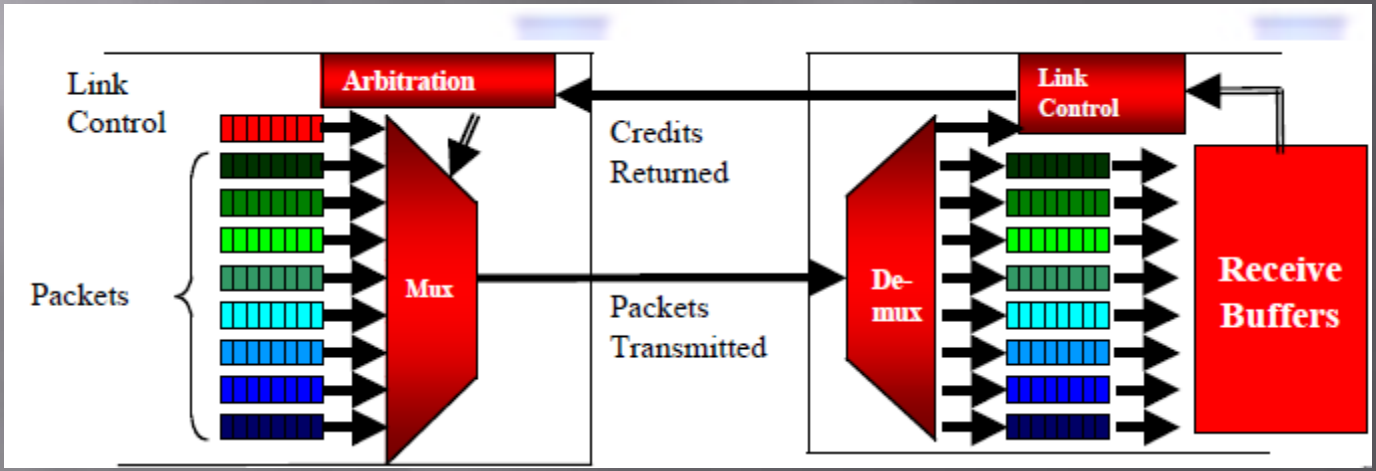
Why IB?

- Head of Line Blocking



Why IB?

- “Magic” Virtual Lanes
  - credit based flow control
  - alleviates HoL blocking



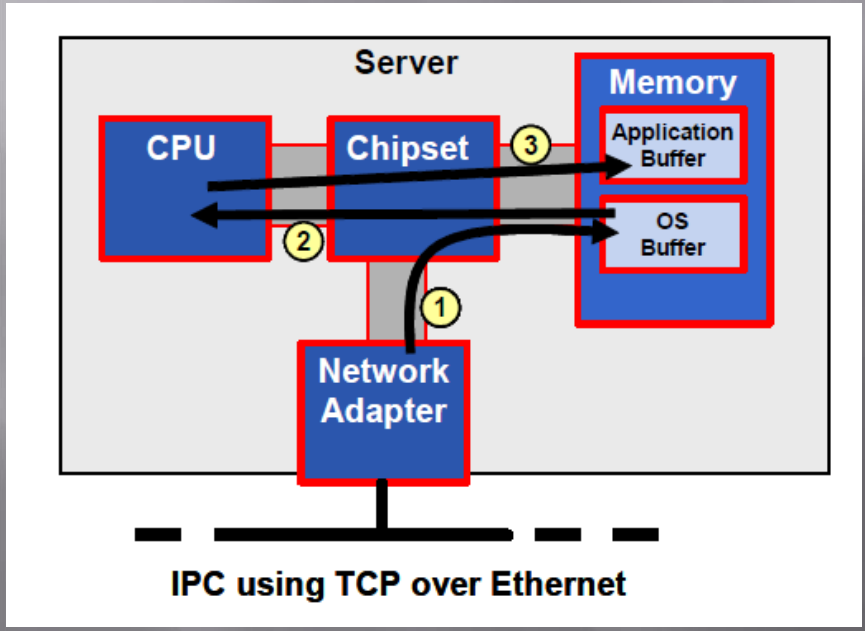
Why IB?

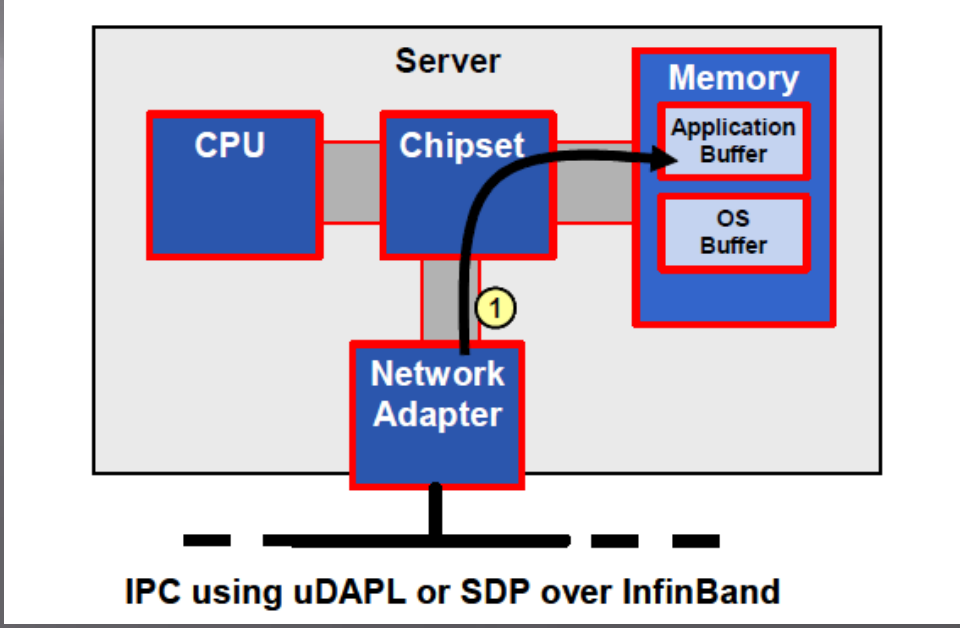
CPU utilization and Scalability

- Heavy TCP headers
  - Bandwidth leak
  - CPU clock leak
- IB protocol is light weight
  - Implements communications stack in hardware
  - RDMA
  - 10X better CPU utilization v.s. Gigabit Ethernet

Why IB?

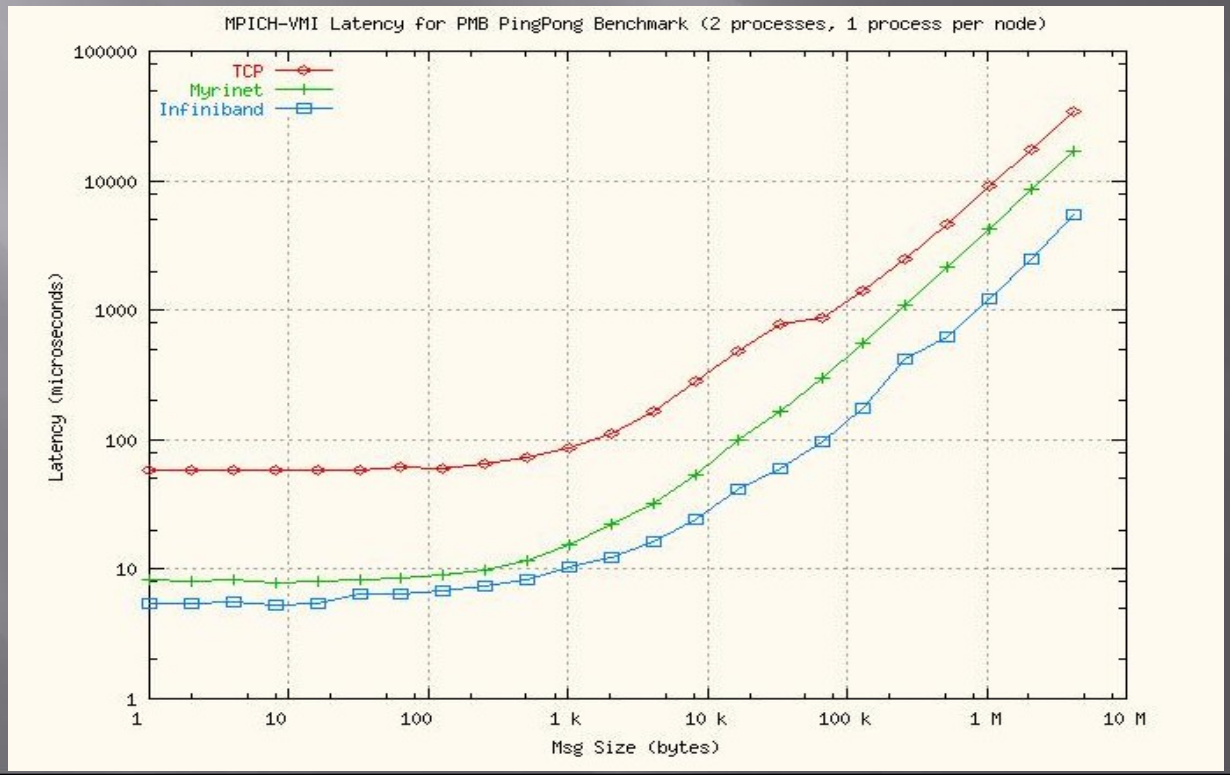
- Remote Direct Memory Access (RDMA)





Why IB?

- Latency, Scalability
  - speed = high avaliable bandwidth + low latency
  - Synchronize ? Latency → scalability



Why IB? _RAS

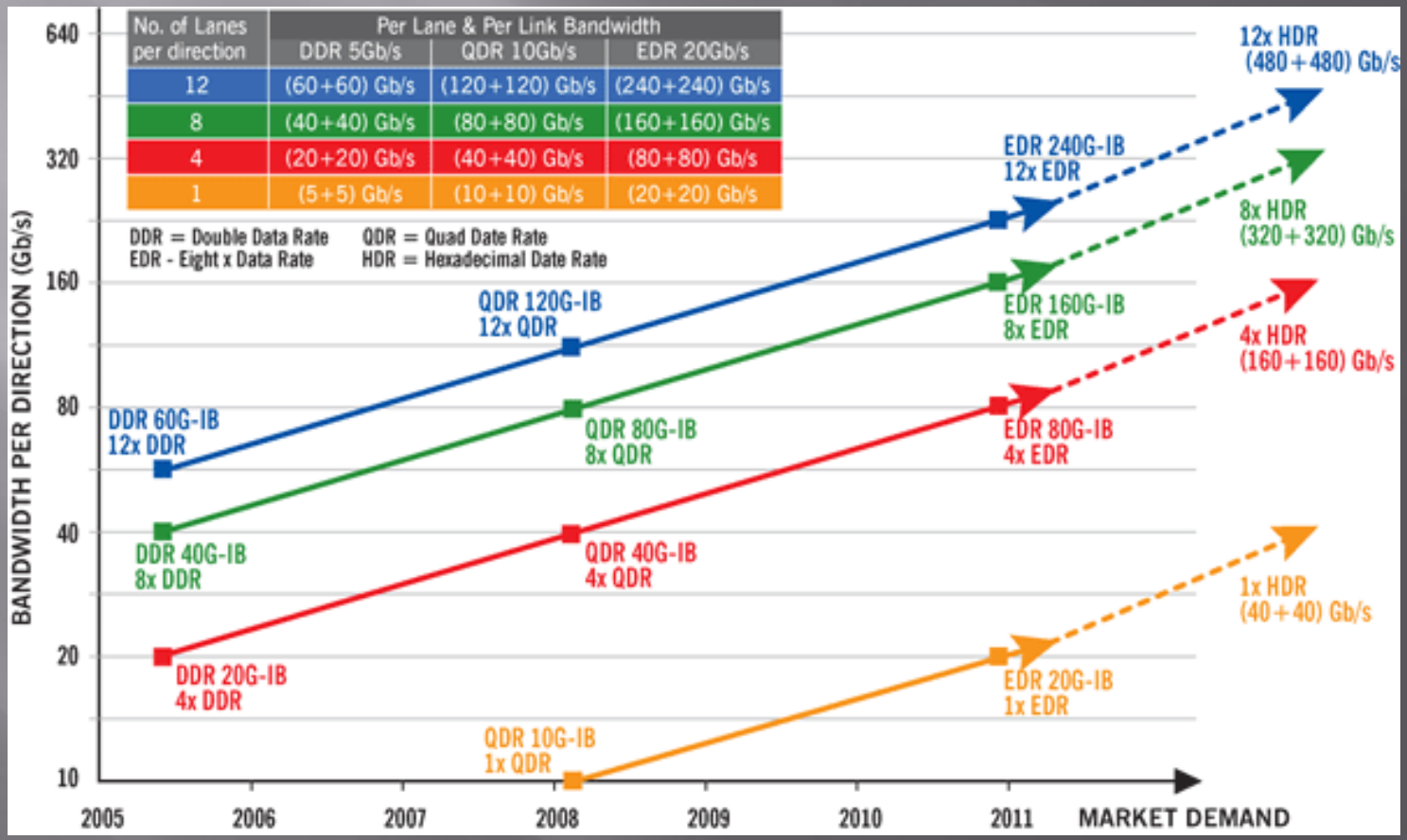
- Reliability
  - 2 CRC packets
  - 16bit Variant CRC(VCRC)
    - Covers whole package
    - Recalculate from hop to hop
  - 32 bit Invariant CRC(ICRC)
    - Covers fields do not change

Why IB? _RAS

- Availability
  - enables redundancy
  - supports failover by switching to an alternative path
- Serviceability
  - hot swappability
  - special management functions

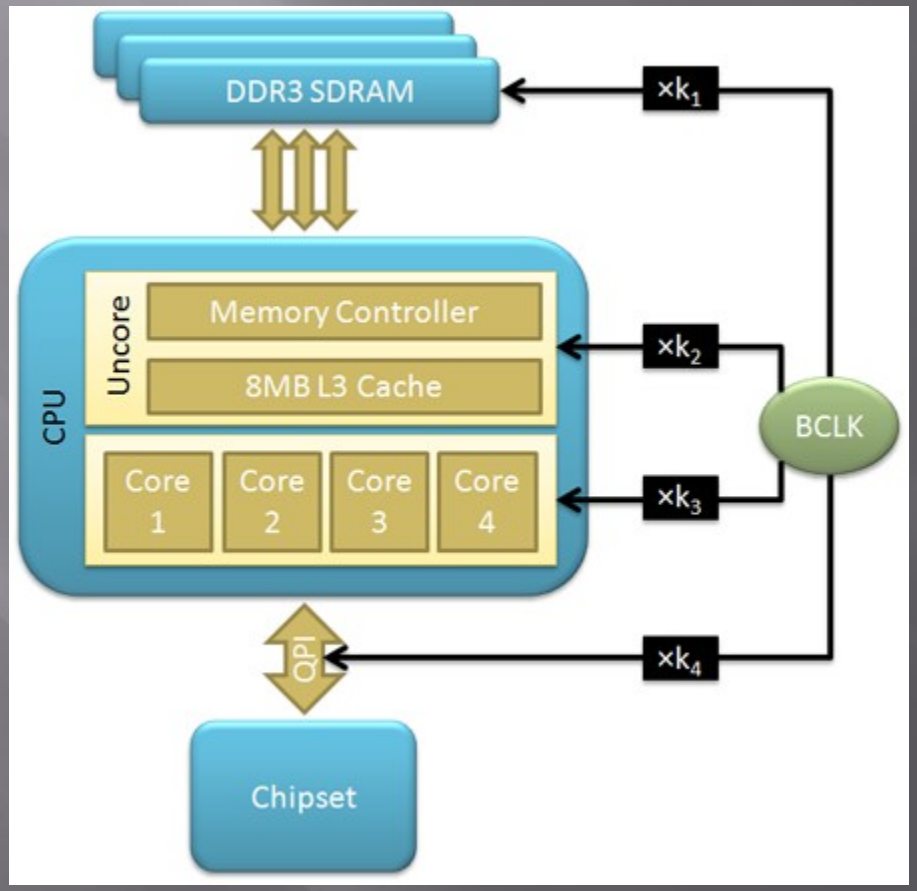
Now & Future

- Bandwidth “out of the box”
  - InfiniBand roadmap



Now & Future

- Bandwidth “inside of the box”
  - Intel i7 processor



References

[1] J.Cowan,C.Madison,G.Still,D.Garcia,M.Bradley & K.Potter, “PI”,*Proceedings of the The 6th International Conference on Parallel Interconnects, p.238, 1999.*

- [2] T.Heil,“InfiniBand Adoption Challenges”, *InfiniBand: A Paradigm Shift From PCI.1 Jun. 2000.*
- [3] Mellanax Technologies Inc,“Introduction to InfiniBand”,*White Paper, 2003.*
- [4]T.C.Jepsen, “InfiniBand”,*Distributed Storage NetworksArchitecture,Protocols and Management,p.159-174,2003.*
- [5]T.M. Ruwart, “InfiniBand – The Next Paradigm Shift in Storage”,*18th IEEE Symposium on Mass Storage Systems and 9th NASA Goddard Conference on Mass Storage Systems and Technologies,17 Apr.2001*
- [6] Wikipedia, “InfiniBand”, *http://en.wikipedia.org/wiki/InfiniBand, 18,Apr.2009*
- [7] G.Huston, “TCP Performance”, *The Internet Protocol Journal - Volume 3, No. 2,2009*
- [8] H.T.Kung&R.Morris, “Credit-Based Flow Control for ATM Networks” *IEEE Network Magazine, March 1995.*
- [9] C.Eddington, “InfiniBridgeTM: An Integrated InfiniBand Switch and Channel Adapter”. *http://mellanox.com/ 19 Apr. 2009*
- [10] Oracle,“Achieving Mainframe-Class Performance on Intel Servers Using InfiniBand Building Blocks” *An Oracle White Paper, April 2003.*
- [11] NCSA, “Latency Results from Pallas MPI Benchmarks”*Virtual Machine Interface 2.1 23 Mar. 2005*
- [12] Mellanax Technologies Inc, “InfiniBand™ Frequently Asked Questions”, *White Paper, 2003.*
- [13] S.Shelvapille&V.Puri, “Encapsulation Methords for Transport of InfiniBand over MPLS Networks*”, Internet-Draft, 12 Mar,2009*
- [14] Cisco Systems, “Quality of Service”,*Internetworking Technology Handbook(http://www.cisco.com/en/US/docs/internetworking/technology/handbook/QoS.html), 21 Apr,2009*

Questions?