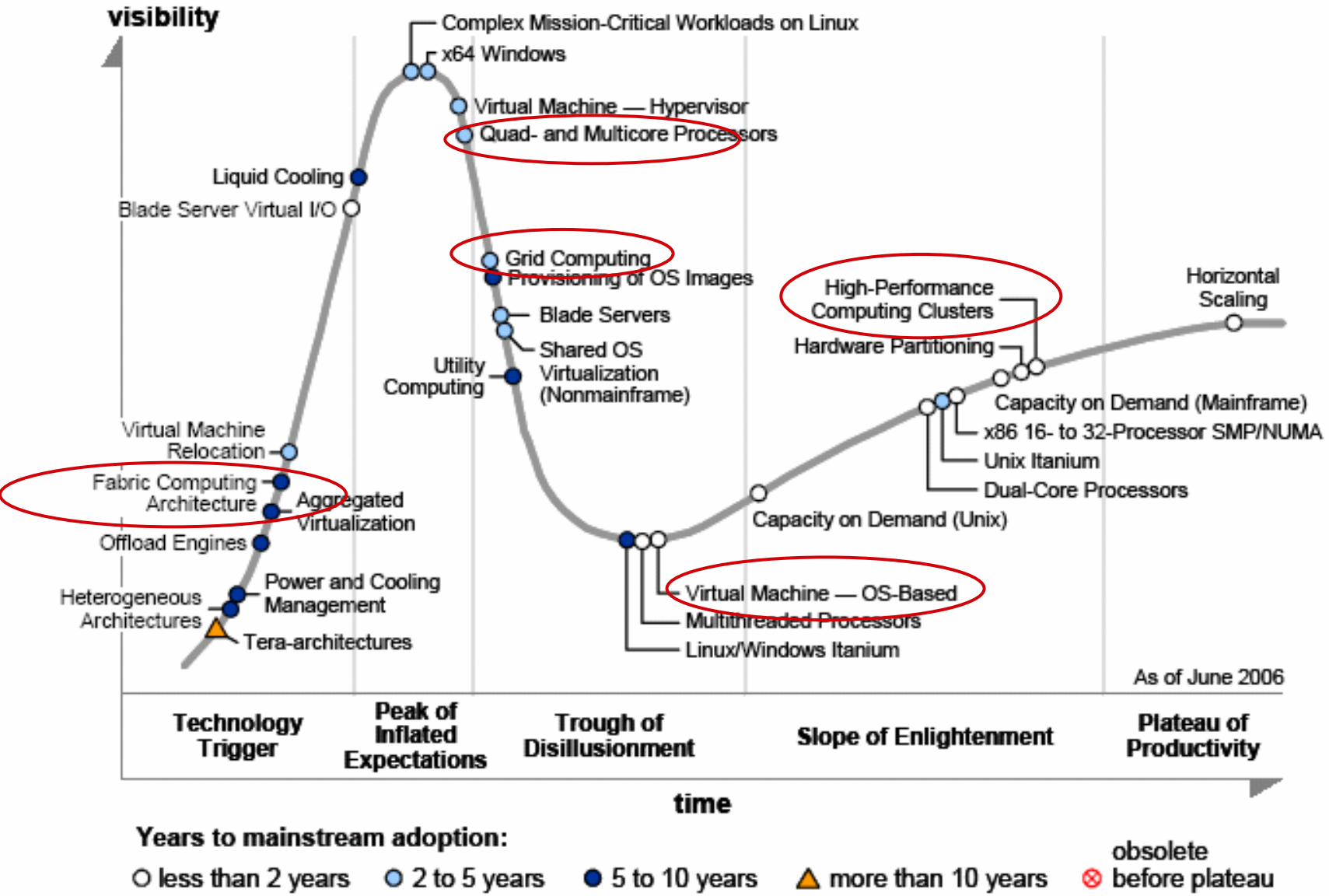
New Low Latency InfiniBand Solutions for Unified Fabrics

Brian Forbes

Director, Technical Alliances– OEM Sales

Figure 1. Hype Cycle for Server Technologies, 2006



Source: Gartner (June 2006)

Gartner

InfiniBand Today

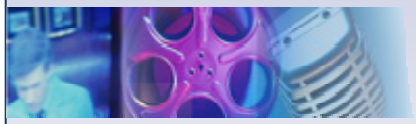
- 20 Gb/s per port, with 40 Gb/s already demonstrated
- End-to-end application latencies approaching 1 microsecond
- Robust cable solutions: copper, active copper, optical, long-haul
- Standard in Linux and Windows, DB2 and Oracle RAC
- Rich set of ULPs: IPoIB, IPoIB-CM, SDP, WSD, RDS, iSER, SRP, NFS/RDMA, uDAPL, multiple MPIs
- Supported by file systems such as Lustre, GPFS, PVS, RapidScale, …
- API convergence with Ethernet RDMA
- Available from many server vendors and integrators
- Hundreds of trained salespeople and field personnel
- Supported by dozens of ISVs
- Several production clusters > 1,000 nodes
- In production use by 100s of commercial customers
- Multiple sites with ≥ 1 Petabyte of IB-attached storage
- Large production clusters using IB for IPC + IP + storage

Solutions For Diverse Vertical Industries

LIFE SCIENCES

SIB
IGIB Institute of Genomics & Integrative Biology
National Institutes of Health
astellas Leading Light for Life

ENTERTAINMENT



MTV Music Television
HBO It's not TV. It's HBO.
Technicolor
VIACOM
SONY

FINANCIAL SERVICES

JPMorgan
NYSE
ISE
Merrill Lynch

GOVERNMENT

National Security Agency
Department of Defense
Los Alamos National Laboratory
Lawrence Livermore National Laboratory
Sandia National Laboratories

RESEARCH

zaik/rrzk
Tokyo Institute of Technology
University of London
CERN
CLEMSON UNIVERSITY

MANUFACTURING

Mitsubishi Motors
Ford
mazda
LG Electronics Inc.
AIRBUS
HONDA
Nippon Steel Corporation

OIL & GAS

Sinopec
PDVSA
CITGO
Schlumberger
Shell
EXXON

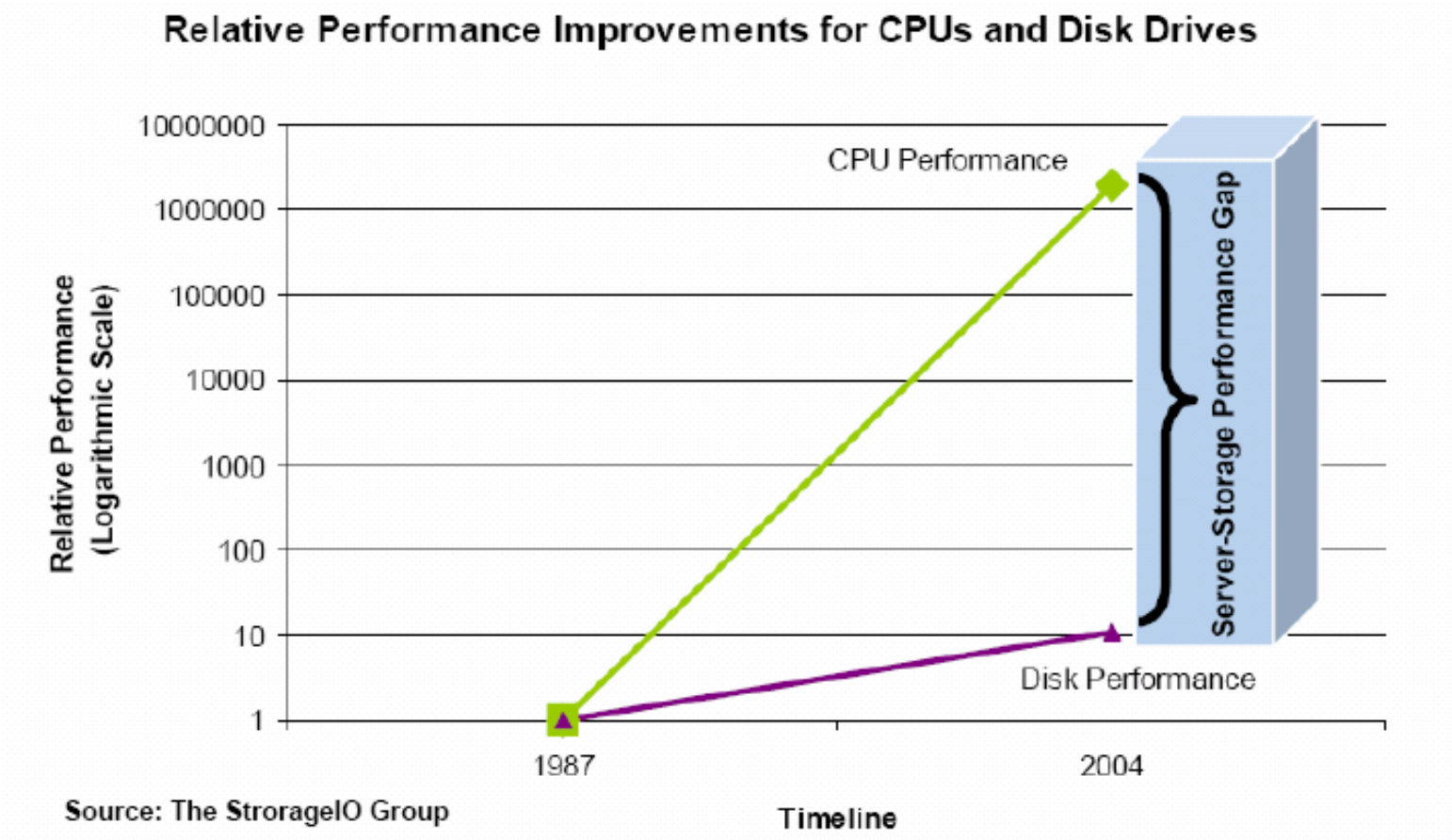
EDA Example (Chip Design)

- Very large EDA company: Synopsys
- Distributed application with significant data set sizes
  - Physical Layout Data up to 300GB
  - Post Optical Proximity Correction 300GB - >1TB
  - Inter-process communication “small” compared to file I/O
  - Uses IP instead of MPI
- Tried several approaches to accelerate performance
  - Non-blocking GigE + dedicated NFS servers
  - Myrinet + GPFS (over IP)
  - InfiniBand for IPC
- 10 Gb/s InfiniBand + Lustre
  - Storage performance: 17x GigE and 6x Myrinet

“…fantastic performance…”

I/O Performance Gap

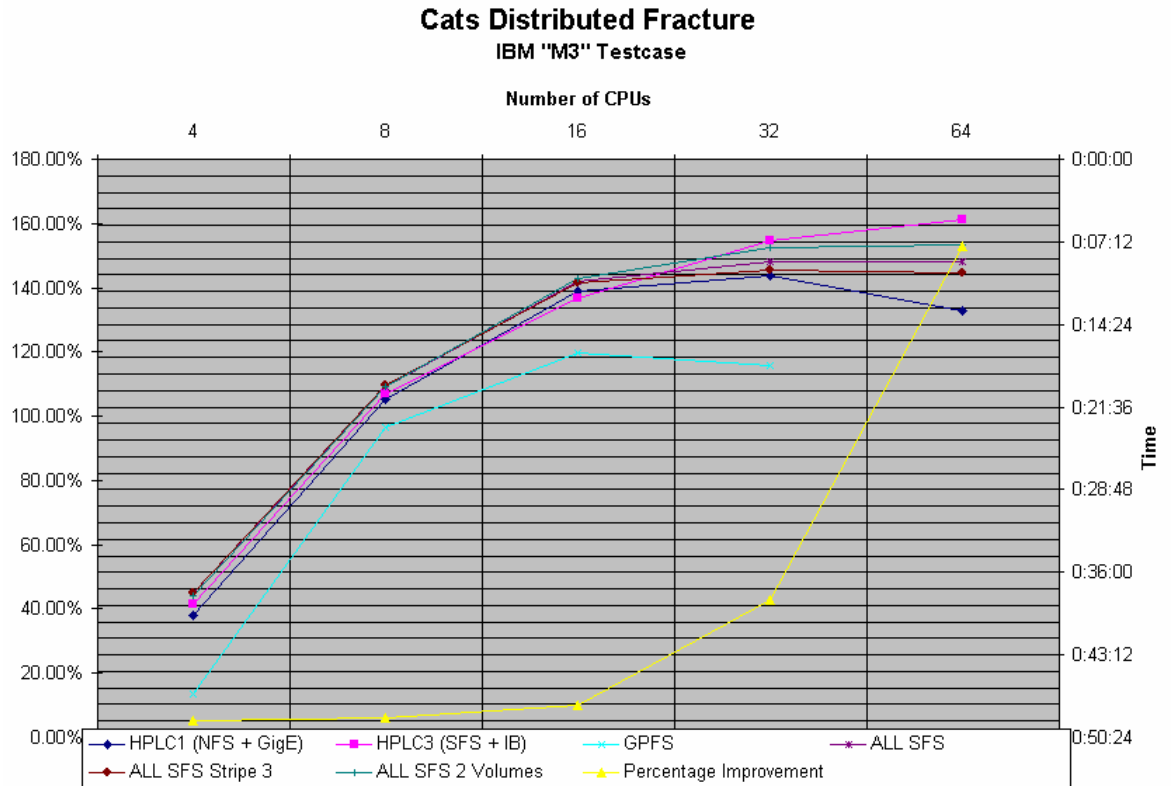
Server-Storage Performance Gap Increases



EDA Example (continued)

- "Our Customer facing Engineers typically see
  - **10x improvement for post layout tools over Fast Ethernet**
  - **3x improvement for post OPC tools over GigE NFS, or direct attached storage…"**

- *Maximum read bandwidth for IB + Lustre is 250 MB/sec, compared to 90 MB/sec with GigE + NFS*



Source: InfiniBand in EDA, Synopsys, OpenFabrics Alliance presentation, May 2007

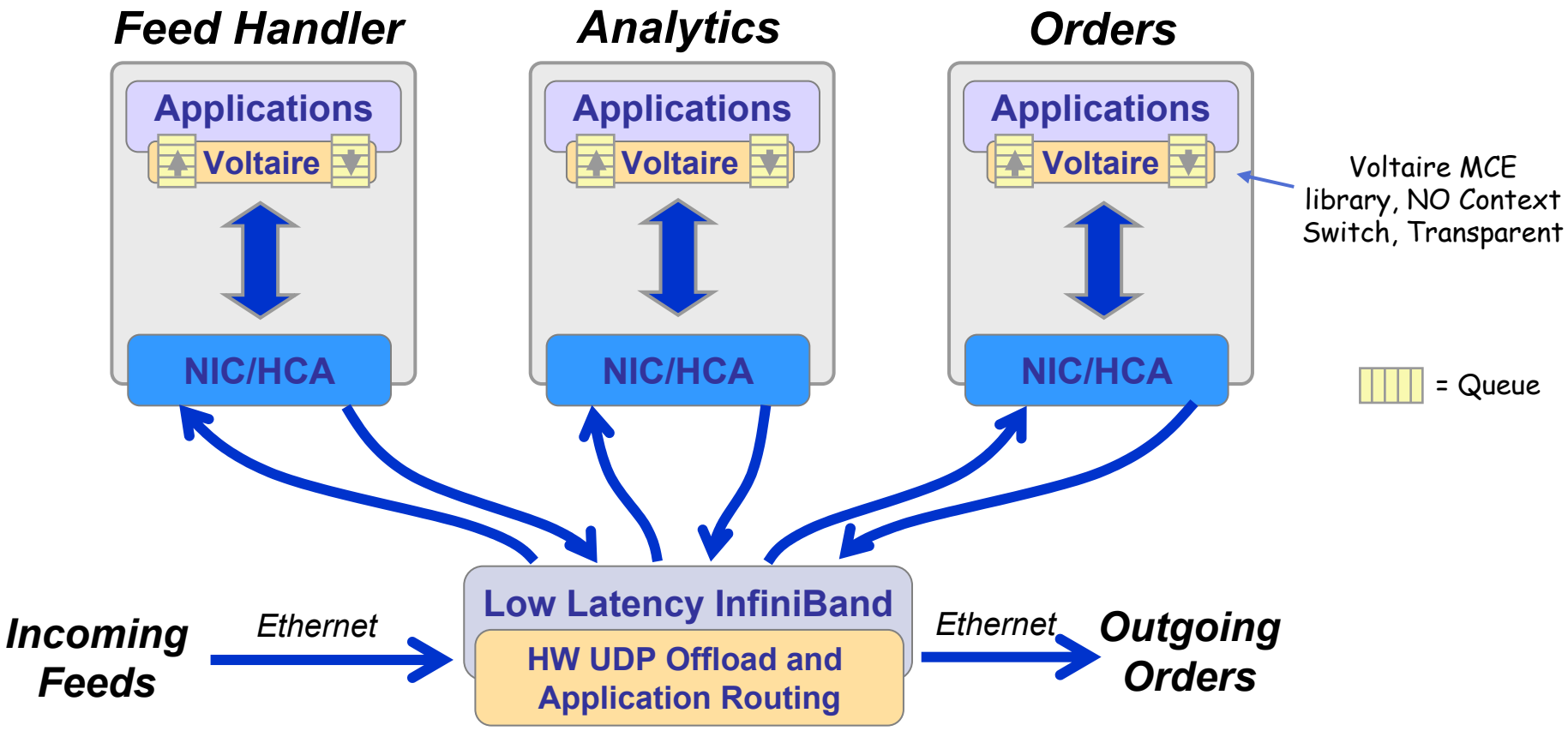
Voltaire High Performance Storage Gateway

- DDR (20 Gb/s) InfiniBand to Fibre Channel
  - iSER to FCP
  - Also supports SAS and SATA
- Four 4 Gb/s FC ports
- ~1.5 GBytes/second aggregate
  - Up to 1 GB/s to a single client
  - ~50K I/Os per second
- Available 4Q07
- Aggregate throughput as high as 2.5 GB/s expected in 2008

Financial Services Example: Market Data Analysis

- Most financial market transactions are generated by software
  - 'Algorithmic trading'
- Time is money – literally
  - Elapsed time between receiving market data and sending a response out is critical
  - Saving a few microseconds can be worth millions of dollars
- Applications have traditionally been sockets-based
  - Ethernet network latency has become a serious limitation
  - CPU cycles per message for IP are high

Voltaire Market Data Solution – Data Flow



Hardware offload, and unique transparent OS Bypass library, gain up to 3-5X better capacity and latency

Improving Messaging Applications With InfiniBand

Benchmark of two applications exchanging multicast messages using standard UDP Multicast sockets APIs (unmodified, using IPoIB MCE)

PPS	Latency	
150K	110 us	GbE — GbE
580K	60 us	GbE — Voltaire — IB
580K	9 us	IB — Voltaire — IB
580K	25 us	IB — Voltaire — GbE — Voltaire — IB

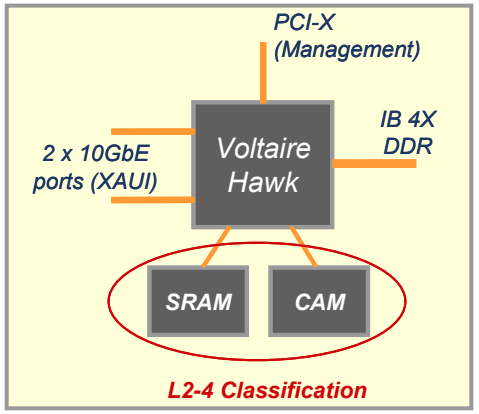
Connecting two IB islands over GbE

Voltaire delivers more than 2X messages, at 10X lower latency without modifying the application

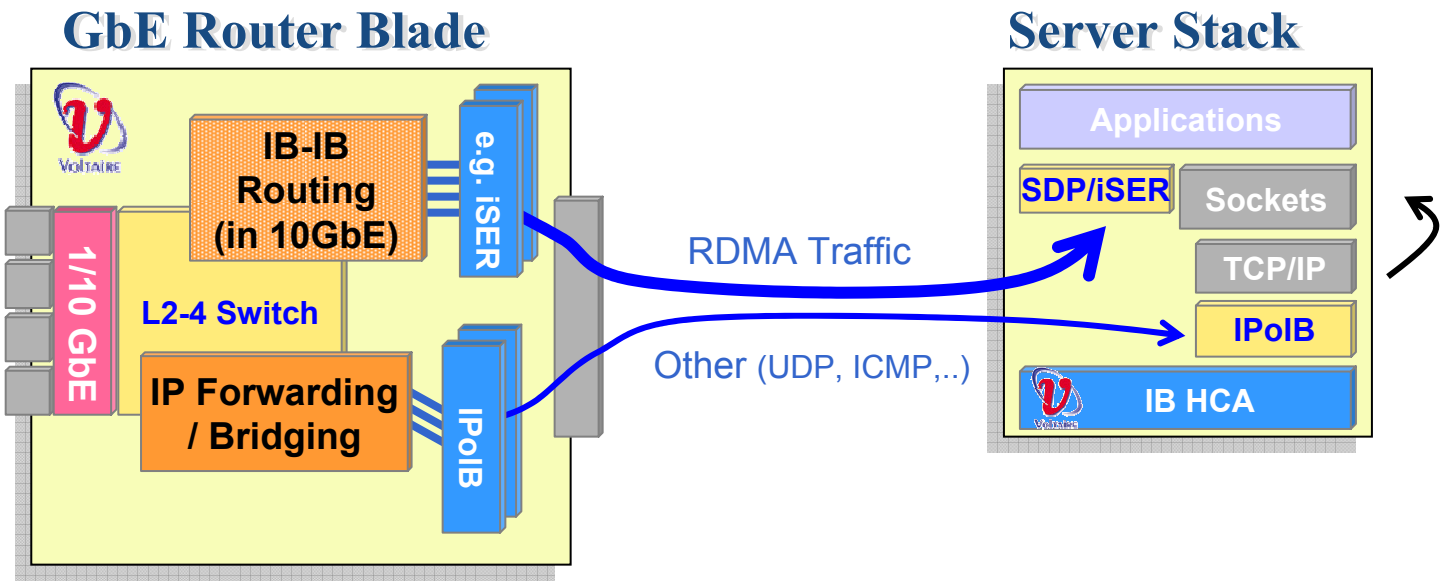
Voltaire 10GbE-IB Integrated Silicon

- Highest density and performance
  - Single chip bridging 2 x 10GbE to IB DDR
  - **15M PPS**, ~ **1.5us latency**
- Most advanced capabilities
  - Transparent **layer 2,3,4+** switching, IO Virtualization
  - Bridging IP/GbE to **IP over IB**
  - and/or tunneling IB packets over IP (**IB-IB routing, Future**)
  - IP & UDP **offloading**
- Completely standard and agent-less
  - Use standard IETF (IP, IPoIB) protocol

Voltaire 10GbE to IB System on a chip



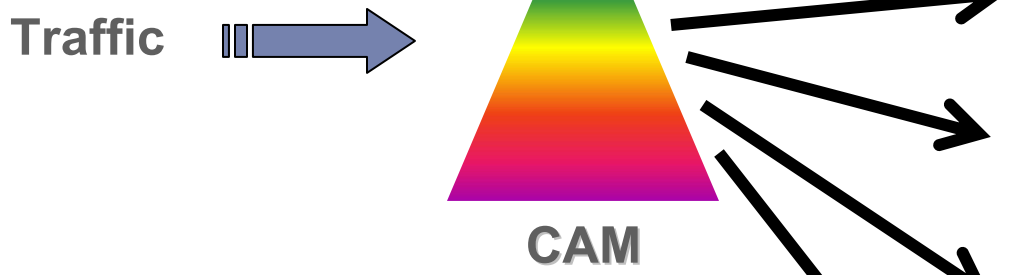
IP Router Logical Architecture



- Layer 2-4 switch core allow flexible classification and routing
- Forwarding (IP Tunneling) mode or IB-IB routing
- Multiple ASICs can be clustered to form larger router
- Viewed as a virtual Ethernet Switch (L2) to the network/user
  - Support 802.3ad, Multicast & IGMP Snooping, VLANs (tagged/untagged), counters, ..

Voltaire ASIC Layer 2-4 Classification

ASIC takes major L2-4 fields to a Ternary wire speed CAM, and gets a result token determining the required action (on that Packet)



Classified Fields (Search Key)

L2: MAC, VLAN, Port#, EtherType

L3: Ver, TOS, Protocol, S/D IP, Options

L4: TCP/UDP ports, State, FLAG's, ICMP Type

Routing Mode

Tunneling / IB Routing * / Offload

Partitioning

Destination Partition / VLAN

Routing

NAT / NAPT / Load-Balancing *

Multi-Pathing

QoS

Priorities, Credits, Tag

Security (Packet Filter)

Drop / Slow-path / Forward

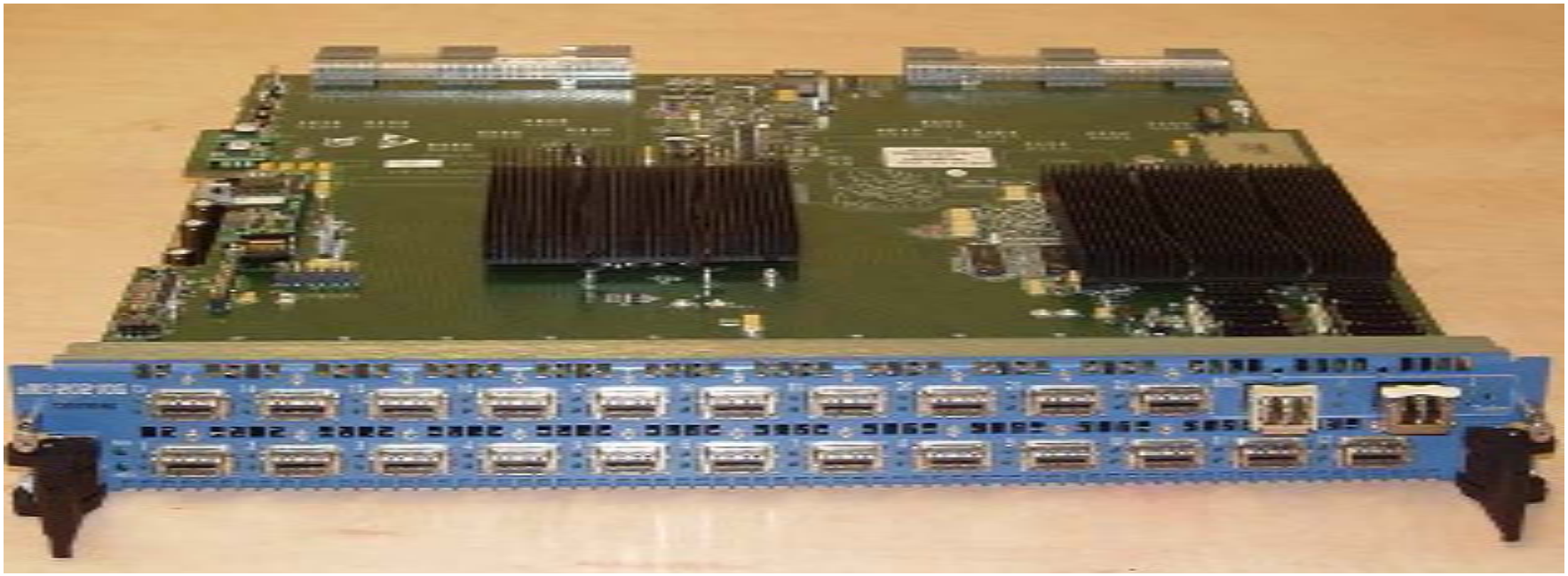
Monitoring

Programmable hardware counters

Event notification

* Future enhancement

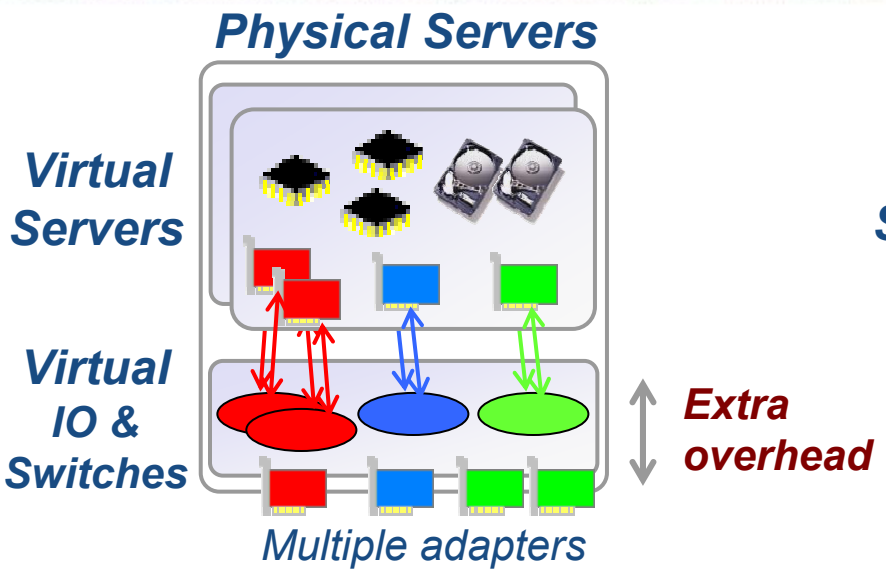
Voltaire sRB-20210G Module

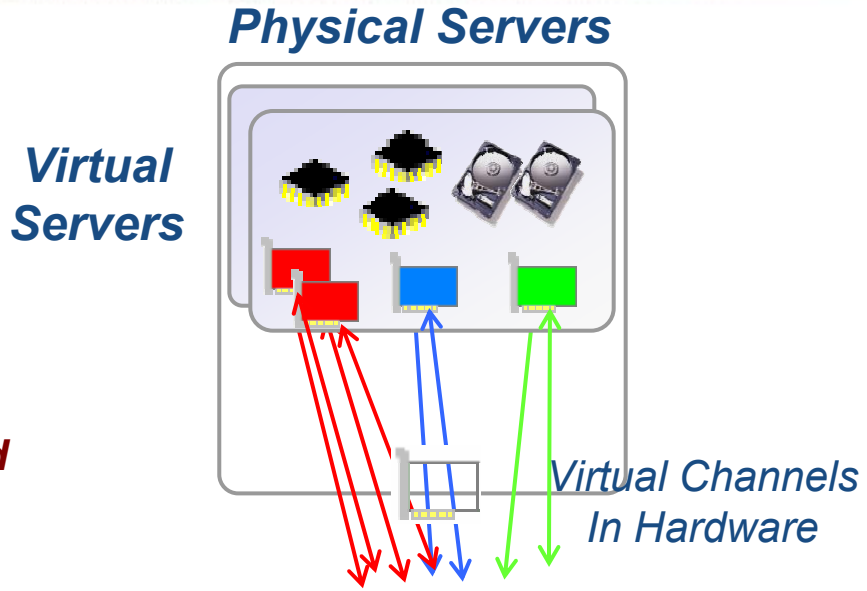


- 22 DDR (20Gb) IB Ports, 2 10GbE XFP ports
- Available 1Q08

THANK YOU

IO & Fabric Virtualization to address VM Bottlnecks

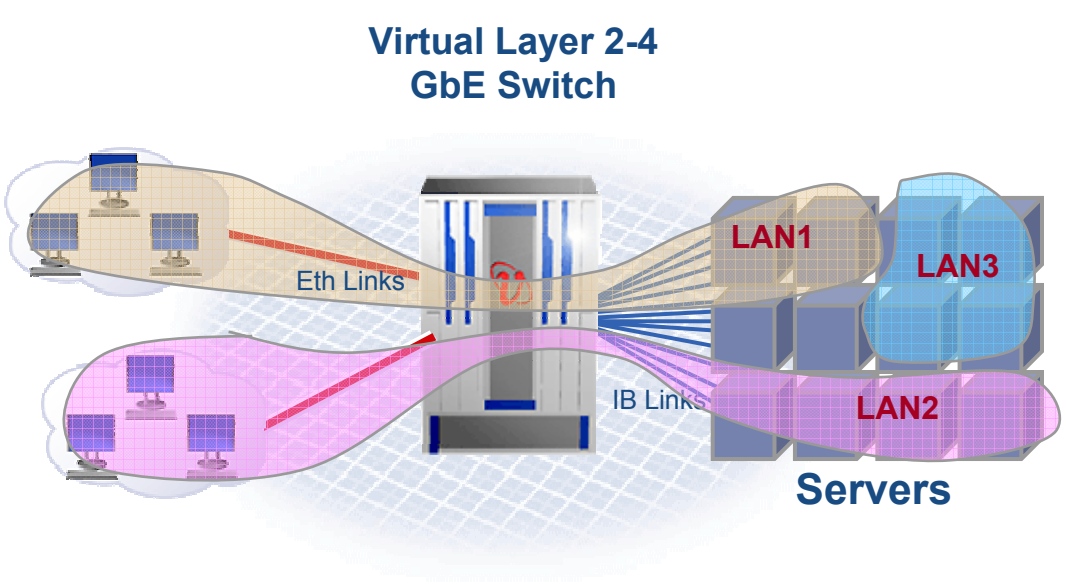


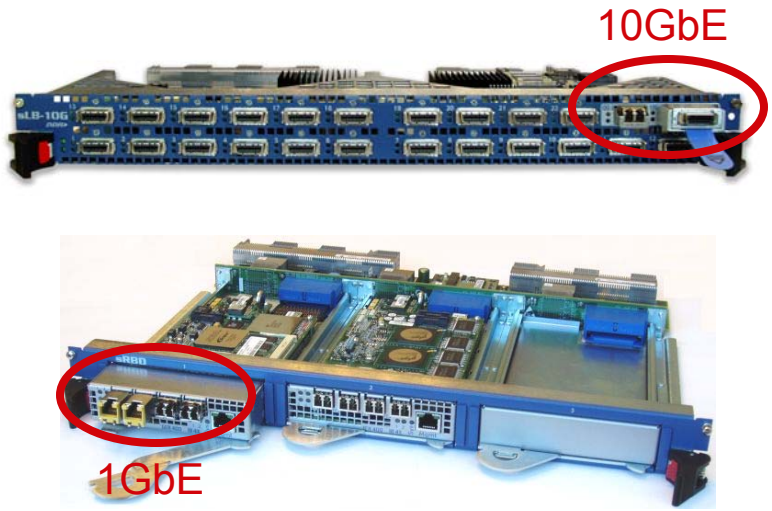


- Slower I/O due to software
- No isolation
- Multiple cards and fabrics
- Not integrated with network/fabric provisioning

- Fast, Direct HW access for I/O
- Single 20Gb/s card for Network, Storage, and IPC
- IO Virtualization in hardware
- Migrate/replicate a server with all its Network environment

Voltaire Network Virtualization with IP Routers





- Enable to build few secured “Ethernet” domains on the same fabric
- Enforced by Hardware, with optional layer 4+ capabilities
- Each node can belong to one or more partitions
- Capable of having full or partial membership (Unique to IB)
  - For communicating with shared resources without compromising security