CIPRICO

Protecting your image

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

April 17, 2001

Presented by

Thomas M. Ruwart

Ciprico, Inc.

Overview

- A brief history of InfiniBand
- A basic overview of InfiniBand
- The IB Paradigm Shift
- The IB Paradigm Shift and Storage
- Summary

Brief history of InfiniBand

- **Future I/O (FIO) was being developed by IBM, Compaq Computer, Hewlett-Packard, 3Com, Adaptec, and Cisco.**
- **Next Generation I/O (NGIO) was being developed by Intel, Dell Computers, Sun, and others.**
- **FIO and NGIO were competing technologies**
- **Neither would “win” so they combined forces to form Serial I/O (SIO) which combined the best of both technologies**
- **The name** *SIO* **could not escape the powerful clutches of the Intel Marketing department and hence was renamed** *InfiniBand Architecture* **or** *IBA* **for short**

What is InfiniBand?

- A technology used to interconnect processing nodes to I/O nodes to form a System Area Network
- Intended to be a replacement for PCI
- Heavily leverages best-of-breed technologies
- For more information or to get the spec for your reading pleasure visit the InfiniBand Trade Association website at:

http://www.infinibandta.org

What InifiniBand is not....

- InfiniBand is not a replacement for Ethernet.
- InfiniBand is not a wide area network. It is intended to be used within a computer room facility (< 100 meters diameter)
- InfiniBand is not a replacement for Fibre Channel

The Problem

- The need for a cost-effective interconnect technology for building *clusters*
- Bus-based architectures (i.e. PCI) are limited to a single host system and cannot easily extend beyond the confines of the "box"
- Bandwidth and Latency between boxes using existing system area networks are limited and/or expensive
- Not clear that bus-based technologies can scale in bandwidth as easily as serial technologies

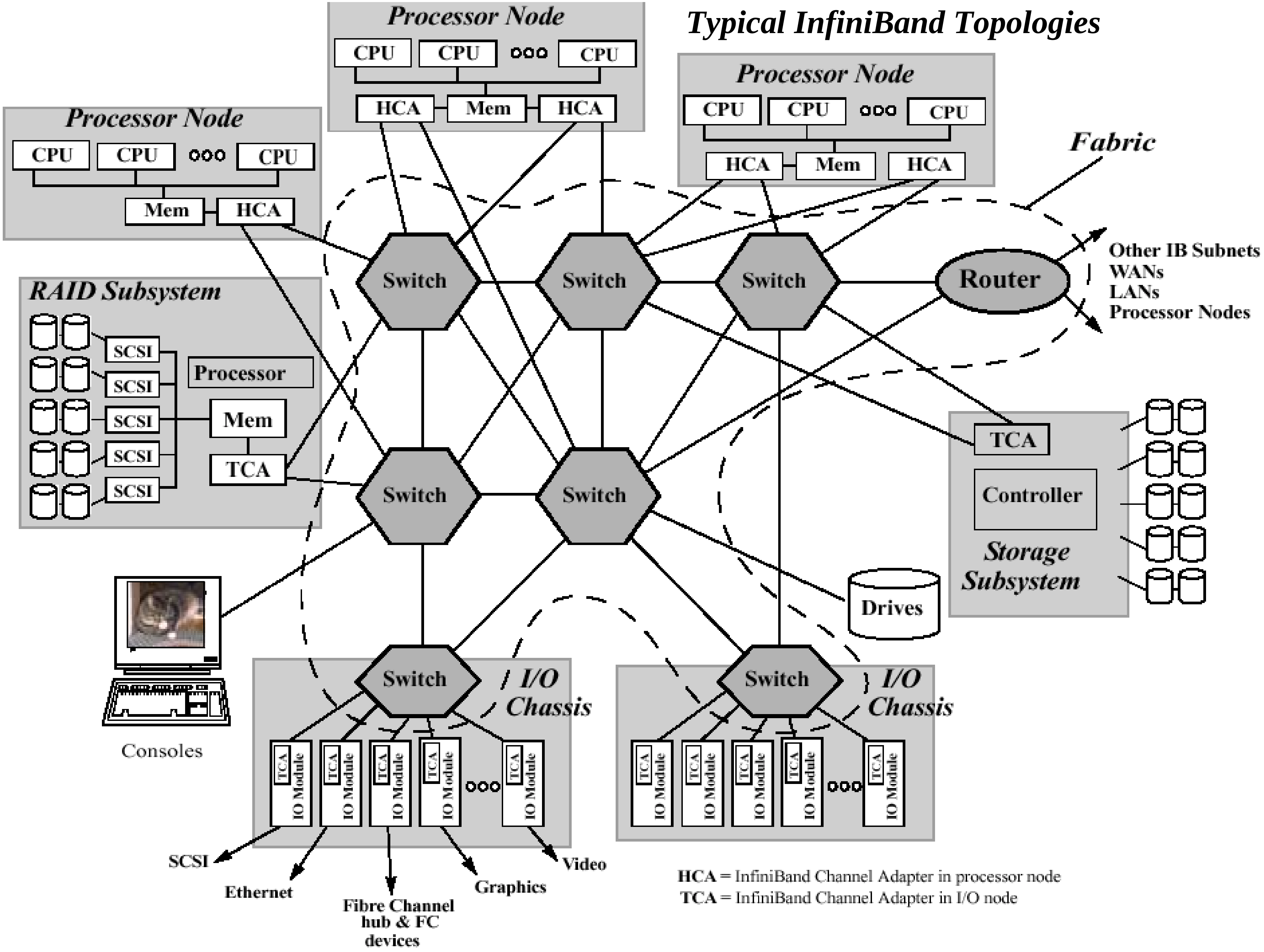
The Solution - InfiniBand

- **Network-based Architecture**
- **Serial communication technology**
- **Supported by a very large consortium – 220 members in the IBTA to date**
- **Targets a volume market in order to take advantage of economies of scale**
- **Not a new technology but rather IB is built on Best-of-Breed technologies**

Best of Breed Technologies

- Fibre Channel levels 0, 1, and 2
  - FC0 – Physical level – cable, connectors, signaling frequency
  - FC1 – 8b/10b Encoding
  - FC2 – Framing protocol, ECC, CRC, Header format
- Addressing – IPV6-type 128-bit addresses
- Software API – Virtual Interface Architecture (VIA)

Typical InfiniBand Topologies



HCA = InfiniBand Channel Adapter in processor node
TCA = InfiniBand Channel Adapter in I/O node

Transport Services

- **Reliable Connection – Acknowledged, connection oriented**
- **Unreliable Connection – Unacknowledged, connection oriented**
- **Reliable Datagram – Acknowledged connectionless, multiplexed transmission**
- **Unreliable Datagram - Unacknowledged connectionless, multiplexed transmission**
- **Raw Datagram – Unacknowledged connectionless**

IBA Features

- Zero Copy data transfers – direct user buffer to user buffer data transfer
- High bandwidth
    - 1x, 4x, 12x, and 32x defined
    - X = 2.5Gbit/sec single link speed on initial release
    - 1x and 4x parts are currently being developed
- Low latency – on the order of 10-40 microseconds initially
- Low overhead – Very little kernel involvement
- Memory protection
- Congestion management
- Hot-plug, auto-discovery and configuration subnet management
- Cost effective

Basic Commands

- Send / Receive operations for low overhead, low latency short messages between nodes
- Memory Binding operations for memory protection and security
- RDMA Write and Read operations that perform low overhead, high bandwidth data movement between nodes
- Atomic operations such as Test-and-Set, Increment, …etc. required primitives for resource sharing techniques such as remote semaphores and distributed lock management

Virtual Lanes and Partitioning

- Virtual Lanes (VL)
    - A VL is a send/receive pair of buffers
    - Up to 16 VL’s on each physical channel
    - VL15 is always present for subnet management
    - 15 remaining VL’s are used for data
- Partitions
    - Defines which nodes can communicate with each other

IBA Components

- Hardware
  - Host Channel Adapters (HCA)
  - Target Channel Adapters (TCA)
  - Switches
  - Routers
- Software
  - Subnet Managers
  - Service Agents
- Glueware: Virtual Interface Architecture (VIA)

Channel Adapters

- Channel adapters in general
  - Only two kinds of Channel Adapters: Host and Target
- Host Channel Adapters (HCA)
  - Very Intelligent
  - Capable of handling large numbers of concurrent connections
  - Typically have a large number of send/receive buffers
- Target Channel Adapters (TCA)
  - Not as much intelligence as HCAs due to the limited scope of their function
  - Need only handle a small number of concurrent connections
  - No as much send/receive buffer space as an HCA

Why do I need InfiniBand?

- Envy - Because everyone else will have it
- Allows for clustering – the next wave in high-performance computing and data-intensive I/O
- Cost effectiveness
- Allows for new and interesting storage architectures and storage/system area network topologies
- Because InfiniBand represents a significant ***paradigm shift*** in the storage industry that will present many new opportunities

The Paradigm Shift

- **In the mid-1990's Fibre Channel brought with it the capability to physically share a storage device between multiple computers.**
- **File Systems of that time were not designed to use that capability and it took 5-6 years for "shared" file systems to appear**
- **InfiniBand provides the capability to physically share (tightly couple) computing and other resources between multiple "nodes" on the IB network.**
- **Operating Systems of today are not designed to use this capability – but at least we can learn from our past experiences with Fibre Channel – can't we?**

InfiniBand and Storage

- **Storage devices will become “peers” within the system fabric instead of peripherals**
- **Lower communication latencies between the storage subsystem and other nodes**
- **Employing IB as the interconnect between a computing subsystem and a storage subsystem however implies new communication protocols between the two subsystems**
- **InfiniBand allows for *Extensible* Storage Architectures**

Extensibility

- **Density – the number of bytes/IOPS/bandwidth per unit volume**
- **Scalability – what does that word really mean?…**
    - **Capacity: number of bytes, number of objects, number of files, number of actuators …etc.**
    - **Performance: Bandwidth, IOPs, Latency, …etc.**
    - **Connectivity: number of disks, hosts, arrays, …etc.**
    - **Geographic: LAN, SAN, WAN, …etc.**
    - **Processing Power**
- **Cost – address issues such as $/MB, $/sqft, $/IOP, $/MB/sec, TCO, …etc.**
- **Adaptability – to changing applications**
- **Capability – can add functionality for different applications**
- **Manageability – Can be managed as a system rather than just a box of storage devices**
- **Reliability – Connection integrity capabilities built into IB**
- **Availability – Fail-over capabilities built into IB**
- **Serviceability – Hot-plug capability built into IB**
- **Interoperability – Supported by many vendors and Interoperability is a key issue being addressed at IB Specification time rather than after product rollout**
- **Power – decrease the power per unit volume**

Timescales

- **InfiniBand Specification 1.0 was released October 2000.**
- **InfiniBand sample parts are available from a variety of manufacturers**
- **Intel demonstrated 1x IB at the Intel Developers Forum in Feb 2001**
- **Production volume shipments of simple IB-enabled systems in 12-18 months**
- **More complex IB systems are likely to be 24 months out**

Summary

- IB is the next paradigm shift in the computing industry
- IB has tremendous potential and presents enormous opportunities for system and storage vendors
- IB has tremendous momentum due to it industry-wide backing
- IB is going to happen – it is just a question of “how soon”

CIPRICO

Protecting your image