

ICCA'03

High Speed I/O Server Computing with InfiniBand

José Luís Gonçalves
zeluis@ipb.pt

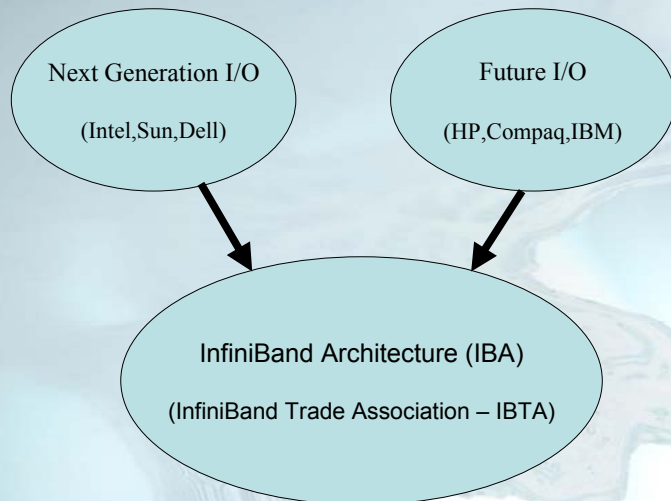
High Speed I/O Server Computing with InfiniBand

Agenda

- Breve descrição da Arquitectura
- Camadas Funcionais
- Protocolos Nativos
- InfiniBand e HPC
- Presente e Futuro da tecnologia
- Conclusão

High Speed I/O Server Computing with InfiniBand

Como surgiu ?



High Speed I/O Server Computing with InfiniBand

Arquitectura InfiniBand

- **O que é?**
 - Nova tecnologia de I/O
 - Novo hardware
 - Novas motherboards
 - Novos chipsets
 - Novos hardware de suporte (backplane)
 - Novos conectores
 - Novo software
 - Ao nível do S.O.
 - Gestão do tecido InfiniBand

Arquitetura InfiniBand

• Que problemas pretende resolver

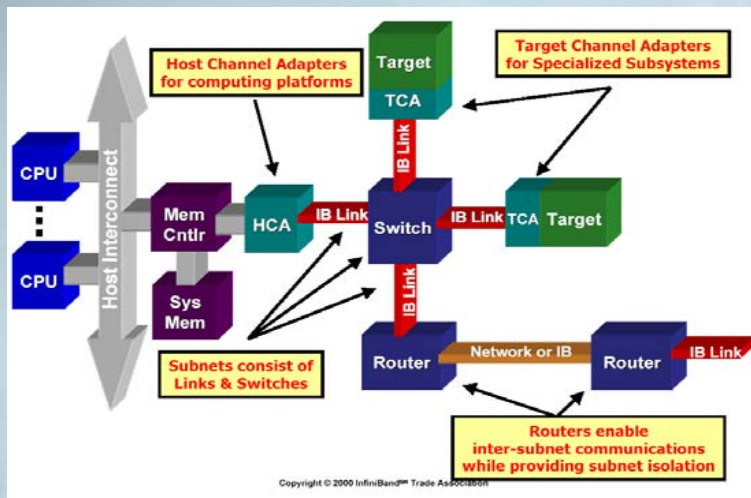
- Sistemas em bus evoluíram muito pouco
- I/O continua a ser problemático
- Dotar os sistemas de uma modularidade eficiente
- Diferentes tecnologias para interligar sistemas (PCI/Ethernet/FC)
- Menos Cablagem

Arquitetura InfiniBand

Onde se pode aplicar

- Data Centers
- Clusters
- Sistemas de alta disponibilidade/performance
- Adapta-se:
 - Inter Process Communication (IPC)
 - Comunicações ponto-a-ponto
 - Storage I/O

Componentes da Arquitectura



Camadas da arquitectura

- Física
 - Fibra óptica e cobre
 - Distância não definida (atenuação de 15 db).
- Ligação
 - Layout dos pacotes
 - Pacotes de dados e gestão
 - Virtual Lanes e QoS
- Transporte
 - Transmissão fiável
 - Transmissão não fiável
 - Sequenciação/Correcção de Erros/retransmissão Implementada em Hardware

Protocolos Nativos da Arquitectura

- Remote Direct Memory Access (RDMA)
 - Elemento chave da arquitectura.
 - Elimina a necessidade de Protocolos “pesados”
 - Possibilita a leitura/escrita directamente na memória
 - Mensagens : Baixa latência
 - Dados : Elevado débito
- Direct Access File System (DAFS)
 - Assenta sobre RDMA
 - Alto débito
 - Potencia/Recicla a utilização de NAS

Protocolos Nativos da Arquitectura

- Socket Direct Protocol (SDP)
 - Manter Compatibilidade
 - Melhor Débito/Latência
 - Ligação LAN/WAN tecido InfiniBand
 - Recorre ao IPoB como suporte ao transporte.
- IP over InfiniBand
 - Transforma tráfego IP em Mensagens InfiniBand
 - Em conjunto ou não com SDP
 - Previsto suporte para QoS em futuras especificações

InfiniBand e outras tecnologias

Feature	InfiniBand™	PCI-X	Fibre Channel	1Gb & 10Gb Ethernet	Hyper-Transport™	Rapid I/O
Bus/Link Bandwidth	2.5/10/30Gb/s ^a	8.51 Gb/s	1/2.1 Gb/s ^b	1 Gb, 10Gb	12.8, 25.6, 51.2 Gb/s ^d	16/32 Gb/s ^c
Bus/Link Bandwidth (Fully Duplexed)	5/20/60Gb/s ^a	Half-Duplex	2.1/4.2 Gb/s ^b	2 Gb, 20Gb	25.6, 51.2, 102 Gb/s ^d	32/64 Gb/s ^c
Pin Count	4/16/48 ^d	90	4	4, Fiber	55,103,197 ^f	40/76 ^e
Maximum Signal Length	Km	Inches	Km	Km	Inches	Inches
Transport Media	PCB, Fiber and Copper Cable	PCB only	Copper and Fiber Cable	Copper and Fiber Cable	PCB only	PCB only
Simultaneous Peer to Peer communication	15 VLS + Mgmt Lane			X		3 Transaction Flows
Native Hwd Transport Support with Memory Protection	X					
In-Band Management	X		Uses out-of-band mgmt	Not native, can use IP		
RDMA Support	X					
Native Virtual Interface Support	X					
End-to-End Flow Control	X			X	X	X
Memory Partitioning ^g	X		X			
Quality of Service	X		X	Limited		X
Reliable	X		X		X ⁱ	X
Scalable	X		X	X	X	X
Maximum Packet Payload	4 KB	Not Packet Based	2 KB	1.5KB (10Gb no jumbo support)	64 bytes	256 bytes

InfiniBand e High Performance Computing

- Largura de Banda

InfiniBand Links and Data Throughput Rates			
Link	Signal Rate (uni-directional)	Useable Bandwidth (80% of signal rate)	Effective Data Throughput (send + receive)
1-wide	2.5 Gb/s	2 Gb/s (250 MB/s)	(250 + 250) MB/s
4-wide	10 Gb/s	8 Gb/s (1GB/s)	(1 + 1) GB/s
12-wide	30 Gb/s	24 Gb/s (3 GB/s)	(3 + 3) GB/s

- Latência

	TCP/IP over Ethernet	TCP/IP with TCP Offload	Mellanox InfiniBand
Useable Bandwidth	~ 400 Mb/s*	~ 850 Mb/s*	~ 1.9 Gb/s
Latency	More than 10s of milli-seconds	More than 10s of milli-seconds	Less than 10s of micro-seconds
Processor Utilization	~90%	~50%	~10%

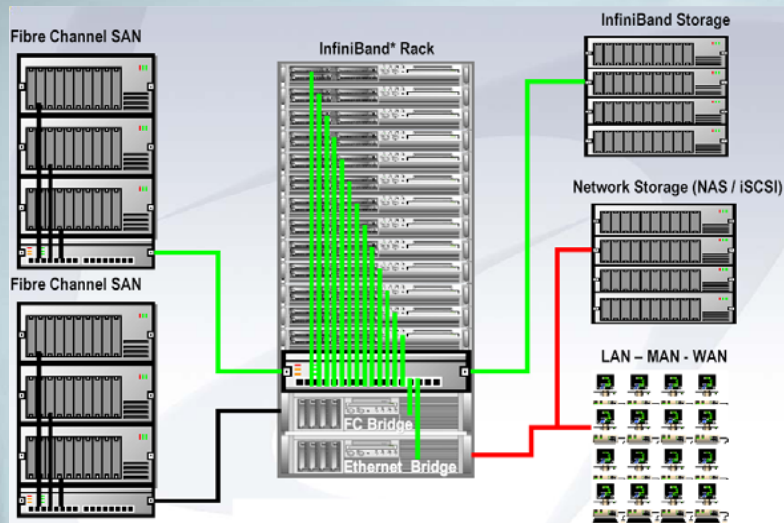
Características relevantes para HPC

- Fail-over automático entre links
- Links redundantes (link inactivo passa activo)
- QoS
- Adicionar links para aumentar largura de banda.
- 64000 nodos por sub rede.

InfiniBand e *Server Blades*



InfiniBand e *Server Blades*



Futuro do InfiniBand

