Na aula passada...

- QoS na Internet
  - Marcação
  - Classificação
  - Policiamento
  - Escalonamento
- Integraded Services
  - Controle por fluxo
  - RSVP
- Differentiated Services
  - Controle por classe
  - PHB

1

Nesta aula...

- MACs de alto desempenho
  - O papel do switching
  - SONET (synchronous optical network)
    - Rede de provedores
  - INFINIBAND
    - Rede de interconexão
- SAN (Storage Area Network)
- Redes em Clusters

2

Provinha – 29.09.2009

Num IDC temos:

• um sistema de webmail com 20 maquinas de front-end e 2 de back-end

•um sistema de máquinas administrativas com 10 máquinas de front-end e 4 de back-end

•um sistema de storage, composto por discos e back-up que serve a todos os back-ends

um cluster de 100 blades, cada uma com 4 processadores

Proponha um sistema de interconexão que atenda as demandas deste ambiente. Coloque redundância entre back-ends e storages. Os servidores poderiam ser aglutinados num mesmo sistema físico de interconexão? Como seria feita a separação lógica? Avalie os problemas de performance que poderão ocorrer na infra-estrutura de conexão, com as redes separadas e juntas.

3

SONET/SDH

Padrão de transmissão de dados criado para alcançar 4 objetivos básicos :

• Interconexão das várias redes de transmissão de dados digitais existentes até então (comprimento de onda, temporização, estrutura de *frames*, etc);

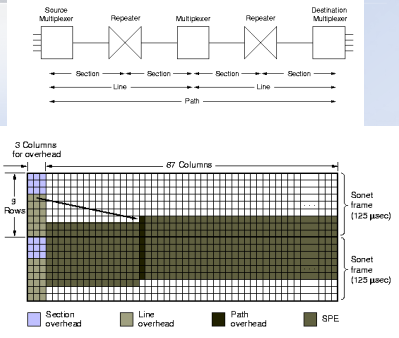
• Unificar os sistemas digitais existentes nos EUA, Europa e Japão;

• Necessidade de multiplexação de vários canais digitais conjuntamente;

• Suporte para operações, administração e manutenção (OAM).

4

SONET/SDH



5

SONET/SDH

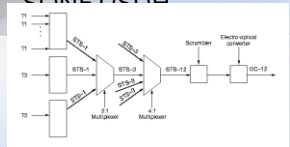


Fig. 2-31. Multiplexing in SONET.

SONET		SDH	Data rate (Mbps)		
Electrical	Optical	Optical	Gross	SPE	User
STS-1	OC-1		51.84	50.112	49.536
STS-3	OC-3	STM-1	155.52	150.336	148.608
STS-9	OC-9	STM-3	466.56	451.008	445.824
STS-12	OC-12	STM-4	622.08	601.344	594.432
STS-18	OC-18	STM-6	933.12	902.016	891.648
STS-24	OC-24	STM-8	1244.16	1202.688	1188.864
STS-36	OC-36	STM-12	1866.24	1804.032	1783.296
STS-48	OC-48	STM-16	2488.32	2405.376	2377.728

Fig. 2-32. SONET and SDH multiplex rates.

6

SONET/SDH

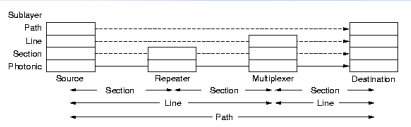


Fig. 2-33. The SONET architecture.

7

SWITCHING

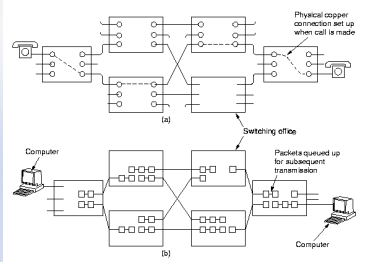


Fig. 2-34. (a) Circuit switching. (b) Packet switching.

8

SWITCHING

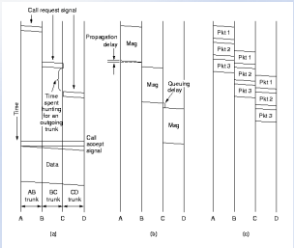


Fig. 2-35. Timing of events in (a) circuit switching, (b) message switching, (c) packet switching.

9

SWITCHING

Item	Circuit-switched	Packet-switched
Dedicated "copper" path	Yes	No
Bandwidth available	Fixed	Dynamic
Potentially wasted bandwidth	Yes	No
Store-and-forward transmission	No	Yes
Each packet follows the same route	Yes	No
Call setup	Required	Not needed
When can congestion occur	At setup time	On every packet
Charging	Per minute	Per packet

Fig. 2-36. A comparison of circuit-switched and packet-switched networks.

SWITCHING

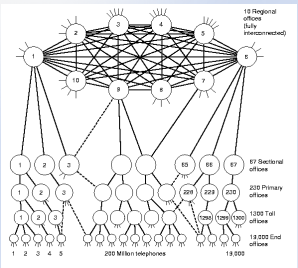


Fig. 2-37. The AT&T telephone hierarchy. The dashed lines are direct trunks.

11

SWITCHING

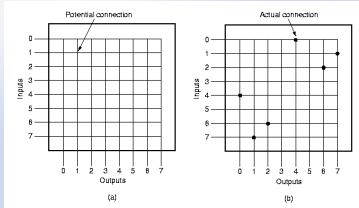


Fig. 2-38. (a) A crossbar switch with no connections. (b) A crossbar switch with three connections set up: 0 with 4, 1 with 7, and 2 with 6.

12

SWITCHING

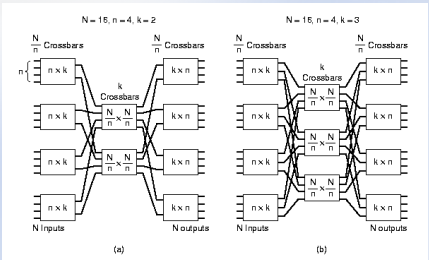


Fig. 2-39. Two space division switches with different parameters.

SWITCHING

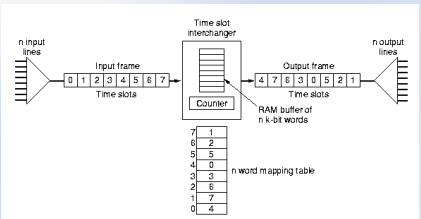


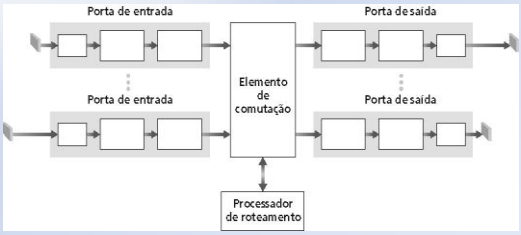
Fig. 2-40. A time division switch.

voltar

Visão geral da arquitetura de um comutador (roteador)

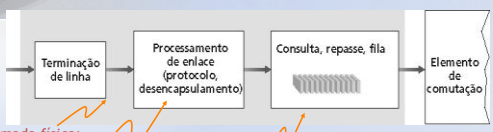
Duas funções-chave do roteador:

- Executar algoritmos/protocolos (RIP, OSPF, BGP)
- **Comutar** os datagramas do link de entrada para o link de saída



15

Funções da porta de entrada



Camada física: recepção de bits

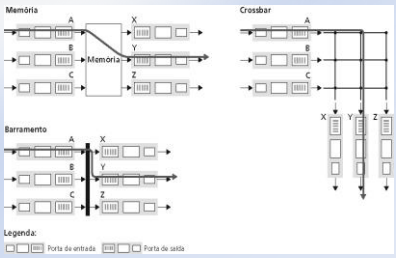
Camada de enlace: ex.: Ethernet (veja capítulo 5)

Comutação descentralizada:

- Dado o destino do datagrama, procura a porta de saída usando a tabela de comutação na memória da porta de entrada
- Objetivo: completar o processamento da porta de entrada na 'velocidade da linha'
- Fila: se os datagramas chegam mais rápido do que a taxa de comutação para o switch

16

Três tipos de estrutura de comutação

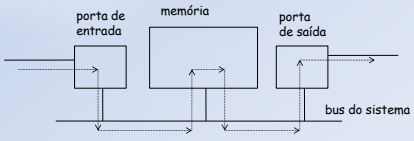


17

Comutação via memória

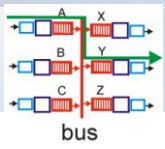
Primeira geração de roteadores:

- Computadores tradicionais com comutação sob controle direto da CPU
- Pacote copiado para a memória do sistema
- Velocidade limitada pela largura de banda (2 bus cruzados por datagrama)



18

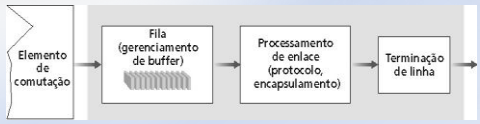
Comutação via bus



- Datagrama da memória da porta de entrada para a memória da porta de saída através de um bus compartilhado
- Contenção do bus: velocidade de comutação limitada pela largura de banda do bus
- Barramento de 1 Gbps, Cisco 1900: velocidade suficiente para roteadores de acesso e de empresas (não para roteadores regionais ou de backbone)

19

Portas de saída

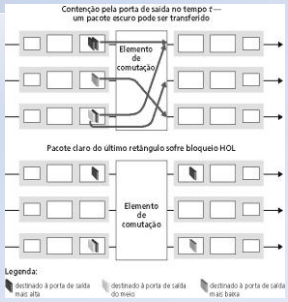


- Buffering necessário quando datagramas chegam do switch mais rápido do que a taxa de transmissão
- Disciplina de agendamento escolhe entre os datagramas na fila para transmissão

20

Enfileiramento na porta de entrada

- Switch mais lento que as portas de entrada combinadas -> pode ocorrer filas na entrada
- Bloqueio Head-of-the-Line (HOL): datagrama na frente da fila impede os outros na fila de se moverem para adiante
- Atraso e perda na fila devido ao overflow no buffer de entrada!



21

Comutação via rede de interconexão

- Supera as limitações de largura de banda do bus
- Redes de Banyan, outras redes de interconexão inicialmente desenvolvidas para conectar processadores em multiprocessamento
- Projeto avançado: fragmentar datagramas em células de tamanho fixo, comutar as células através do switch.
- Cisco 12000: comuta Gbps através da rede de interconexão

22

Universidade de São Paulo
Instituto de Ciências Matemáticas e de Computação

"ARQUITETURA DE COMUNICAÇÃO EM SISTEMAS *HIGH PERFORMANCE COMPUTING*"

- Seminário – Arquitetura de Computadores
- Prof. Eduardo Marques e Prof. Alexandre C. B. Delbem
- Dagoberto Carvalho Junior

Sumário

Introdução e Contextualização

- Os Mainframes utilizavam políticas de time-sharing ou batch para executar processos
- Houve uma evolução dos computadores pessoais
- Dobravam sua capacidade de processamento a cada 18 meses (Lei de Moore)
- Computadores pessoais foram ligados em rede e os processos foram distribuídos (Cluster)
- Os clusters são caracterizados como sistemas computacionais de alto desempenho (HPC)

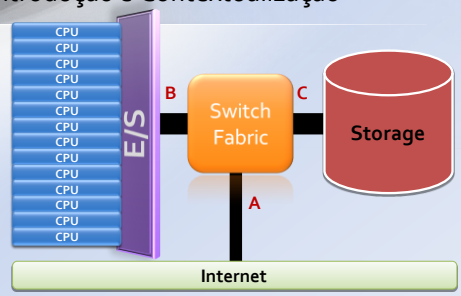
Introdução e Contextualização

- Os clusters normalmente são instalados em ambientes com infra-estrutura adequada
- Os Data Centers (DCs) abrigam os sistemas de HPC
- Desempenho é um fator importante em HPC
- Operações de I/O (Entrada e Saída) é um ponto importante de degradação de desempenho em HPC

Introdução e Contextualização

- Tecnologias específicas para a comunicação dos nós em HPC foram desenvolvidas
- Myrinet e Infiniband são duas tecnologias específicas
- O Ethernet propicia aplicações mais amplas

Introdução e Contextualização



Arquitetura de um Fabric em um DC

Introdução e Contextualização

- A programação paralela é o instrumento para explorar o poder de processamento do cluster
- MPI (Message Passive Interface) é a principal forma de gerenciamento de processos (Gropp et al., 1999).
- MPI colaborou para a exploração de novas tecnologias de comunicação
- Infiniband recentemente foi proposta como arquitetura não proprietária de próxima geração (Infiniband Trade Association, 2009)

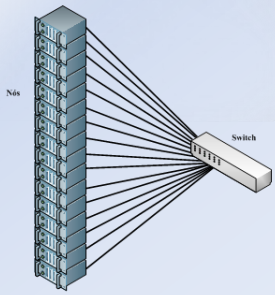
Introdução e Contextualização

CARACTERÍSTICAS DOS "10 MAIS" VEICULADOS NA LISTA DO TOP500 DE NOVEMBRO DE 2008.
[http://www.top500.org/lists/2008/11]

	Sistema	Familia	Processador	Fabricante	S.O.	Interconexão
1	Roadrunner	IBM Cluster	PowerXCell 8i 3200 MHz (12.8 GFlops)	IBM	Linux	Infiniband
2	Jaguar	Cray XT	AMD x86_64 Opteron Quad Core 2300 MHz (9.2 GFlops)	Cray Inc.	CNL	XT4 Internal Interconnect
3	Pleiades	SGI Altix	Intel EM64T Xeon E54xx (Harpertown) 3000 MHz (12 GFlops)	SGI	SLES10	Infiniband
4	BlueGene/L	IBM BlueGene	PowerPC 440 700 MHz (2.8 GFlops)	IBM	SLES 9	Proprietary
5	BlueGene/L	IBM BlueGene	PowerPC 450 850 MHz (3.4 GFlops)	IBM	SLES 9	Proprietary
6	Ranger	Sun Blade System	AMD x86_64 Opteron Quad Core 2300 MHz (9.2 GFlops)	Sun	Linux	Infiniband
7	Franklin	Cray XT	AMD x86_64 Opteron Quad Core 2300 MHz	Cray Inc.	CNL	XT4 Internal Interconnect
8	Jaguar	Cray XT	AMD x86_64 Opteron Quad Core 2100 MHz (8.4 GFlops)	Cray Inc.	CNL	XT4 Internal Interconnect
9	Red Storm	Cray XT	AMD x86_64 Opteron Quad Core 2300 MHz	Cray Inc.	SUSE Linux	XT3 Internal Interconnect
10	Dawning 5000A	Dawning Cluster	AMD x86_64 Opteron Quad Core 1900 MHz (7.6 GFlops)	Dawning	Windows HPC 2008	Infiniband

Arquitetura de Comunicação HPC e o Barramento PCI-E

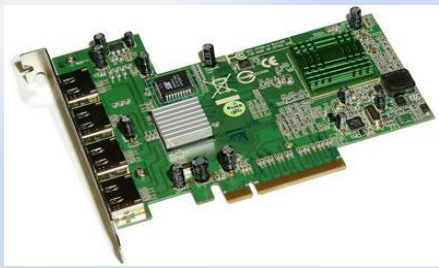
- A arquitetura HPC utiliza uma topologia centralizada de comunicação
- Os nós são interligados através de um barramento de I/O



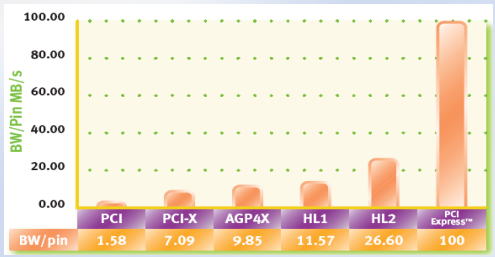
Arquitetura de Comunicação HPC e o Barramento PCI-E

- Atualmente o PCI-Express (PCI-E) é a tecnologia de barramento mais utilizada na construção de cluster
- As interfaces Infiniband utilizam o barramento PCI-E
- O PCI-E 8x possibilita um slot com capacidade de 32 Gbps (16 Gb/s em cada direção) (Ajay V., 2002).
- Alguns sistemas de HPC já estão utilizando a versão 2 do PCI-E, atingindo na interface de 8x velocidades de até 80 Gbps

Arquitetura de Comunicação HPC e o Barramento PCI-E



Arquitetura de Comunicação HPC e o Barramento PCI-E



Barramento PCI-Express comparado a outros tipos. (Ajay V., 2002)

A Tecnologia Infiniband (IB)

Estatística de utilização de tecnologias de interconexão para os 500 mais. (http:// www.top500.org/stats/list/32/connfam)

Interconnect Family	Count	Share %	Rmax Sum (GF)	Rpeak Sum (GF)	Processor Sum
Myrinet	10	2.00 %	350290	488934	56576
Quadrics	4	0.80 %	122220	147507	21040
Gigabit Ethernet	282	56.40 %	4948233	9795163	941748
Infiniband	141	28.20 %	6549813	8721697	841730
Crossbar	1	0.20 %	35860	40960	5120
Mixed	1	0.20 %	66567	82944	13824
NUMAlink	3	0.60 %	122554	137625	21504
SP Switch	10	2.00 %	229541	273754	34208
Proprietary	42	8.40 %	4143049	5243830	1108169
Cray Interconnect	6	1.20 %	359197	469470	73004
Totals	**500**	**100%**	**16927325.79**	**25401883.80**	**3116923**

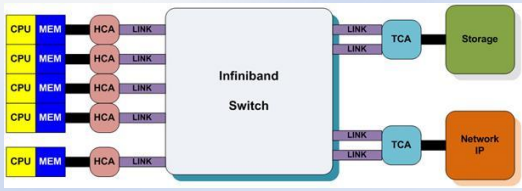
A Tecnologia Infiniband (IB)

- Cada elemento da rede Infiniband é caracterizado por suas funções
  - Na visão mais central da topologia da rede temos o Switch IB (Liu, J., et al., 2004) (Infiniband Trade Association, 2009) (Rashti, M.J. e Afsahi, A., 2007).
  - O Switch desempenha a função de centralização e controle de comutação dos quadros entre os nós

A Tecnologia Infiniband (IB)

- Nos nós finais (end-point) ficam os Host Channel Adapters (HCA)
- O HCA realiza a comunicação do nó com o Switch IB central
- Além do HCA, o Target Channel Adapter (TCA) é responsável pela comunicação entre os nós e os elementos de comunicação IP ou ainda com o Storage
- O TCA promove interoperabilidade entre tecnologias diferentes (e.g. Fibrechannel e Ethernet)

A Tecnologia Infiniband (IB)



Elementos de hardware da arquitetura IB.

A Tecnologia Infiniband (IB)

- O IB é apresentado em várias velocidades de comunicação
- SDR (Single Data Rate) ou 1x
- DDR (Double Data Rate)
- QDR (Quadruple Data Rate)

A Tecnologia Infiniband (IB)

Nomenclaturas e Velocidades do IB.
[Implementing InfiniBand on IBM System, September 2007]

Name	Speed	Data rate	Fully duplexed rate
1X	2.5 Gbps	2 Gbps	4 Gbps
4X	10 Gbps	8 Gbps	8 Gbps
12X	30 Gbps	24 Gbps	48 Gbps
1X DDR	5 Gbps	4 Gbps	8 Gbps
4X DDR	20 Gbps	16 Gbps	32 Gbps
12X DDR	60 Gbps	48 Gbps	96 Gbps
1X QDR	10 Gbps	8 Gbps	16 Gbps
4X QDR	40 Gbps	32 Gbps	64 Gbps
12X QDR	120 Gbps	96 Gbps	192 Gbps

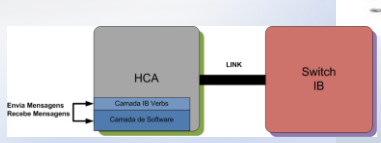
A Arquitetura Infiniband

Host Channel Adapters (HCA) e Target Channel Adapters (TCA)

- O HCA entende todos os "verbs" definidos no padrão
- Verbs são termos semânticos que definem como a arquitetura deve agir (Liu, J., et al., 2004)
- Estas mensagens são enviadas e recebidas pelos nós
- Uma interface de gerenciamento (Verb-Based – VB) destas mensagens semânticas realiza o controle
- Principalmente quando as mensagens são funções que alteram o estado do RDMA (Liu, J., et al., 2004) (Infiniband Trade Association, 2009) (Rashti, M.J. e Afsahi, A., 2007)

A Arquitetura Infiniband

Host Channel Adapters (HCA) e Target Channel Adapters (TCA)



Camada Verbs no controlador HCA

Placa HCA Infiniband
[http://www.sun.com/products/networking/infiniband/]

A Arquitetura Infiniband
Host Channel Adapters (HCA) e Target Channel Adapters (TCA)

- O TCA é um tipo especializado de HCA
- Ele não possui todas as funcionalidades do HCA, sendo assim, ele não entende todos os "verbs"
- Normalmente é utilizada para interligar um Storage de armazenamento de informações à rede IB
- O TCA também pode interligar uma rede IB a um backbone IP

A Arquitetura Infiniband
Switches IB

- Os Switches são elementos fundamentais na arquitetura de comunicação IB
- Eles concentram grandes quantidades de portas IB e consequentemente todas as interligações em uma topologia estrela de comunicação
- Este conceito faz o Switch IB fundamental na arquitetura
- Ele não consome largura de banda, ele apenas gerencia o tráfego em os dispositivos

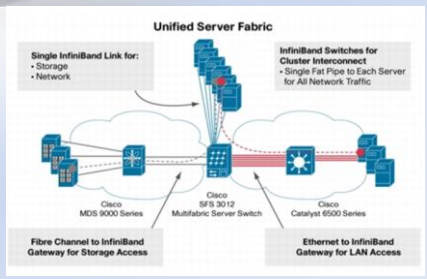
A Arquitetura Infiniband
Switches IB

Switch Mellanox de 19U com 324-port de 20 e 40Gb/s. (InfiniBand Chassis Switch - MTS3610)

A Arquitetura Infiniband
Switches IB

- O Switch permite que diversos dispositivos possam conversar através de suas portas
- Este conceito também é conhecido por fabric (Cisco Systems, 2005)
- Um fabric nada mais é do que uma estrutura de alto desempenho que mantém as conexões em alta velocidade
- Não cria gargalos em sua essência

A Arquitetura Infiniband
Switches IB



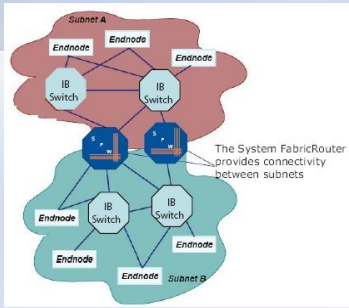
Único Switch Fabric. (Cisco Systems, 2005)

A Arquitetura Infiniband
Routers

- Equipamentos destinados a encaminhar pacotes entre subnets diferentes
- Igualmente aos Switches, eles não consomem largura de banda e não são dispositivos de destino na comunicação
- Diferenciam dos Switches no quesito lógica de encaminhamento
- O roteador (router) IB lê as informações de rota através dos cabeçalhos dos pacotes Ipv6 e os encaminha para a subnet de destino apropriada (Cisco Systems, 2006)
- Para isto ele possui uma tabela que converte as informações de IP para as informações de controle de link
- Cada link possui um identificador conhecido por Logical IDentifier (LID), esta tabela é gerida pela entidade Subnet Manager

A Arquitetura Infiniband

Routers

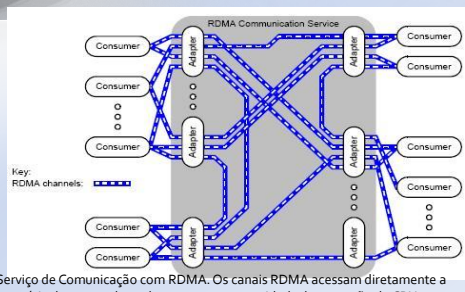


Topologia extendida, duas fabrics em subnets diferentes unidas através dos roteadores IB. [http://www.systemfabricworks.com/fabricRouter.html]

RDMA – REMOTE DIRECT MEMORY ACCESS

- As interfaces tradicionais de acesso a arquivos e informações não provêem características apropriadas para paralelizar as operações de I/O (Input/Output)
- Os sistemas HPC construídos com interfaces tradicionais não conseguem desempenho apropriado
- Ocorre um gargalo por conta do kernel do sistema operacional
- O Remote Direct Memory Access (RDMA) traz características que satisfaz as necessidades de controle paralelo de I/O
- RDMA move os dados de diferentes processos que estão posicionados na memória da CPU e carrega estes dados para a memória da interface IB
- Minimiza o overhead causado pelo sistema operacional em um tradicional acesso (Velusamy, V. et al., 2004).

RDMA – REMOTE DIRECT MEMORY ACCESS

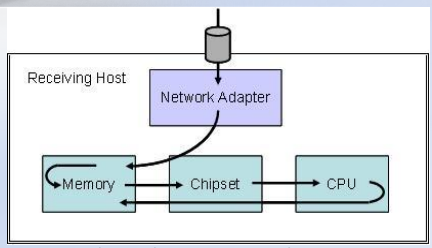


Serviço de Comunicação com RDMA. Os canais RDMA acessam diretamente a memória do outro adaptador, sem ter a necessidade de operação da CPU.

RDMA – REMOTE DIRECT MEMORY ACCESS

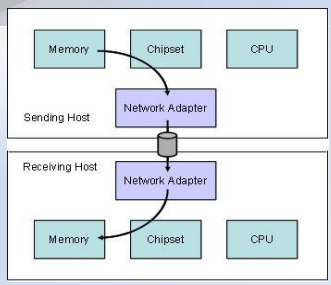
- Segundo Gilad Shainer (Mellanox Technologies, 2006) as Universidades de Princeton e Cornell iniciaram em 1990 o estudo em comunicação de mapeamento de memória
- Em 1997 um grupo formado por Compaq (HP recentemente), Intel e Microsoft criou um draft baseado nas pesquisas iniciadas em 1990
- Este draft resultou em uma interface programada chamada de Virtual Interface Architecture (VIA)
- Uma das principais ações deste modelo de comunicação foi a diminuição do overhead causado pelo sistema operacional

RDMA – REMOTE DIRECT MEMORY ACCESS



Processo de acesso e cópia de informações para a memória da CPU. Processo normalmente instanciado em tecnologias comuns de comunicação. (Gilad Shainer, 2006)

RDMA – REMOTE DIRECT MEMORY ACCESS



Arquitetura da RDMA, o acesso à memória não depende da CPU. (Gilad Shainer, 2006)

RDMA – REMOTE DIRECT MEMORY ACCESS

- InfiniBand utiliza mensagens semânticas (enviar e receber)
- Por exemplo, um nó pode escrever diretamente na memória buffer de outro nó, ou um nó pode ler os dados diretamente da memória buffer remota de outro nó
- Quando um nó deseja enviar informações para um nó remoto, antes ele realizar uma solicitação através de uma mensagem semântica

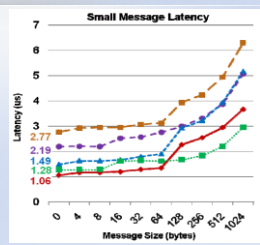
DESEMPENHO

- O RDMA com as características de mensagens semânticas colabora para a eficiência do MPI
- Em MPI,existem dois protocolos de comunicação entre os nós pertencentes ao Cluster

DESEMPENHO

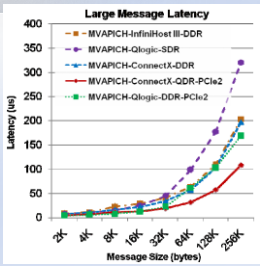
- Eager e tem como principal característica a de envio de pequenas (curtas) mensagens
- Rendezvous que tem como característica o envio de mensagens extensas (longas).

DESEMPENHO



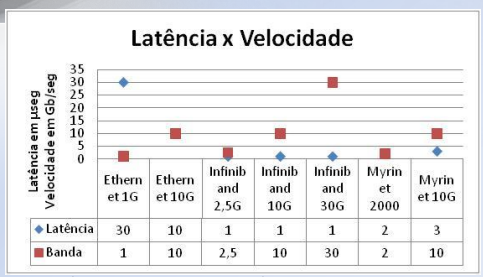
Latência das mensagens curtas sobre IB. (Panda, D.K., 2008)

DESEMPENHO



Latência das mensagens longas sobre IB. (Panda, D.K., 2008)

DESEMPENHO



	Ethernet 1G	Ethernet 10G	Infiniband 2,5G	Infiniband 10G	Infiniband 30G	Myrinet 2000	Myrinet 10G
Latência	30	10	1	1	1	2	3
Banda	1	10	2,5	10	30	2	10

Análise comparativa da Ethernet, InfiniBand e Myrinet
(Informações dos fabricantes)

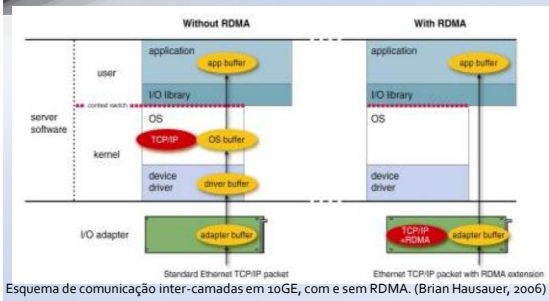
DESEMPENHO DO 10GE VERSUS IB

- Novo padrão de comunicação Ethernet foi criado, o 10-Gigabit Ethernet
- Este novo padrão de comunicação trouxe algumas melhorias para minimizar o overhead existente na tecnologia Ethernet
- 10GE (10-Gigabit Ethernet) concorre com outras tecnologias destinadas à HPC (e.g. Infiniband, Myrinet, e Quadrics) (Rashti, M.J. e Afsahi, A., 2007).

DESEMPENHO DO 10GE VERSUS IB

- Os tópicos avançados sobre 10GE foram incorporados no hardware do adaptador
- Uma camada de suporte RDMA sobre TCP/IP realiza o kernel bypass
- Estas especificações são chamadas de iWARP

DESEMPENHO DO 10GE VERSUS IB



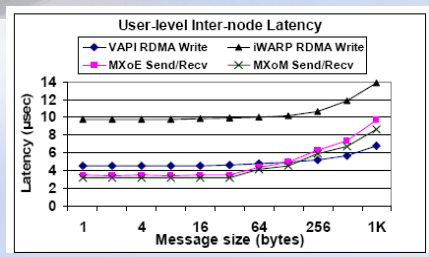
Esquema de comunicação inter-camadas em 10GE, com e sem RDMA. (Brian Hausauer, 2006)

DESEMPENHO DO 10GE VERSUS IB

Comparação das taxas de latência e largura de banda para 4 interfaces de comunicação no nível do usuário

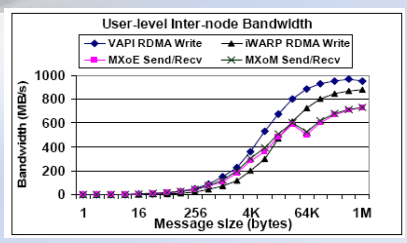
- NetEffect iWARP verbs 1.4.3 (10GE)
- Mellanox VAPI 4.1.1 (Infiniband)
- MX-10G over Ethernet (MXoE) preliminary version 1.2.1 (Myrinet)
- MX-10G over Myrinet (MXoM) version 1.2.0. (Myrinet)

DESEMPENHO DO 10GE VERSUS IB



Teste de desempenho no nível do usuário. (Rashti, M.J. e Afsahi, A., 2007)

DESEMPENHO DO 10GE VERSUS IB

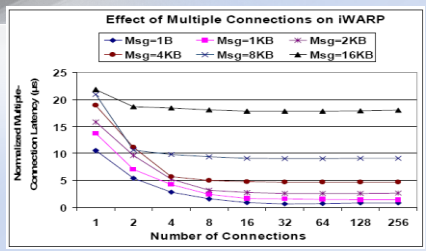


Teste de desempenho no nível do usuário. (Rashti, M.J. e Afsahi, A., 2007)

DESEMPENHO DO 10GE VERSUS IB

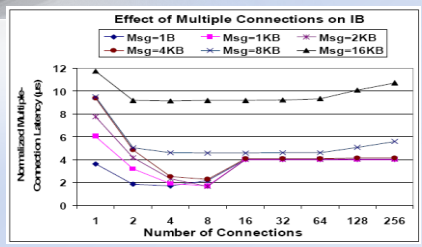
- Comparação das taxas de latência e vazão quando há concorrência (Rashti, M.J. e Afsahi, A., 2007)
- Mensagens de 1B, 1kB, 2kB, 4kB, 8kB e 16kB
- Concorrência de 1 a 256 conexões simultâneas

DESEMPENHO DO 10GE VERSUS IB



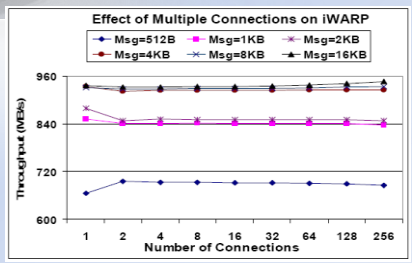
Latências obtidas para múltiplas conexões. Interface iWARP. (Rashti, M.J. e Afsahi, A., 2007)

DESEMPENHO DO 10GE VERSUS IB



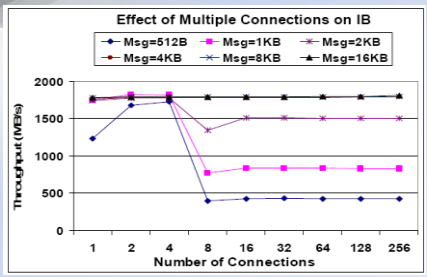
Latências obtidas para multiplas conexões. Interface IB. (Rashti, M.J. e Afsahi, A., 2007)

DESEMPENHO DO 10GE VERSUS IB



Vazões obtidas para multiplas conexões. Interface iWARP. (Rashti, M.J. e Afsahi, A., 2007)

DESEMPENHO DO 10GE VERSUS IB



Vazões obtidas para múltiplas conexões. Interface IB. (Rashti, M.J. e Afsahi, A., 2007)

Comparação de Preço – IB x Ethernet

	Gigabit Ethernet	10 Gigabit Ethernet	InfiniBand 10Gb/s
Vendor	Extreme Networks	Foundry	Voltaire
Product	Summit 7i	Fes-X	ISR9024
Price	$16,495	$12,500	$8,850
Number of Ports	32	2	24
Price per Port	$515	$6,250	$369

Comparação de Custo entre GE, 10GE e IB (Mellanox Technologies – White Paper, 2005)

ESTUDO DE CASO – INFINIBAND

National Center for Supercomputing Applications (NCSA)

O NCSA fica na Universidade de Illinois e a duas décadas provê recursos para Computação de Alto Desempenho

Muitos setores utilizam os recursos computacionais: áreas da ciência, engenharia e do setor privado

Uma empresa comercial que explora óleo e gás precisava de uma plataforma HPC de alto desempenho

Aplicações paralelas que explorassem os problemas sísmicos.

ESTUDO DE CASO – INFINIBAND

National Center for Supercomputing Applications (NCSA)

Visão ampla do cluster.
(Cisco Systems, http://www.cisco.com/en/US/prod/collateral/ps6418/ps6419/ps6421/prod_ case_study0900aecd8033e808.html)

ESTUDO DE CASO – INFINIBAND

National Center for Supercomputing Applications (NCSA)

- Após análise do problema pelos analistas da NCSA
- Sistema aberto Linux e processador Intel EM64T
- A interconexão dos nós através de arquitetura Infiniband
- Os equipamentos IB foram adquiridos da Cisco Systems
- Fabric totalmente padronizado

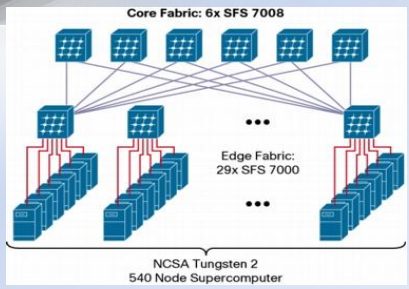
ESTUDO DE CASO – INFINIBAND

National Center for Supercomputing Applications (NCSA)

- 540 computadores adquiridos da Dell, PowerEdge 1850 servers, 2 processadores EM64T de 3.6 GHz
- Para o Fabric, 6 switches IB de alta performance da Cisco Systems modelo Core Fabric SFS 7008
- 29 switches IB de média performance da Cisco Systems modelo Edge Fabric SFS 7000
- O supercomputador foi batizado com o nome Tungsten 2

ESTUDO DE CASO – INFINIBAND

National Center for Supercomputing Applications (NCSA)

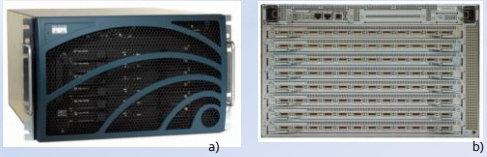


Topologia lógica completa do cluster Tungsten 2.
(Cisco Systems, http://www.cisco.com/en/US/prod/collateral/ps6418/ps6419/ps6421/prod_ case_study0900aecd8033e808.html)

ESTUDO DE CASO – INFINIBAND

National Center for Supercomputing Applications (NCSA)

Servidor PowerEdge 1850 – Dell Inc. [Dell, 2009]



a) b)

Switch IB de alto desempenho da Cisco Systems, modelo Core Fabric SFS 7008. a) visão frontal; b) visão traseira.

ESTUDO DE CASO – INFINIBAND

National Center for Supercomputing Applications (NCSA)

a)

b)

Switch IB de médio desempenho da Cisco Systems, modelo Edge Fabric SFS 7000. a) visão frontal; b) visão traseira.

Provinha – 29.09.2009

Num IDC temos:

• um sistema de webmail com 20 maquinas de front-end e 2 de back-end

•um sistema de máquinas administrativas com 10 máquinas de front-end e 4 de back-end

•um sistema de storage, composto por discos e back-up que serve a todos os back-ends

um cluster de 100 blades, cada uma com 4 processadores

Proponha um sistema de interconexão que atenda as demandas deste ambiente. Coloque redundância entre back-ends e storages. Os servidores poderiam ser aglutinados num mesmo sistema físico de interconexão? Como seria feita a separação lógica? Avalie os problemas de performance que poderão ocorrer na infra-estrutura de conexão, com as redes separadas e juntas.

CONCLUSÃO

- IB oferece alto desempenho para HPC
- Baixa latência e elevada largura de banda
- Infiniband é uma tecnologia em ascensão
- 28,20% dos supercomputadores (Top500) utilizam Infiniband como tecnologia de comunicação
- Alcança latências de comunicação menor que 10 microssegundos entre os elementos da rede IB

CONCLUSÃO

- O 10GE é uma boa opção com iWARP
- O custo de IB comparado ao 10GE iWARP é relativamente bom

REFERÊNCIAS

R. J. Creasy, "The origin of the VM/370 time-sharing system", IBM Journal of Research & Development, Vol. 25, No. 5 (September 1981), pp. 483–90, PDF ,perspective on CP/CMS and VM history by the CP-40 project lead, also a CTSS author.

Geppert, L.; Sweet, W., "Breakthroughs Will Leave Their Mark On Many Key Technologies," Spectrum, IEEE , vol.35, no.1, pp. 19-22, Jan. 1998

Jiuxing Liu; Vishnu, A.; Panda, D.K., "Building Multirail InfiniBand Clusters: MPI-Level Design and Performance Evaluation," Supercomputing, 2004. Proceedings of the ACM/IEEE SC2004 Conference , vol., no., pp. 33-33, 06-12 Nov. 2004

Nemertes Research, "Data Center I/O Consolidation", http://www.nemertes.com/products_services/research/issue_papers/nemertes_issue_paper_data_center_i_o_consolidation2005. Acessado em março de 2009.

Fibrechannel Industry Association, "Fibre Channel - Overview of the Technology", http://www.fibrechannel.org/technology/overview.html. Acessado em março de 2009.

REFERÊNCIAS

Liu, J.; Balasubramanian Chandrasekaran; Yu, W.; Wu, J.; Buntinas, D.; Sushmitha Kini; Panda, D.K.; Wyckoff, P., "Microbenchmark performance comparison of high-speed cluster interconnects," Micro, IEEE , vol.24, no.1, pp. 42-51, Jan.-Feb. 2004

Infiniband Trade Association, "InfiniBand Architecture Specification", http://www.infinibandta.org/specs. Acessado em março de 2009.

Rashti, M.J.; Afsahi, A., "10-Gigabit iWARP Ethernet: Comparative Performance Analysis with InfiniBand and Myrinet-10G," Parallel and Distributed Processing Symposium, 2007. IPDPS 2007. IEEE International , vol., no., pp.1-8, 26-30 March 2007

W. Gropp, E. Lusk, and A. Skjellum. Using MPI: Portable Parallel Programming with the Message Passing Interface, 2nd edition. MIT Press, Cambridge, MA, 1999.

Ajay V. Bhatt, Technology And Research Labs - Intel Corporation, White Paper, "Creating a PCI Express Interconnect", 2002.

http://www.pcisig.com/specifications/pciexpress/resources/PCI_Express_White_Paper.pdf, acessado em abril de 2009.

REFERÊNCIAS

Dino Quintero, Norbert Conrad, Rob Desjarlais, Marc-Eric Kahle, Jung-Hoon Kim, Hoang-Nam Nguyen, Tony Pirraglia, Fernando Pizzano, Robert Simon, Shi Lei Yao, Octavian Lascu. "Implementing InfiniBand on IBM System" – Red Book IBM, First Edition (September 2007). 330 p.

Cisco Systems. "Unified Fabric: Benefits and Architecture of Virtual I/O", White Paper, 2005. http://www.cisco.com/en/US/prod/collateral/ps6418/ps6423/ps6429 /prod_white_paper0900aecd80337bb8.html

Cisco Systems. "Cisco Server Fabric Switch InfiniBand Fabric", White Paper, 2006. http://www.cisco.com/en/US/prod/collateral/ps6418/ps6423/ps6429/prod_white_paper0900aecd805cd9c6.pdf

Velusamy, V.; Skjellum, A.; Kanevsky, A., "Employing an RDMA-based file system for high performance computing," Networks, 2004. (ICON 2004). Proceedings. 12th IEEE International Conference on , vol.1, no., pp. 66-70 vol.1, 16-19 Nov. 2004

Marazakis, M.; Papaefstathiou, V.; Kalokairinos, G.; Bilas, A., "Experiences from Debugging a PCIX-based RDMA-capable NIC," Cluster Computing, 2006 IEEE International Conference on , vol., no., pp.1-10, 25-28 Sept. 2006

REFERÊNCIAS

Gilad Shainer , "Why Compromise?", Mellanox Techologies, White Paper, 2006. http://www.hpcwire.com/features /17888274.html. Acessado em abril de 2009.

Jiuxing Liu, Jiesheng Wu, Dhabaleswar K. Panda, "High Performance RDMA-Based MPI Implementation over Infiniband", Journal-Papers, Ohio State University, 2004. P 13.

Panda, D.K., "Designing next generation clusters with InfiniBand and 10GE/iWARP: Opportunities and challenges," Cluster Computing, 2008 IEEE International Conference on , vol., no., pp.202-202, Sept. 29 2008-Oct. 1 2008

Rashti, M.J.; Afsahi, A., "10-Gigabit iWARP Ethernet: Comparative Performance Analysis with InfiniBand and Myrinet-10G," Parallel and Distributed Processing Symposium, 2007. IPDPS 2007. IEEE International , vol., no., pp.1-8, 26-30 March 2007

Dalessandro, D.; Devulapalli, A.; Wyckoff, P., "iWarp protocol kernel space software implementation," Parallel and Distributed Processing Symposium, 2006. IPDPS 2006. 20th International , vol., no., pp.8 pp.-, 25-29 April 2006

REFERÊNCIAS

J. Hilland, P. Culley, J. Pinkerton and R. Recio. "RDMA protocol verbs specification" (v1.0), 2003. http://www.rdmaconsortium.org/. Acessado em abril de 2009.

Brian Hausauer. "iWARP Ethernet: Eliminating Overhead In Data Center Designs", White Paper, NetEffect Inc., 2006, 8 p.