****

**Aplicações distribuídas usando o protocolo** **Infiniband**

Kemoel Lima G. Silva

kemoel@msn.com

1Sistemas de Informação – Faculdade Estácio de Alagoas.
2Redes de Computadores – Faculdade Estácio de Alagoas.

***Resumo.*** *Este artigo propõe uma abordagem que visa uma insignificativa visão no que vem ser Infiniband um* *link de comunicação de rede usado em computação de alto desempenho com elevado rendimento de rede e no desempenho na comunicação de aplicações distribuídas para aglomerados. A proposta inclui a implementação de sistemas distribuídos usando InfiniBand que pode ser ideal para tarefas de computação de grande porte que necessitam de dezenas, centenas ou mesmo milhares de servidores, ou nós, que trabalham em conjunto para resolver um único problema. InfiniBand proporciona uma elevada taxa de transferência de dados entre os nós, bem como uma baixa taxa de latência.*

Palavras chave: *Infiniband, Volume de dados, Alto desempenho, Baixa Latência, Nós de computação.*

**1 Introdução.**

A medida que mais organizações tira proveito da computação em nuvem, há uma busca correspondente para alcançar os mais altos níveis de desempenho a partir da nuvem. Um dos gargalos em termos de infra-estrutura em nuvem é a velocidade com que as ligações de rede entre os servidores podem executar. Quando é necessário um desempenho absoluto, mesmo um pequeno atraso nesses links podem ter impacto sobre os níveis de serviço. Este é o lugar onde Infiniband entra.

InfiniBand originou em 1999 a partir da fusão de dois projetos concorrentes: Future I/O e Next Generation I/O. Isso leva à formação da Associação de Comércio InfiniBand, incluindo Compaq, Dell, Hewlett-Packard, IBM, Intel, Microsoft e Sun.

Infiniband é uma tecnologia de rede de alto desempenho que é uma arquitetura de rede padrão da indústria. Infiniband oferece alta velocidade de transferência de até 56Gbps entre nós de computação.

Umas das curiosidades e irrelevante, mas para dar um sentido de perspectiva, que é o equivalente a transferência de dados de mais de 30.000 discos Blu-ray em um único dia.

Infiniband é usado em mais de 200 dos 500 supercomputadores mais poderosos em todo o mundo de boa companhia, de fato!

Existe uma grande iniciativa da indústria para a utilização e validação da tecnologia Infiniband.

Dentro da engenharia de produção, esta tecnologia vem sendo aplicada em duas frentes, que são a computação em centros de dados que manipulam com grandes volumes de informações e em redes especializadas para processamento de aplicações numéricas e simulações que de mandam por alto desempenho.

O presente artigo tem por objetivo apresentar a arquitetura de interconexão Infiniband, que

desponta como um novo padrão de tecnologia para a comunicação eficiente entre computadores.

O artigo também mostra a contribuição de Infiniband para a computação em grandes centros

de dados e a sua aplicabilidade em ambientes que realizam a computação numérica cuja

velocidade de rede de interconexão é um fator de extrema relevância.

**2 Informações Gerais**

Aplicações distribuídas que requerem alto desempenho na comunicação, necessitam que o atraso na comunicação seja o menor possível. O atraso na comunicação entre dois processos remotos contém duas componentes: tempo gasto no processamento da mensagem e atraso da rede. Em redes de baixa velocidade, tais como Ethernet a 10Mbps, o gargalo da comunicação é o atraso da rede. Isto levou ao desenvolvimento de mecanismos de comunicação baseados em sockets, onde o processamento das mensagens no núcleo do sistema operacional causa muitas cópias e muitas trocas de contexto, o que aumenta muito o atraso fim-a-fim.

Perante o cenário de desempenho cada vez melhor para mensagens grandes, pequenas mensagens estão se tornando cada vez mais importante em muitas aplicações. Por exemplo, em sistemas distribuídos. A tecnologia é encontrar orientada a objetos generalizada adoção e é naturalmente estendido em toda a rede permitindo a transferência de objetos e o controle remoto execução de métodos. Os objetos são geralmente pequenos, em relação para a mensagem de tamanhos necessária para alta largura de banda, e, assim, a comunicação desempenho sofre menos que mensagem sobrecarga é baixo. O local de trabalho eletrônico depende muito de conjuntos de serviços distribuídos complexos que se destinam a ser transparente para o utilizador. A maioria de tais serviços invocações são solicitações para servidores de banco de dados simples que implementam mecanismos como a nomeação de objetos, localização do objeto, autenticação, proteção, etc. A o tamanho da mensagem observado nestes sistemas variam de 20 a 80 bytes para os pedidos e as respostas em geral podem ser encontrados na gama de 40-200 bytes. Para limitar a travessia de rede de maior distribuição objetos, técnicas de caching tornaram-se um elemento fundamental parte da maioria dos modernos sistemas distribuídos. Manter as cópias consistente introduz um grande número de mensagens pequenas coerência. Os tempos de ida e volta são importante que o solicitante é normalmente bloqueada até que o a sincronização é conseguida. Algoritmos de tolerância a falhas de software e comunicação em grupo ferramentas muitas vezes exigem protocolos multi-round, cujo desempenho é limitado latência. Lowlatency (baixa latência) comunicação impede que tais protocolos para ser usados ​​hoje em aplicações de controle de processo, financeira sistemas de negociação, ou aplicativos de multimídia de groupware.

Projetando para o futuro, em geral sistemas podem beneficiar substancialmente assim: Protocolos de fluxo de dados confiável como o TCP tem buffer requisitos que são diretamente proporcionais ao de ida e volta latência de ponta a ponta. Por exemplo, o TCP tamanho da janela é o produto da largura de banda da rede e o tempo de ida e volta. Alcançar baixa latência vai manter o consumo de buffer dentro da razão e assim torná-lo viável para alcançar a largura de banda máxima.

A arquitetura Infiniband tem por objetivo ultrapassar os limites de desempenho impostos pela organização em barramento compartilhado, onde somente um componente pode usar o meio de interconexão em um dado instante. As tecnologias que empregam a organização em barramento mais utilizadas são a PCI *(Peripheral Interconnect Componen*t) e a Ethernet. A PCI é utilizada para interconexão de componentes dentro de um computador e a Ethernet é um tipo de tecnologia de rede amplamente utilizada para a comunicação entre computadores. Neste contexto, Infiniband surge como uma solução para substituir as tecnologias de interconexão padrão e, prover os requisitos necessários para manipular centros de dados e servidores com alto desempenho e gerenciabilidade.

Com a organização apresentada na figura 1, IBA fornece escalabilidade, qualidade de serviço, melhor desempenho para a troca de dados pela rede, tolerância à falhas e alta disponibilidade. Na área referente a tolerância a falhas, Infiniband oferece a detecção e recuperação a falhas, redundância de informações através de múltiplos caminhos pela rede, determinação dinâmica de caminhos, migração automática de caminho e isolamento. Para atingir alto desempenho, Switch-IB, a sétima geração de comutação IC de (Mellanox, 2014), é a ISIC de soluções EDR de comutação da empresa pela primeira vez com a largura de banda líder de mercado, baixa latência e escalabilidade. O dispositivo suporta até 36 portas EDR 100Gb/s InfiniBand. Todas as portas de acordo com a especificação InfiniBand para negociação automática SDR para EDR. Com esses índices, a Infiniband proporciona uma alta taxa de transferência de dados se comparada com os índices alcançados pelas tecnologias tradicionais.

Mesmo com os avanços tecnológicos ocorridos nos últimos tempos, existe uma gama de aplicações numéricas que demandam por alto poder de processamento para que seja encontrada uma solução em um tempo satisfatório. Para a execução desse tipo de aplicações foram planejadas arquiteturas computacionais específicas para esse fim. Uma das mais populares arquiteturas desse gênero são os aglomerados de computadores para a computação numérica de alto desempenho (WILKINSON, 1999). Sua popularidade advém da sua relação de custo/benefício, visto que pode ser formada por computadores comuns, fabricados em larga escala, interconectados por uma rede dedicada. Nessa arquitetura, os computadores cooperam através da rede para a solução mais rápida de um problema em comum. Nesse sentido, a arquitetura Infiniband padroniza uma série de tecnologias de interconexão para aglomerados de computadores e é capaz de oferecer altas taxas de desempenho para as operações de transferência de dados, além de gerência de recursos para este ambiente. Com a integração da arquitetura de aglomerados de computadores e Infiniband, pode-se resolver em um tempo reduzido problemas numéricos complexos de engenharia de produção, como por exemplo os problemas do caixeiro viajante e de simulações de produção (CHISTOFIDES,1979). Além da redução do tempo, esta integração proporciona tolerância à falhas e escalabilidade, de modo a aumentar a produtividade do sistema e a qualidade dos resultados. As seções que seguem apresentam, respectivamente, a utilização de Infiniband em um supercomputador e a proposta de uma ferramenta para a escrita de programas paralelos e distribuídos para executar sobre um hardware Infiniband.

**3 Supercomputador Dell Texas Advanced Computing Center**

A arquitetura de aglomerados de computadores também é empregada no supercomputador

Texas Advanced Computing Center (TOP500, Junho 2014). Este aglomerado tem por objetivo específico suportar a execução de aplicações numéricas paralelas e distribuídas que demandam por alto poder de processamento, também chamados de aplicações com problemas de grande desafio. A tecnologia Infiniband é utilizada como tecnologia de interconexão nesse aglomerado, que está classificado em sétimo lugar segundo o site que organiza os 500 supercomputadores com maior desempenho no mundo (TOP500, Junho 2014). Este aglomerado é formado por 462,462 cores, Stampede - PowerEdge C8220, Xeon E5-2680 8C 2.700GHz, Infiniband FDR, Intel Xeon Phi SE10P, 192,192 GB de memória e atinge um pico de desempenho de 5,168.11 TFlop/s.

Entre o conjunto de áreas de aplicações que são executadas no Texas Advanced Computing Center, podem-se citar a química quântica, a nano ciência, o cálculo de otimizações baseado em heurísticas, a bioquímica computacional, a otimização da aerodinâmica de sólidos, a modelagem de ciclos de células, a modelagem de sistemas sem fio *(wireles*s), a acústica molecular e a emulação de redes de larga escala.

Muitas simulações em grande escala pode produzir saída de dezenas ou hundereds de terabytes. Com muito espaço de armazenamento disponível, os pesquisadores são capazes de executar simulações muito maiores. E porque Stampede tem grande memória e subsistemas de visualização build-in, os pesquisadores são capazes até mesmo de pós-processar e visualizar os seus dados in situ, sem ter que se preocupar sobre a movimentação de todos os terabytes para um sistema diferente.

Um computador desktop típico tem dois ou quatro núcleos de processamento, e também pode incluir um acelerador, como uma placa gráfica. Com tantos mais núcleos de processamento disponíveis, Stampede é capaz de executar códigos ordens de magnitude mais rápido do que é possível em um computador desktop.

Stampede suporta todos os domínios da ciência, incluindo as ciências humanas. A seguir estão os sucessos iniciais de utilizador no sistema.

Stampede é um dos supercomputadores mais poderosos do mundo. Mas, o que isso significa e por que é importante?

Supercomputadores complementar a teoria científica e da observação através da modelagem e análise de qualquer coisa que é muito grande (os planetas), muito pequenas (moléculas da droga), ou (testes de colisão para carros) muito caros ou perigosos para testar em laboratório. Determinar onde e quando terremotos greve; explorar o que nano materiais irá converter a luz solar em energia; e compreender como os tumores cerebrais rápidas crescer - estes problemas sociais importantes e complexas requerem computadores poderosos como Stampede, que fornece um desempenho de pico de cerca de 10 petaflops (FP), ou cerca de 10 quatrilhões de operações matemáticas por segundo.

Stampede é uma parte importante do portfólio da NSF para infra-estrutura de computação avançada que permite a investigação de ponta fundamental para a ciência e engenharia computacional e uso intensivo de dados. A capacidade da sociedade para enfrentar os desafios globais de hoje depende o avanço cyberinfrastructure.

1 petaflop (PF) = 1 quatrilhão de operações matemáticas por segundo. Os sistemas mais poderosos do mundo são sistemas petascale, onde um grande número de computadores trabalham em paralelo para resolver o mesmo problema. Dada a velocidade atual de progresso, especialistas do setor estimam que os supercomputadores vai chegar (operações de um quintilhões por segundo) um exaflop em 2018.

**4 Images**

Figura 1



Figura 2 - switch infiniband Figura 3 – conector infiniband

 

**5 Conclusão**

Hoje em dia, pode se observar uma grande de manda por gerenciabilidade de informações e tecnologias de rede com alto desempenho capazes de proporcionar um rápida comunicação entre computadores. No contexto de um centro de dados, estas duas áreas são de extrema importância e existe grandes investimentos da indústria para projetar tecnologias de ponta para suprir estas necessidades. Nesse sentido, foi desenvolvida a tecnologia de interconexão Infiniband, agregando as qualidades de uma série de tecnologias de rede, arquiteturas de computadores e tecnologias da informação. Infiniband desponta como uma tecnologia capaz de reduzir a sobrecarga inerente aos sistemas baseados em barramento compartilhado, através de um conexão que permite comunicação concorrente e rápida. Ela fornece alta escalabilidade, qualidade de serviço, altas taxas de transferência de dados, tolerância à falhas e gerenciabilidade de recursos e de informações em uma rede. Devido a todas estas características, Infiniband é capaz de efetuar a transmissão de dados em tempo real, permitindo a realização de vídeo conferências com qualidade, além de tráfego de voz e de grandes volumes de informações. Para o ambiente de centro de dados, Infiniband também oferece as vantagens conhecidas como RAS (Confiabilidade, Disponibilidade e Habilidade de Servir). Além da aplicabilidade em centros de dados, Infiniband também é apropriada para a interconexão em aglomerados de computadores (clusters of computers) destinados ao processamento de análise numérica com alto desempenho e precisão. Esta arquitetura é um iipo de rede especializada formada por computadores que cooperam, através de troca de informações pela rede, para a solução de uma aplicação em comum. É digno de nota que a computação em tal arquitetura é uma realidade e vem sendo aplicada por grandes empresas, instituições de ensino e centros de pesquisa. Os principais benefícios proporcionados por

Infiniband para este tipo de aglomerado são a alta taxa de transferência de dados e a alta escalabilidade. Supercomputador Stampede Texas Advanced Computing Center validam o funcionamento da tecnologia Infiniband. Ela se demonstrou uma boa opção de interconexão para redes de organizações que trabalham com um grande volume de informações, além de obter alto desempenho na comunicação em redes especializadas para o processamento de análise numérica.

A evolução de tecnologias para redes leva a utilização de arquiteturas de interconexão mais eficientes, como é o caso da Infiniband, que é capaz de proporcionar várias funcionalidades, além da alta taxa de transferência de dados, que são requeridas por gerentes de tecnologia da informação de grandes organizações. A tendenciada arquitetura Infiniband é ser utilizada em conjunto com as tecnologias padrão existentes atualmente e, dessa forma, construir ambientes que melhorem a vida dos usuários e dos administradores de ambientes computacionais.

**6 References**

[1] T.E. Anderson, D.E. Culler, D.A. Patterson, et. al. *A*

*Case for NOW (Networks of Workstations).* IEEE Micro,

Feb. 1995, pages 54-64.

[2] M. Blumrich, C. Dubnicki, E.W. Felten and K. Li. *Virtual-*

*Memory-Mapped Network Interfaces*. IEEE Micro,

Feb. 1995, pages 21-28.

[3] D. Borman, R. Braden, and V. Jacobson. *TCP Extensions*

*for High Performance*. RFC 1323, May 1992.

[4] D. E. Culler, A. Dusseau, S. C. Goldstein, A. Krishnamurthy,

S. Lumetta, T. von Eicken, and K. Yelick. *Introduction*

*to Split-C.* In Proc. of Supercomputing '93*.*

[5] D. E. Culler, A. Dusseau, R. Martin, K. E. Schauser. *Fast*

*Parallel Sorting: from LogP to Split-C.* In Proc. of

WPPP '93, July 93.

[6] D.E. Culler, et. al. *Generic Active Message Interface*

*Specification, version 1.1* http://now.cs.berkeley.

edu/Papers/gam\_spec.ps

[7] P. Druschel and L.L. Peterson. *Fbufs: A High-Bandwidth*

*Cross-Domain Transfer Facility*. In Proc. of the 14th

SOSP. pages 189-202. December 1993.

[8] Wikipedia Dezember 2011 - http://en.wikipedia.org

[9] "InfiniBand Cards - Overview". Mellanox. Retrieved 30 July 2014.

[10] Texas Advanced Computing Center - https://www.tacc.utexas.edu/stampede/

[11] Top 500 Supercomputing Sites - <http://www.top500.org/>

[12] InfiniBand Trade Association IBTA (2002) - Inifiniband Architecture Specification, Vol. 1, Release 1.1, 2002,

Nov. Disponível em <http://www.infinibandta.org/specs>.

[13] IBM Corporation (2003) - IBM DB2 Integrated Cluster Environment for Linux – Solution blueprint, Disponível

em http://www-306.ibm.com/software/data/pubs/papers/ linuxcluster/linuxcluster.pdf