Estudo realizado no trabalho: "Uma família de Abstração Myrinet para EPOS"

Autor: Fernando Roberto Secco

Aluno: Leonardo Winkelmann

Universidade Federal de Santa Catarina – UFSC

Leonardo@bshd.com.br

RESUMO

O trabalho em estudo apresenta um sistema de comunicação dedicado para redes de baixa-latência myrinet que utiliza o sistema operacional EPOS. Este sistema operacional foi desenvolvido, guiado pela metodologia denominada Projeto de Sistemas Orientados à Aplicação (AOSD – *Application-Oriented System Design)* tendo como objetivo utilizar apenas funções de SO que realmente são necessárias para o seu perfeito funcionamento. Associando o EPOS e técnicas de comunicação (protocolo leve) é possível conseguir uma melhor performance em dispositivos de interconexão com alta taxa de transmissão.

1. INTRODUÇÃO

O surgimento de dispositivos de interconexão com alta taxa de transmissão e a popularização dos computadores pessoais tornou viável a implementação de sistemas distribuídos como os agregados computacionais. É uma opção de baixo custo se comparado às máquinas paralelas convencionais. Nestes tipos de sistema de computação, a rede de interconexão, o software gerenciador e o software de comunicação são componentes decisivos para a formação de um sistema de alto desempenho. Recentemente, com a redução de custo, dispositivos de interconexão diferenciados vêm ganhando espaço no mercado e na pesquisa científica. Exemplos

clássicos são: Myrinet, Quadrics e Infiniband. Poderosos agregados computacionais optam por estas tecnologias de interconexão mais eficientes.

Buscando solucionar problemas de aplicações que necessitam de um alto poder de processamento, intensificou as pesquisas na área de computação de alta performance. Essas buscas em aumentar o poder de processamento é uma constante na evolução da computação.

Os primeiros esforços foram para aumentar o poder de processamento do próprio processador, aumentado a velocidade do relógio, aumentando a quantidade de instruções por clico de relógio, ou ainda, fazer com que as instruções trabalhem em paralelo. Desta forma foram criadas arquiteturas com o *Pipeline*, processadores vetoriais entre outras.

Num segundo momento, buscando a otimização dos recursos, que em alguns momentos do tempo encontravam-se ociosos, começou a pensar em aproveitá-los. Com isso surgem os multi-procesadores e multi-computadores. Com isso a preocupação também começa a recair sobre os sistemas de interconexão. A evolução destas arquiteturas permitiu o surgimento de *Clusters* computacionais dedicados ao processamento paralelo.

Uma arquitetura existente para *Clusters* é a Myrinet, desenvolvida pela empresa Myricom, é um sistema de interconexão que procura deixar a comunicação de rede independente do sistema Operacional, evitando assim interrupções no mesmo para esta comunicação.

Os pacotes de uma rede Myrinet podem ser de tamanho variável, desta forma podem encapsular outros tipos de pacotes, incluído pacotes IP, sem precisar de nenhuma camada de adaptação [Myrinet].

A arquitetura Myrinet, trabalha com o conceito de processador de rede. Utilizando um processador LANai e um software que roda no próprio NIC chamado Myrinet Control Program, que é praticamente um firmware, é possível que a própria

placa trate as interrupções decorridas da recepção de mensagens dando liberdade ao NIC.

Relacionando um SO que busque a otimização em sistemas dedicados, como é o caso do EPOS e melhorando a comunicação, aplicando um protocolo leve, por exemplo, na camada de transporte, é possível conseguir resultados ainda melhores em agregados computacionais que utilizam arquiteturas de dispositivos de interconexão com alta taxa de transmissão, por exemplo, a Myrinet.

2. VISÃO GERAL

2.1. Cluster Computing

Atualmente agregados de PCs, supercomputadores construídos através da interconexão de computadores pessoais convencionais, apresentam desempenho comparável ao de arquiteturas como processadores massivamente paralelos e computadores vetoriais, conquistando um lugar de destaque na área de computação de alto desempenho e sendo considerados a melhor opção em infra-estrutura computacional quando se leva em consideração a relação custo/desempenho.

2.2. User Level Networking

Num sistema de comunicação, tem-se como meta a máxima utilização da largura de banda da rede com o mínimo atraso. Basicamente existem duas possibilidades de comunicação:

- Kernel-Level – Que é a mais utilizada e comum. Neste tipo de comunicação, para qualquer comunicação realizada, é necessário que a solicitação passe pelo Kernel do sistema operacional, causando assim muitas interrupções no mesmo. A vantagem é que o sistema operacional

possuí um maior gerenciamento e controle sobre as ações de comunicação, retirando assim essa responsabilidade do User;

- User-Level – Já o user-level procura diminuir a interação entre aplicação e o SO, afetando assim a quantidade de cópias feitas diminuindo-as ao máximo para que o tempo necessário para essas operações seja reduzido. Do contrário do Kernel-level, o controle e gerenciamento, fica na maior parte para o User.

Com os conceitos citados acima é possível visualizar que dispositivos de interconexão com alta taxa de transmissão utilizam o tipo User-Level. Por isso a importância de um sistema dedicado eficiente, como é caso do EPOS.

2.3 Processador de Rede

O processador de rede surgiu em meados da década de 1990, e tem a finalidade de deixar o mais independente possível do sistema operacional, a comunicação através da rede. O NP ainda hoje é motivo de segredo em muitas empresas. Alguns *datasheets* são difíceis de encontrar tornando o processo de aquisição de informação muito restrita. É possível citar algumas empresas que são fabricantes de processadores de rede: Intel (IXP1200); IBM (NP4GS3); Cisco (Toates 2); Lexra (Netvortex); Chameleon.

Assim com o surgimento dos *Network Processors – NPs* ou Processadores de Rede uma nova concepção de desenvolvimento de sistemas de comunicação em geral. Através de um processador rápido próprio, é possível aumentar à autonomia de acessos a memória através de DMA – *Directory Memory Access,* funcionalidades importantes à comunicação disponíveis no hardware é possível aos desenvolvedores criarem pilhas inteiras de protocolos e implantá-las diretamente no hardware.

Os processadores de rede possuem outras vantagens, como a alta largura de banda (1.2GB *full-duplex*), baixo atraso (menos de 1 micro segundo), mais de um canal

independente de DMA, um custo relativamente barato comparando com os *Clusters* Computacionais altamente agregados.

Neste contexto, os NPs tem sido utilizados para implementar, de forma mais eficiente, mecanismos ligados ao controle de congestionamentos, tratamento de erros, reenvio de pacotes, balanceamento de carga, entre outros.

2.4 Myrinet

Baseada na arquitetura de NPs, a Myricom desenvolveu um dispositivo de interconexão com alta taxa de transmissão chamado Myrinet.

A ATOMIC buscando novas tecnologias em hardware, em 1993 lançou uma nova proposta de como modelar hardware de alto desempenho para MPPS – *multi-**process** phytoremediation system*. A Myricom foi fundada por membros da equipe de desenvolvimento da ATOMIC, com isso toda a tecnologia foi levada para a Myricom e em 1995 foi lançada a proposta para intercomunicação entre *Clusters.*

A Myrinet foi desenvolvida para se tornar uma tecnologia de interconexão, baseada em chaveamento e comunicação por pacotes, de baixo custo e de alta performance. Principalmente para a interconexão de clusters de computadores, de diferentes tipos como PCs, estações de trabalho, entre outros.

Ainda considera-se que a utilização de cluster é uma forma econômica de se conseguir:

- **alta performance**: através da distribuição da computação através das diversas máquinas. Mas para se conseguir uma boa performance é necessário que se tenha uma comunicação com uma alta taxa de transmissão de dados e ainda com baixa latência.
- **alta disponibilidade**: através da possibilidade de realizar as computações em um subconjunto dos equipamentos. Assim é necessário

que a interconexão seja capaz de detectar e isolar as falhas, fornecendo caminhos alternativos para os hosts.

A montagem de clusters pode ser feita através da utilização de redes como Gigabit Ethernet, Fast Ethernet, mas redes mais comuns como a própria Ethernet é incapaz de fornecer a performance e as características necessárias para se conseguir esta alta disponibilidade e alto desempenho. As características que distinguem a Myrinet das demais são especialmente:

- portas e interfaces full-duplex alcançando 1.28 Gb/s para cada link;
- controle de fluxo, de erro, e monitoramento contínuo dos links;
- baixa latência, switches crossbar com monitoramento para aplicação de alta disponibilidade;
- suporte a qualquer configuração de topologia;
- interfaces das estações possuem programa de controle para interagir diretamente com os processos para realizar comunicação com baixa latência.

2.4.1 Link Lógico

Um link Myrinet é composto por um par de canais full-duplex, onde a conexão desse link com o sistema é denominada porta.

Os canais full-duplex que formam um link Myrinet são responsáveis por transmitir pacotes de dados que não necessitam possuir um tamanho fixo de bytes. Além disso, é responsável por manter a ordem de envio dos pacotes e realizar um controle de fluxo, de modo que o fluxo de informação em um canal pode ser interrompido temporariamente pelo receptor.

Ainda é feito um controle do tempo em que o transmissor fica bloqueado, de modo que quando esse tempo é ultrapassado o receptor recebe informação para

reinicializar a comunicação, dessa forma falhas de transmissão podem ser contornadas através da retransmissão.

O circuito da porta detecta o estado de utilização da porta, se ela não está sendo usada ou se ela está conectada a um terminal que está desligado. Neste tipo de condição o transmissor não é bloqueado como o cenário acima, mas sim seus pacotes são descartados.

Os pacotes de uma rede Myrinet podem ser de tamanhos variados, desta forma podem encapsular outros tipos de pacotes, incluindo pacotes IP, sem precisar de nenhuma camada de adaptação. Cada pacote é identificado por um tipo, de tal forma que a Myrinet, assim como Ethernet, pode carregar pacotes de vários tipos de protocolos concorrentemente, suportando desta forma, vários tipos de interfaces de software.

Assim sendo, vários tipos de softwares são utilizados, como o próprio TCP/IP (ou UDP/IP), os quais conseguem taxas de transmissão, aproximadamente, de 250 a 1147 Mbits/s. Entretanto, implementações mais específicas para Myrinet conseguem desempenhos melhores, tais como MPI e VIA desenvolvidos pela própria Myricom.

2.4.2. Roteamento

O roteamento dos pacotes ao longo de uma rede Myrinet é feito através do seu cabeçalho. Quanto um pacote é injetado na rede por uma interface ele contém um cabeçalho de 1 ou mais bytes.

Ele tem apenas um byte no caso de uma rede composta apenas de duas interfaces e nenhum switch. Para redes com 3 ou mais nodos, para as quais é necessária a presença de um switch, o cabeçalho de possuir 2 ou mais bytes.

Somente o primeiro byte do cabeçalho é examinado quando um pacote é recebido por um switch ou uma interface. Uma vez que esse byte é utilizado para selecionar uma

porta do switch, ele é descartado, e retirado do cabeçalho antes do envio do pacote pela porta selecionada.

A tag contida no último byte do cabeçalho é utilizada pelo programa de controle para distinguir entre diferentes tipos de pacotes, tais como pacotes de usuário e pacotes utilizados pelo protocolo de mapeamento da rede. Ela é lida pelo programa de controle quando o pacote é recebido e utilizada para selecionar a rotina de tratamento do tipo de pacote indicado.

Todos os bytes que indicam a rota a ser seguida possuem seu bit mais significativo marcado em 1. Sendo que o último byte, a tag, possui esse bit em 0. Quando um switch encontra um pacote com esse bit em 0, indica que o pacote não alcançou o seu destino final por algum problema de criação do cabeçalho, por exemplo, neste caso o pacote é descartado e é informado um erro.

2.5 Sistemas Operacionais dedicados

Como em todas as áreas da computação, o sistema operacional se encontra no centro da execução de tarefas. Nos *Clusters*, o sistema operacional não é diferente, sendo responsável pelo interfaceamento entre o usuário e o hardware.

Um sistema operacional dimensionado para um *Clusters* deve identificar quais recursos serão necessários para a aplicação, buscando assim uma melhor performance, já que não serão disponíveis todos os recursos existentes num SO. Esses recursos devem ser disponíveis de forma eficiente.

Conforme relata o autor, não é uma tarefa fácil, pois existem muitas variáveis que precisam ser levadas em conta, garantindo assim a disponibilidade de recursos, garantindo também o desempenho e a confiabilidade.

2.5.1 EPOS

Em 2001, Fröhlich propôs um sistema operacional adaptável que procura através das técnicas de *Application Oriented System Design* (AOSD) aproximar ao máximo o SO das necessidades da aplicação.

O resultado desta proposta citada acima é o EPOS, que segundo o autor demonstra ser um sistema adequado para ser utilizado em aplicações e sistemas dedicados, como é o caso do *Clusters*, reduzindo assim o preço de sobrecarga de um sistema genérico, garantindo o acesso apenas aos recursos necessários, além de garantir uma maior confiabilidade do sistema.

O EPOS é codificado em C++, e atualmente é para rodar nativamente em plataformas ix86, ou guest-level no Linux. Pode ser configurado tanto para ser embutido na aplicação, como no Kernel. O sistema Linux-guest é implementado através de uma biblioteca e um módulo carregável do Kernel. Ambas as versões dão apoio a Myrinet.

O EPOS permite desabilitar ou mesmo configurar o sistema às necessidades apresentadas por cada aplicação, diminuindo assim o *Overhead* (sobrecarga) do sistema.

Por suas características, o EPOS foi selecionado como plataforma ideal para *Clusters* do projeto SNOW que se encontra no laboratório de Bioinformática da UFSC.

3. CONCLUSÃO

Clusters computacionais são uma ótima opção para substituir os supercomputadores no que se refere a custo/desempenho. Mas para que se consiga alcançar um melhor desempenho, é necessário levar em consideração alguns pontos, como o hardware que compõe o sistema, softwares adequados com a aplicação e uma estrutura de comunicação apropriada.

Foi constatado que através de um sistema operacional específico para sistemas distribuídos, que seja ajustável com a aplicação, aumentam a performance do sistema. A AOSD e o EPOS mostram que é possível desenvolver software básico de altíssimo

nível e que se, desenvolvido de forma correta, podem garantir a qualidade do produto final. Utilizando a organização em famílias é possível que haja a escolha somente dos membros de famílias que satisfaçam as necessidades da aplicação e dessa forma é que a AOSD garante a qualidade do software gerado.

Outro ponto em que melhora a performance é a utilização do protocolo de comunicação correto. Os protocolos padronizados traz a vantagem da facilidade, mas em contrapartida reduz a performance da comunicação. A escolha correta de um protocolo leve ou até o desenvolvimento de um protocolo específico para a aplicação pode trazer ótimos resultados, aumentando assim a performance da comunicação.

Combinando hardware de alta qualidade, sistemas operacionais, protocolos e técnicas de engenharia de software é possível obter um ganho maior no que se refere à performance do sistema, tornando ainda mais a opção de *Cluters* competitiva com os supercomputadores.

4. REFERÊNCIA

FRÖHLICH, A. A.; TIENTCHEU, G. P.; SCHRÖDER-PREIKSCHAT W. *EPOS and Myrinet: Effective Communication Support for Parallel Applications Running on Clusters of Commodity Workstations*. Acessado em agosto de 2008. Disponível em: <http://www.lisha.ufsc.br/~guto/publications/hpcn2000.pdf >

LISHA. Laboratory for Software and Hardware Integration. Acessado em agosto de 2008. Disponível em <www.lisha.ufsc.br >

MYRICOM. *Overview of Myrinet*. Acessado em agosto de 2008. Disponível em <www.myri.com>

SANCHES, A. L. G. *Sistema de Comunicação de Alto Desempenho Baseado em Programação Genérica*. Acessado em agosto de 2008. Disponível em <www.lisha.ufsc.br/~guto/teaching/theses/gobbi.pdf>

SECCO, F. *Uma Família de Abstrações Myrinet para EPOS*. Acessado em agosto de 2008. Disponível em <www.lisha.ufsc.br/~guto/teaching/theses/secco.pdf>