Fundação Educacional do Município de Assis
Instituto Municipal de Ensino Superior de Assis
Campus "José Santilli Sobrinho"

WERNER ANOBILE SCHWARZ

COMPUTAÇÃO DE ALTO DESEMPENHO UTILIZANDO CLUSTER DE COMPUTADORES

Assis
2013

WERNER ANOBILE SCHWARZ

COMPUTAÇÃO DE ALTO DESEMPENHO UTILIZANDO CLUSTER DE COMPUTADORES

Trabalho apresentado para o Programa de Iniciação Científica (PIC) do Instituto Municipal de Ensino do Município de Assis – IMESA e à Fundação Educacional do Município de Assis – FEMA
Orientador: Fábio Eder Cardoso
Linha de Pesquisa: Rede de Computadores

Assis
2013

FICHA CATALOGRÁFICA

ANOBILE SCHWARZ, Werner
Computação De Alto Desempenho Utilizando Cluster De Computadores/Werner Anobile Schwarz. Fundação Educacional do Município de Assis – Assis, 2013.

22 pg.
Orientador: Msc. Fábio Eder Cardoso.
Programa de Iniciação Científica (PIC) – Instituto Municipal de Ensino Superior de Assis

1. Cluster. 2.Alto-Desempenho.

001.61
Biblioteca da FEMA

COMPUTAÇÃO DE ALTO DESEMPENHO UTILIZANDO CLUSTER DE COMPUTADORES

WERNER ANOBILE SCHWARZ

Trabalho apresentado para o Programa de Iniciação Científica (PIC) do Instituto Municipal de Ensino do Município de Assis – IMESA e à Fundação Educacional do Município de Assis – FEMA analisado pela seguinte comissão organizadora:

Orientador: _____

Analisador 1:_____

Assis
2013

RESUMO

COMPUTAÇÃO DE ALTO DESEMPENHO UTILIZANDO CLUSTER DE COMPUTADORES

Diversos computadores, independentemente de suas configurações de hardware podem ser conectados em rede formando um Cluster. No presente trabalho serão apresentados os componentes necessários para a construção de um cluster, como os componentes mais básicos de hardware e software assim como o conceito de cluster, baseando-se em softwares livres.

Palavras-chave: Cluster; Alto Desempenho.

ABSTRACT

HIGH PERFORMANCE COMPUTING USING CLUSTER COMPUTERS

Many computers, regardless of your hardware settings can be networked to form a cluster. Necessary for the construction of a cluster, as the most basic components of hardware and software as well as the concept of clustering, based on free software components will be presented in this work.
Keywords: Cluster, High Performance.

.

SUMÁRIO

1. INTRODUÇÃO

Uma crescente demanda por poder de processamento sempre foi um grande desafio na informática. Apesar de com o passar dos anos ter sido possível aumentar consideravelmente a velocidade dos processadores há limitações técnicas que impedem que este aumento siga indefinidamente utilizando-se a tecnologia atual.

Uma solução a este problema é organizar muitos CPUs de forma a fazê-los executar tarefas em conjunto, somando seu poder computacional. Os multicomputadores, também conhecidos como Aglomerados de Computadores ou *Clusters* são computadores completos, cada qual com seu processador, memória e placa de rede conectados entre si. Não necessitando de hardwares específicos para sua construção, os Clusters são baratos e fáceis de implementar.

Existem diversas tecnologias para conectar um computador ao outro formando um Cluster. De acordo com a quantidade de nós a serem utilizados, algumas alternativas tornam-se mais eficientes. Neste trabalho, será implementado o projeto bidimensional de Malha ou Grade, amplamente utilizado em sistemas comerciais e de fácil escalamento para tamanhos grandes.

2. SISTEMAS OPERACIONAIS

Sistemas operacionais são programas responsáveis por gerenciar os recursos do sistema, fornecendo uma interface entre o hardware o usuário.

As tarefas de um sistema operacional podem ser resumidas em comunicação do usuário; carregar, executar, pausar e parar programas; gestão e alocação de uso do processador; gestão de memória interna para aplicações, gestão e exploração dos equipamentos conectados e funções de proteção, por exemplo, por restrições de acesso.

Atualmente, os sistemas operacionais mais utilizados no mundo são o Windows, o OS X e o Linux, sendo este último de código aberto e disponível em diversas distribuições.

2.1. DEBIAN

A distribuição Debian GNU/Linux inclui o kernel do sistema operacional Linux e centenas de aplicações pré-empacotadas. Fundado por Ian Murdock em 16 de Agosto de 1993, o Projeto Debian reúne voluntários que produzem um sistema operacional composto inteiramente por software livre.

Para alcançar e manter um alto padrão de qualidade, o Debian adotou um rico conjunto de políticas e procedimentos para empacotamento e distribuição de software. Backups são automatizados através de ferramentas e a documentação detalha todos os elementos chaves do Debian de uma forma aberta e visível.

2.2. OPENMOSIX

A extensão OpenMosix é instalada no núcleo do Linux para a configuração de *clusters,* possibilitando a conversão de uma rede clássica de computadores em um super-computador para aplicações Linux.

Ao ser instalado, o OpenMosix faz com que os nós do *cluster* mantenham comunicações entre si, transmitindo informações acerca da disponibilidade dos recursos (processador e memória), permitindo com estas informações disponibilizar os seus recursos de maneira adequada para cada nó. Desta forma, se um nó com vários processos detecta que outro nó tem disponibilidade superior, então o OpenMosix encarrega-se de migrar um desses processos para esse nó, dando origem ao processamento distribuído.

O OpenMosix tenta continuamente classificar os custos de trasladação e fazer previsões sobre a viabilidade da mesma e utiliza o seu próprio sistema de arquivos, o *OpenMosix Filesystem* (oMFS) para permitir as trocas de dados entre vários processos.

Este mecanismo suporta algumas das funcionalidades de *Inter Process Communication* (IPC) mais simples, como *canalizações*, FIFOs, e redirecionamento de ficheiros. Utilizando oMFS e uma configuração adequada, é ainda possível permitir aos processos remotos o acesso diretos a arquivos e dados, ainda que estes não existam no nó anfitrião do processo.

As principais características do OpenMosix são a escalabilidade, ou seja, a capacidade de permitir que sejam adicionados novos nós à rede, que se adapta automaticamente sem a necessidade de requisitos ou softwares adicionais; a adaptabilidade, que permite nós com diferentes arquiteturas desde que a versão OpenMosix instalada em todos eles seja a mesma; e a possibilidade de ser acessado por qualquer aplicação sem a necessidade de alterações em seus códigos.

A maior desvantagem do OpenMosix é a dificuldade em migrar processos que utilizem chamadas IPC (*Inter Process Communication*), nomeadamente semáforos, memória partilhada e filas de mensagens. Esta característica é uma situação normal para programas que falhariam ao serem movimentados pelo OpenMosix. Estes programas devem rodar como planejado no nó onde foram iniciados.

2.3. MÁQUINAS VIRTUAIS

Existente desde a década de 1960, a tecnologia de máquinas virtuais (VM) possibilita instalar diversos sistemas operacionais em um único hardware. A VM emula normalmente um ambiente físico fazendo com que o uso de CPU, memória, disco, rede e outros recursos de hardware sejam gerenciados pela camada de virtualização, traduzindo essas solicitações para o hardware físico subjacente.

A máquina virtual é executada sob o controle do *Virtual Machine Monitor* (VMM) ou *hypervisor*. “Com a tecnologia das máquinas virtuais, o único software que funciona no módulo núcleo é o hipervisor, que tem duas ordens de magnitude, menos linhas de código que um sistema operacional e, portanto, menos erros.”.

O Sistema operacional não virtual, instalado na máquina é chamado de *host*, ou hospedeiro. Já os sistemas instalados em VM são denominados convidados. Normalmente, os sistemas operacionais convidados e programas não podem detectar se eles estão rodando em uma plataforma virtual, podendo ser utilizados como se fossem instalados no hardware de servidor físico. Por exemplo, um disco físico pode ver a partir do sistema operacional convidado. Os pedidos reais de E/S são traduzidos pela camada de virtualização e redirecionados a um arquivo para o qual o sistema operacional hospedeiro tem acesso.

As máquinas virtuais oferecem muitas vantagens em relação à instalação direta de sistemas operacionais e software no hardware físico. Uma delas é o isolamento proporcionado, garantindo que o sistema operacional hospedeiro ou outras VMs não interfiram nos aplicativos e serviços que são executados dentro de uma máquina virtual. As máquinas virtuais podem ser facilmente movidas, copiadas e re-alocadas entre os servidores hospedeiros para otimizar a utilização de recursos de hardware. Os administradores também podem tirar proveito de ambientes virtuais para simples backups ,recuperação de desastres , novas implantações e tarefas de administração do sistema básico de uso.

3. HARDWARE

A estrutura geral dos computadores atuais foi concebida em meados dos anos 1940 pelo matemático John von Neumann. Podemos classificar como principais componentes de um computador o processador, a memória RAM, a placa mãe e o disco rígido, responsáveis pela execução e armazenamento de tarefas.

Com a tarefa de conectar diversos computadores, podemos destacar a tecnologia ethernet e o switch.

3.1. PROCESSADOR

O processador, também abreviado como *CPU*, do inglês *Central Processing Unit* (Unidade Central de Processamento), tem um papel central na construção de um computador. Ele consiste de um sistema de transistores, que funcionam como minúsculos interruptores elétricos trabalhando no sistema binário, permitindo ou não a passagem de corrente elétrica, sendo que a posição 'off' corresponde ao estado 0 e a posição 'on' ao o estado 1.

A função do processador é executar instruções advindas de programas armazenados na memória principal.

Um processado é composto de registros, uma unidade aritmética lógica, uma unidade de controle e as linhas de dados (bus) que permitem a comunicação com outros componentes.

Um processador com velocidade do *clock* de um Hertz é capaz de processar com precisão uma operação por segundo. Isto significa que o período de duração da vibração é de um segundo. Se for realizado um overclock para dois hertz, este seria capaz de lidar com duas operações por segundo, uma vez que o período da oscilação tem apenas metade do tempo e a assim por diante.

Ocorre que quanto mais ciclos por segundo a CPU é capaz de realizar, mais calor ela produz. Ao mesmo tempo, em algum momento, haverá limitações físicas para o aumento na velocidade das CPUs.

3.2. MEMÓRIA RAM

Comumente conhecida como memória de acesso aleatório, a memória RAM (*random-access memory*), é a memória primária do computador. Ela lê e grava informações através de sinais elétricos sendo que a ausência de carga é interpretada com a representação do dígito *0* e a presença de carga com o dígito *1*. As informações são guardadas na memória apenas enquanto há corrente elétrica, sendo perdidas quando esta corrente cessa. Por esta característica, a memória RAM é considerada uma memória volátil.

3.3. PLACA MÃE

A placa-mãe é o soquete permite a conexão de todos os elementos essenciais do computador. Na forma de uma placa grande de circuitos, possui conectores para placas de expansão, cartões de memória, processador, entre outros. Existem vários formatos de placa mãe sendo que o mais utilizado atualmente é o ATX.

A placa-mãe é constituída por diversos elementos que estão integrados na placa de circuitos. O *Chipset* coordena os diversos componentes do computador como processador, memória, etc. e pode possuir um chip gráfico ou um chip de áudio integrado. O *RTC* (*Real Time Clock*) é responsável pela sincronização dos sinais do sistema sendo que *Frequência de Clock* (representado em *MHz*) é um sinal que oscila em determinada frequência que, quanto maior, mais informações pode editar o sistema. O *CMOS* (*Complementary Metal-Oxide-Semiconductor*) é alimentado continuamente por uma bateria, sendo responsável por manter informações sobre o sistema, como a hora do sistema, data do sistema, e alguns parâmetros importantes do sistema. A *BIOS* (*Basic Input/Output System*) é um programa básico que serve como uma interface entre o sistema operacional e a placa-mãe. Os Slots permitem

que sejam conectados à placa mãe dispositivos como placas de vídeo, som, modem e rede ao barramento.

3.4. DISCO RÍGIDO

Também conhecido como memória secundária, o Disco Rígido o faz de maneira mais permanente através de um padrão magnético. Constituído por um disco de metal coberto com material magnetizável, ele gira em alta velocidade sob um cabeçote, também conhecido como cabeça de leitura/gravação, lendo ou gerando padrões magnéticos.

3.4.1. Memória Virtual

Dividindo-se a memória em partes iguais de tamanho fixo, conhecidos como blocos, os processos também podem ser divididos da mesma forma, sendo que cada parte destes processos é chamada de página. Essas páginas podem ser alocadas nos blocos de memória disponíveis, otimizando o uso da memória. Este conceito aqui descrito resumidamente serve como base para outro que proporciona o uso ainda mais eficiente da memória conhecido por Memória Virtual.
Com a memória virtual, a paginação é feita *sob demanda*, ou seja, cada página do processo é trazida à memória apenas quando é necessária. Desta forma, como não é necessário carregar um processo inteiro na memória, este processo pode ser maior do que a área total da memória eliminando a necessidade, por parte do programador, de saber a quantidade de memória que ele tem disponível, pois, esta tarefa fica a cargo do hardware e do sistema operacional.

3.5. ETHERNET

Em 1973, Robert Metcalfe trabalhava no Centro de Pesquisa de Palo Alto da Xerox (PARC) quando entrou em contato com um artigo que descrevia o Sistema *Aloha* como uma alternativa para comunicação entre computadores. O artigo descrevia o

desenvolvimento de uma rede baseada em rádio que veio a ser conhecida como ALOHANET.

Inspirado pelo papel ALOHANET e com a ajuda de David R. Boggs, Metcalfe começou a elaborar suas ideias e escreveu um memorando no qual esboçou um esquema rápido que mudaria para sempre a comunicação em rede, era o início do Ethernet.

O primeiro protótipo Ethernet, um CSMA/CD sistema de 2,94 Mbps conectava mais de 100 estações de trabalho em um cabo de 1 km. Em 1979, Metcalfe deixou o PARC para fundar uma nova empresa chamada 3Com, convencendo a Xerox, Intel e Digital Equipment Corporation (DEC) a promover o Ethernet como um padrão.

Foi formado um comitê para desenvolver padrões de rede de área local definindo e especificando as camadas de software para Ethernet com fio, sendo que em 23 de junho de 1983, o IEEE 802.3 foi aprovado como padrão.

Atualmente, a tecnologia Ethernet mais utilizada é a Fast-Ethernet, capaz de transmitir dados a uma taxa nominal de 100 mbps.

3.6. SWITCH

Em uma rede Ethernet totalmente formada por switches, cada segmento é dedicado a um nó, ou seja, seu ponto de conexão, normalmente um computador. Estes segmentos se conectam a um switch, que suporta múltiplos segmentos dedicados, sendo que o número destes pode chegar a centenas.

O switch captura cada transmissão de um nó antes que esta atinja outro nó e encaminha então o pacote para o segmento apropriado. Uma vez que cada segmento contém apenas um único nó, o pacote só chega ao destinatário pretendido. Este arranjo permite múltiplas trocas simultâneas em uma rede que usa um switch.

Redes formadas por switchs empregam cabo par trançado ou fibra óptica. Tanto sistemas de cabo par trançado como cabos de fibra óptica utilizam condutores independentes para enviar e receber dados. Neste tipo de ambiente a rede dedica uma pista separada de tráfego que flui em cada direção dos nós. Esta dedicação permite que os nós possam transmitir ao switch, ao mesmo tempo em que o switch transmite para os nós. Transmissão em ambos os sentidos, também pode efetivamente dobrar a velocidade da rede quando dois nós trocam informações. Por exemplo, se a velocidade da rede é de 10 Mbps, cada nó pode transmitir a 10 Mbps, ao mesmo tempo.

Três métodos de encaminhamento de tráfego são utilizados pelos switches: o cut-through, o store-and-forward e o fragment−free.

Switches cut-through fazem a leitura do endereço MAC logo que um pacote é detectado pelo switch. Depois de armazenar os seis bytes que compõem a informação de endereço, os switches começam imediatamente a enviar o pacote para o nó de destino, mesmo que o resto dele ainda esteja entrando no switch.

Um switch que usa store-and-forward salva todo o pacote no buffer e verifica os erros ou outros problemas no pacote usando o CRC (Cyclic Redundancy Check). Se o pacote tem um erro, é descartado, caso contrário, o switch procura o endereço

MAC e envia o qpacote para o nó de destino. Muitos switches de combinam os dois métodos, utilizando cut-through até um que certo nível. Muito poucas opções são estritamente cut-through porque este não fornece correção de erros.

Um método menos comum é o fragment−free. Neste, o switch são armazenados os primeiros 64 bytes do pacote antes de enviá-lo. A razão para isto é que a maior parte dos erros e todas as colisões ocorrem durante os primeiros 64 bytes de um pacote.

4. CLUSTER BEOWULF

Desde o início dos anos 80, mais e mais estações de trabalho RISC baseados em Unix começaram a povoar os departamentos e institutos. Sua eficiência crescente levou à ideia de reuni-los e, à noite e nos fins de semana, quando eles não eram utilizados de forma a aproveitá-los para aplicações de computação intensiva. Nasceu então o conceito de Rede de Estações de Trabalho.

O passo seguinte foi simplificar utilização desses agregados para que eles parecessem, para o usuário, como um único computador paralelo virtual. Isso foi feito usando o paradigma de programação de troca de mensagens, que permite uma comunicação relativamente fácil de usar entre processos em diferentes computadores. O primeiro representante de destaque foi a biblioteca de troca de mensagens PVM (Parallel Virtual Machine), eficaz para aplicações heterogêneas que exploram os pontos fortes específicos de cada máquina em uma rede.

A relação custo/desempenho favorável de PCs e hardwares de rede abriu a possibilidade de construir computadores paralelos de PCs hardware completos. Nesta categoria, o primeiro computador conhecido foi o 486-cluster “Beowulf” NASA idealizado por Thomas Sterling e Donald Becker em 1994. O nome Beowulf, para descrever esta categoria de computação, naturalizou e primeiro computador deste tipo apareceu no TOP500 lista dos computadores mais rápidos do mundo.

O Cluster Beowulf é um tipo de aglomerado que pode ser construído a partir de componentes de hardware facilmente encontrado em lojas, gerenciado por um sistema operacional Open Source, ou seja, de código fonte aberto, como o Linux e interconectados por uma rede de alta velocidade privada.

Este tipo de cluster funciona como se fosse apenas uma máquina, utilizando todos os recursos disponíveis em seus nós para executar tarefas de alto desempenho, sendo apenas um dos nós dedicado à conexão com o mundo exterior, possuindo periféricos de entrada e saída.

Apesar de poder ser construídos que uma ampla variedade de peças, os sistemas Beowulf também podem ser constituídos por máquinas específicas visando o aumento de seu desempenho. Este tipo de configuração, entretanto pode aumentar o orçamento para a montagem e manutenção do cluster.

A tecnologia de rede Ethernet, principalmente da categoria Fast Ethernet, possibilitou a construção de sistemas de memória distribuída larguras de banda relativamente altas, latência razoavelmente baixa a baixo custo.

Sistemas operacionais livres, como o Linux, amplamente disponíveis, confiáveis e com bom suporte ao usuário, são distribuídos com o código-fonte completo, o que incentiva o desenvolvimento de ferramentas adicionais como controladores de baixo nível, sistemas de arquivos paralelos, e bibliotecas de comunicação.

O poder e os baixos preços dos computadores pessoais disponíveis atualmente e a grande disponibilidade de conexões de rede Ethernet de 100/1.000 Mbps, tornam viável a construção de computadores de alto desempenho e ambientes de computação paralela.

5. CONCLUSÃO

A construção de um cluster é uma alternativa viável em se tratando de custo/benefício. É possível aproveitar computadores que, de outra forma, teriam pouca utilidade, para realizar tarefas que exigem alto poder de processamento.

As informações sobre como construir clusters estão altamente disponíveis na internet, em comunidades bastante ativas, o que torna acessível a todos testar as possibilidades abertas por este recurso.

Além do openMosix, é possível utilizar outras extensões para clustering como a MOSIX, OpenSSI e a Rocks. Cada qual com suas características, todas open-source.

O presente trabalho foi de suma importância para a construção do conhecimento em torno da tecnologia. Este projeto foi publicado em dois eventos de relevante significância, sendo o primeiro no VI fórum científico e I mostra de software que ocorreu durante 21 a 23 de outubro de 2013 na Fundação Educacional do Município de Assis - FEMA, a segunda publicação se deu no XX fórum de iniciação científica e I fórum de desenvolvimento tecnológico e inovação da Universidade do Sagrado Coração localizada em Bauru/SP, no período de 18 a 23 de novembro.

REFERÊNCIAS

KUROSE, James; ROSS, Keith. **Redes de computadores e a Internet: uma abordagem top-down**. 5ª ed. Tradução Opportunity translations. São Paulo: Editora Addison Wesley, 2010.

PITANGA, Marcos. Construindo Supercomputadores com Linux. 3ª ed. Rio de Janeiro: Editora Brasport, 2008.

RIBEIRO, Uriá. **Sistemas Distribuídos: Desenvolvendo Aplicações de Alta Performance no LINUX.** 1ª ed. Rio de Janeiro: Editora Axcel Books, 2005.

SLOAN, Joseph D. **High Performance Linux Clusters with OSCAR, Rocks, OpenMosix, and MPI.** 1 ed. Sebastopol: O'Reilly Media, 2004.

STALLINGS, Willian. **Arquitetura e Organização de Computadores.** 8ª ed. Tradução Opportunity translations. São Paulo: Editora Pearson Prentice Hall, 2010.

TANEMBAUM, Andrew. **Sistemas Operacionais Modernos.** 3ª ed. Tradução de Ronaldo A.L. Gonçalves. São Paulo: Editora Pearson Prentice Hall, 2009.

TANEMBAUM, Andrew. **Redes de Computadores.** 4ª ed. Tradução de Vanderberg D. de Souza. Rio de Janeiro: Editora Elsevier, 2003.