Instalação e Configuração de *Clusters* de Estações de Trabalho: Experiência do Laboratório de Computação de Alto Desempenho do Departamento de Informática da UFES

Sérgio N. Simões, Soterio F. De Souza, Leonardo Muniz, Dijalma Fardin Jr.,
Alberto F. De Souza, Neyval C. Reis Jr., Andrea Maria P. Valli, Lúcia Catabriga
Departamento de Informática – Universidade Federal do Espírito Santo
{sns, soterio, lmuniz, dijalma, alberto, neyval, avalli, luciac}@inf.ufes.br

Resumo

Devido à crescente demanda por capacidade computacional na resolução de problemas, agregados de computadores (Clusters) vêm se tornando uma das alternativas mais viáveis à utilização de Supercomputadores. Este artigo descreve a instalação, configuração e manutenção do Cluster Enterprise do Laboratório de Alto Desempenho (LCAD) do Departamento de Informática da Universidade Federal do Espírito Santo. Enterprise possui 64 nós de processamento e desempenho máximo teórico de 195,8 GFLOPS. Montado integralmente pela equipe do LCAD, Enterprise passou por várias análises de hardware e software que são discutidas neste artigo.

1. Introdução

Em diversas áreas existe a necessidade de se obter alto poder computacional na resolução de problemas cada vez maiores em períodos de tempo cada vez mais curtos. Processamento paralelo é uma das alternativas para se obter alto desempenho computacional. Dentre alguns problemas que podem se beneficiar de Processamento Paralelo podemos citar: previsão do tempo, exploração de petróleo, dinâmica dos fluidos, visão computacional, aerodinâmica, mapeamento do genoma humano, etc.

A aquisição de supercomputadores paralelos para resolver tais problemas nem sempre constitui a melhor alternativa, por serem máquinas usualmente caras. Atualmente, a utilização de *Clusters* de estações de trabalho [1, 2] tem permitido que muitas empresas tenham acesso a um alto poder computacional a um baixo custo. Devido a isso, os *Cluster*s têm se tornado cada vez mais uma alternativa aos supercomputadores. Um *Cluster* de Computadores é um grupo de computadores interligados em rede e configurados para trabalharem de forma paralela com um objetivo comum.

Este artigo descreve o procedimento usado para a montagem, instalação, configuração e manutenção do *Cluster* do Laboratório de Computação de Alto Desempenho (LCAD) do Departamento de Informática da Universidade Federal do Espírito Santo (UFES).

Totalmente operacional em janeiro de 2003 e com desempenho teórico máximo de 195,8 GFLOPS, o *Cluster* do LCAD é composto por 64 nós de processamento e uma máquina *host*, todos utilizando sistema operacional Linux *Red Hat 7.1* e interligados através de dois *switches Fast Ethernet* de 100Mb/s. O *Cluster* foi batizado de *Enterprise* (Figura 1) e figurava na lista dos clusters mais poderosos do mundo (http://clusters.top500.org) em 48^{o}. lugar em Janeiro de 2003 (hoje o *Enterprise* figura em 51^{o}. lugar nesta lista).

Figura 1. O *Cluster Enterprise*

2. Análise Inicial da Configuração

Na especificação de um *Cluster*, as configurações de nó de processamento e de rede são os itens mais importantes. Usualmente, dispõe-se de um montante de recursos e sabe-se qual será o principal tipo de aplicação que se deseja executar. Assim, para decidir-se sobre as configurações de nó e rede, deve-se examinar qual a razão processamento por comunicação (MFLOPS/Bytes

transferidos entre nós) das aplicações mais freqüentes. Se esta razão for baixa, deve-se aumentar a velocidade da rede de comunicação e colocar mais processadores em um mesmo nó. Outros indicadores que merecem consideração são o modelo de programação e as características dos algoritmos das aplicações mais utilizadas.

Para o *Enterprise*, escolhemos nós de processamento com um único processador e uma rede de baixa velocidade, *Fast Ethernet* – 100Mb/s, porque estamos interessados em desenvolver algoritmos para extrair desempenho de máquinas de baixo custo – uma rede de 1000Mb/s custaria aproximadamente 3 vezes mais que a rede usada e significaria 29 máquinas a menos, dado as limitações de recursos iniciais. Portanto, decidimos ter mais máquinas operando para poder atender a uma quantidade maior de usuários, mesmo que com uma rede mais lenta.

Como regra geral para nós de processamento deve-se buscar escolher máquinas com a maior quantidade possível de memória *cache* e memória principal, barramentos internos e externos ao processador de maior velocidade possível, tudo isso dentro dos recursos disponíveis. Quanto ao processador, as medidas SPECint e SPECfloat (www.specbench.org) permitem avaliar qual o melhor modelo em um determinado momento tecnológico (estas medidas avaliam, em certo grau, todos os parâmetros mencionados).

Para a escolha da rede de interconexão e, em particular, dos *switches*, pode-se usar como características mais importantes: taxa de transferência, latência, capacidade de comutação de pacotes e o melhor custo por porta.

3. Configuração do Hardware de *Enterprise*

Com os recursos disponíveis (R$140.000,00) conseguimos adquirir 64 nós de processamento, uma máquina servidora (*host*) e 2 *switches*. Cada um dos 64 nós de processamento adquiridos possui 256MB de memória e 20GB de capacidade de armazenamento, totalizando, assim, 16GB de memória RAM e 1,2TB (Terabytes) de capacidade de armazenamento. Alguns dispositivos vieram com drivers específicos para o sistema operacional Red Hat 7.1 (interface Gigabit, por exemplo) e, visto que a maioria dos computadores do departamento de informática já utilizavam este sistema, o mesmo foi escolhido para ser utilizado por todas as máquinas do cluster.

Uma vez decidida a arquitetura dos nós e da rede de interconexão, outros aspectos importantes são a compatibilidade entre Hardware/Software e a homogeneidade dos nós. Ou seja, de nada adianta adquirir uma placa mãe mais barata se o *chipset* desta não for compatível com o sistema operacional escolhido. Além disso, é muito importante que todos os nós tenham a mesma configuração de hardware, pois isto pode ajudar na instalação de um novo nó apenas pela replicação da imagem de um nó previamente instalado.

A seguir é apresentda a especificação dos nós e da rede de interconexão utilizada no *Enterprise*.

3.1 Nós de Processamento

Cada nó de processamento é composto por um processador ATHLON XP 1800+ fabricado pela AMD, versão Palomino de 0,18 mícrons, com cache L1 de 64Kb e L2 de 256Kb, barramento de 133MHz (ou 266MHz com dois acessos por ciclo de maquina) e memória RAM de 256MB, 133MHz e tempo de acesso de 7ns. Visto que o processador possui duas unidades de ponto flutuante e a freqüência de operação é de 1,53 GHz, tem-se um desempenho teórico de pico igual a 2 x 1,53 = 3,06 GFLOPS por nó de processamento. O desempenho teórico de pico do sistema como um todo é então igual a 64 x 3,06 = 195,8 GFLOPS.

A placa mãe de cada nó é do fabricante Soyo, modelo SY-K7VTA PRO, FSB (*front side bus*) 200/266MHZ, 5 *slots* PCI de 32 bits e 2 barramentos PCI internos. Um detalhe muito importante na escolha da placa mãe é o *chipset*, que deve ser compatível com o sistema operacional escolhido.

A placa de rede de cada nó é do fabricante 3Com, modelo 3C905TX-NM, *Fast Ethernet* 10/100 Mb/s, enquanto que o disco rígido é fabricado pela Samsung e tem 20GB de capacidade, 5.400 RPM, compatível com tecnologia SMART, padrão Ultra ATA 100, buffer de 2MB, tempo de acesso 8,9ms, latência 5,56ms, modelo Spintpoint V40.

3.2 Rede de Interconexão

A rede de interconexão é composta, basicamente, por dois *switches* 3Com modelo 4300. Cada *switch* possui 48 portas *Fast-Ethernet* (100Mb/s) para comunicação com os nós de processamento e são interligados entre si através de um módulo *Gigabit Ethernet* (1000Mb/s). Para uma melhor utilização dos *switches*, optou-se por dividir igualmente a quantidade de nós de processamento entre os dois *switches*. Além disso, um dos *switches* possui uma porta *Gigabit* a mais para conexão com o servidor (*host*). Assim, a máquina servidora consegue enviar dados para os nós de processamento de forma rápida. Alguns *switches*, como o que usamos, possuem opções de configuração que podem ser otimizadas para o *Cluster* a fim de evitar tráfego desnecessário na rede.

3.3 Servidor

A máquina Servidora (*host*) é composta por um processador ATHLON XP 1800+ com 512 MB RAM, 80 GB de disco, duas placas de rede, sendo uma 3COM *Gigabit Ethernet* para conexão com os nós de

processamento e uma 3COM *Fast-Ethernet* para conexão externa.

Esta é a única máquina que possui acesso aos nós de processamento. Assim, todas requisições devem passar por ela, sendo esta máquina o *Front-end* entre os usuários e os nós de processamento. Com isso, é possível tornar mais segura a rede do *Cluster*. Esta máquina também é responsável pela distribuição dos processos nos nós de processamento, pelo armazenamento das contas dos usuários, além das atividades de configuração, atualização e monitoramento dos nós de processamento. Com o emprego do *host* desta forma obtém-se um maior controle na administração dos nós de processamento, além de tornar a rede do *Cluster* mais segura.

3.4 Instalação Física, Refrigeração e Rede Elétrica

Uma vez determinadas a quantidade de nós, o tamanho dos gabinetes e a potência dissipada, deve pensar no espaço físico que será ocupado pelos nós e qual a melhor disposição deles. Os nós de processamento foram instalados em quatro estantes, 16 nós por estante, organizadas na forma de "U" (Figura 2). Com isso, a manutenção das máquinas é facilitada, pois qualquer máquina pode ser manuseada sem muita dificuldade.

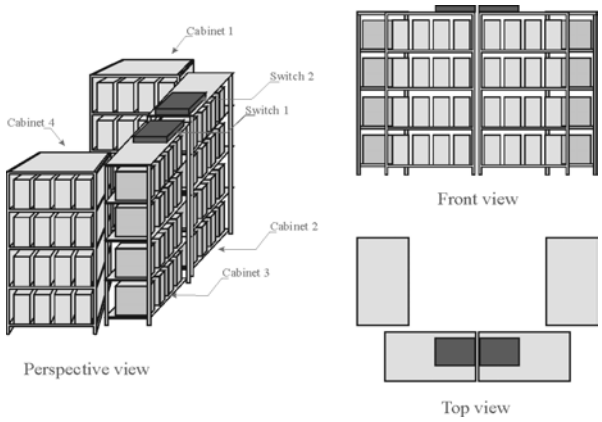

Figura 2. Disposição física dos nós

A refrigeração do sistema é de extrema importância. Esta é feita por dois condicionadores de ar que totalizam 39.000 BTUs e um ventilador que permite direcionar o fluxo de ar frio para o centro do "U", fazendo com que o ar quente suba, refrigerando assim todo o ambiente. Esta capacidade de refrigeração foi superestimada para o caso de haver falha em um dos condicionadores de ar.

Outro aspecto importante a ser levado em consideração é o planejamento da rede elétrica. A rede elétrica utilizada é uma rede com três fases distintas, cada uma alimentando um grupo de máquinas. A potência total da rede elétrica projetada foi de 50A x 127V x 3fases = 19.050Watts. Para este cálculo, estimou-se o consumo máximo de um determinado nó, multiplicou-se pela quantidade de nós, acrescido de uma margem de segurança.

3.5 Proteção do Hardware

Enterprise dispõe de um dispositivo de proteção que desliga automaticamente em casos de superaquecimento, sobre-tensão, sub-tensão, perda de neutro ou de terra. Este sistema, de baixo custo, foi projetado, montado e instalado pela equipe que desenvolveu *Enterprise*. Infelizmente, este sistema somente foi desenvolvido após a ocorrência de uma sobre-tensão na rede elétrica que ocasionou a queima de 29 fontes dos nós de processamento. O custo de um *short-break* comercial capaz de proteger *Enterprise* custaria pelo menos 10% do total investido em todo sistema.

4. Configuração de Software de *Enterprise*

Devido à disponibilidade de software livre de boa qualidade, optou-se por usar todo o recurso disponível para aquisição de hardware, conforme já mencionado. Assim, as escolhas de software, descritas a seguir, contém apenas software livre e gratuito.

4.1 Sistema Operacional

A mais importante escolha de software é a do sistema operacional. Como visto anteriormente, esta escolha não pode ser feita sem considerar o hardware; ou seja, deve-se escolher um software apropriado para o hardware escolhido. O sistema operacional escolhido foi o Linux Red Hat 7.1. Como a compatibilidade do sistema operacional com o hardware já havia sido verificada, não houve problemas de falta de *driver* ou outros semelhantes.

Inicialmente, foram feitas a instalação e configuração do sistema operacional da máquina servidora. Após isto, foi feita a configuração dos *switches* para trabalharem de forma otimizada. O procedimento adotado para instalação dos nós de processamento foi instalar apenas quatro máquinas inicialmente como teste e replicá-las apenas após verificar que estavam funcionando perfeitamente. Assim, a replicação dos demais nós foi feita apenas quando as primeiras haviam sido exaustivamente testadas.

Para que os nós de processamento realizassem a carga do sistema normalmente, foi necessário habilitar a opção de boot sem teclado na BIOS de todas as máquinas.

É importante mencionar que muitas dificuldades encontradas foram superadas após atualização dos pacotes de software relacionados.

4.2 Sistema de Arquivos

O sistema de arquivos escolhido foi o NFS. Este sistema permite que os arquivos dos usuários residam em apenas uma máquina, apesar de estarem disponíveis automaticamente em qualquer máquina do cluster

utilizada pelos mesmos via NIS. O conjunto NIS+NFS possibilita, ainda, a administração centralizada das contas dos usuários. Mas somente estas estão compartilhadas via NFS.

Visto que, na maioria dos casos, a rede de interconexão constitui um gargalo do sistema, os pacotes de software foram instalados em todos os nós de processamento para evitar tráfego desnecessário na rede. A desvantagem desta configuração é que cada atualização de pacote deve ser feita em todos os nós. Para minimizar este problema, foram criados Scripts para auxiliar nas tarefas de manutenção e instalação dos nós, permitindo alterar a configuração de todas as máquinas a partir de um único comando.

4.3 Autenticação de Usuários

Como mencionado, optou-se por instalar o sistema de autenticação de usuários NIS. Para minimizar problemas com a segurança, os nós de processamento e a máquina servidora foram configurados com IPs reservados. Deve-se ter em mente que sistemas de autenticação mais simples, como o NIS, nem sempre são suficientemente seguros, ao passo que sistemas de autenticação mais complexos podem não funcionar apropriadamente com um determinado software de gerência de *Cluster*.

4.4 Ferramentas de Desenvolvimento

Todos os softwares de desenvolvimento utilizados no laboratório são softwares de domínio público e, dentre eles, destacam-se: os compiladores gcc, g77, VF90 e HPF, as bibliotecas de troca de mensagens LAM-MPI, MPICH e PVM, as bibliotecas para cálculos numéricos BLAS & Atlas BLAS, além dos editores de texto emacs e vim que se destacam por serem facilmente utilizáveis remotamente.

4.5 Comandos Remotos

Para uma melhor manutenção, configuração e até mesmo o uso de ferramentas de gerência é necessário configurar todos os nós para aceitarem comandos remotos. Utilizou-se o pacote rsh, de conexão remota, tendo em vista que a ferramenta de gerenciamento de *Jobs* escolhida também utiliza este pacote. Além disso, *scripts* podem ser criados para automatizarem tarefas de manutenção por executarem um determinado comando em todas as máquinas simultaneamente.

4.6 Sistema de Gerenciamento de *Jobs*

Para tirar o máximo de proveito de um Cluster e permitir uma melhor utilização dos recursos é necessário algum sistema de gerenciamento de *Jobs*. Dentre as ferramentas de gerenciamento de *Jobs* disponíveis para a plataforma Linux, merecem destaque: SGE (Sun Grid Engine) [3] e OpenPBS (Portable Batch System) [4]. Optou-se pelo SGE porque este foi desenvolvido como produto e disponibilizado gratuitamente, além de existirem vários grupos que o utilizam e desenvolvem atualizações para esta ferramenta.

5. Operação e Administração

Manter dezenas de máquinas funcionando pode constituir um desafio. Quando não se possui um sistema de *no-break*, a probabilidade de máquinas apresentarem defeito aumenta consideravelmente. Isto pode acabar inviabilizando a utilização de todas as máquinas simultaneamente. Assim, o ideal é ter máquinas reservas para que haja reposição imediata em caso de falhas.

Constituem tarefas básicas de operação e administração: criação de contas, atualização dos pacotes, monitoramento / manutenção dos nós de processamento, reinicialização do sistema em caso de queda de energia, entre outras.

6. Análise de desempenho

Uma aplicação bastante utilizada para avaliação desempenho de supercomputadores paralelos é a *High-Performance Linpack benchmark*, que realiza basicamente uma fatoração LU de uma matriz densa. Nos últimos testes realizados com a *Linpack*, o *Enterprise* alcançou um desempenho de 52,3 GFLOPS. Dentre os fatores que impediram que o desempenho teórico máximo (195,8 GFLOPS) fosse alcançado podemos citar: latência da hierarquia de memória, tamanho das caches, latência e taxa de transferência da rede de interconexão, paralelismo no nível de instrução disponível na aplicação, entre outros. Estes fatores somados explicam a distância entre o desempenho máximo teórico e o medido com o *Limpack*.

7. Considerações finais

Como foi visto, a construção de um *Cluster* utilizando soluções não integradas pode representar um desafio, o que exige pessoal capacitado. Mas, uma vez tomadas as decisões corretas, pode-se obter alto desempenho a um custo realmente baixo.

8. Bibliografia

[1] BELL, G., Beowulf *Cluster* computing with Linux, The MIT Press, 2002.

[2] BUYYA, R., High performance *Cluster* computing: architectures and systems, Prentice Hall, 1999.

[3] Sun, Sun™ Grid Engine: Enterprise Edition 5.3 Administration and User's Guide, Sun Microsystems, Inc., 2002.

[4] BAYUCAN, A., et al. Portable Batch System Administration Guide, Veridian System, v2.3, 2000.