Computação de Alto Desempenho usando Clusters

César Castelo Fernández*
Instituto de Computação, Universidade Estadual de Campinas, Unicamp
Av. Albert Einstein 1251, Cidade Universitaria, CEP 13083-970
Campinas, SP, Brasil
ccastelo@liv.ic.unicamp.br
RA: 089028

RESUMO
Neste artigo são apresentados todos os conceitos relativos aos Clusters, sendo que eles são configurações que oferecem um desempenho muito bom e são muito baratos. O cluster é formado por muitos computadores que são interligados para funcionar como um computador único, para o qual são necessários muitos componentes que garantem o correto funcionamento do cluster. No artigo são descritos todos esses componentes e ao final é apresentado um exemplo muito importante que descreve o uso da arquitetura de cluster no mundo real, através de uma aplicação usada no mundo inteiro.

Palavras chave
Clusters, Computação Alto Desempenho, Cluster do Google, Multi-processamento, Redes de Computadores, Software Paralelo, Sistemas Operacionais

1. INTRODUÇÃO
Um cluster é um conjunto de computadores interligados, instalados e programados de tal forma que os seus usuários tenham a impressão de estar usando um recurso computacional único [5]. São usados para fazer alguma tarefa específica onde é necessário muito processamento, o que sería inviável fazê-lo com um computador só.

Uma das principais motivações para a construção dos clusters é o alto custo de um computador com arquitetura multiprocessador. Atualmente, o custo de implementar um cluster de computadores é muito mais baixo do que comprar um computador com arquitetura multiprocessador com o mesmo poder de processamento, já que os clusters podem ser construídos com estações de trabalho muito baratas, ou até com computadores pessoais. Outro motivo importante é o uso cada vez maior do processamento paralelo em atividades acadêmicas, científicas e empresariais. As universidades ou centros de pesquisa montam os clusters para fazer as provas nas pesquisas desenvolvidas por eles. Os clusters são usados em muitas áreas da ciência, como física, química, mecânica, computação, matematica, medicina, biologia, etc, incluindo problemas como predição numérica do clima, simulação de semicondutores, oceanografia, modelagem em astrofísica, sequenciamento do genoma humano, análise de elementos finitos, aerodinânica computacional, inteligencia artificial, reconhecimento de padrões, processamento de imagens, visão computacional, computação gráfica, robótica, sistemas espertos, exploração sísmica, análise de imagens médicas, mecânica quântica, etc.

Atualmente, os clusters são muito usados no mundo inteiro. Pelo fato de serem muito fáceis de implementar, são usados tanto por empresas muito grandes, quanto por empresas menores, estudantes ou até pessoas nas suas casas. Existem muitas implementações de clusters que são muito conhecidas porque realmente tem um desempenho ótimo, como o cluster do Google, da Sun, do Oracle, etc, e o fato de serem usados por empresas tão importantes como essas ajudou também a torná-los mais populares.

Na construção de um cluster é importante levar em conta aspectos muito importantes incluíndo Hardware e Software, porque são essenciais para conseguir o melhor desempenho possível para uma dada arquitetura. No Hardware destacam-se elementos como as interfaces de comunicação entre computadores, o tipo de processador usado em cada computador, a quantidade de memória, etc; e no software é muito importante a escolha do Sistema Operacional, software para a sincronização, compiladores especiais; assim como também é fundamental o desenvolvimento de aplicações paralelas que possam aproveitar ao máximo o poder de processamento paralelo do cluster.

O trabalho está organizado da seguinte maneira: na secção 2 são apresentados os principais componentes de um cluster, considerando o software e o hardware; na seção 3 são descritas as vantagens de implementar um cluster; na seção 4 são descritas as opções para implementar um cluster. A seguir, é aprensentado um exemplo real sobre implementação de um cluster na seção 5; e na seção 6 são apresentadas as conclusões do trabalho.

2. ELEMENTOS DE UM CLUSTER
Os clusters tem a vantagem de dar para o usuário muita liberdade na hora de escolher os componentes que formarão o cluster, sendo que ele pode escolher um cluster fabricado

*B. Eng César Castelo Fernández

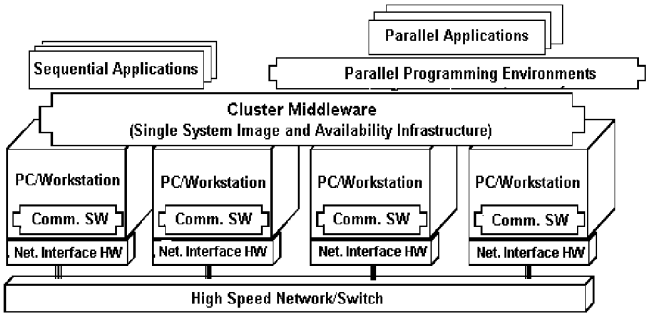

Figura 1: Arquitetura de um cluster.

completamente por um fabricante só, ou pode escolher as peças de diferentes marcas segundo as suas preferências.

Em geral, um cluster é composto por quatro elementos principais [9]: os computadores que são os nós, as redes de interligação (hardware), o conjunto de ferramentas para fazer programas a serem executados no cluster e o conjunto de programas para a administração do cluster (software). Esses quatro elementos são compostos de outros elementos, que serão estudados nas próximas seções [4]. Na Figura 1 pode-se observar a relação entre esses componentes.

2.1 Hardware

2.1.1 Computadores

O computador contém todos os dispositivos de hardware necessários para a execução dos programas, para o qual tem que armazenar a informação e usar interfaces para mostrar os resultados. Os principais componentes do computador são:

- Microprocessador: é o elemento que fornece o poder de processamento ao computador. É muito importante escolher um com bom desempenho, destacando também a quantidade de memória cache que tem. Os mais usados são as familias: Intel x86, Compaq Alpha systems, IBM Power PC Chip e SUN SPARC, porém, é possível implementar um cluster quase com qualquer microprocessador.
- Memória principal: aqui é onde está armazenada a informação usada pelos programas que estão em execução. A tecnologia mais usada atualmente é a DRAM e os fabricantes de memórias sempre estão aumentando a capacidade e velocidade de acesso. Elas são mais lentas do que a memória cache, mas tem muito mais capacidade.
- Motherboard: é o chip onde estão conectados todos os componentes do computador para formar um sistema único. Tem interfaces para a comunicação entre os componentes e cada vez estão sendo desenvolvidas interfaces mais rápidas e compatíveis com todo tipo de componentes, como o USB, PCI, etc.
- Armazenamento secundário: é onde a informação fica armazenada de maneira permanente. Sempre é muito maior do que a memória RAM, mas também é muito mais lenta. Este tipo de armazenamento não é fundamental para o cluster, porque está mais focado no processamento; até alguns nós no cluster tem apenas o espaço necessário para executar o sistema operacional, reduzindo os custos.

2.1.2 Multi-Processadores Simétricos

Além do paralelismo ser implementado distribuindo o processamento entre os diferentes computadores, também é importante usar em cada computador processadores paralelos, aumentando assim a capacidade de processamento total do sistema. Um multi-processador simétrico é formado por um conjunto de processadores e tem como característica principal o fato de todos os processadores terem acesso à mesma memória, ou seja, compartilhar a informação na memória. A sua principal vantagem é que todos os processadores tem a possibilidade de usar a memória completa, mas também é necessário que tenham mecanismos de escolha no hardware para o acesso à memória.

Atualmente este tipo de processadores não tem um custo muito elevado e, dependendo do processamento a ser feito pelo cluster, podem ser usados em todos os computadores do cluster, ou apenas em um servidor que é parte do cluster.

Também pode-se considerar um tipo de processadores muito parecidos, que são os Microprocessadores Múltiplos, que estão implementados como um único chip contendo vários processadores. Estes tem como vantagem que a comunicação entre eles é mais rápida porque estão no mesmo chip, além de poderem compartilhar a memória cache do chip, que é a memória mais rápida do computador. Atualmente este tipo de processadores é cada vez mais acessível e o seu poder de processamento é maior, assim como o número de processadores inclusos no mesmo chip.

2.1.3 Redes

Uma rede é uma combinação de meios físicos de transporte e mecanismos de controle para o transporte. A informação é enviada de um computador para outro usando mensagens, que podem ser muito pequenas ou muito grandes. Uma mensagem é um conjunto de bits que estão organizados segundo um formato, o qual permite que a mensagem seja corretamente interpretada. Quando a mensagem é enviada, é usada uma interface que faz a comunicação entre a aplicação e a rede. Essa interface aumenta na mensagem algumas informações que são usadas para encontrar o destino correto da mensagem quando é enviada pela rede.

Os parâmetros utilizados para medir o desempenho de uma rede são:

- Taxa de transferência: que é medida como a quantidade de informação que pode ser transmitida por unidade de tempo.
- Latência: que é o tempo total necessário para transmitir uma mensagem, a qual vai depender da taxa de transferência.

Existem muitas tecnologias para fazer a interligação entre os computadores na rede, as mais importantes são:

- Ethernet: é a tecnologia mais popular e barata, destacando-se dois: Fast Ethernet, que permite uma velocidade de 10 Mbps até 100 Mbps e ainda é a tecnologia mais usada nos clusters; e Gigabit Ethernet, que permite velocidades na ordem dos Gigabits por segundo e que são usadas somente quando é necessário uma alta taxa de transferência em aplicações como processamento de imagens de alta qualidade, consultas em tempo real, aplicações distribuídas, etc, sendo que também tem compatibilidade com as redes Fast Ethernet.

- Myrinet: é uma rede que usa interruptores de baixa latência e atinge velocidades de 640 Mbps até 2.4 Gbps. Uma de suas grandes limitações é que os interruptores que são usados para desenhar o caminho a ser percorrido pelos pacotes podem ser bloqueados, causando assim um desempenho mais baixo, mas essa dificudade tende a ser superada com a alta velocidade da rede.

- cLAN: é um conjunto de interruptores para clusters de alto desempenho, que oferece um desempenho muito bom na taxa de transferência e na latência, conseguindo velocidades de até 2.5 Gbps por cada um dos 13 portos disponíveis. O problema é que não tem uma especificação dos sinais que são enviados, nem das interfaces de comunicação, o que faz com que não sejam muito usadas.

- Scalable Coherent Interface: é uma interface desenvolvida pela IEEE para conectar sistemas de memória compartilhada com caches coerentes. É muito pouco usada porque as motherbords atuais não tem suporte para os mecanismos de coerência que são necessários para os SCI.

- Infiniband: Mistura conceitos usados na interface cLAN, com um conjunto de especificações detalhadas para conseguir produtos mais compatíveis. É uma interface que está sempre em desenvolvimento para atingir velocidades cada vez maiores.

2.2 Software

2.2.1 Sistema Operacional

É o software básico para a administração dos recursos do cluster, é fundamental na correta operação do cluster. Existem várias tarefas que tem que ser feitas pelo sistema operacional:

- Consulta: os trabalhos iniciados por diferentes usuários em diferentes lugares e com requisitos diferentes tem que ser armazenados pelo sistema operacional em uma fila de prioridades até que o sistema esteja na possibilidade de fazer os trabalhos para cada usuário.

- Planificação: Talvez seja o componente mais complexo do sistema, pois tem que administrar as prioridades aos trabalhos iniciados pelos usuários, levando em conta os recursos do sistema e as políticas estabelecidas pelo administrador do sistema. O planejador deve fazer um equilíbrio entre as aplicações que vão ser executadas, já que há aplicações que devem usar todos os nós do cluster. Embora outras apenas usem alguns nós, há outras que tem que ter respostas e visualizações em tempo real; algumas precisam ter a maior prioridade para terminar sua execução antes do que as outras.

- Controle de recursos: As aplicações deste tipo colocam as aplicações nos nós atribuidos pelo planificador, disponibilizam os arquivos necessários, começam, terminam e suspendem trabalhos. Notificam ao planificador quando os recursos estão disponíveis.

- Monitoração: Para um adequado controle do desempenho do cluster é necessário que sempre seja reportado o estado geral do cluster a algum lugar de controle centralizado, ou a algum dos nós do cluster. Os relatórios tem que incluir disponibilidade de recursos, estado das tarefas em cada nó e quão "saludáveis" estão os nós. Quando esta informação está disponível são executadas certas ações já programadas.

- Contabilidade: Em todo cluster sempre é necessário armazenar informações quantitativas sobre o funcionamento do sistema, seja para calcular os custos de operação para depois serem faturados aos clientes, ou simplesmente para medir o desempenho do sistema, através do análise da utilização do sistema, disponibilidade, resposta efetiva às exigências.

Tradicionalmente, os sistemas operacionais mais usados para implementar clusters são Linux, Windows e Solaris. A seguir serão explicadas as suas características mais importantes tomando como exemplo uma distribução de cada um deles [4]:

- Linux Redhat 7.2 (2.4.x kernel) [8]: É uma das distribuções mais populares do Linux, e atualmente a IBM investe muito dinheiro no desenvolvimento do sistema operacional, sendo que existe uma versão livre e outra que é vendida pela IBM, chamada Redhat High Availability Server, que é muito mais poderosa e foi desenvolvida para trabalhar com clusters. As principais vantagens do Redhat são: o código é livre para modificar, suporte para multi-tasking, suporte para muitos tipos de sistemas de arquivos.

  Oferece um sistema especial de arquivos chamado Parallel Virtual File System, que foi desenvolvido especialmente para oferecer alto desempenho em execuções em paralelo. A versão de Rredhat para clusters desenvolvida pela IBM é uma opção muito boa para fazer um cluster, sendo que oferece um custo baixo em comparação com outros productos no mercado, mas é muito bom para trabalhar com Servidores Web, Servidores FTP, firewalls, VPN gateways, etc. As características principais oferecidas são: alto desempenho e escalabilidade, flexibilidade, maior segurança, disponibilidade, baixo custo, suporte.

- Microsoft Windows 2000 [8]: É o sistema operacional que domina o mercado dos computadores pessoais atualmente e é baseado na arquitetura Windows NT, que é implementada para 32 bits, e é estável, suporta multi-tasking e multi-thread, tem suporte para diferentes arquiteturas de CPU (Intel x86, DEC Alpha, MIPS, etc), tem um modelo de segurança baseado em objetos, tem o sistema de arquivos NTFS e protocolos

de rede TCP-IP, IPX-SPX, etc. Um grande problema do Windows é o custo, porque para construir o cluster é necessário comprar licenças para cada nó.

Para o trabalho com cluster é usado o Windows 2000 Cluster Server (chamado de Wolfpack), que é formado pelos seguintes componentes: Database Manager (contém informação dos nós do cluster, tipos de recursos, grupos de recursos), Node Manager (controla o correto funcionamento de cada nó, fazendo testes entre diferentes nós para ativar ou desativar o nó quando alguma comunicação entre eles não pode ser completada), Event Processor (é o centro de comunicações do cluster e é responsável por comunicar aplicações e componentes do cluster quando é solicitado), Communication Manager (é responsável pela comunicação entre o Cluster Service de um nó com algum Cluster Service de outro nó) e Global Update Manager (é responsável por mudar o estado dos recursos do cluster e enviar notificações quando ocorrem as mudanças).

- SUN Solaris [1]: É um sistema operacional baseado em UNIX, que é multi-thread e multi-user. Suporta as arquiteturas Intel x86 e SPARC, possuindo suporte aos protocolos de rede TCP-IP e Remote Procedure Calls (RPC). Também tem recursos já integrados no kernel para suporte a Cluster.

  Para trabalhar com cluster existe um sistema operacional distribuído que é baseado no Solaris, que é o SUN Cluster, sendo que o sistema inteiro é oferecido ao usuário como uma imagem única, ou seja como se fosse um sistema operacional único e não um cluster. A arquitetura do SUN Cluster é formada pelos seguintes componentes: Object and Communication Support (é usado o modelo de objetos CORBA para definir os objetos e o mecanismo para RPC), Process Management (administra as operações dos processos tal que sua localização é transparente ao usuário), Networking (o sistema implementa sistema de rede para dar suporte à implementação do cluster) e Global Distributed File System (é implementado um sistema global de arquivos, para simplificar as operações de arquivos e a administração de processos).

2.2.2 *Middleware*

É o software que é aumentado no Sistema Operacional para conseguir que todos os computadores no cluster funcionem como um único sistema (chamado de Single System Image). Também tem uma capa de software que permite acesso uniforme aos nós, mesmo tendo diferentes sistemas operacionais. Este software é o responsável por prover alta disponibilidade do sistema. Destacam-se dois software Middleware:

- Bibliotecas para Comunicações Paralelas

  A única maneira de fazer a comunicação entre computadores no cluster é enviando mensagens entre eles, e é por isso que o maior objetivo do Middleware é conseguir portabilidade entre computadores que usam diferentes sistemas operacionais (i.e. conseguir uma interpretação correta das mensagens). Existem duas bibliotecas que são as mais populares: PVM (que foi criada para redes de workstations) e MPI (que é um padrão suportado por IBM, HP, SUN, etc e tem mais funcionalidade do que o PVM).

  PVM (Parallel Virtual Machine) [2] é um framework para o desenvolvimento de aplicações paralelas de maneira fácil e eficiente, administrando com transparência o envio de mensagens entre os computadores do cluster. O PVM é um modelo simples que oferece um conjunto de interfaces de software que podem ser facilmente implementadas para fazer tarefas como executar e finalizar tarefas na rede e sincronizar e comunicar estas tarefas. As tarefas que estão sendo executadas podem executar ou finalizar outras tarefas. Apresenta também características que o fazem tolerante a falhas.

  MPI (Message Passing Interface) [7] é um sistema padronizado e portável de troca de mensagens, que foi projetado para trabalhar com muitos tipos de arquitecturas de computadores e oferencendo interfaces para programação em Fortran, C e C++. Pode-se se dizer que é um padrão melhor do que o PVM e existem algumas razões: Possui muitas implementações livres de boa qualidade, oferece comunicação assíncrona completa, administra eficientemente buffers de mensagens, suporta sincronização com usuários de software proprietário, é altamente portável, é especificado formalmente, é um padrão. Existe uma nova versão do MPI chamada de MPI2, que incrementa suporte para entrada e saída paralelas, operações remotas de memória, administração dinâmica de processos e suporta threads.

- Bibliotecas para Desenvolvimento de Aplicações

  Um dos principais problemas para o desenvolvimento dos clusters é a dificuldade de construção de software paralelo, mas felizmente cada vez é mais comun seu desenvolvimento, mesmo para computadores que não sejam paralelos, pois igual representam uma melhoria no desempenho. Como as tecnologias de hardware estão constantemente mudando, as aplicações paralelas desenvolvidas devem ter a capacidade de executar-se em diferentes arquiteturas e por isso é recomendado desenvolver aplicações trabalhando com a camada baixa do middleware.

  Os modelos e linguagens de programação paralela devem ter algumas características para serem considerados adequados [6]: Facilidade de programação, Ter uma metodologia de desenvolvimento de software, independência de arquitetura, facilidade de entendimento, capacidade de ser eficientemente implementados, oferecer informação precisa sobre o custo dos programas.

  Um problema que está sempre presente nas aplicações paralelas é que algumas vezes são muito abstratas para conseguir maior eficiência, ou ao contrario, não são muito eficientes para ser menos abstratas; por isso podem ser classificadas segundo esses criterios [6]: nada explícito e paralelismo implícito; paralelismo explícito e descomposição implícita de tarefas; descomposição explícita de tarefas e mapeamento implícito de memória; mapeamento explícito de memória e comunicação implícita; comunicação explícita e sincronização implícita; e tudo explícito.

  Dois pacotes para desenvolvimento de software paralelo que são muito conhecidos são BSP e ARCH, os

quais são muito usados atualmente.

2.2.3 *Software de Redes [9]*

Os elementos mais importantes no software de redes são os seguintes:

- TCP-IP: A diferença dos supercomputadores, os clusters não tem protocolos de comunicações proprietários, senão que usam protocolos padrão como o TCP-IP, que é o padrão de fato nos sistemas de cluster. O protocolo IP é dividido conceitualmente em diferentes capas lógicas que tem como objetivo dividir a mensagem que vai ser transmitida em pacotes individuais de dados, chamados de Datagramas, que tem o seu tamanho limitado pela longitude do meio de transmisão. Depois de serem transmitidas individualmente, as mensagens são reconstruidas no computador de destino, o qual é indicado em cada pacote mediante o endereço IP do host de destino. Atualmente existem as versões IPv4 e IPv6 que tem 4 bytes e 16 bytes para representar o endereço do host de destino. Os serviços suportados pelo protocolo IP são TCP (Transmission Control Protocol) e UDP (User Datagram Protocol).

- Sockets: Os sockets são as interfaces de baixo nível entre as aplicações de usuário e o sistema operacional e o hardware no computador. Eles oferecem um mecanismo adequado para a comunicação entre as aplicações, tendo suporte para diferentes formatos de endereços, semântica e protocolos. Foram propostos no sistema operacional BSD 4.2 e agora existe uma API no Linux que oferece suporte aos sockets. Alguns programadores não gostam de trabalhar com sockets diretamente, mas com outras capas que fazem uma abstração dos sockets. A idéia básica nos sockets é a implementação de um arquivo que permita fazer operações read e write de um computador a outro como se fosse um arquivo que está no mesmo computador. Pode-se usar muitos tipos de implementações de sockets, mas na prática os mais usados são Connectionless Datagram Sockets (que são baseados no protocolo UDP) e outro tipo baseado no protocolo TCP, que é melhor do que o baseado em UDP porque oferece uma comunicação de ida e volta entre os elementos ligados; os sockets para o protocolo UDP tem o problema que algumas vezes as mensagens não são enviadas corretamente. Os sockets deixam para o conhecimento do programador como é o funcionamento dos mecanismos de troca de mensagens, pelo qual algumas vezes é comparado com a linguagem Assembler.

- Protocolos de Alto Nível: Os sockets não são muito usados pelos programadores que desenvolvem aplicações para os clusters, mas são usados pelos programadores de sistemas operacionais, sendo que os programadores de aplicações tem preferência por usar protocolos do mais alto nível, que até algumas vezes não precisam usar sockets. Os protocolos de alto nível mais usados são Remote Procedure Calls (RPC) e Distributed Objects (CORBA e Java RMI). O RPC evita que o programador tenha que indicar explicitamente a lógica para a troca de mensagens, sendo que o seu objetivo é que os programas distribuídos possam se programar como se fossem programas sequenciais, ou seja, que um processo é chamado em uma função e o compilador tem a função de executá-lo em outro computador. RPC foi criado mais para trabalhar com programação distribuída do que com programação paralela. Os objetos distribuídos são baseados no conceito de programação baseada em objetos, mas usado para chamar métodos remotos, sendo que os serviços de rede representam aos objetos, que tem implementados os métodos a serem executados remotamente; a idéia também é executar os métodos sem considerar em que serviço estão implementados, ou seja, que os serviços estão disponíveis para todos os objetos da rede.

- Sistema de Arquivos distribuído: Todos os computadores no cluster tem o seu próprio sistema de arquivos local, mas também os outros computadores podem precisar acessar arquivos em outras máquinas, sendo que estas operações tem que ser feitas sem fazer diferença com acessos a arquivos locais. Existem dois sistemas de arquivos que são os mais usados: NFS e AFS.

  O NFS (Network File System) foi proposto pela Sun para ser um padrão aberto e foi adotado em sistemas UNIX e Linux. Está estruturado como uma arquitetura cliente-servidor e usa chamadas RPC para fazer a comunicação entre clientes e servidores. O servidor envia os arquivos solicitados pelo cliente para que sejam lidos por ele como se estivessem armazenados localmente nele. Quando são feitas as solicitações por arquivos pelo cliente, não é armazenada nenhuma informação no servidor e, desta maneira, cada uma é independente das outras, mesmo que sejam do mesmo cliente.

  O AFS (Andrew File System) foi proposto pela IBM e a Universidade Carnegie Mellon, para solucionar os problemas de outros sistemas de arquivos como o NFS. Ele consegue reduzir o uso do CPU e o tráfego de rede mantendo um acesso eficiente aos arquivos para grandes quantidades de clientes. Depois o AFS tornou-se um sistema proprietário e por muito tempo não foi incluído no Linux, ficando recentemente disponível para ser usado no Linux.

3. VANTAGENS DO CLUSTER

Implementar um cluster tem muitas vantagens em muitos aspectos como baixo custo, bom desempenho, etc. As principais vantagens são [9]:

- Oferece muita facilidade para alcanzar escalabilidade, ou seja, para poder incrementar o número de computadores que estão ligados ao cluster.

- A arquitectura baseada em cluster tornou-se na configuração de fato quando é necessário fazer tarefas paralelas muito grandes.

- Oferece uma relação custo-benefício muito boa, pois é uma arquitetura muito barata para ser implementada e tem um desempenho muito bom.

- Tem flexibilidade para configuração e atualização, pois é implementado com sistemas operacionais comerciais que tem muito suporte.

- Permite trabalhar com as últimas tecnologias.
- Alta disponibilidade, oferencendo tolerancia a falhas, porque pela sua configuração o sistema ainda funciona quando um dos computadores tem problemas.
- Pode ser montado por uma pessoa com bons conhecimentos, mas não depende de um fabricante específico porque os computadores podem estar configurados com peças de diferentes fabricantes.
- Baixo custo e curto tempo de produção, porque os computadores que são usados para montar clusters não são fabricados exclusivamente para esse fim, mas para uso cotidiano, motivo pelo qual são mais populares.

4. ALTERNATIVAS DE CONSTRUÇÃO DE UM CLUSTER

Existem várias alternativas para a construção do cluster, sendo que pode ser oferecido como uma solução integral, ou pode ser construído comprando independentemente as partes. As diferentes opções são [5]:

4.1 Soluções Integradas Proprietárias

Como foi explicado, os cluster foram propostos como uma alternativa para ter computadores com um alto desempenho mas com um custo muito baixo em comparação aos supercomputadores. Hoje, os computadores mais poderosos são clusters, o que quer dizer que os fabricantes mais conhecidos apresentam soluções baseadas em clusters, embora seja na área dos supercomputadores (Cray, NEC), como também na área dos grandes servidores corporativos (IBM, HP-Compaq, Sun). Estas soluções são construídas com arquiteturas próprias, as quais podem ser melhores do que as arquiteturas genéricas, mas também tem o inconveniente de criar dependência a estas arquiteturas, como redes com altas taxas de transferência, memórias com mais capacidade, etc. Estas arquiteturas, embora sejam melhores do que as outras, também são mais caras e é por isso que só algumas organizações tem condições de comprar este tipo de equipamento.

4.2 Soluções Integradas Abertas

Estas soluções são parecidas às Soluções Proprietárias porque também são oferecidas como um equipamento já pronto para ser usado, mas a diferença entre elas é que não usam exclusivamente peças fabricadas por um fabricante só, ou arquiteturas proprietárias que difícilmente podem ser substituídas por outras. Estas soluções usam peças encontradas comúmente no mercado tecnológico e também é por isso que o seu preço é menor do que as Soluções Proprietárias. São oferecidas para dar solução a problemas genéricos que necessitam um desempenho alto, mas não são problemas específicos como os solucionados pelas Soluções Proprietárias.

4.3 Soluções Não Integradas

Este tipo de soluções são as mais adequadas quando o objetivo principal é gastar a menor quantidade possível de dinheiro, porque elas são montadas pelo próprio usuário segundo as suas necessidades e as suas possibilidades econômicas, sendo que são montadas com peças genéricas que não tem uma arquitetura específica e que são mais baratas. Um problema é que necessitam de pessoal especializado para fazer a escolha, instalação e manutenção do cluster, considerando todos os componentes de hardware e software. Este tipo de cluster também apresenta o problema que o software nem sempre consegue se integrar com o hardware escolhido, porque nem o software nem o hardware foram projetados para trabalhar juntos.

5. EXEMPLO: O CLUSTER DO GOOGLE

Nesta seção vai ser apresentado um exemplo de um cluster real, que tem muito sucesso atualmente e é muito conhecido, porque suporta uma ferramenta usada por muitas pessoas no mundo inteiro: o Google [3].

O volume de informação disponível na Internet vem apresentando uma taxa de crescimento elevada, e por isso o Google foi proposto como um motor de pesquisa que seja capaz de suportar esse crescimento, oferecendo sempre um desempenho muito bom quanto ao tempo de pesquisa, sem importar se o volume de dados aumenta de maneira exponencial como ocorre atualmente. Também foi desenvolvido para estar sempre disponível, sendo usado por milhões de pessoas no mundo inteiro, a qualquer hora do día. Quanto ao tempo, os engenheiros que o projetaram tinham a meta de todas as consultas serem atendidas em tempo menor do que 0.5 segundos; objetivo que atualmente é conseguido pelo sistema.

Quanto à arquitetura do cluster do Google, temos que no ano 2001 tinha aproximadamente 6000 processadores e 12000 discos rígidos, conseguindo assim uma capacidade de armazenamento total de 1 petabyte, tornando-se assim o sistema com maior capacidade de armazenamento no mundo. Em vez de oferecer a disponibilidade do sistema através de configurações RAID nos discos, Google trabalha com armazenamento redundante, tendo milhares de discos e processadores distribuídos no Vale de Silício e em Virginia, nos EUA e ao mesmo tempo tem também um repositório de páginas em cache, que tem alguns terabytes de tamanho. Usando essa configuração, está quase garantido que o sistema vai conseguir atender sempre as consultas, porque no caso de cair tem armazenamento redundante e além disso tem páginas já armazenadas no cache.

São usados dois switches redundantes para conectar os Racks de PCs, sendo que em cada switch podem ser conectados 128 linhas com velocidade de 1 Gbit/s e depois os racks são conectados também com os switches, o qual no total oferece uma capacidade de 64 racks de PCs. Cada rack pode ter até 40 computadores, que usam interfaces de 100 Mbits/s e algumas de 1 Gbit/s.

Cada um dos computadores que formam os racks são computadores com configurações simples, que são usados cotidianamente por todas as pessoas, com capacidades padrão no processador, disco rígido e memória RAM e que atualmente são muito baratas. Por exemplo: Intel Pentium IV 3.0 Ghz ou Intel Core 2 Duo 2.0 Ghz, com disco rígido de 160 Gb e 2.0 Gb de memória RAM. O custo associado atualmente a um desses computadores não é maior do que $600.

Como já foi explicado anteriormente, um problema muito comum nos clusters montados pelo usuário é que apresenta

falhas devido a que o software não foi projetado para o hardware escolhido. No caso do Google, também se apresenta esse problema, sendo que muitas vezes os computadores tem que ser reinicializados manualmente, ou também acontece que os discos rígidos e as memórias tem que ser substituídos frequentemente.

Uma característica que é impressionante do cluster do Google é que o objetivo que foi estabelecido no começo, de atender as consultas em menos tempo do que 0.5 segundos, até hoje ainda é atendido, mesmo que a informação na Internet tenha um crescimento exponencial e que o cluster cada vez tenha mais computadores, o que significa que a informação deve ser procurada em muitos meios de armazenamento e viajar por muitas vias de comunicação, distribuídas em lugares muito distantes.

6. CONCLUSÕES

- Os clusters são uma excelente opção para conseguir desempenho muito alto, sem ter que gastar muito dinheiro, porque eles são construídos com componentes computacionais de uso cotidiano, que são baratos e podem ser configurados sem muito problema.
- Os clusters são usados atualmente em muitas aplicações no mundo inteiro, tanto no ambiente empresarial, como no ambiente científico e tecnológico, sendo que são usados para resolver problemas de todas as áreas da ciência.
- Graças ao baixo custo e facilidade de implementação, um cluster pode ser usado tanto por empresas muito grandes, como por empresas medianas ou até por estudantes ou pessoas em universidades ou particularmente.
- O desempenho do cluster depende de todos os seus componenetes, sendo que tem igual importância ter um adequado sistema operacional, configurado corretamente, como também ter o software de rede e o software middleware.
- Quanto ao hardware, é muito melhor poder usar processadores que suportem multi-processamento, que tenham discos rígidos com boa capacidade e uma boa quantidade de memória RAM, assim como também é vital contar com um adequado hardware de rede, porque é o responsável pela transmissão da informação entre os computadores do cluster.
- Quanto às alternativas para a implementação do cluster, algumas vezes é bom comprar uma Solução Integrada, que já está pronta para funcionar, porque ela não vai requerer a escolha das peças que vão integrar o cluster, mas tem a desvantagem de gerar dependência com o fabricante das peças e também porque é mais caro do que as soluções Não Integradas.

7. REFERÊNCIAS

[1] R. Buyya. *High Performance Cluster Computing.* Prentice-Hall International, Inc, 1999.

[2] A. Geist, A. Beguelin, J. Dongarra, W. Jiang, R. Manchek, and V. Sunderam. *PVM: Parallel Virtual Machine, A Users' Guide and Tutorial for Networked Parallel Computing.* MIT Press, 1994.

[3] J. Hennessy and D. Patterson. *Computer Architecture: A Quantitative Approach, Third Edition.* Elsevier, 2002.

[4] R. Morrison. *Cluster Computing: Architectures, Operating Systems, Parallel Processing and Programming Languages.* GNU General Public Licence, 2003.

[5] M. Pasin and D. Kreutz. Arquitetura e administração de aglomerados. *3ra Escola Regional de Alto Desempenho*, pages 3–34, 2003.

[6] D. B. Skillikorn and D. Talia. *Models and Languages for Parallel Computation.* 1996.

[7] M. Snir, S. Otto, S. Huss-Lederman, D. Walker, and J. Dongarra. *MPI: The Complete Reference.* MIT Press, 1995.

[8] W. Stallings. *Operating Systems: Internals and Design Principles.* Prentice-Hall International, Inc, 2001.

[9] T. Sterling. *Beowulf Cluster Computing with Linux.* MIT Press, 2001.