Agrupamentos computacionais

Raphael Marcos Menderico
RA 009702
rmm@ic.unicamp.br

ABSTRACT

Agrupamentos computacionais (*clusters*) são grandes máquinas paralelas montadas a partir de computadores comerciais interligados por uma rede. Utilizados como alternativa ao elevado custo das máquinas paralelas SMPs, outros fatores como alto custo de gerenciamento e programas não desenhados para execução paralela podem fazer com que não se obtenha o ganho de desempenho esperado. O objetivo desse trabalho é apresentar diversos aspectos envolvidos na montagem, gerenciamento e uso de um aglomerado computacional, passando por algumas análises de custo e de desempenho dos mesmos como forma de mostrar as vantagens e desvantagens desse tipo de arquitetura.

1. INTRODUÇÃO

Muitas aplicações necessitam de uma capacidade de processamento superior àquela oferecida por um único processador, seja porque é preciso executar uma tarefa muito complicada, seja para atender um determinado pedido no menor tempo possível. A solução encontrada para resolver esse problema passa pela replicação do recurso computacional de que precisamos (processador, disco, memória, etc..) e a divisão das tarefas em pequenos pedaços que podem ser processados por esse recurso agora replicado.

Um desafio ao replicarmos determinados componentes é o custo em que isso pode implicar, que nem sempre traduz-se no ganho de desempenho esperado. Por exemplo, colocar dois processadores em uma mesma máquina sem nenhuma alteração no *software* ou no *hardware* praticamente não apresentará ganhos de performance para a execução de um único programa, por outro lado, o computador terá um ganho ao executar dois programas, se eles não competirem pelo acesso aos demais dispositivos.

Por conta disso, máquinas com multiprocessamento simétrico (SMP) possuem um custo elevado para permitir que os processadores possam aproveitar ao máximo a sua performance, reduzindo conflitos que tais processadores possam ter ao acessar os demais elementos da máquina, como memória, discos e rede. Uma máquina com muitos processadores pode ter um custo proibitivo, e algumas vezes, devido à natureza de cada aplicação, bastaria replicar um dos recursos para que a máquina já apresentasse um ganho considerável de performance.

O objetivo desse trabalho é apresentar uma arquitetura que tem sido usada como alternativa à máquinas multiprocessadas: os agrupamentos computacionais (*clusters*). Esses agrupamentos são formados por computadores comerciais (ou seja, que originalmente não foram adaptados para trabalharem em conjunto) interligados por um rede e que realizam sua tarefa como se fossem um único e grande computador paralelo. Analisaremos exatamente qual é a definição de um agrupamento, aspectos envolvidos na sua construção e gerenciamento, tipos de aplicações de tais máquinas paralelas, estudos quanto ao desempenho obtido com tais máquinas e casos de uso desses conjuntos de computadores.

O restante do documento apresenta a seguinte estrutura: A Seção 2 apresenta diversas definições de agrupamentos encontradas na literatura e as características que serão exploradas no restante do documento; a Seção 3 mostra uma estrutura básica de *software* e *hardware* para montagem de um agrupamento; a Seção 4 detalha os três tipos mais comuns de agrupamentos, aplicações típicas e desafios na hora de implementar cada um dos tipos; a Seção 5 mostra resultados obtidos comparando diferentes configurações de agrupamentos computacionais; a Seção 6 detalha a arquitetura de alguns agrupamentos existentes; por fim, a Seção 7 encerra esse documento resumindo os resultados apresentados nas seções anteriores.

2. DEFINIÇÃO

Para iniciar nossa discussão sobre o que é um agrupamento, vamos apresentar diversas visões de como máquinas podem ser agrupadas para realizar uma mesma tarefa. De acordo com Baker e Buyya [3], um agrupamento é um tipo especial de processamento paralelo ou distribuído, que é formado por uma coleção de computadores conectados que funcionam como um único e integrado recurso computacional. Eles podem ser classificados quanto a:

Aplicações alvo: Um agrupamento pode ser direcionado para aplicações de alto desempenho ou de alta disponibilidade

Propriedade dos nós: Os nós de um agrupamento podem ser dedicados (usados exclusivamente para executar tarefas enviadas ao agrupamento) ou compartilhados (pode ser usado para outras tarefas além daquelas enviadas para o agrupamento).

Tipo de *hardware* dos nós: Um agrupamento pode ser formado por PCs, Estações de trabalho (*workstations*) ou máquinas com multiprocessamento simétrico (SMPs)

Sistema operacional: Um agrupamento pode utilizar diferentes sistemas operacionais, como Linux, Windows, Solaris, etc ...

Configuração dos nós: Os nós podem possuir todos a mesma configuração (homogêneo) ou cada nó pode ter uma configuração própria (heterogêneo)

Nível de agrupamento: Um agrupamento pode ter desde poucos computadores ligados por redes de alta velocidade até milhões de computadores espalhados pelo mundo.

Já de acordo com Gray [7], a diferença entre um conjunto de computadores trabalhando juntos (a descrição de Baker e Buyya) e um agrupamento de computadores é que agrupamentos possuem *hardware* e *software* homogêneo, uma interface de administração única e uma rede de comunicação próxima da ideal, com latência baixa.

De fato, diversos autores [12, 5, 8, 4] possuem uma definição de agrupamento muito próxima a de Gray. Ao apresentarem seus supercomputadores, eles sempre apresentam um conjunto homogêneo de computadores comerciais (não desenhados especificamente para esse fim) dedicados, ligado através de uma rede de alta velocidade e fisicamente próximos.

As diferenças entre os agrupamentos apresentados ficam por conta do sistema operacional, da aplicação alvo e do tipo do tipo de *hardware* dos nós.

Quando os nós estão à uma distância maior que uma sala (desde um prédio até um conjunto de universidades) e possuem computadores não dedicados e heterogêneos, mas ainda estão ligados por uma rede confiável e de alta velocidade, eles são chamados de grades. Quando essa rede não está disponível, chama-se esse conjunto de computadores desempenhando uma mesma tarefa de "computação pública" [1]. Nesse trabalho, não abordaremos grades nem computação pública, embora eles se encaixem na definição de Baker e Buyya, ficando restritos à definição de agrupamento de Gray e dos demais autores.

3. ESTRUTURA

Do ponto de vista do *hardware*, um agrupamento é formado por diversos computadores comerciais e adequados para a tarefa a ser realizada (computadores rápidos quando for necessário alta performance, computadores com tempo de resposta adequado a aplicação, mas não necessariamente alta performance de pico, quando desejarmos alta disponibilidade), uma rede de alta velocidade (*Gigabit ethernet* ou outros protocolos especializados em alta velocidade), comutadores de rede e demais equipamentos para integrar os computadores. Alguns outros equipamentos, como teclado, *mouse* e monitor de vídeo podem ser necessários para realizar tarefas administrativas.

O principal desafio ao montar-se um agrupamento está no *software* que será empregado [7]. Programas devem ser instalados em cada nó para que os usuários possam aproveitar o poder computacional do supercomputador sem a necessidade de determinar em qual máquina os processos executarão, e quais as características de cada uma das máquinas. Tal característica é chamada de transparência [6].

Especificamente para agrupamentos, além da transparência deve-se fornecer também uma visão unificada do sistema (*single system image, ou SSI*) [3], ou seja, ele deve transformar o conjunto de computadores em um único computador, cujos recursos computacionais sejam a somatória dos recursos de cada computador. Essa imagem simplificada do sistema:

- Facilita a administração das máquinas, uma vez que todos os recursos são administrados como um único.
- Reduz o risco de um operador cometer um erro, uma vez que boa parte das decisões que poderiam provocar erros cabe à SSI
- Facilita a localização de recursos computacionais ligados ao agrupamento, uma vez que sua localização física não importa para o usuário desde que ele esteja conectado a um dos nós.

Essa imagem pode ser implementada em *hardware*, através de mecanismos que permitam enxergar toda a memória existente nos computadores como uma única memória compartilhada com diferentes velocidades de acesso (NUMA) [8] ou pelo sistema operacional através de sistemas de escalonamento, balanceamento de carga e tolerância a falhas.

Uma terceira forma de fornecer uma imagem unificada do agrupamento é fornecer um *middleware*, ou seja, um conjunto de funções utilizado pelas aplicações para emular recursos que originalmente deveriam ser disponibilizados pelo *hardware* ou pelo sistema operacional, mas que por algum motivo não estão disponíveis, ou simplesmente para criar uma única chamada para um serviço que possa ser atendido de diversas formas, dependendo da existência ou não de determinadas propriedades em cada nó. Um exemplo de *middleware* é o MPI [10], que fornece um serviço de envio e recepção de mensagens entre máquinas heterogêneas, utilizando memória compartilhada ou envio de mensagens pela rede, dependendo de qual dos serviços está disponível entre os nós envolvidos na comunicação.

4. TIPOS DE AGRUPAMENTOS

Considerando somente os agrupamentos que atendem aos requisitos apresentados na seção 2, temos três tipos principais [2]:

Agrupamentos de Balanceamento de Carga: O objetivo de um agrupamento de balanceamento de carga é

simplesmente dividir tarefas entre diversos computadores, mas nesse caso as tarefas devem ser executadas integralmente por cada um dos nós, sem necessidade de consultar outros (e portanto, sem sincronia entre os estados de nós distintos). O excessivo número de pedidos torna proibitiva a execução de todas as tarefas pelo mesmo nó sem um atraso considerável, por isso a necessidade de dividir as tarefas.

É o tipo de agrupamento mais simples de ser montado, uma vez que só necessita de pequenos programas para fazer esse balanceamento de carga, função essa que muitos sistemas operacionais já realizam [15]. É muito utilizado em servidores de internet, para atender *sites* com um tráfego elevado de requisições.

Um exemplo de agrupamento para balanceamento de carga é o Google [4], pois apesar dos servidores de pesquisa possuírem um estado (a situação atual de cada página), ele pode ser alterado lentamente e de forma assíncrona entre os servidores sem comprometer a qualidade da resposta ao usuário final. Discutiremos mais sobre o Google na seção 6.1, bem como sobre os desafios enfrentados no projeto de um agrupamento real de balanceamento de carga.

Agrupamentos de Alta Disponibilidade: Além de prover o balanceamento de carga, esse tipo de agrupamento também disponibiliza o acesso aos dados de forma a garantir a maior tempo de acesso possível para o mesmo. Para tal, é necessário criar um sistema de replicação de dados, seja através de discos replicados através da rede [6], seja utilizando um servidor de armazenamento [8] ligado a uma rede de alta velocidade.

Esse tipo de agrupamento é muito utilizado por servidores de banco de dados, que devem atender a uma transação no menor tempo possível e também tornar a informação disponível ao usuário final a todo instante.

Agrupamentos de Alto Desempenho: Esse tipo de agrupamento tenta concentrar o maior poder computacional possível, aliado com algoritmos paralelos eficientes, para realizar uma tarefa o mais rápido possível. Um exemplo típico são os agrupamentos computacionais científicos, onde uma tarefa complexa é dividida através dos nós e a comunicação é feita através de troca de mensagens, utilizando um *middleware*.

É o tipo de agrupamento mais complexo de ser criado, pois sua implementação envolve tanto desafios de *hardware* quanto de *software*, para nem sempre obter o desempenho e a escalabilidade desejada. Na seção 5, apresentaremos uma discussão mais detalhada dos desafios enfrentados ao tentar-se criar um agrupamento de alto desempenho.

5. DESEMPENHO DE UM AGRUPAMENTO

Tipicamente o desempenho de um agrupamento é medido de duas formas: calculando-se o número de instruções executadas por segundo (FLOPS), como faríamos com qualquer outro computador, e custo de cada instrução (quantidade de dinheiro gasto para obter cada FLOPS). As principais vantagens dos aglomerados em relação a outras máquinas com o mesmo desempenho estão no custo por FLOPS, o qual é muito mais baixo que o obtido por um supercomputador montado com *hardware* proprietário.

Essa seção apresenta três análises sobre o desempenho de um agrupamento, realizada por grupos diferentes. Dois estudos, um no National Center for High-Performance Computing (NCHC), em Taiwan [5], e um no Departamento de Engenharia Elétrica e de Computação da Wayne State University [12], comparam agrupamentos com máquinas paralelas de desempenho equivalente, somente considerando o custo do *hardware* de cada máquina. O terceiro estudo, realizado por Patterson e Henessy [8], mostra as diferenças de custo entre diferentes configurações de agrupamentos, considerando todos os custos envolvidos, inclusive o custo de propriedade.

5.1 Estudo realizado no NCHC

No primeiro estudo, realizado no National Center for High-Performance Computing (NCHC), em Taiwan [5], compara-se o desempenho de um agrupamento de computadores contra computadores paralelos. O agrupamento era formado por computadores comerciais com processador Pentium II 400Mhz, enquanto o computador paralelo utilizava processadores IBM SP2 160 Mhz, mas com uma unidade de ponto flutuante mais rápida que a existente no Pentium. Além disso, a máquina paralela possui uma memória principal mais eficiente, pois todos os processadores acessam toda a memória utilizando um barramento de dados.

O primeiro conjunto de testes executado foi um conjunto proprietário desenvolvido pelo próprio NCHC, de diferentes domínios (física, química e mecânica dos sólidos e dos fluídos). O resultado obtido indica que o desempenho do agrupamento é dependente das características de cada aplicação.

Somente duas aplicações obtiveram uma melhora de desempenho nas máquinas Pentium: uma que utiliza prioritariamente operações com números inteiros, e outra cuja entrada cabe praticamente inteira na memória cache dos processadores Pentium. Para todas as outras, o desempenho obtido foi pior no agrupamento que o obtido nas demais máquinas paralelas.

Utilizou-se também dois *benchmarks* comerciais para comparar o mesmo conjunto de computadores, o Linpack e o NAS. O Linpack é um conjunto de testes desenvolvido para avaliar a performance de supercomputadores, enquanto que o NAS utiliza equações comumente utilizadas em dinâmica dos fluídos para criar um conjunto de testes facilmente paralelizável. Novamente, a máquina paralela baseada em processadores IBM foi mais rápida.

A grande vantagem obtida pelo agrupamento é quando compara-se o custo de cada milhão de instruções processada por segundo (MFLOPS). Com a mesma quantidade de dinheiro, obtém-se uma quantidade de MPLOPS muito superior utilizando-se o agrupamento computacional proposto.

Além da comparação entre máquinas paralelas e agrupamentos, o artigo também apresenta uma comparação entre os dois diferentes agrupamentos existentes no centro: um com máquinas com um único processador e outro com máquinas com dois processadores, ambos os agrupamentos totalizando 16 processadores cada. Os computadores com SMP apresentaram um custo menor por MFLOPS, mas por outro lado seu desempenho foi inferior aos obtidos pelos computadores

com processadores isolados, devido ao compartilhamento de recursos dentro de uma máquina SMP.

5.2 Estudo realizado na Wayne State University

Novamente, compara-se um agrupamento com uma máquina paralela de poder de processamento equivalente. Os testes foram realizados na Wayne State University [12] e o agrupamento consistia em um conjunto de computadores utilizando processadores Intel Pentium II 350 Mhz. A máquina paralela era formada por até 10 processadores UltraSparcII de 250 Mhz. Foram utilizados dois diferentes *benchmarks*: o NAS, que utiliza o MPI para envio e recebimento de mensagens, e o Splash-2, com programas que utilizam memória compartilhada.

A primeira comparação realizada foi quanto a largura de banda necessária. Testou-se duas redes diferentes de interconexão: uma utilizando-se canais de 100 Mbps (Fast Ethernet) e outra, com canais de 1Gbps (Gigabit ethernet). A maioria dos programas dos dois *benchmarks* obteve pouco ganho utilizando a rede mais rápida em relação ao tempo de execução obtido com a rede mais lenta. Segundo os autores, isso deve-se ao fato da rede de 100 Mbps ter a largura de banda necessária para realizar a troca de informações entre os programas.

Em seguida, comparou-se a escalabilidade de ambas as configurações (agrupamento e máquina paralela). A máquina paralela apresentou ganhos de performance conforme aumentava-se o número de processadores para os dois conjuntos de testes. Já o agrupamento não escalou corretamente para o *benchmark* Splash-2. Esses programas utilizam memória compartilhada, que é muito mais rápida na máquina paralela que no aglomerado, pois no último é necessário utilizar um mecanismo de distribuição da memória compartilhada através da rede, provocando atrasos.

Por fim, comparou-se o desempenho de agrupamentos formado por máquinas com um único processador com o desempenho daqueles constituido por nós com dois processadores, sempre com o mesmo número total de processadores. Para o conjunto de testes com memória compartilhada (Splash-2), houve um pequeno ganho ao utilizarmos quatro computadores com dois processadores cada ao invés de oito computadores com um único processador. Esses computadores são, inclusive, mais baratos, pois o custo de um único processador e da placa mãe é menor que o custo de um computador inteiro (que inclui uma placa de rede, cabeamento, espaço físico, etc ...)

Por outro lado, com o *benchmark* NAS, o desempenho do agrupamento com oito computadores foi melhor. Segundo os autores, isso deve-se principalmente a problemas com a cache.

A conclusão final dos autores é que a escolha de qual máquina deve-se utilizar para realizar uma computação de alta performance é totalmente dependente das características da aplicação a ser executada, e que nem mesmo podemos afirmar que os agrupamentos SMPs apresentarão melhor performance em função do custo. A tabela 1 apresenta a máquina onde cada um dos programas dos *benchmarks* obteve o seu melhor desempenho e custo dessa máquina.

Benchmark NAS (Baseado em troca de mensagens)

Programa	Melhor performance	Custo (US$)
BT	Agrupamento com 16 CPUs e Gigabit Ethernet	31000
CG	SMP com 8 CPUs e compilador especial	60480
EP	Agrupamento com 16 CPUs e Gigabit Ethernet	31000
LU	Agrupamento com 16 CPUs e Gigabit Ethernet	31000
SP	SMP com 4 CPUs	35280
MG	Agrupamento com 16 CPUs e Gigabit Ethernet	31000

Benchmark Splash-2 (Baseado em memória compartilhada)

Programa	Melhor performance	Custo (US$)
LU-c	Agrupamento com 16 CPUs e Fast Ethernet	17000
Water-n	SMP com 8 CPUs	60480
Water-sp	SMP com 8 CPUs	60480
Volrend	SMP com 8 CPUs	60480
Raytrace	SMP com 8 CPUs	60480

Table 1: Computadores com melhor desempenho para os programas do *benchmark* NAS e Splash-2

5.3 Estudo realizado por Patterson e Henessy

Patterson e Hennesy, em seu livro *Computer Architecture: A Quantitative Approach* [8], realizam uma análise sobre o custo de diferentes configurações de agrupamentos, todas com computadores IBM xSeries. Ele não leva em conta o tipo de aplicação a ser executado em cada um dos agrupamentos e as vantagens e desvantagens de cada tipo de máquina.

Os computadores da série xSeries utilizam processadores Intel Pentium III. o xSeries 300 possui somente um processador de 1000 Mhz e será a máquina monoprocessada. o xSeries 330 pode receber até dois processadores, também de 1000 Mhz e o xSeries 370 tem espaço para até 8 processadores, mas agora de 700 Mhz.

A primeira análise utiliza somente computadores IBM, sem qualquer dispositivo de armazenamento de dados centralizado (embora cada máquina possua pelo menos um disco). O objetivo é totalizar 32 processadores, 32 GB de memória RAM e mais de 2T de capacidade de armazenamento. O agrupamento montado com o xSeries 330, de duas vias, foi o que apresentou menor custo total de montagem, 161 mil dólares, aproximadamente. Em segundo, temos o agrupamento com o xSeries 300, de uma via, com custo total de 180 mil dólares. E por fim, ficou o xSeries 370, com custo total de 253 mil dólares.

Um dos motivos para que o 330 fosse mais barato que o 370 foi a redução no custo da rede (uma vez que são menos computadores, e portanto, são necessários menos pontos de rede) e a menor necessidade de sistemas de suporte (fontes de alimentação, controladores de disco, cabos, etc ..). Já o agrupamento com somente quatro computadores com oito

processadores cada teria um custo muito mais elevado pois, embora exista uma economia considerável no custo da montagem da rede, existe um custo considerável na alocação de processadores, memórias e discos adicionais (alguns justificáveis, como o gasto com disco, mas outros nem tanto, como um custo maior para colocar memórias equivalentes a utilizadas nas outras máquinas).

A segunda análise apresentada leva em conta esse mesmo conjunto de agrupamentos, mas agora com uma controladora de disco externa ligada através de uma rede de alta velocidade específica para dispositivos de armazenamento (Esse tipo de tecnologia é chamada de SAN - *Storage Area Network* ou rede de armazenamento). Os autores apresentam essa solução como uma forma de evitar que falhas em um dos discos comprometa a capacidade de armazenamento e simplifique a administração do sistema.

O custo total de montagem para o agrupamento com o xSeries 300 (32 nós com um processador) foi de 281 mil dólares, 230 mil dólares para o agrupamento com o xSeries 330 e 289 mil dólares para o xSeries 370. Novamente, a opção com dois processadores foi mais econômica, pois os custos para a instalação da rede SAN é proporcional ao número de nós e a diferença de custo entre a solução com 16 nós e a solução com 4 nós era grande o suficiente para impedir que a economia gerada na instalação da SAN tornasse a solução com 4 nós a mais econômica. O maior impacto no custo foi sentido pelo agrupamento com um único processador, que agora possui um custo equivalente ao custo do agrupamento com 4 nós.

O terceiro teste reúne todos esses equipamentos e calcula o custo total de propriedade desses agrupamentos ao longo de 3 anos. Ele parte das seguintes premissas:

- São necessários 6 conjuntos de fitas para os *backups* mensais (de 2TB cada conjunto), 4 conjuntos de fitas para os *backups* semanais (novamente de 2TB) e 14 fitas para os *backups* incrementais diários. Considera-se que no máximo 8% da informação existente nos discos será alterada a cada dia
- Levam-se em conta os custos envolvidos na criação e uso do espaço fisico necessário para cada configuração, bem como os custos de conexão das máquinas com a Internet.
- Calcula-se o custo necessário para manter os equipamentos em funcionamento.
- Considera-se que todos os agrupamentos executarão transações de bancos de dados. O sistema operacional utilizado é o Windows 2000 e o banco de dados, o SQL Server. Não é feita uma comparação entre soluções baseadas em *software* proprietário e outra baseada em *software* livre. Como a análise é de custo total de propriedade, essa é uma variável que deveria ser levada em conta.
- Pressupõe-se que a solução utilizando a SAN acarretará em uma redução pela metade do custo de administração do sistema(de 100 mil dólares ao ano para 50 mil dólares ao ano), sem considerar outras opções como sistemas de arquivos distribuídos [6] ou sistemas que simplesmente são tolerantes a falhas e cuja administração é simples o suficiente para que não haja necessidade de unificarmos todo o espaço de armazenamento [4].

Levando esses fatores em conta, a solução mais econômica foi o agrupamento montado com computadores xSeries 330 (dois processadores) com a SAN. A principal vantagem apresentada foi o custo do operador, que foi metade do o custo de administração dos agrupamentos sem a SAN e foi o maior custo envolvido na operação dos agrupamentos, em alguns casos superando inclusive o custo de construção dos agrupamentos. O segundo custo mais relevante foi o aluguel do espaço, responsável por uma despesa de 36 mil dólares durante três anos para a máquina mais econômica e 72 mil dólares para os agrupamentos com 1 e 8 processadores por nó.

O custo total de propriedade é uma parte significativa de qualquer sistema computacional. Nos nossos exemplos, ele acaba sendo equivalente ao custo de montagem do computador. Para a máquina mais econômica, o custo total de montagem foi de 230 mil dólares e o custo de propriedade de 267 mil dólares. A tabela 2 apresenta um resumos dos custos de cada agrupamento dessa seção.

	Custo (milhares de dólares)		
Agrupamento	Montagem	Propriedade	Total
32 xSeries300	180	429	609
16 xSeries330	161	415	576
8 xSeries370	253	454	707
32 xSeries300 + SAN	281	317	598
16 xSeries330 + SAN	230	267	497
8 xSeries370 + SAN	289	305	594

Table 2: Custos dos agrupamentos propostos por Patterson e Henessy

6. ESTUDOS DE CASO

Nessa seção estudaremos principalmente a arquitetura do agrupamento de computadores do Google, onde mostraremos algumas das decisões de projeto dos engenheiros que montaram e realizam a manutenção [8, 4]. Esse agrupamento difere dos demais apresentados por tratar-se de um sistema de balanceamento de carga simples, que é a aplicação mais facilmente paralelizável. Por esse motivo o Google consegue tirar tanto proveito de computadores comerciais. Veremos mais detalhes sobre ele na seção 6.1.

Existem outros agrupamentos de alta performance voltados para a pesquisa científica. Um deles, o Virginia Tech TeraScale Computing Facility [14], é formado por 1100 computadores Apple XServe G5, cada um com 2 processadores Power PC de 2.3 GHz, 4 GB de memória e 80 GB de disco, 3 nós especiais para compilação dos programas e uma unidade de armazenamento Apple XServe RAID com 2.7 TB de armazenamento. Ele também possui duas redes: Uma Gigabit Ethernet para manutenção e outra InfiniBand (de 10GBits por segundo), para ligar os nós à unidade de armazenamento. O Sistema operacional de todas as máquinas é o Mac OS X.

Já o TeraGrid [11] é na verdade uma grade computacional gerenciada por quatro instituições, nos Estados Unidos, formado por quatro agrupamentos. Um dos agrupamentos, no National Center for Supercomputing Applications (NCSA), em Chicago, Estados Unidos, possui 887 nós formados por computadores IBM com dois processadores Intel Itanium de 1.3 a 1.5 GHZ cada. Parte dos nós possui 12 GBytes de memória para realizar o processamento de grandes aplicações paralelas. A conexão entre os nós é formada por três tipos de rede: Mirinet e FiberChannel para os discos e Gigabit Ethernet para manutenção e acesso. Esse agrupamento utiliza dois discos centralizados ligados por redes SAN para armazenar os dados, totalizando 130 TB de dados armazenados. Todos os nós utilizam o sistema operacional Suse Linux.

O supercomputador Lisa [9], mantido pela universidade de Amsterdã e por outros centros de pesquisa holandeses, foi montado com 272 nós cada um com dois processadores Intel Xeon 3.4 GHz e 2GBytes de memória RAM. Um dispositivo de armazenamento centralizado foi utilizado também nesse caso, novamente ligado por uma SAN e com capacidade total de armazenamento de 10 GBytes. O sistema operacional empregado é o Debian Linux.

No ranking das 500 máquinas mais rápidas do mundo publicado em junho de 2005 [13] o computador do VirginiaTech está em 14^o lugar, com a marca de 12250 GFLOPS no teste Linpack (o teste oficial do ranking), de um máximo teórico de 20240 GFLOPS; o TeraGrid está em 38^o lugar, com 7215 GFLOPS no teste Linpack de um máximo teórico de 10208 GFLOPS e por fim o Lisa está em 140^o lugar, com um limite teórico de 3740 GFLOPS e um máximo obtido no *benchmark* de 2371 GFLOPS.

Uma característica interessante dos três computadores apresentados e que não existe no agrupamento do Google é a presença de servidores de armazenamento (SANs) ligadas por redes de altíssima velocidade. Infelizmente nenhum dos casos estudados aqui apresentam o custo total de construção de seus supercomputadores, e somente o Google, como será visto adiante, apresenta um estudo comparativo entre máquinas paralelas SMP e agrupamentos.

6.1 Google Inc.

O Google possui atualmente o maior mecanismo de busca de páginas na Internet. Em 1998 ele possuía aproximadamente 1.3 bilhões de páginas indexadas e 70 milhões de pesquisas por dia, uma média de 1000 consultas por segundo. O desafio dos engenheiros era escolher equipamentos que fossem adequados a tarefa e atendessem ao requisito de atender cada consulta com uma latência máxima de 0,5 segundos [8].

A solução encontrada foi utilizar PCs comerciais, mas adaptados para caber em apenas uma unidade dos *racks* e transferir a tarefa de ser tolerante a falhas para o *software*. Para entendermos como isso foi feito, vamos analisar como cada consulta do Google é processada [8, 4].

A primeira etapa é dividir as consultas entre vários prédios, e dentro de cada prédio, entre vários servidores. A divisão entre prédios espalhados pelo mundo serve para permitir tolerância a falhas causadas por desastres naturais ou quedas bruscas de energia, além de reduzir o número de transações processadas por cada prédio. O balanceamento da carga para permitir que mais transações sejam processados por segundo é feito dentro de cada prédio. Todas essas soluções de distribuição de carga são realizadas por *software*.

Uma vez que foi escolhido qual servidor irá realizar a consulta, a próxima etapa é realizar a pesquisa nos índices e nos servidores de conteúdo para poder montar a página para o usuário final. Se considerarmos o índice inteiro, teríamos muitos *terabytes* de dados para serem gerenciados e manipulados a cada vez, e isso tornaria uma pesquisa extremamente custosa. A solução foi dividir o índice em pequenos pedaços e espalhar esses pedaços entre os computadores. Assim, uma consulta ao índice precisa acessar todas as páginas para retornar uma resposta. Cada uma dessas páginas é replicada em diversas máquinas e um balanceador de carga novamente seleciona qual é o melhor nó para atender a consulta ao índice para cada um dos pedaços. Um processo semelhante é realizado para coletar as informações sobre cada página após a consulta ao índice.

O balanceamento de carga e a replicação permitem que muitas transações sejam atendidas por segundo, embora cada computador individualmente atenda um número baixo de transações, além de prover tolerância a falhas, pois caso um nó pare de responder aos pedidos de consulta, ele é transferido para outro nó, que possui uma cópia dos dados.

Um problema enfrentado por muitas aplicações onde existem consultas e atualizações é a replicação dos dados (de fato, essa é a diferença entre agrupamentos de balanceamento de carga e de alta disponibilidade). Entretanto, no caso do Google as atualizações são muito menos freqüentes que as consultas e as atualizações podem ser feitas como um conjunto de comandos enviados para cada uma das réplicas (relaxando as garantias de consistência existentes em um banco de dados comum). Por essa característica é que o Google pode ser considerado um agrupamento de balanceamento de carga e não um de alta disponibilidade.

Além do balanceamento de carga e das replicações de dados, o Google não utiliza nenhum outro sistema de tolerância a falhas. Quando uma falha ocorre em um nó, simplesmente ele é desativado até que um operador reinicie o computador, ou realize os reparos necessários, inclusive substituição. Uma média de 20 máquinas são reiniciadas por dia por problemas relativos a *software* e cerca de 2 a 3 por cento das máquinas têm que ser substituídas por ano por problemas com *hardware*. Para o Google, é mais barato simplesmente trocar computadores que comprar componentes mais caros, mas tolerantes a falhas.

Uma das características do sistema de busca do Google é ser bastante paralelizável (como todo sistema de consulta a dados). Seus engenheiros apresentam uma comparação entre um *rack* com 176 CPUs, 176 GB de memória RAM e 7 TBytes de disco, que poderia ser comprado por 278 mil dólares. Do outro lado, um servidor com 8 CPUs idênticas as 176 do *rack*, 64 GBytes de memória e 8 TBytes de disco custaria 758 mil dólares, ou seja, três vezes mais por 22 vezes menos processamento, um terço da memória e um pouco mais de espaço em disco. Segundo os engenheiros, a aplicação do Google não conseguiria obter vantagens dos

recursos extras disponíveis no servidor, e portanto o custo extra dessa máquina seria injustificável.

Como vimos nas comparações de desempenho, existem situações em que a máquina paralela é mais rápida que um conjunto de computadores, mas não quando as aplicações são facilmente paralelizaveis e com baixa taxa comunicação, como no caso do Google. Além disso, o número de transações processadas por segundo é muito mais importante que o tempo que cada transação, individualmente, leva para processar, e o objetivo final é minimizar o custo total de propriedade, seja na aquisição dos computadores, seja na sua administração. Por isso, para prover o serviço de busca, a melhor opção acaba sendo o agrupamento de computadores.

7. CONCLUSÃO

Analisando os diversos agrupamentos apresentados ao longo desse trabalho, bem com as análises de custo e desempenho, observamos que a performance obtida a partir de um agrupamento é completamente dependende da aplicação à qual ele se destina, e as decisões a serem tomadas durante o projeto desse tipo de supercomputador devem considerar as características individuais de cada aplicação. Quando esse fator é considerado, os agrupamentos podem representar uma solução de baixo custo para se obter uma grande capacidade de processamento ou atender a muitas transações concorrentemente.

8. REFERENCES

[1] D. P. Anderson. Public computing: Reconnecting people to science. In *Conference on Shared Knowledge and the Web*, 2003.

[2] M. Baker. Cluster computing white paper. http://citeseer.ist.psu.edu/baker00cluster.html. Último acesso em 01/11/2005.

[3] M. Baker and R. Buyya. Cluster computing: the commodity supercomputer. *Software Practice and Experience*, 29(6):551–576, 1999.

[4] L. A. Barroso, J. Dean, and U. Hölzle. Web search for a planet: The google cluster architecture. *IEEE Micro*, 23(2), 2003.

[5] H.-Y. Chang, C.-Y. Shen, C.-Y. Chou, S.-C. Tcheng, and K.-C. Huang. Benchmark and Performance Evaluation of NCHC PC Cluster. In *Proceeding of APSCC'2000*, 2000.

[6] G. Coulouris, J. Dollimore, and T. Kindberg. *Distributed systems: concepts and design — 3rd edition.* Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2001.

[7] J. Gray. Super servers: Commodity computer clusters pose a software challenge. In *BTW*, pages 30–47, 1995.

[8] J. L. Hennessy and D. A. Patterson. *Computer Architecture: A Quantitative Approach — 3rd edition.* Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2003.

[9] Lisa Cluster. Disponível em http://www.sara.nl/userinfo/lisa/description/index.html. Último acesso em 01/11/2005.

[10] MPI-2: Extensions to the Message Passing Interface. Technical report, Message Passing Interface Forum, 1997.

[11] NCSA IA-64 TeraGrid Cluster. Disponível em http://teragrid.ncsa.uiuc.edu/TGIA64LinuxCluster.html. Último acesso em 01/11/2005.

[12] D. Thaker, V. Chaudhary, G. Edjlali, and S. Roy. Cost-Performance Evaluation of SMP Clusters.

[13] 25th TOP500 list. Disponível em http://www.top500.org/lists/2005/06/. Último acesso em 01/11/2005.

[14] Virginia Tech TeraScale Computing Facility. Disponível em http://www.tcf.vt.edu/index.html. Último acesso em 01/11/2005.

[15] W. Zhang, S. Jin, and Q. Wu. Creating linux virtual servers. Technical report, National Laboratory for Parallel & Distributed Processing.