1

Arquitetura e Administração de Aglomerados

Marcelo Pasin (LSC-UFSM, *pasin@inf.ufsm.br*)[1]
Diego Luís Kreutz (LSC-UFSM, *kreutz@inf.ufsm.br*)[2]

Resumo:

Este texto apresenta os sistemas computacionais de alto desempenho atuais, baseados em grupos de computadores tradicionais interligados por uma rede de alto desempenho. São estudadas as principais características dos elementos (computadores e rede) atualmente utilizados. São por fim discutidos aspectos práticos de implementação, como seleção, montagem, instalação e administração deste tipo de sistema.

[1]Doutor em Informática pelo Instituto Nacional Politécnico e Grenoble (França, 1999) na área de Processamento Paralelo. Mestre em Ciência da Computação pela Universidade Federal do Rio Grande do Sul (1994) e Engenheiro Eletricista pela Universidade Federal de Santa Maria (1988). Professor da Informática da Universidade Federal de Santa Maria, coordenador do curso de Bacharelado em Ciência da Computação.

[2]Bacharel em Ciência da Computação pela Universidade Federal de Santa Maria (2003).

LSC-UFSM, Universidade Federal de Santa Maria, 97105-900, Santa Maria, RS

1.1. Introdução

Motivados pela impossibilidade tecnológica de aumentar indiscriminadamente a freqüência do relógio dos circuitos eletrônicos usados nos computadores, os arquitetos de sistemas de computação buscam constantemente outras alternativas para o aumento do desempenho. Uma alternativa encontrada foi o paralelismo, onde uma tarefa a ser executada é decomposta em partes menores, as quais são executadas de forma concorrente, diminuindo o tempo total de execução da tarefa completa.

Os primeiros computadores a fazer uso de paralelismo, os supercomputadores, foram construídos no intuito de executar em tempo hábil algumas das aplicações que requerem um enorme potencial de processamento. Tais computadores possuíam um custo bastante elevado e eram privilégios de organizações com alto poder aquisitivo. Somente este tipo de organização se dava o trabalho de escrever aplicações paralelas, pois elas eram as únicas a possuir um equipamento capaz de executá-las.

É senso comum atualmente que a evolução dos computadores, até mesmo dos mais simples e baratos, passará por um caminho onde cada vez mais será explorado o paralelismo. Nos dias de hoje, computadores que exploram o paralelismo não são mais tão caros. Até mesmo os microcomputadores pessoais comuns encontrados atualmente no comércio apresentam algum tipo de paralelismo. Aplicações que façam uso de paralelismo começam a serem desenvolvidas para serem usadas em computadores comuns.

No contexto tecnológico atual, os **aglomerados**[3] **de computadores** representam a forma mais popular de computador paralelo. Na verdade, um aglomerado não é um computador, mas 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. Boa parte disto é obtido graças a camadas de software, dentro ou fora do sistema operacional. O objetivo deste texto é apresentar as características arquiteturais destes aglomerados, dando um enfoque um pouco mais prático e tecnológico que nas publicações tradicionais.

A segunda seção deste texto faz uma brave introdução ao tema da programação paralela e define alguns termos. A seção seguinte, núcleo principal do texto, apresenta as arquiteturas paralelas mais populares atualmente. A seguir, são apresentadas algumas formas de construção de aglomerados, de um ponto de vista tecnológico. Em seguida são discutidas questões da instalação física de aglomerados e algumas questõés da sua administração.

1.2. Programas Paralelos

Não é intenção deste texto estudar em profundidade as questões pertinentes dos programas paralelos. Entretanto, faz-se necessário definir alguns termos que serão usados no decorrer do texto, relacionados com a programação das máquinas paralelas.

As aplicações para as quais tipicamente se utiliza uma máquina paralela são chamadas de **aplicações paralelas**. São geralmente aplicações seqüenciais consideradas muito grandes. Algumas delas, por apresentarem características excepcionalmente grandes, são chamadas de Aplicações dos Grandes Desafios[4] (GCAs). Algumas delas neces-

[3] *clusters*

[4] *grand challenge applications*

sitam de enormes quantidades de memória, outras possuem um tempo de processamento grande demais, outras ainda fazem uso elevado de comunicação. Em geral, elas resolvem problemas de áreas fundamentais de conhecimento e que possuem grande impacto econômico ou científico [KIT 90]. Algumas delas são tidas como impossíveis de se resolver sem o uso de modernos computadores paralelos, por causa do tamanho de suas necessidades em matéria de tempo de processamento, memória ou comunicação. Exemplos de aplicaçõés paralelas são as simulações de problemas estruturais, de cristalografia, tomografia, dinâmica de proteínas, química quântica, meteorologia global, eventos discretos, etc.

Todo computador paralelo possui diversos **elementos processadores**, cada um capaz de acessar uma certa porção de memória (que pode eventualmente ser a mesma dos outros). Com o propósito de executar uma aplicação em paralelo, partes do programa, chamadas **tarefas elementares**, e dos dados são atribuídos a cada um dos elementos processadores. O conjunto de tarefas elementares de uma aplicação é chamado de **programa paralelo**.

As aplicações dos programas paralelos geralmente apresentam um tempo de processamento bastante elevado. A maioria delas é impraticável se executada em um computador comum, pois o tempo de sua execução pode exceder àquele que se está disposto a esperar pelo resultado. O desafio nesta área é encontrar o limite de simplificação que se pode aplicar às soluções, mantendo uma qualidade aceitável do resultado. Computadores mais velozes e ambientes de desenvolvimento de programas paralelos mais eficientes impoem menos restrições às aplicações e permitem encontrar resultados mais realistas. Não será entretanto tratado aqui o desenvolvimento das aplicações paralelas propriamente ditas.

1.3. Máquinas Paralelas e Aglomerados

Uma das primeiras formas vislumbradas de construção de máquinas capazes executar programas paralelos foi conectar um grande conjunto de unidades de processamento a uma memória em comum. São exemplos deste tipo de máquina os processadores vetoriais[CAL 79], os grandes *pipelines* [PAT 98], os processadores superescalares [PAT 98] e os multiprocessadores simétricos[5] [HAM 85]. Com o aumento do número de elementos processadores, este tipo de arquitetura apresenta problemas de congestionamento de acesso à memória, o que termina por limitar drasticamente o seu paralelismo.

Com novas tecnologias de interconexão de processadores, principalmente no domínio das redes de computadores, surgiu uma nova direção na evolução das máquinas paralelas. Com circuitos de interconexão suficientemente rápidos os computadores paralelos passaram a trabalhar usando módulos de memória independentes para cada elemento processador (ou grupo de elementos processadores). O compartilhamento de dados passou a se dar por mensagens enviadas pela rede de interconexão. O tempo de espera necessário para a entrega de mensagens, que era um fator crítico neste tipo de arquitetura, passa a não ser mais tão importante graças ao alto desempenho dos circuitos das redes. Não obstante, esta técnica complica ainda mais a já difícil tarefa de programar em paralelo, pois mensagens devem ser previstas para efetivar o compartilhamento de dados entre tarefas paralelas elementares.

[5] *symmetric multi-processors, SMPs*

Esta seção abordará o tema da construção dos computadores paralelos atuais. Inicialmente serão estudados os multiprocessadores simétricos e os seus caminhos de evolução, incluindo uma discussão sobre os processadores baseados em múltiplos fluxos de execução. Logo após serão rapidamente descritos os computadores vetoriais. A seguir, serão apresentados os aglomerados[6] de multiprocessadores como solução prática e efetiva em matéria de custo. Finalmente, esta seção será concluída com a apresentação das redes com capacidade de endereçamento de memória distante, um paradigma que possibilita a simplificação da programação dos aglomerados por permitir o uso de memória compartilhada.

1.3.1. Grandes Multiprocessadores

Multiprocessadores são computadores que possuem mais de um processador. Os primeiros multiprocessadores apareceram na década de 80, na época ainda muito caros. Este tipo de arquitetura já se tornou hoje uma coisa comum, pois se encontram no mercado até computadores pessoais com mais de um processador. No intuito de diminuir os preços, acelerar as trocas de dados, diminuir o consumo, etc. certos fabricantes já começam a anunciar o lançamento de microprocessadores com dois processadores independentes embutidos em seu interior [TEN 2002].

Geralmente todos os processadores de um multiprocessador acessam uma região de memória comum. Quando todos os processadores têm a capacidade de exercer qualquer função indistintamente, estes computadores são chamados de **multiprocessadores simétricos**. Para isto, toda a memória acessível por um processador é igualmente acessível por todos os outros. Por este motivo este tipo de computador também é dito como tendo **acesso uniforme à memória**,[7] ou UMA.

A popularidade destes sistemas não é devida somente à sua capacidade de suprir a demanda por alto desempenho. Estes sistemas são excelentes para execução de sistemas baseados em multiprogramação, como é o caso da maioria dos servidores corporativos atuais. Dentro de certos limites, o desempenho de tais computadores cresce de forma linear com relação ao seu número de processadores. Eles apresentam as seguintes vantagens [HWA 98]:

- **Simetria**: qualquer processador pode acessar qualquer parte da memória ou qualquer dispositivo de entrada e saída;
- **Espaço de endereçamento único**: esta característica se desmembra em duas:
  - Imagem única do sistema seja qual for o processador usado, já que somente uma cópia do sistema operacional e da aplicação reside em memória. O sistema escolhe em qual processador do sistema um dado processo deve ser executado, dependendo da carga do sistema, obtendo de forma simples um equilíbrio dinâmico de carga.
  - Compartilhamento global de dados. Os processos não precisam se preocupar em transferí-los de um espaço de endereçamento para outro.

---

[6] *clusters*

[7] *uniform memory access*

- **Baixa latência**: a troca de dados de um processo a outro é feita por acessos ordinários (*load*/*store*) à uma porção de memória em comum. Não é necessário copiar dados.

- **Replicação e coerência**: a localidade de dados é fornecida pelas *caches* dos processadores, cuja coerência é garantida pelo *hardware*, com o uso de *caches* bisbilhoteiras[8] [ARC 86].

A principal vantagem de sistemas deste tipo é a possibilidade que qualquer processador tem de manipular toda a memória e todos os processadores. Eles precisam entretanto de algum dispositivo para decidir entre eles quem faz o quê e quando. Isto significa que o hardware deve possuir mecanismos de árbitro. Geralmente este problema é resolvido através de acessos de memória com leitura e modificação atômicos.

Atualmente diversos fabricantes vendem multiprocessadores simétricos de alto desempenho, como as séries eSeries e xSeries IBM, a série Alphaserver da Compaq/HP, o T3E da Cray e as famílias Enterprise e Fire da Sun. Muitos deles são SMPs com acesso uniforme à memória que podem ser aglomerados para formar sistemas maiores, com acesso não uniforme à memória.

1.3.1.1. Arquitetura de multiprocessadores simétricos

A figura 1.1 mostra a arquitetura típica de um SMP. O circuito de interconexão unindo os elementos (processadores, memória e dispositivos de entrada e saída) mais simples que pode ser construído é um barramento. Ele pode ser visto como um conjunto de fios paralelos que ligam todos os elementos, um protocolo de comunicação e eventualmente algum dispositivo auxiliar (árbitros, relógios, etc.). Por ele deve obrigatoriamente passar todo o tráfego entre todos os elementos existentes.

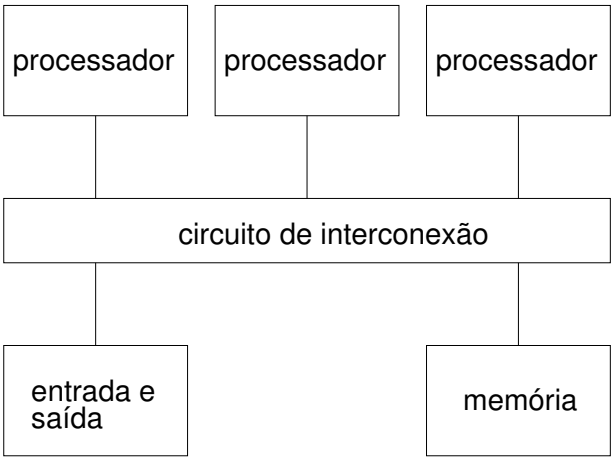

Figura 1.1: Arquitetura típica de um SMP.

Por ser implementado com circuitos lógicos seqüenciais, um barramento funciona segundo uma cadência de relógio. Com isto, um barramento suporta uma taxa de transferência[9] limitada de dados. O aumento do número de elementos ativos (processadores

[8] *snoopy caches*
[9] *bandwidth*

principalmente) pode transformá-lo em um gargalo, limitando o desempenho do sistema. Além disto, um barramento único é um ponto crítico de falha: se ele falhar o sistema pára. Por estes motivos, pode-se optar por construir um circuito de interconexão com mais caminhos entre os elementos, permitindo que mais de uma transferência ocorra ao mesmo tempo. Três maneiras são geralmente usadas para atingir este objetivo: barramentos comutados,[10] arquiteturas CC-NUMA (seção 1.3.1.2.) e aglomerados (seção 1.3.3.).

A figura 1.2 mostra um exemplo de circuito de interconexão usando um comutador cruzado [LAN 92]. Nele podem haver três comunicações simultâneas, desde que os processadores não queiram acessar o mesmo elemento (memória ou entrada e saída). Construções deste tipo são naturalmente mais caras e mais complicadas de se construir que barramentos simples. Por este motivo, somente os computadores das gamas mais altas apresentam barramentos comutados.

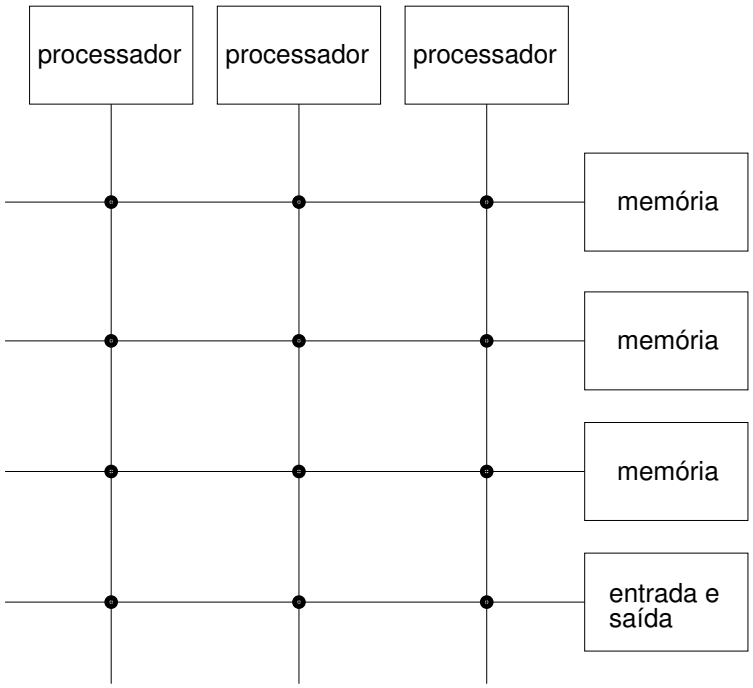

Figura 1.2: Um circuito de interconexão comutado.

1.3.1.2. Sistemas CC-NUMA

São chamados CC-NUMA[11] os sistemas que apresentam acesso à memória não uniforme e coerência de *cache*. Estes sistemas são construídos servindo-se de elementos independentes, aos quais são compostos de um bloco de memória e um ou mais processadores. Um processador pode acessar tanto o bloco de memória de seu elemento — local — quanto um bloco de memória de outros elementos — remotos. O circuito de interconexão é feito em dois níveis, de tal maneira que o tráfego de dados de e para um bloco de memória não interfira no tráfego de um outro bloco, permitindo assim aumentar consideravelmente o número de processadores do sistema.

---

[10] *crossbar*

[11] *cache-coherent non-uniform memory access*

A figura 1.3 mostra um exemplo de arquitetura de sistema CC-NUMA. Nele se nota um barramento local ligado a cada bloco de memória, nos quais dois processadores e um diretório também estão conectados. Estes barramentos locais, juntamente com seus processadores e memória, formam o que se chama de elemento. Um barramento vertical interliga os elementos, criando assim uma hierarquia de barramentos.

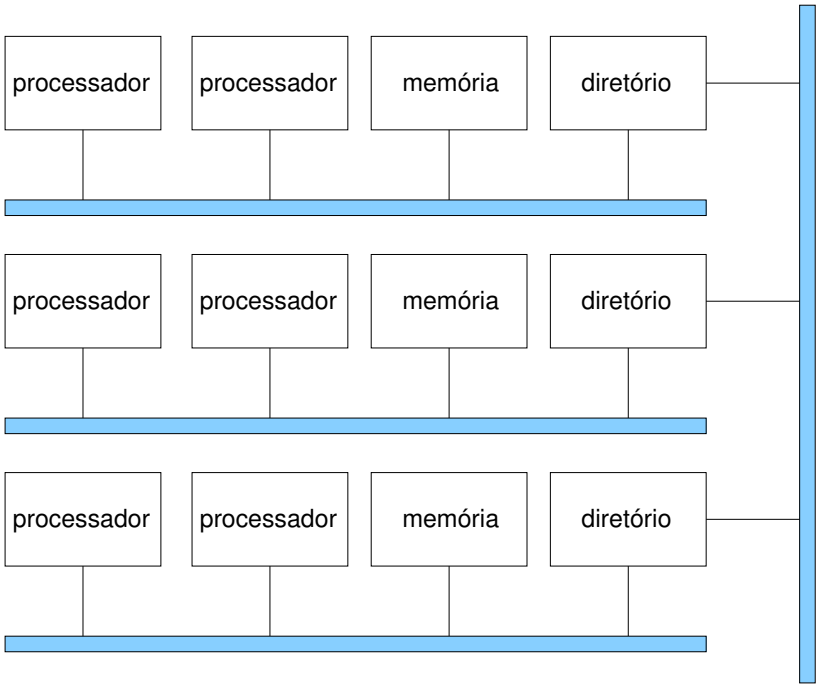

Figura 1.3: Exemplo de arquitetura CC-NUMA.

Os processadores de um sistema CC-NUMA continuam usando *caches* bisbilhoteiras e tendo coerência entre elas no barramento local. Os diretórios, além de servirem de roteadores de acessos remotos a memória, devem manter a coerência global do sistema. Eles se inserem nos conjuntos locais de *caches* bisbilhoteiras e se comunicam entre si, de maneira a invalidar as linhas de *cache* locais que forem modificadas em alguma *cache* distante.

Cada acesso a um bloco de memória local passa somente pelo barramento local enquanto que acessos a blocos remotos passam por dois diretórios e pelo barramento vertical. Isso significa dizer que acessos a blocos remotos de memória são mais lentos que acessos ao bloco local, daí a razão dos acessos à memória não serem uniformes. Ainda assim, estes sistemas são uma extensão dos SMP, pois apresentam características muito similares de programação. Programas de CC-NUMAs devem apenas evitar acesso a blocos remotos de memória devido ao seu maior tempo de acesso.

Ao utilizar um sistema de memória compartilhada distribuída, os sistemas CC-NUMA apresentam melhor escalabilidade que os sistemas SMP. A quantidade de processadores, memória e dispositivos de entrada e saída pode ser aumentada simplesmente adicionando mais elementos ao sistema. Os problemas de desempenho no barramento inter-elementos pode ser resolvido com o uso de barramentos comutados ou ainda por hierarquias sucessivas. Além de maior escalabilidade, eles podem apresentar maior disponibilidade, se o barramento inter-elemento for construído de forma a permitir a substituição de elementos sem a interrupção do funcionamento do sistema.

1.3.1.3. Microprocessadores Múltiplos

Com a crescente capacidade de integração de transístores em um só circuito, pode-se imaginar várias formas de integração agressiva de paralelismo nos futuros microprocessadores [BUR 97]. Algumas destas formas podem requerer uma transformação do modelo tradicional de programação. Vários grupos de pesquisa trabalham atualmente em diversas direções, averiguando as suas vantagens e desvantagens.

Atualmente os processadores superescalares possuem *pipelines* múltiplos, permitindo a execução simultânea de três ou quatro instruções. Um dos fatores que limitam atualmente esta execução simultânea é a taxa de despacho de instruções. Para aumentar esta taxa pode-se usar *caches* de rastros de execução[12] [PAT 97], que armazenam o fluxo de instruções executado em vez de guardar os blocos de memória que contém as instruções executadas, reduzindo o seu tamanho e otimizando o seu acesso. Outra maneira de se aumentar a multiplicidade de execução de instruções é a especulação massiva. Um modelo que use fracas dependências de dados e que faça a busca de instruções em blocos maiores (128 por vez, por exemplo) pode permitir uma execução especulativa em maior escala. Nenhuma destas soluções implica em mudar os conjuntos de instruções atuais.

Baseado na idéia de que os sistemas operacionais atuais são multiprogramados, sabe-se que um processador atual invariavelmente vai executar diversos fluxos de execução em tempo compartilhado. Um processador com diversos *pipelines* independentes poderia executar simultaneamente diversos fluxos de execução [SMI 97, SMI 2001]. A vantagem aqui é que não existem dependências de dados entre dois fluxos de execução diferentes e as instruções não precisam ser reordenadas. Além disso, cada vez mais programas vêm sendo escritos para utilizar múltiplos fluxos de execução cooperativos. Neste caso, um processador com multiplos fluxos de execução terá ainda acessos otimizados à sua *cache*, pois fluxos cooperativos compartilham dados [TUL 95].

Fazendo extensões ao conceito de multiprocessamento simétrico, há quem pesquise a construção de circuitos integrados contendo múltiplos processadores [HAM 97, TRE 2000, TEN 2002]. Isto reduz os custos e otimiza o tráfego no barramento, já que todos os processadores do circuito integrado poderiam compartilhar a mesma *cache*. Estes processadores podem cooperar para a execução de uma única tarefa paralela ou executar processos independentes. Embora este tipo de multiprocessador possa executar tarefas independentes em cada um dos processadores, o compartilhamento da *cache* pode impôr severas perdas em desempenho. Já pelas propriedades de localidade, o acesso à *cache* não seria prejudicado por tarefas cooperativas. Os defensores desta arquitetura acreditam que no futuro será corriqueiro o uso de tarefas paralelas cooperativas, o que permitirá obter um bom desempenho deste tipo de multiprocessador.

1.3.2. Computadores Vetoriais

Em vez de executar diversas tarefas paralelas de forma concomitante, os computadores vetoriais propõem uma abordagem diferente. Processadores vetoriais apresentam conjuntos de instruções atípicos se comparados a processadores comuns. Neles, um único fluxo de execução define o programa paralelo, mas este possui instruções especiais que operam sobre um conjunto de dados a cada vez.

---

[12] *trace caches*

Os algoritmos que operam vetores de dados são aqueles que mais tipicamente tiram melhor proveitos dos computadores baseados em procesadores vetoriais. Em vez de iterar em um laço de repetição, operando um elemento do vetor por vez, uma instrução opera com diveros elementos do vetor simultaneamente. Compiladores especiais reconhecem os laços de repetição que iteram as operações com vetores e os convertem em instruções vetoriais [ZIM 91].

A figura 1.4 mostra um trecho de programa que faz a soma de dois vetores (a), compilado para ser executado em um processador tradicional (b) e em um processador vetorial (c). O programa do processador tradicional soma elemento por elemento do vetor (instrução `add`) e conta o número de iterações com o registrador `r2`. O programa vetorial usa somente uma instrução (`addv`) para fazer, de uma só vez, a soma de dez elementos de `b[]` em `a[]`.

```
for i = 1 to 10
    a[i] = a[i] + b[i]
```

(a)

```
    ld  r2, 0
loop:
    add   (r3+r2), (r4+r2)
    incr  r2
    comp  r2, 10
    blt   loop
```

(b)

```
    addv  r3, r4, 10
```

(c)

Figura 1.4: Programa vetorizável.

As instruções vetoriais aumentam enormemente a complexidade dos processadores. Mais que isso, por depender de vetores freqüentemente armazenados em memória e por necessitar efetuar o acesso aos elementos destes vetores em posições distintas de memória, os processadores vetoriais terminam por serializar as micro-operações dos diferentes elementos dos vetores. Assim sendo, processadores com tecnologia RISC e com *caches* e *pipelines* avantajados podem obter desempenhos similares aos dos processadores vetoriais.

Não obstante, muitos processadores modernos incorporam instruções vetoriais no seu rol de instruções [PEL 97, OBE 99, MIC 2002, DIE 2000]. Estas instruções foram incorporadas na maioria dos microprocessadores comerciais, devido à necessidade de tratamento intensivo de sinais ao trabalhar com dados de multimídia. Esta instruções geralmente operam sobre um número reduzido de registradores, ao invés de operar diretamente com vetores na memória.

Cronologicamente as arquiteturas baseadas em processadores vetoriais apareceram antes que aquelas baseadas em multiprocessadores. Há quem diga hoje que os computadores vetoriais ainda são superiores em desempenho para certas aplicações maciçamente numéricas, onde as tarefas paralelas fazem intensivo compartilhamento de da-

dos. Fora dos pequenino grupo de fabricantes de supercomputadores (excetuando-se as instruções de multimídia) pouco investimento se faz atualmente no sentido de desenvolver este tipo de arquitetura. Ainda assim, os computadores de mais alto desempenho atualmente são uma combinação agressiva de diversos modelos arquiteturais, como processadores vetoriais, multiprocessadores simétricos e aglomerados.

1.3.3. Aglomerados

Um aglomerado[13] é um sistema de computação paralela composto por um conjunto de computadores independentes (elementos) trabalhando de forma integrada como se fosse um recurso computacional único [BUY 99]. A grande idéia por tras dos aglomerados é que alguém pode construir um equipamento com alto poder computacional simplesmente empilhando diversos computadores comuns (até mesmo computadores pesoais). Basta aumentar o tamanho da pilha quando mais poder computacional for necessário.

Os elementos componentes de um aglomerado podem residir em um único gabinete ou estar fisicamente separados e conectados por uma rede local. O que determina a posição física dos elementos é o tipo de rede usado na sua interconexão, principalmente quanto ao comprimento máximo dos cabos usados. Em geral, como as redes usadas tendem a se encontrar no limiar máximo do seu desempenho e são construídas com componentes comuns (baratos), a distância entre os elementos raramente ultrapassa alguns metros.

Os aglomerados são mecanismos eficientes em matéria de custo na obtenção de altos desempenho e disponibilidade. Por serem construídos com peças comuns, encontradas com facilidade no mercado, a relação custo benefício de um aglomerado é mais baixa que aquela encontrada em grandes computadores. Um aglomerado construído com uma dúzia de computadores pessoais, que execute um certo programa de simulação numérica em um dia, será provavelmente mais barato que um computador único, dotado de um ou vários elementos processadores (eventualmente vetoriais), que execute este mesmo programa no mesmo tempo.

Exemplos de aglomerados são os projetos Beowulf [STE 99] e NOW[14] [PAT 95], e o IBM SP [STU 95]. Na verdade, todos grandes recursos computacionais da atualidade são aglomerados. Mesmo fazendo uso de multiprocessadores e de processadores vetoriais, somente foi posível utilizar milhares de processadores ao construir sistemas sob a forma de um enorme aglomerado. Isto pode ser visto pelas estatísticas atuais dos 500 sistemas computacionais mais poderosos do mundo (`http://www.top500.org`), onde, dos 10 maiores atualmente, todos são aglomerados de SMPs, um formado com processadores vetoriais da série SX-6 da NEC, quatro da série AlphaServer da HP/Compaq, três são IBM SP e dois são montados com processadores Pentium Xeon.

As principais características de um aglomerado são:

- **Independência**: Cada elemento possui um ou mais processadores, memória, dispositivos de entrada e saída e executa o seu próprio sistema operacional. Este elemento geralmente pode ser desligado e até mesmo removido do aglomerado sem que isto afete o funcionamento dos outros.

---

[13] *cluster*

[14] *network of workstations*

- **Imagem única do sistema**: um aglomerado é um recurso computacional único, ao contrário de um sistema distribuído onde os elementos são recursos individuais.

- **Conexão especializada**: Os elementos de um aglomerado são geralmente conectados por algum tipo de rede rápida de tecnologia aberta. Apesar disso, geralmente são usados protocolos de comunicação especializados.

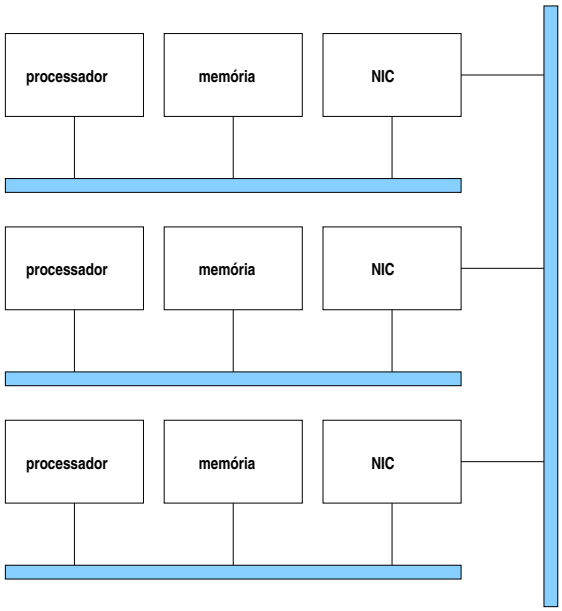

Figura 1.5: Exemplo de aglomerado.

A figura 1.5 mostra um exemplo de arquitetura de um aglomerado. Nele, a interface de rede[15] (NIC) contém um processador de comunicação para transferir pacotes de dados através de um circuito de conexão existente entre os elementos do aglomerado. Esta idéia tem a vantagem de levar em consideração a evolução das arquiteturas das máquinas paralelas atuais segundo dois eixos:

- A crescente integração de múltiplos processadores dentro de uma mesma máquina (SMPs);

- A interconexão sistemática de máquinas usando redes cada vez mais rápidas.

Freqüentemente as interfaces de rede de um aglomerado são interfaces comuns de mercado, em suas versões mais rápidas, como Gigabit Ethernet ou Myrinet por exemplo, mas certos aglomerados comerciais apresentam uma interface de rede proprietária. Eles usam protocolos de comunicação rápidos, como por exemplo mensagens ativas [EIC 92]. Em geral estes protocolos permitem deixar de lado o sistema operacional durante a comunicação, evitando a sua sobrecarga implícita e permitindo acesso direto, a nível de usuário, às interfaces de rede.

Um aglomerado pode trabalhar de forma coletiva como um sistema computacional único ou como um grupo de computadores individuais. Uma camada de software situada

[15] *network interface card*

entre o sistema operacional e a aplicação[16] é responsável pela ilusão de que o sistema tem uma imagem única. Ela desobriga o usuário de saber onde as aplicações estão sendo executadas e onde se encontram os recursos usados. Ela muitas vezes também faz com que o sistema se recupere automaticamente no caso de falhas.

O funcionamento integrado de um aglomerado permite que ele atinja altos índices de desempenho em certas aplicações. Um exemplo é um grande servidor corporativo, onde cada um dos elementos atende um subconjunto dos clientes do sistema, compartilhando os recursos de todo o aglomerado através da sua rede local dedicada. Outro exemplo é um sistema dedicado ao processamento paralelo, onde tarefas elementares de uma aplicação cooperam entre si usando a rede local dedicada.

Dependendo de como um aglomerado é construído, ele pode apresentar poucos ou muitos componentes de fabricação genérica. Alguns fabricantes constróem aglomerados a partir do nada, que apresentam alto desempenho mas seus elementos não são compatíveis com a maioria dos encontrados no mercado. Isto diminui a flexibilidade conceitual do aglomerado. No outro extremo, certos aglomerados são construídos inteiramente pelos seus proprietários, com componentes padronizados encontrados em diferentes fornecedores. Desta forma obtém-se alta flexibilidade mas não se pode contar com suporte de nenhum fabricante específico. A seção 1.4. apresentará com mais detalhe as soluções atualmente existentes, contrastando as questões como estas.

1.3.4. Arquiteturas com memória distribuída compartilhada

Existem dois grandes paradigmas de programação paralela atualmente em voga: programação com memória compartilhada e programação por troca de mensagens. O primeiro paradigma pressupõe que exista memória compartilhada entre os processadores que estão executando as tarefas elementares do programa paralelo. O segundo paradigma não conta com esta memória compartilhada. Nele, cada dado pertence a uma só tarefa elementar e somente ela tem acesso a ele. Faz-se uso de troca de mensagens sempre que for necessário passar um dado de uma tarefa a outra. Há quem advogue que o paradigma com memória compartilhada é mais simples, pois nele todos os dados pertencem a todas as tarefas elementares e não é necessário se preocupar com a troca de mensagens. Esta seção apresenta uma classificação dos computadores paralelos atuais, com relação a sua capacidade de oferecer uma memória compartilhada entre os elementos processadores.

Computadores paralelos contém múltiplos elementos, cada um tendo um ou mais processadores e uma memória local. Com relação a um dado elemento, a memória dos outros elementos é chamada de memória remota. Dependendo de como é feita a interligação destes elementos, o acesso à memória remota pode ter diferentes níveis de complexidade. Com relação a este aspecto, tais sistemas podem seguir um dos seguintes modelos:

- **NORMA**:[17] sem acesso à memória remota
- **NCC-NUMA**:[18] sem acesso uniforme à memória, sem coerência de *cache*.
- **CC-NUMA**:[19] sem acesso uniforme à memória, com coerência de *cache*.

---

[16] *middleware*
[17] *no remote memory access*
[18] *non-cache-coherent, non-uniform memory access*
[19] *cache-coherent, non-uniform memory access*

- **COMA**:[20] arquitetura de memória somente de *cache*.
- **SC-NUMA**:[21] sem acesso uniforme à memória, com coerência feita por *software*.

Os sistemas do tipo NORMA não possuem dispositivos físicos capazes de realizar um acesso remoto à memória. Cada elemento possui um espaço de endereçamento separado e desconhecido dos outros elementos. A única forma de acessar dados remotos é através da troca de mensagens. É a arquitetura de memória mais comumente encontrada em aglomerados.

Os outros quatro modelos possuem algum mecanismo físico que permite o acesso às memórias remotas. Este mecanismo agrupa os espaços de endereçamento de cada elemento formando um único espaço de endereçamento. Qualquer processador do sistema pode acessar qualquer módulo de memória do sistema. Entretanto, dependendo do modelo escolhido, suas formas de acesso diferem.

Em uma NCC-NUMA, quando um acesso de leitura é feito a um endereço remoto, uma cópia é feita em uma porção local de memória. Sob comando, um conteúdo da memória local pode ser escrita remotamente. Normalmente os acessos se fazem em blocos maiores que linhas de *cache*. A responsabilidade de manter a coerência das cópias em memória local com os dados armazenados em memórias remotas é dos programas em execução. Este tipo de arquitetura é comumente construída com placas de rede com capacidade de endereçamento de memória remota, como por exemplo a *Scalable Coherent Interface* [OMA 97] e a *Memory Channel* [GIL 96].

Sistemas CC-NUMA já foram descritos anteriormente (seção 1.3.1.2.), como sendo uma extensão típica de multiprocessadores. Nestes sistemas os acessos a endereços remotos são transferidos por uma rede de interconexão dedicada a este fim. As placas de rede deste tipo estão ligadas diretamente ao barramento de memória do elemento. A coerência das *caches* dos processadores com os módulos de memória dos elementos remotos é feita pelas placas de rede, que mantêm localmente diretórios das linhas de *cache* correspondentes a memórias remotas.

Sistemas COMA [BUR 92] e sistemas CC-NUMA são muito similares. Ambos apresentam acesso direto à memória remota e placas de rede com diretórios de linhas de *cache*. A diferença reside na forma na qual são feitas as migrações dos dados. No caso dos sistemas CC-NUMA, embora o sistema operacional possa decidir pela migração de uma dada página se ela for muito acessada remotamente, não há nenhum dispositivo físico capaz de efetuar esta migração automaticamente. Já nos sistemas COMA, toda a memória é organizada como linhas de *cache*. Neste caso, as próprias placas de rede fazem a migração da memória ao manter a coerência das *caches*.

Quando não existe meio físico de acesso (NORMA) ou quando o acesso remoto faz uma cópia em memória local (NCC-NUMA), pode-se optar por efetuar a coerência de memória por *software*. São assim chamados os sistemas SC-NUMA ou ainda os sistemas com memória compartilhada distribuída[22] (DSM). Nestes sistemas, o acesso a um dado remoto causaria a sua transferência para a memória local (usando mensagens em uma NORMA ou fazendo cópia à distância em uma NCC-NUMA). Dependendo do nível de software onde se inserem os mecanismos de coerência de memória, ela pode ser feita por páginas (no sistema operacional) ou por objetos (na linguagem de programação).

---

[20] *cache-only memory architecture*

[21] *software-coherent, non-uniform memory access*

[22] *distributed shared memory*

1.4. Construção de Aglomerados

Falar de conceitos tão recentes como aglomerados sem apresentar exemplos práticos é muito difícil, pois muitos destes novos conceitos surgiram ao se resolver problemas industriais ou comerciais específicos. Esta seção se devotará a mostrar um panorama atual (novembro de 2002) da construção de aglomerados, tentando abordar seus aspectos conceituais sem se proibir de entrar em detalhes tecnológicos.

Para melhor apresentar as características tecnológicas atuais, os aglomerados foram divididos em três grupos principais. O primeiro a ser apresentado será o grupo dos aglomerados construídos inteiramente por um fabricante. Este tipo de aglomerado foge um pouco da idéia tradicional, por ser construdo usando peças de projeto proprietário, em benefício do desempenho. A seguir, será apresentado um grupo que reune aglomerados montados por grandes fabricantes, desta vez usando-se de peças de projeto aberto. Estas máquinas têm a vantagem de ser compatibilizadas (a nível de software e hardware) pelo fabricante. Por último, será apresentado o grupo dos aglomerados não integrados. Estes, por serem feitos pelos seus próprios proprietários são as soluções mais baratas.

1.4.1. Soluções Integradas Proprietárias

Os inventores dos primeiros aglomerados preconizavam que esta seria uma maneira barata de se construir sistemas computacionais de alto desempenho. Pode-se assim pensar que todos os aglomerados são construídos com componentes de baixo custo. Não obstante, uma característica muito importante dos aglomerados não tem relação com o preço: a escalabilidade. É impossível construir sistemas computacionais gigantescos mantendo-se a memória comparilhada entre os elementos processadores, paradigma básico dos computadores vetoriais e dos SMPs. Em outras palavras, os sistemas computacionais de mais alto desempenho que são concebidos atualmente são aglomerados.

Todos grandes fabricantes de computadores, tanto na área de supercomputadores (Cray, NEC) como na área de grandes servidores corporativos (IBM, HP/Compaq, Sun), apresentam soluções baseadas em aglomerados. Como nas soluções antigas dos computadores centrais,[23] as novas soluções integradas proprietárias para processamento de alto desempenho continuam pouco flexíveis e mantendo o usuário vinculado diretamente ao fabricante do sistema. Isso se deve ao fato de estas soluções integradas não se preocuparem com detalhes como compatibilidade e flexibilidade.

O objetivo final dos grandes fabricantes continua sendo o desenvolvimento de sistemas que forneçam o máximo de desempenho aos usuários finais. Esta característica implica comumentemente em sistemas com barramento de memória específicos, processadores de arquitetura própria, redes com altas taxas de transferência exclusivamente projetadas para o sistema em questão e componentes projetados de acordo com as necessidades da nova arquitetura. Estas decisões arquiteturais permitem obter altos desempenhos, mas a um custo alto. Somente organizações que têm demandas muito altas por desempenho e que não tenham muitas restrições orçamentárias é que podem se oferecer este tipo de equipamento.

Outro grande inconveniente destas soluções é a dependência criada entre fornecedor e cliente. Ao comprar um equipamento baseado em um computador de arquitetura proprietária, fabricada e vendida por um único fabricante, estabelece-se uma relação de

[23] *mainframes*

dependência com este fabricante. Atualizações, espansões, reparações, todas terminam por ser contratadas do mesmo fornecedor inicial, por er o único capaz de fornecer peas compatíveis com o seu sistema. Isto impossibilita o uso da concorrência para se obter melhores preços.

Algumas outras vantagens, além do alto desempenho, podem ser observadas neste tipo de sistema. Da maneira como eles são montados, com componentes homogêneos e com software proprietário, é praticamente impossível enfrentar problemas de compatibilidade. O próprio fabricante resolve estes problemas, lançando novos modelos que continuam sendo compatíveis com os ateriores. Em alguns casos apenas é necessário fazer-se uma atualização no software. Geralmente se tem a garantia do fabricante de se preservar o investimento.

Por serem soluções inteiramente concebidas e montadas por um único fabricante, é fácil obter-se suporte para eventuais problemas. Como será visto mais adiante, aglomerados montados com peças fornecidas por diversos fabricantes apresentam sérios problemas de suporte. Se uma certa interface de rede não funciona em um certo barramento, mas funciona com as outros teoricamente compatíveis, quem deve providenciar a solução? O fabricante da interface de rede ou o do barramento? Geralmente nenhum dos dois resolve. Soluções integradas não tem este problema.

Atualmente, aglomerados de computadores e estações de trabalho normais não são capazes de bater o desempenho destas soluções se comparados utilizando números iguais de processadores ou de quantidade de memória, por exemplo. Soluções como a SGI Origin 3900 da Silicon Graphics (SGI) disponibilizam até 128 processadores MIPS em um único armário de 19 polegadas. Entre as soluções da Sun Microsystems podemos citar a sua série Fire, uma cc-NUMA que suporta até 106 processadores UltraSPARC III, atualmente 1.05 GHz e possui componentes como CPUs e discos rígidos, que podem ser instalados ou trocados com o sistema em funcionamento. Já a IBM apresenta soluções para processamento de alto desempenho com a série SP, com até 2048 processadores.

O grande diferencial destas soluções são suas arquiteturas e dispositivos desenvolvidos especificamente para prover alto desempenho. Pode-se citar como exemplo a arquitetura de memória do IBM pSeries 690 que é capaz de fornecer taxas de largura de banda de até 200 GB/s. Já o sistema modular de memória NUMAflex da SGI é capaz de fornecer alto desempenho combinando até 512 processadores em um único sistema de memória compartilhada. A solução Sun Fire 6800 merece destaque pela rede de interconexão Fireplane Interconnect que fornece uma taxa de vazão de 9.6 GB/s. Como pode ser observado, as soluções apresentadas possuem características próprias e incomuns aos demais sistemas disponíveis no mercado. Esse é um fator marcante no desempenho final do sistema.

1.4.2. Soluções Integradas Abertas

Atualmente os grandes fabricantes de computadores estão investindo pesado em soluções abertas para aglomerados de computadores, se configurando como uma tendência clara de mercado. Estas soluções se baseiam fundamentalmente em aglomerados, com a montagem de grandes sistemas computacionais usando elementos comumente encontrados no mercado. Teoricamente, por ser de tecnologia aberta, qualquer peça de um aglomerado deste tipo pode ser substituída por uma peça similar feita por um outro fabri-

cante. Estas soluções são freqüentemente denominadas **soluções de prateleira**,[24] fazendo uma alusão aos bens comprados em estabelecimentos populares, de auto-serviço, como supermercados e lojas de departamentos.

As soluções integradas abertas são soluções que incorporam, entre outros atrativos, eficiência, homogeneidade e custos mais baixos que suas soluções proprietárias. Elas são especialmente interessantes quando se entra na questão de homogeneidade e suporte. As empresas que desenvolvem e disponibilizam estas soluções tem a preocupação centrada no cliente, ou seja, procuram fornecer soluções para problemas genéricos e específicos que atendam as necessidades dos usuários e, principalmente, tornem a tarefa de instalação, manutenção, configuração e utilização extremamente simples e prática. Para tanto, as soluções são normalmente compostas de conjuntos de arquiteturas e equipamentos homegêneos, bem como, um suporte completo de software.

Soluções integradas, apesar de possuirem um custo imediato mais elevado, possuem uma série de vantagens. Estas soluções são normalmente bastante elegantes. Geralmente, possuem uma arquitetura compacta e mais eficiente que arquiteturas de computadores pessoais ou estações de trabalho normais. Isso se deve ao fato de serem arquiteturas desenvolvidas, epecificamente, para suprir as necessidades de consumo do mercado de aglomerados de computadores.

Uma unidade de processamento é comumentemente constituída de um gabinete específico, otimizado e bastante compacto, com tamanho de 1U (unidade de altura de módulos para armários de 19 polegadas). Uma arquitetura deste gênero é extremamente atrativa. Em primeiro lugar, estas soluções já vem otimizadas de forma a amenizar problemas de organização da grande quantidade de cabeamento gerado por um aglomerado de computadores. Em segundo lugar, e possivelmente uma característica determinante para uma grande leva de consumidores, é a otimização do espaço, ou seja, num pequeno espaço físico é possível ter dezenas, ou até centenas, de nós de processamento. O mesmo espaço poderia armazenar apenas alguns poucos computadores com gabinetes normais.

Outra grande vantagem destas soluções integradas abertas é o alto nível de suporte disponibilizado pelos fornecedores. Estes fornecem software desenvolvido especialmente para estas arquiteturas, o que resulta em uma eficiente e fácil utilização dos rescursos disponibilizados por um aglomerado de computadores deste gênero. Além de software específico, os fabricantes fornecem ainda soluções completas para o gerenciamento, monitoração, manutenção e instalação do aglomerado. Esta composição de hardware e software resulta em um produto de maior qualidade e menor custo de instalação, gerenciamento e manutenção. Estes são alguns dos fatores que devem ser levados em consideração quando de uma análise de custo-benefício. O custo imediato pode até ser maior, mas, a médio e longo prazo este custo pode ser compensado pela qualidade, eficiência a fácil manutenibilidade destas soluções integradas.

Principais Soluções Integradas Abertas

Os grandes fabricantes de computadores procuram cada vez mais apresentar produtos atrativos de alto desempenho baseados em aglomerados de arquitetura aberta. Estas, normalmente, buscam fornecer um sistema integrado, estável e que disponibilize o melhor desempenho pelo menor custo. Contudo, estas soluções perfazem um custo direto de aquisição maior que as soluções não integradas (ver seção 1.4.3.). O texto que segue apre-

[24] *off-the-shelf*

senta as principais características e vantagens destas soluções levando em consideração os processadores, barramentos e memórias geralmente utilizados.

Processador As soluções integradas abertas utilizam microprocessadores comerciais comuns, como Pentium da Intel, Athlon da AMD, PowerPC da Motorola, UltraSPARC da Sun, etc. Apesar disso, os fabricantes têm em geral a preocupação de integrar sistemas usando os modelos de mais alto desempenho destes processadores. A linha xSeries da IBM, por exemplo, faz uso de processadores Pentium 4 Xeon da Intel. Já a família xServer da Apple e a utiliza processadores PowerPC G4. Alguns modelos da família PowerEdge da Dell utilizam dois processadores Xeon de 2.4 GHz. Por fim, a série 1UPA da Integrated Solution Systems utiliza processadores Athlon MP da AMD.

Alguns fabricantes optam por manter proprietários os seus processadores e oferecer barramentos abertos para periféricos (tipo PCI [SHA 95], por exemplo) e memória. Isto permite a expansão do sistema com periféricos e memórias de prateleira. Isto no fim das contas tem o mesmo efeito de se usar processadores de tecnologia aberta, já que raramente se troca a marca do processador de um computador sem trocar também sua placa principal. Neste veio se encontra a linha pSeries da IBM, com vários modelos mono e multiprocessados baseados em processadores POWER3 e POWER4, que são fabricados somente pela própria IBM. Da mesma forma a linha Netra da Sun inclui processadores UltraSPARC, de fabricação própria. A HP, ao unir-se com a Compaq, que por sua vez havia adquirido a Digital, oferece igualmente seus AlphaServers baseados em processadores Alpha mas com barramentos PCI para seus paeriféricos.

Barramento de expansão As soluções integradas abertas apresentam em sua vasta maioria barramentos do tipo PCI. Vale talvez ressaltar que, por geralmente incorporar as versões de mais alto desempenho ou preço de componentes de prateleira, estas soluções possuem igualmente os barramentos de mais alto desempenho do padrão PCI. Muito possuem barramentos PCI-X, que permitem a retirada de placas de interface sem desligar o equipamento. Geralmente funcionam a 100 ou 133 MHz, com largura de 64 bits (o PCI tradicional funciona com 32 bits a 66 MHz).

Memória Um dos aspectos fundamentais para se atingir alto desempenho é o sistema de memória. Pouco adianta possuir velozes processadores operando a 2 GHz se a memória trabalhar a 100 MHz. Para amenizar esta disparidade encontrada em computadores comuns, algumas soluções integradas apresentam sistemas de memória mais sofisticados e velozes. Na grande maioria das vezes, o custo agregado para tornar um sistema mais eficiente é compensado pelo desempenho final do mesmo, como um todo. Neste contexto, duas soluções destacam-se pela utilização da tecnologia DDR SDRAM. Os xServer da Apple e os PowerEdge da Dell fazem uso desta tecnologia. Isto quer dizer que estas duas soluções utilizam barramentos de memória de 266 MHz ao invés de um barramento de 133 MHz, que é o barramento utilizado pelas demais soluções e pela maioria dos atuais computadores de prateleira disponíveis no mercado.

1.4.3. Soluções não integradas

A melhor maneira para se garantir um gasto inicial mínimo é construir o seu próprio aglomerado. Ao selecionar no mercado peças comuns para compor um aglome-

rado, pode-se buscar aquelas que apresentem o menor preço possível para o desempenho esperado. Em geral somente optam por esta solução as organizações que possuem pessoal especializado capaz de garantir o suporte necessário para a montagem e o funcionamento de complexos sistemas computacionais como os deste tipo.

A seleção do software adequado não é menos problemática que a seleção do hardware. Cabe também ao usuário a tarefa de encontrar os melhores componentes, encontrar os melhores e mais adaptado software para a instalação, manutenção e administração do aglomerado. É comum que devido a isto surjam diversos problemas. Por exemplo, é possível que não haja suporte em software para os dispositivos escolhidos, ou que o suporte exista em um sistema operacional diferente do que será instalado. Este problema é particularmente comum no caso de aglomerados não integrados, que são geralmente feitos com computadores pessoais executando o sistema operacional Linux enquanto a maioria das interfaces vêm com pilotos [25] para o sistema Windows.

Um problema mais sutil relacionado ao software dos aglomerados integrados pelo usuário é o baixo desempenho derivado da sua adaptação. Isto ocorre porque nem sempre os programas que controlam algum dispositivo são feitos pelos seus fabricantes. A comunidade de desenvolvedores de código livre, baseado na descrição do funcionamento do dispositivo escreve estes programas. Em geral estes desenvolvedores não têm o mesmo conhecimento a nível interno dos dispositivos e não sabem como obter o seu melhor desempenho. Pior ainda, sistemas com diversos componentes diferentes, feitos por fabricantes diferentes, controlados por programas feitos por organizações diferentes, raramente têm condições de extrair mais desempenho que no caso das soluções integradas. Um fabricante integrador pode fazer ajustes finos permeando todas as camadas de sofware de forma a obter as melhores condições de funcionamento de um dado conjunto de dispositivos.

Também derivado do fato que um aglomerado seja montado com peças compradas no mercado, um inconveniente que ocorre com freqüência é a falta de homogeneidade. O mercado de peças avulsas evolui muito mais rapidamente que as linhas de produtos de fabricantes tradicionais. Devido a isso, peças de reposição nem sempre são as mesmas usadas na construção inicial. Da mesma forma, expansões raramente fazem uso de peças idênticas. Termina-se tendo um aglomerado montado com muitas peças diferentes e por vezes incompatíveis entre si. Isso obriga o administrador do sistema a gerar versões particulares de sistema para as máquinas diferentes, o que se transforma em um problema grave quando o aglomerado for muito grande.

Escolha dos componentes

A escolha do hardware é um dos fatores determinantes no desempenho final de um aglomerado. Devem ser observados, por exemplo, descrições detalhadas dos componentes por parte dos fabricantes, eficiência do componente, uma avaliação dos similares de mercado e a compatibilidade destes com os demais componentes que serão adquiridos. Como pode ser observado, esta avaliação é um processo bastante complexo, o que pode ser considerado como um custo indireto e que geralmente não é contabilizado.

Arquiteturas tradicionais A possibilidade de aquisição de computadores de uso pessoal torna a solução não integrada bastante atraente quanto ao aspecto custo. A evolução

[25] *drivers*

das aplicações executadas por computadores pessoais (automação de escritório, multimídia, jogos, etc.) forçou os fabricantes de procesasdores a elevar drasticamente o seu desempenho. Unitariamente, um processador de computador pessoal tem um desempenho bastante similar ao dos procesaodores dos grandes computadores paralelos.

Arquiteturas tradicionais de computadores pessoais incluem processadores como os da família Pentium da Intel e Atlon da AMD ou PowerPC da Motorola e IBM. Estas aquiteturas vem, de um modo geral, com barramento principal e de memória padrões de 133 MHz, interfaces e barramentos PCI de 66 MHz e memória cache igual ou inferior a 512 KBytes. A preocupação principal destas arquiteturas não é o desempenho e sim o custo, já que o objetivo é atender principalmente as necessidades de usuários domésticos. O custo de computadores pessoais deste tipo é bastante inferior se comparado as unidades de processamento de uma solução integrada. Evidentemente que a eficiência e o desempenho normalmente acompanham esta mesma relação de proporções.

Entretanto, a maioria dos usuários que pretende montar um aglomerado de computadores utilizando arquiteturas tradicionais o fará escolhendo os melhores componentes de forma a atingir um desempenho um pouco superior ao daquele dos computadores de uso pessoal. Neste caso, a escolha passa pela opção da placas-mãe que ofereça os melhores recursos e o melhor barramento, de quantidades de memória maiores, de processadores com o maior desempenho, de interfaces de componentes que apresentam os melhores vantagens.

Arquiteturas multiprocessadas O uso de computadores multiprocessados em aglomerados permite fazer uso de paralelismo com memória compartilhada, necessária para dar alto desempenho em algoritmos com muito compartilhamento de dados entre as tarefas paralelas. A técnica de construção de computadores SMP já é tecnologia dominada pelos fabricantes de peças de computadores pessoais produzidos em laga escala (pelo menos com até dois processadores na mesma placa). É possivel encontrar no mercado placas multiprocessadas de preço bastante baixo. Da mesma forma que os fabricantes de peças já produzem pequenos SMs em larga escala, os sistemas operacionais mais populares são capazes de controlar estes computadores.

Apesar dos multiprocessadores serem uma solução atrativa e aparentemente excelente, existem alguns aspectos que precisam ser levados em consideração. O mais importante é certamente a latência da memória. Em sistemas multiprocessados baratos a latência de memória é consideravelmente alta. Se isto já é um problema em computadores monoprocessados, torna-se um problema severo em sistemas SMPs. Dependendo da aplicação, o seu desempenho pode ser insignificantemente superior a execução desta mesma aplicação em uma máquina monoprocessada. Este é o caso de aplicações que necessitam de uma diversidade muito grande de dados da memória, onde a memória cache não é suficiente para manter estes dados, o que transforma o barramento da memória e a memória no gargalo de uma máquina multiprocessada.

As arquiteturas multiprocessadas comumentemente encontradas para comercialização são o Pentium III e o Xeon da Intel e o Athlon MP da AMD. O número máximo de processadores interligáveis vária de acordo com o modelo. O fator determinante normalmente é a velocidade do barramento de memória, a arquitetura de memória cache e a complexidade do controlador de barramento e de acesso a memória e a dispositivos. Quando é aumentado o número de processadores um outro fator que passa a ser determinante é o custo. Devido a grande complexidade e quantidade de circuitos e barramentos necessários quando do aumento do número de processadores o preço final acaba sendo

elevado. Por esta razão os modelos mais baratos (Pentium III e Athlon MP) somente podem ser montados dois a dois.

Redes rápidas Um dos principais fatores que levaram a rápida e crescente construção e utilização de algomerados de computadores é a maior largura de banda e o menor tempo de latência disponibilizadas pelas novas tecnologias de rede. Neste contexto, emergiram novos padrões de rede como cLAN, SCI, e a Myrinet. Infelizmente somente a última vem mantendo uma presença forte no mercado, basicamente para a construção de aglomerados. Mas acima de todos os padrões novos está um bem mais antigo, ainda que por vezes com mais baixo desempenho: a Ethernet e suas versões modernas com velocidades de até 10 Gbps.

Para todas estas tecnologias existem adaptadores de rede que são conectáveis ao barramento padrão de entrada e saída de um computador pessoal. As redes Myrinet, Gigabit e cLAN são estritamente ponto-a-ponto e necessitam de um comutador para conectar momentanemente cada par de adaptadores envolvidos em uma comunicação. Já uma rede SCI pode ser formada por anéis de múltiplos nós, somente sendo necessário um comutador para montar redes grandes.

A **Gigabit Ethernet** é uma estensão dos bem sucedidos padrões Ethernet IEEE 802.3 de 10Mbps e 100Mpbs. Esta tecnologia é definida no padrão IEEE 802.3 e é manufaturada por inúmeras companhias. Um dos fatores que torna esta tecnologia bastante atraente é que suas tendências de preços acompanham os da tecnologia Fast Ethernet, provendo conexões eficientes em custo para taxas de transmissão vizinhas ao gigabit por segundo. Existem hoje alguns fabricantes que produzem redes Ethernet de 10 gigabits por segundo, mas a um preço centenas de vezes maior que a Gigabit Ethernet comum.

Uma das principais vantagens em favor da utilização da tecnologia Gigabit Ethernet é o aproveitamento da infraestrutura de redes Ethernet, já disponível em institutos de pesquisa e organizações. A tecnologia Gigabit Ethernet está cada vez mais se tornado a melhor opção custo-benefício do mercado. Esta tecnologia tende a ser a sucessora da tecnologia Fast Ethernet, atualmente utilizada na maioria das instalações de rede do mundo.

A **cLAN** da Giganet é a primeira implementação nativa da Arquitetura de Interface Virtual[26] (VIA) [SPE 99]. A única funcionalidade da cLAN é prover acesso direto a aplicações, deixando para trás as interfaces do sistema operacional atingindo uma taxa de vazão extremamente alta. O comutador cLAN 5000 da Giganet fornece vazão de 1.25 Gb/s em dois sentidos de forma simultânea[27] e o tempo de latência porta-a-porta é de apenas 800 nanosegundos. Graças ao processador de comunicação incorporado ao adaptador, esta rede preserva o processador principal para o processamento da aplicação.

A tecnologia **Myrinet** tem suas raízes no projeto de multiprocessamento massivo do Caltech Mosaic e sua rede Atomic [COH 93], do qual ela herdou, entre outras características, a estratégia de roteamento *wormhole*. Tanto a interface de hardware quanto a de software Myrinet, bem como os protocolos, estão publicados e são abertos, o que tem encorajado inúmeros projetos.

Os pacotes Myrinet podem ser de qualquer tamanho, e, sendo assim, pode encapsular outros tipos de pacotes, incluindo pacotes IP, sem a adoção de uma camada. Cada pacote é identificado por tipo, para que a Myrinet, assim como a Ethernet, possa carregar de forma concorrente pacotes de vários tipos ou protocolos. Por isso, Myrinet suporta

---

[26] *virtual interface architecture*

[27] *full-duplex*

diversas interfaces de software. Pacotes especializados, desenvolvidos por usuários Myrinet para aplicações destinadas a aglomerado de computadores, alcançam latências baixas entre processos de usuário Unix menor que 5 microsegundos e taxas de transmissão, em um sentido, que excede um gigabit por segundo durante períodos longos.

O SCI (*Scalable Coherent Interface*) é uma interface especificada no padrão IEEE 1596 de 1992. Este padrão define hardware e protocolos para conectar conjuntamente mais de 64000 nós. A mais remarcável diferença entre redes tradicionais para uma construída utilizando a tecnologia SCI é que a troca de dados entre os nós é realizada com comunicação implícita, ou seja, através de acesso remoto a memória, ao invés de utilizar passagem de mensagem explícita. Implementações atuais da SCI disponibilizam larguras de banda crua de 1 gibabit por segundo e estão disponíveis para barramentos PCI e SBus.

SCI define um espaço de endereçamento global, compartilhado por todos os nós do aglomerado de computadores, de 64 bits, e acessável através de operações de escrita, leitura e movimentação semelhantes a um barramento de memória normal. Esta transações são atômicas, livres de interbloqueios e preservam coerência de memória e de cache pelo aglomerado de computadores. Mais de 64 transações podem estar sendo realizadas em cada nó, o que permite uma melhor utilização da largura de banda disponível. Neste contexto, nós podem exportar segmentos de memória de seu espaço de endereçamento físico local através do mapeamento destes espaços no espaço de endereçamento global SCI. Em suma, SCI especifica um hardware baseado em memória compartilhada distribuída.

A integridade dos dados é mantida em hardware através de somas de verificação[28] gerados e verificados para cada pacote transferido. Caso um pacote ruim tenha sido recebido, ele é simplesmente descartado. Cada transação é confirmada e associada a um determinado prazo. Se por erro de transmissão uma transação não recebe resposta de confirmação, o pacote é retransmitido. Esta combinação de somas de verificação e prazo para confirmação garante transações confiáveis. Não obstante, protocolos SCI não garanter entregas ordenadas.

1.5. Instalação Física

Invariavelmente a montagem de um aglomerado implicará na instalação de diversos equipamentos computacionais. A implantação de um sistema deste tipo apresenta problemas tradicionais de instalações físicas de centros de processamento de dados e de salas de servidores de redes. O texto que segue tem por objetivo apresentar os principais problemas especificamente relacionados a instalação física de aglomerados de computadores e possíveis soluções. O objetivo principal é abordar aspectos como disposição física de aglomerados, energia, topologia e arquitetura de rede, cabos, gabinetes e dispositivos de instalação a quente.

1.5.1. Espaço físico e ventilação

O espaço físico necessário para a alocação de um aglomerado de computadores é talvez um dos primeiros problemas enfrentados por usuários que desejam adquirir ou realocar um aglomerado de computadores. Este espaço físico necessário depende de vários elementos. Entre eles, o número de nós do aglomerado, o tamanho físico de cada nó, a

[28] *checksums*

disposição dos nós (em prateleiras com altura e largura personalizadas, espalhados pelo chão, etc.). O planejamento de alocação dos nós de um aglomerado é um dos pontos de partida a partir da existência de uma aglomerado ou do planejamento de compra. O aspecto relacionado ao espaço físico utilizado por um aglomerado de computadores por ser, inclusive, um fator decisivo na hora de escolha da arquitetura e dos componentes do conjunto de máquinas que formarão o aglomerado.

A decisão sobre a disposição dos equipamentos é uma importante e trabalhosa tarefa, porque é necessário não somente se preocupar com a estrutura e disposição, mas sim, também com problemas como a ventilação. Deve-se sobretudo planejar a alocação dos nós de processamento de forma a diminuir problemas de ventilação. Um aglomerado de computadores destinados a processamento de alto desempenho normalmente necessita de um sistema de refrigeração capaz de evitar o superaquecimento dos nós de processamento. Como se sabe, a temperatura dos processadores é maior quanto maior a carga de processamento e a quantidade de horas consecutivas de processamento. Além disso, atualmente processadores com altas freqüências de relógio necessitam de sistemas extras de ventilação.

Muitas vezes, a simples utilização de condicionadores de ar não é o suficiente para garantir uma adequada refrigeração do aglomerado. Uma grande quantidade de nós em uma única sala pode necessitar de um sistema de refrigeração especial. Neste caso, a aquisição de armários com ventilação forçada e com suportes para a disposição e estruturação do aglomerado de computadores, garantindo a passagem de ar refrigerado por todos os lados nos elementos, é fundamental para se obter boa refrigeração. É comum até mesmo a construção de pisos em forma de grades perfuradas, com a circulação forçada de ar de baixo para cima.

1.5.2. Planejamento da energia

Um pequeno aglomerado de computadores pode ser instalado e posto em funcionamento sem grandes planejamentos de alimentação em um circuito único, prevendo algo como um ou dois Ampères por computador. Um aglomerado maior necessita um adequado e apropriado circuíto de alimentação. Em aglomerados estruturados em forma de pilhas, utilizando prateleiras, a instalação de réguas de energia capazes de suportar o consumo máximo atingido pelos nós da pilha é uma das soluções mais eficazes e práticas. Cuidado deve ser tomado para não ligar réguas em cascata, pois seus cabos de alimentação não têm capacidade de transportar energia para duas ou mais réguas.

O primeiro passo para o planejamento da energia é a determinar as necessidades de energia através da catalogação de todos os dispositivos e seu consumo de energia. Deve-se sempre calcular o consumo de energia do sistema através do pico de consumo de cada dispositivo, ao invés da média de consumo. Durante o inventário é muito importante incluir qualquer equipamento ligado, não somente os nós de processameto. Isto pode incluir placas especiais dentro dos nós de processamento, monitores, luzes, dispositivos de fita e mesmo as barras e réguas de energia.

Todas as tomadas e réguas de alimentação devem possuir o terceiro pino de aterramento devidamente instalado para evitar problemas como queima de dispositivos por causa de eletricidade estática acumulada. É ainda relevante um planejamento que estime também possíveis expansões ou troca de equipamentos pois, o que pode parecer apropriado no momento da instalação poderá não ser o suficiente no futuro.

Uma solução especialmente atrativa em termos de baixo consumo de energia, alto

desempenho e baixa manuntenção é a arquitetura apresentada pela RLX Tecnologies. Este arquitetura consiste na utilização de nós de processamento que possuem processadores Crusoe Transmeta. Estes processadores perfazem praticamente o mesmo desempenho de processadores Pentium III de mesma frequência, mas com um consumo de energia muito menor.

1.5.3. Gabinetes

A maioria dos fabricantes atuais a vende computadores para aglomerados em gabinetes compactos e otimizados em tamanho. O objetivo é fornecer a maior quantidade de nós e desempenho no menor espaço físico possível. Os gabinetes desenvolvidos por estas empresas são geralmente feitos para ser instalados em armários de 19 polegadas e têm alturas padronizadas em múltiplos de 44 milímetros, uma unidade chamada de "U". A figura 1.6 mostra a aparência típica de gabinetes tradicionais e de gabinetes para armários de 19 polegadas. Estes últimos muitas vezes são chamados de gabinetes industriais. Quando são muito finos, com altura de 1U por exemplo, é comum chamá-los de lâminas.[29]

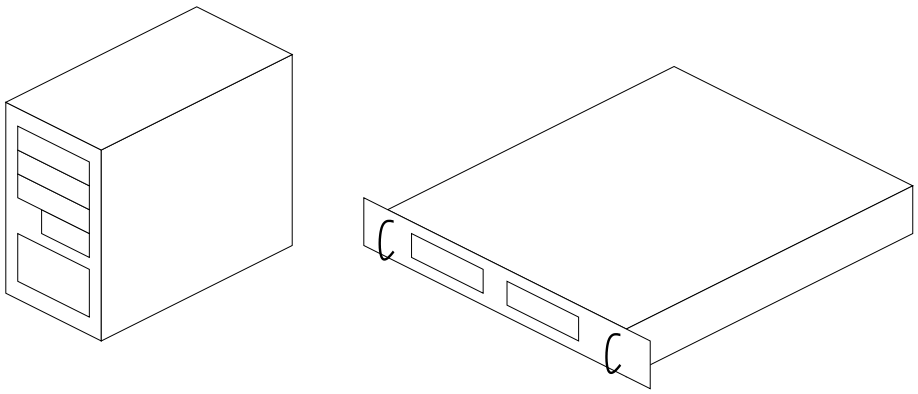

Figura 1.6: Gabintes típicos.

Outro ponto importante a ser averiguado é a montagem, a facilidade de manuseio e a disposição dos conectores do gabinete. Um gabinete que não faz uso de parafusos pode ser uma opção prática para os administradores do sistema. Os conectores traseiros e dianteiros devem ser observados de forma a garantir que seja possível instalar todos os cabos necessários. Estes detalhes geralmente não são importantes quando se instala um computador isolado, mas pode complicar bastante ao instalar algumas dezenas em uma pilha única.

Por fim, se os elementos processadores forem montados pelo usuário (solução não integrada), antes de escolher o tamanho e a estrutura dos gabinetes é necessário realizar uma análise dos componentes internos de cada nó. Um gabinete inadequado pode causar problemas de instalação dos componentes internos, que às vezes apresentam disposições inusitadas. Ao dispor os gabinetes em prateleiras, é necessário verificar o espaço reservado entre os níveis disponíveis para a acomodação dos nós do aglomerado. Há que se deixar espaço entre os andares da prateleira e o topo dos gabinetes de forma a garantir a circulação de ar.

[29] *blades*

1.5.4. Comutadores

Mesmo em aglomerados de pequeno porte não é incomum a utilização de duas redes de comunicação entre os nós de processamento. Uma rede normal, como a Fast Ehternet por exemplo, e uma rede de alta velocidade tipo Gigabit Ethernet, a Myrinet ou a SCI. A primeira é normalmente destinada a trafegar pacotes de serviços internet (transferência de arquivos, terminal remoto, etc.). A segunda é reservada às aplicações que necessitam de alto desempenho e comunicação eficiente entre processos executando em diferentes nós.

Dependendo do tamanho do aglomerado e da tecnologia de rede usada, pode ser necessário montar uma rede estruturada, aumentando ainda mais o número ou o tamanho dos comutadores. A idéia geral é que a montagem destes comutadores deve ser feita de forma que cada intervenção física seja prática, sem a necessidade de movimentar outros equipamentos. Geralmente são usadas as mesmas técnicas de redes tradicionais, com painéis de conexão[30] próximos aos comutadores.

Além dos comudadores de rede, comutadores de teclado, vídeo e mouse (comutadores KVM) são comuns. A seqüência de inicialização de muitos computadores pessoais baratos não se completa se este não possuir um teclado conectado. Um comutador KVM permite que o administrador opere qualquer nó através de um único monitor apenas escolhendo o nó por meio de botões ou comandos de teclado. Um comutador KVM é interligado a cada um dos nós por três cabos comuns de vídeo, teclado e mouse. Em um aglomerado com dezenas de nós isso pode gerar um problema de cabeamento, não só pela quantidade de cabos mas também por restrições de distância.

A fim de resolver o problema de cabeamento dos comutadores KVM surgiram tecnologias como o C2T da IBM e o AutoView da Avocent. A solução C2T da IBM foi desenvolvida para a série de arquiteturas xServer. O C2T é baseado em um cascateamento, usando um único cabo especial para conectar simultaneamente os sinais o vídeo, o tecla e o mouse. Cada nó é ligado somente ao seguinte e um terminal console especial é ligado em uma das pontas da cadeia. Como os servidores tem o formato de uma lâmina e todos os conectores C2T estão na mesma posição, isso elimina muitos problemas de cabeamento. Para grandes aglomerados de computadores, onde o administrador deseja ter acesso ao terminal de cada nó, esta pode ser uma boa solução. Já a solução da Avocent é baseada em chaveamento sobre IP para acesso local de qualquer parte da rede Internet.

1.5.5. Cabeamento

A organização do cabeamento em um aglomerado de computadores é um ponto fundamental para garantir o fácial acesso, manutneção, remoção e inclusão de nós por parte dos administradores do sistema. Um aglomerado com duas redes de comunicação e um comutador KVM normal gera uma grande quantidade de cabos. A disposição estruturada desta grande quantidade de cabos é algo essencial.

Certos aglomerados, pelo seu grande tamanho, podem precisar fazer uso de piso falso para transportar os cabos de um armário a outro. Existem também prateleiras e armários com suporte lateral para o escoamento dos cabos, o que pode vir a ser útil. Não menos importante é a identificação dos cabos, em ambas as extremidades, para o caso de ser ter diversos deles sendo desconectados ao mesmo tempo.

---

[30] *patch panels*

1.5.6. Dispositivos de instalação a quente

Dispositivos que podem ser retirados e colocados enquanto o sistema está em funcionamento são úteis em sistemas que não podem parar. Estes sistemas, normalmente denominados de sistemas de alta disponibilidade, podem possuir vários componentes que podem ser trocados ou instalados a quente.

Os discos rígidos e as placas interfaces (tipicamente de rede) são os dispositivos mais comuns que podem ser instalados a quente. Diversas soluções integradas para aglomerados de computadores apresentam encaixes para discos rígidos e interfaces que podem ser retirados ou inseridos sem desligar o sistema. Alguns possuem também fonte de alimentação redundante. Neste último caso, quando uma das fontes falhar ela pode ser trocada com a máquina em funcionamento.

Além de discos rígidos e interfaces, talvez o componente mais interessante no caso de aglomerados de processamento de alto desempenho seja o processador. Esta característica somente é útil em máquinas multiprocessadas muito grandes e que não podem parar. O objetivo principal destas soluções é disponibilizar alto desempenho (a qualquer custo) para aglomerados que rodam aplicações críticas onde a falha de um nó pode comprometer a execução total de uma aplicação.

1.6. Administração de Aglomerados

Com o advento e a expansão dos aglomerados de computadores, tanto em número quanto em tamanho, surge cada vez mais a necessidade de ferramentas para a instalação, gerenciamento e manutenção destas máquinas paralelas. Estas ferramentas devem prover, além de uma solução para os principais problemas, soluções eficientes e escaláveis.

Esta seção trata sobre os problemas que surgem no processo de administração de aglomerados de computadores, bem como possíveis soluções. Inicialmente é apresentado um dos problemas mais comuns em aglomerados de computadores: a instalação e configuração de sistemas operacionais. Em seguida o texto mostra problemas relacionados a atualização, configuração e propagação de programas em um aglomerado. Por fim, são apresentadas algumas ferramentas que tem por objetivo monitorar a operação, com a finalidade de detectar e sanar problemas.

1.6.1. Instalação do Sistema Operacional

Pode se considerar que o passo inicial da instalação do sistema operacional de um aglomerado é a escolha do próprio sistema operacional. Evidentemente que a escolha do sistema vai depender da arquitetura do aglomerado, pois não existem versões de todos os sistemas operacionais para qualquer tipo de computador. Em geral é a aplicação que será executada do aglomerado é que determinará não só o sistema operacional mas também a arquitetura do mesmo. Não obstante, a inexistência de recursos de instalação e administração adequados para aglomerados pode transformar a tarefa do administrador em um grande pesadelo. Estes recursos devem obrigatoriamente estar disponíveis para o sistema operacional escolhido antes da escolha do mesmo.

A grande maioria dos aglomerados existentes tem como base alguma versão de um sistema operacional tipo Unix. Isto se deve a uma série de motivos. Primeiramente, excetuando-se os poucos sistemas operacionais proprietários que existem e que executam

em algumas máquinas específicas (e proprietárias), duas opções básicas existem: sistemas tipo Unix ou alguma versão de Windows da Microsoft. Basicamente existem versões de Windows para computadores construídos com processadores que seguem a arquitetura IA-32 da Intel e versões de Unix (proprietárias e livres) para praticamente todo tipo de computador.

Os computadores usados para montar aglomerados são geralmente produzidos em larga escala, e por esta razão, seguem a arquitetura IA-32 da Intel. Poder-se-ia pensar que por isso o Windows seria um forte candidato para controlar aglomerados. Ocorre que, apesar de serem produzidos em larga escala, geralmente os computadores de um aglomerado não têm um usuário interagindo diretamente com ele através de uma console gráfica. Ao contrário, todo o processamento feito em um aglomerado é causado por uma requisição de execução que chega pela rede. Isso equivale a dizer que os computadores de um aglomerado são servidores, mesmo quando eles são dedicados ao processamento paralelo científico. No mercado de servidores, o sistema operacional dominante não é o Windows, mas o Unix e todas as suas variantes. Além disso, a maioria dos grandes aglomerados está instalada em instituições de ensino e pesquisa. Estas instituições preferem (alguns diriam que por questões filosóficas) os sistemas operacionais de código aberto. Por estes e por outros motivos, a grande maioria do software escrito para aministrar aglomerados foi escrito para funcionar com sistemas tipo Unix e não funciona do Windows.

O primeiro problema prático na implantação de um aglomerado é a instalação do sistema operacional em todos os seus elementos. O procedimento de instalação do sistema é complexo porque pressupõe que nenhum programa esteja executando previamente no computador. Isso obviamente impede o uso de atuação remota, via rede. À medida que o número de nós de um aglomerado cresce este problema cresce em proporções semelhantes. Uma das técnicas mais conhecidas para solucionar o problema de instalação de um grande número de computadores é a utilização de carga[31] remota de sistema operacional, explicada a seguir.

1.6.1.1. Carga remota de sistema operacional

A carga remota é uma técnica simples e eficiente para admiinstrar a instalação de um grande número de computadores idênticos interligados por uma rede. Para fazer uso dela um computador precisa ter uma placa de rede dotada de um programa de carga de sistema pela rede. Este programa é lançado ao inicializar o computador e, usando protocolos padronizados, faz a requisição pela rede de um novo programa para executar. Este novo programa geralmente é o núcleo de um sistema operacional ou eventualmente um programa que, após gravar um novo sistema operacional no disco rígido local, carrega o mesmo de forma tradicional.

A estrutura necessária para implementar a carga remota consiste em uma ou mais máquinas conectadas à mesma rede que executem servidores dos protocolos de carga remota de sistema operacional[32] (BOOTP) ou de configuração dinâmica[33] (DHCP) e de transferência trivial de arquivos [34] (TFTP). As máquinas que recebem o sistema pela rede (máquinas clientes da carga remota) precisam preferencialmente ter interfaces rede que

---

[31] *boot*
[32] *bootstrap protocol*
[33] *dynamic host configuration protocol*
[34] *trivial file transfer protocol*

possuam ambiente de execução pré-carga[35] (PXE). Hoje, praticamente todas as interfaces de rede disponíveis no mercado possuem esta tecnologia embutida.

O processo consiste em inicialmente instalar o sistema em uma das máquinas de um conjunto de máquinas homegêneas. A partir desta instalação é gerada uma imagem de sistema operacional, que é instalada no servidor. A partir daí as demais máquinas serão inicializadas através da rede, ou seja, carregarão o sistema da máquina servidora automaticamente. Este processo economiza tempo e esforço por parte dos administratores para a disponibilização e manutenção dos sistemas das rede. Além disso, este sistema de carga remota possibilita a manutenção de diversos sistemas, sendo que, a cada inicialização poderá ser escolhido um novo sistema para cada máquina.

Algumas ferramentas foram desenvolvidas para operar de forma associada à carga remota do sistema operacional. O Ka-admin, por exemplo, é um sistema que acelera a carga remota para aglomerados de computadores compostos por centenas de nós, fazendo a carga do sistema de forma múltipla e hierárquica. A distribuição Rocks do sistema Linux apresenta uma solução automatizada para o gerenciamento e manutenção de diferentes versões do sistema operacional. A empresa Sixfold propõe um servidor de rede pré-instalado com um sistema dedicado de carga remota simultânea. O sistema ClusterWorx da Linux NetworX possui um gerenciador de imagens para o controle e manipulação da instalação de sistemas operacionais no aglomerado.

1.6.2. Atualização e configuração de programas

Não menos complexo que instalar e configurar o sistema operacional é fazer o mesmo com os pacotes de programas necessários para a operação do aglomerado. Essa tarefa pode ser extremamente difícil quando se está trabalhando com aglomerados de computadores com algumas dezenas ou centenas de nós. Neste ponto começam a aparecer problemas de escalabilidade e eficiência.

Uma saída muitas vezes utilizada é o compartilhamento pela rede de um sistema de arquivos contendo os pacotes. Esta solução tem baixa escalabilidade por trazer congestionamento à rede quando do acesso dos sistemas de arquivos compartilhados por um número muito grande de clientes. Neste caso, caches de sistemas de arquivos remotos [DAH 94] podem ser bastante úteis, já que os sistemas de arquivos que contém estes pacotes sofrem poucas atualizações.

Sistemas como xCAT da IBM e SCMS da Universidade Kasetsart possuem funcionalidades para realizar o gerenciamento, instalação e atualização de pequenas unidades de software. O SCMS disponibiliza recursos como os comandos paralelos Unix. Através da utilização deste sistema de comandos paralelos é possível ativar uma tarefa em um conjunto de nós, ou em todos os nós, do aglomerado simultaneamente. O sistema de gerenciamento xCAT é baseado em conjuntos de *shell scripts* que automatizam alguns processos de gerenciamento e instalação de aglomerados de computadores. Tarefas como a de configuração e reinicialização de processos que prestam serviços nos nós de um aglomerado podem ser facilmente feitas através da utilização dos recursos disponibilizados por sistemas tipo o xCAT ou o SHOC [TAN 2002].

A grande vantagem de sistemas deste gênero é a facilidade em o administrador adaptar ou acrescentar funcionalidades que atendam as necessidades específicas com pouco esforço. Além disso, estes sistemas podem ser facilmente portados e inclusive

[35] *pre-boot execution environment*

rodar simultaneamente sobre plataformas distintas devido a difusão e vasta utilização de *shell scripts* em sistemas operacionais tipo Unix.

1.6.3. Monitoramento de processos e subsistemas

A tarefa de monitoramento talvez seja uma das mais importantes no processo de gerenciamento do aglomerado. Ferramentas eficientes de monitoramento auxiliam, por exemplo, na tomada de decisão para a migração de algum processo que esteja sobrecarregando algum nó de processamento, ou ainda a migração de algum recurso que esteja sendo pouco utilizado em sua localização atual. O monitoramento pode fornecer dados úteis para um eventual balanceamento de cargas para uma melhor utilização dos componentes e recursos de processamento disponíveis.

Um bom monitoramento também disponibiliza informações referentes a utilização dos recursos disponíveis por cada usuário. Neste ponto, pode-se identificar os usuários que mais utilizam o aglomerado e os que necessitam mais de determinados recursos. Com isso, é possível alocar ou reservar recursos específicos a determinados usuários de forma a fornecer o máximo de desempenho aos usuários durante determinados períodos.

Devido a grande utilidade das informações disponibilizadas por ferramentas de monitoramento e a crescente disseminação dos aglomerados de computadores, surgiram algumas ferramentas que tem por objetivo apresentar dados referentes a utilização dos nós que compõem o aglomerado de forma mais prática, eficiente e realística. Dentre as ferramentas, específicas para este fim, podemos citar: RVision [FER 2002], PARMON [BUY 2000], BWatch e o Big Brother. Além disso, sistemas de gerenciamento como o SCMS, o ClusterWorX e o Ka-admin possuem seus próprios recursos e ferramentas para o monitoramento de aglomerados.

1.6.4. Diagnóstico, prevenção e conserto de falhas

Após a instalação, configuração e operação do aglomerado de computadores o administrador não fica menos livre da necessidade de boas ferramentas para o diagnóstico e conserto de possíveis falhas e problemas. Entre os exemplos mais simples são a falhas em bibliotecas de comunicação, falhas de um sistema de arquivo e uma falha geral do sistema. Para identificar estas falhas são necessárias ferramentas práticas e eficientes.

Dentro deste contexto uma solução merece destaque. A solução integrada ICE Box da Linux NetworX é disponibilizada como uma solução de gerenciamento de hardware pela Linux NetworX. Esta solução trabalha em conjunto com o sistema de gerenciamento do ClusterWorX, pertencente a mesma companhia. O objetivo principal desta solução é prover os recursos e ferramentas necessárias para realizar um monitoramento e controle do hardware dos nós de um aglomerado de forma prática e eficiente.

Além destes problemas, normalmente de origem no software, existem ainda os problemas de hardware. A detecção e diagnóstico de falhas de hardware normalmente necessitam, além de sistemas específicos em software, recursos auxiliares de monitoramento em hardware. Normalmente, no caso de falhas de componentes de hardware os administradores não têm muito o que fazer a não ser ir até o nó e fazer uma verificação manual. Para resolver problemas deste tipo a IBM propõe uma interface de supervisor remoto[36] que prove, além de gerenciamento avançado, recursos para realizar testes de

[36] *remote supervisor adapter*

hardware nos nós do aglomerado. Para atingir este objetivo a interface possuí até um conector de energia próprio. Segundo a IBM, esta interface permite o controle total sobre o hardware e o software dos servidores da linha xSeries.

1.7. Conclusão

Este texto apresentou algumas tendências atuais das arquiteturas paralelas. Em especial, foi mostrado que os aglomerados de multiprocessadores simétricos produzidos em larga escala têm um papel dominante neste terreno. Algumas possibilidades de desenvolvimento se encontram na área de redes, com a crescente interligação de mais e mais processadores, com diversas hierarquias de interconexão. Outro filão de melhoria é na área de microprocessadores, com a maior integração de elementos processadores em um só circuito integrado.

Quanto à construção de aglomerados, foi visto as soluções integradas por algum fabricante apresentam melhor desempenho e melhor suporte mas têm maior custo. Os maiores desempenhos ainda estão reservados aos aglomerados construídos com tecnologia proprietária. Aglomerados com tecnologia totalmente aberta já podem ser adquiridos da maioria dos grandes fabricantes. O mercado de peçás produzidas em larga escala já apresenta soluçõés muito baratas e com alto desempenho. A tarefa de montar o seu próprio aglomerado pode parecer barata mas necessita de uma equipe com conhecimento no assunto e com tempo para resolver os problemas de integração.

A maioria dos problemas de instalação e administração de aglomerados vem do fato de as técnicas usadas não possuirem muita escalabilidade. Existem entratanto algumas metodologias elegantes e simples para automatizar o usimplificar tarefas como conectar cabos no painel traseiro de um computador ou instalar nele um sistema operacional. Entretanto, por serem muito recentes, a maioria dos sistemas de administração de aglomerados se situa no domínio acadêmico e soluções industriais ainda devem ser desenvolvidas.

1.8. Bibliografia

[ARC 86] ARCHIBALD, J.; BAER, J.-L. Cache coherence protocols: evaluation using a multiprocessor simulation model. **ACM Transactions on Computer Systems**, v.4, n.4, p.273–298, nov 1986.

[BUR 97] BURGER, D.; GOODMAN, J. R. Billion transistor architectures. **IEEE Computer**, v.30, n.9, p.46–48, Sept. 1997.

[BUR 92] BURKHARDT, H. et al. **Overview of the ksr1 computer system**. [S.l.]: Kendall Square Research, 1992. (KSR-TR-9202001).

[BUY 99] BUYYA, R. **High performance cluster computing**: systems and architectures. [S.l.]: Prentice Hall PTR, 1999. v.1.

[BUY 2000] BUYYA, R. PARMON: a portable and scalable monitoring system for clusters. **Software: Practice and Experience**, v.30, n.7, p.723–739, June 2000.

[CAL 79] CALAHAN, D.; AMES, W. Vector processors: Models and applications. **IEEE Trans. Circuits and Syst.**, v.CAS-26, p.715–776, 1979.

[COH 93] COHEN, D. et al. ATOMIC: A low-cost, very-high-speed, local communication architecture. In: INTERNATIONAL CONFERENCE ON PARALLEL PROCESSING. VOLUME 1: ARCHITECTURE, 1993., 1993, Syracuse, NY. **Proceedings...** CRC Press, 1993. p.39–46.

[DAH 94] DAHLIN, M. D. et al. **A quantitative analysis scalability for network file systems**. [S.l.]: University of California, Berkeley, 1994. Technical Report. (CSD-94-798).

[DIE 2000] DIEFENDORFF, K. et al. AltiVec extension to PowerPC accelerates media processing. **IEEE Micro**, v.20, n.2, p.85–95, Mar./Apr. 2000.

[EIC 92] EICKEN, T. von et al. **Active messages**: a mechanism for integrated communication and computation. [S.l.]: University of California, Berkeley, 1992. Technical Report. (CSD-92-675).

[FER 2002] FERRETO, T. C.; ROSE, C. A. F. D. Rvision: uma ferramenta aberta e configurável para a monitoração de clusters. In: SEGUNDA ESCOLA REGIONAL DE ALTO DESEMPENHO, 2002. **Anais...** [S.l.: s.n.], 2002. (Anais da Segunda Escola Regional de Alto Desempenho).

[GIL 96] GILLETT, R. B. Memory channel network for PCI. **IEEE Micro**, v.16, n.1, p.12–18, Feb. 1996.

[HAM 85] HAMMOND, L.; NAYFEH, B.; OLUKOTUN, K. Essentials issues in multiprocessor systems. **IEEE Computer**, v.18, n.6, p.70–79, June 1985.

[HAM 97] HAMMOND, L.; NAYFEH, B.; OLUKOTUN, K. A single-chip multiprocessor. **IEEE Computer**, v.30, n.9, p.79–85, Sept. 1997.

[HWA 98] HWANG, K.; XU, Z. **Scalable parallel computing**: technology, architecture, programming. [S.l.]: McGraw-Hill, 1998.

[KIT 90] KITCHENS, T. The U.S. Department of Energy's "grand challenge" program. **The International Journal of Supercomputer Applications**, v.4, n.3, p.3–5, Fall 1990.

[LAN 92] LANDIS, A.; HAGERSTEIN, E.; HARIDI, S. Ddm - a cache-only memory architecture. **IEEE Computer**, 1992.

[MIC 2002] MICROSYSTEMS, S. **Ultrasparc III Cu user's manual**. [S.l.: s.n.], 2002.

[OBE 99] OBERMAN, S.; FAVOR, G.; WEBER, F. AMD 3DNow! technology: architecture and implementations. **IEEE Micro**, v.19, n.2, p.37–48, Mar./ Apr. 1999.

[OMA 97] OMANG, K.; PARADY, B. Scalability of SCI workstation clusters: A preliminary study. In: INTERNATIONAL PARALLEL PROCESSING SYMPOSIUM (IPPS-97), 11., 1997, Los Alamitos. **Proceedings...** IEEE Computer Society Press, 1997. p.750–755.

[PAT 97] PATT, Y. N. et al. One billion transistors, one uniprocessor, one chip. **IEEE Computer**, v.30, n.9, p.51–57, Sept. 1997.

[PAT 95] PATTERSON, D. A.; CULLER, D. E.; ANDERSON, T. E. A case for NOW (networks of workstations)—abstract. In: FOURTEENTH ANNUAL ACM SYMPOSIUM ON PRINCIPLES OF DISTRIBUTED COMPUTING, 1995, Ottawa, Ontario, Canada. **Proceedings...** [S.l.: s.n.], 1995. p.17.

[PAT 98] PATTERSON, D. A.; HENNESY, J. L. **Computer organization and design, the hardware/software interface**. [S.l.]: Morgan Kaufmann, 1998.

[PEL 97] PELEG, A.; WILKIE, S.; WEISER, U. Intel MMX for multimedia PCs. **Communications of the ACM**, v.40, n.1, p.24–38, Jan. 1997.

[SHA 95] SHANLEY, T.; ANDERSON, D. **PCI system architecutre**. 3.ed. New York: Addison-Wesley, 1995.

[SMI 2001] SMITH, B. Cray MTA: multithreading for latency response. **Computer**, v.34, n.4, p.59–65, Apr. 2001.

[SMI 97] SMITH, J. E.; VAJAPEYAM, S. Trace processor: moving to fourth-generation microarchitectures. **IEEE Computer**, v.30, n.9, p.68–74, Sept. 1997.

[SPE 99] SPEIGHT, E.; ABDEL-SHAFI, H.; BENNETT, J. K. Realizing the performance potential of the Virtual Interface Architecture. In: INTERNATIONAL CONFERENCE ON SUPERCOMPUTING, 1999., 1999, Rhodes, Greece. **Proceedings...** [S.l.: s.n.], 1999. p.184–192.

[STE 99] STERLING, T. L. et al. **How to build a Beowulf**: A guide to the implementation and application of PC clusters. Cambridge, MA, USA: MIT Press, 1999. xxi + 239p. (Scientific and Engineering Computation).

[STU 95] STUNKEL, C. B. et al. The SP2 high performance switch. **IBM Systems Journal**, v.34, n.2, p.185–204, 1995.

[TAN 2002] TAN, C.; TAN, C.; WONG, W. Shell over a cluster (shoc): towards achieving single system image via the shell. In: IEEE INTERNATIONAL CONFERENCE ON CLUSTER COMPUTING (CLUSTER 2002), 2002, Chicago, IL. **Anais...** [S.l.: s.n.], 2002. p.28–36.

[TEN 2002] TENDLER, J. et al. Power4 system microarchitecture. **IBM Journal of Research and Development**, v.46, n.1, p.5–25, Jan. 2002.

[TRE 2000] TREMBLAY, M. et al. The MAJC architecture: A synthesis of parallelism and scalability. **IEEE Micro**, v.20, n.6, p.12–25, Nov./Dec. 2000.

[TUL 95] TULLSEN, D. M.; EGGERS, S. J.; LEVY, H. M. Simultaneous multithreading: maximizing on-chip parallelism. In: ANNUAL INTERNATIONAL SYMPOSIUM ON COMPUTER ARCHITECTURE, 22., 1995, Santa Margherita Ligure, Italy. **Proceedings...** [S.l.: s.n.], 1995. p.392–403.

[ZIM 91] ZIMA, H.; CHAPMAN, B. **Supercompilers for parallel and vector computers**. [S.l.]: Addison-Wesley, 1991.