Arquiteturas paralelas versáteis e de baixo custo para a pesquisa e o ensino na área de processamento paralelo e distribuído

Giancarlo C. Mai*

César A. F. De Rose†

Pontifícia Universidade Católica do Rio Grande do Sul
Pós-Graduação em Informática
Caixa Postal 1429
CEP 90619-900 – Porto Alegre, RS, Brasil
Tel.: (051) 320-3611 Fax: (051) 320-3621

Resumo

Arquiteturas baseadas em agregados vem se destacando nos últimos anos como uma alterntiva para a construção de máquinas paralelas versáteis e de baixo custo. Esta versatilidade permite seu uso tanto para o ensino como para a pesquisa na área de processamento paralelo e distribuído. Este trabalho apresenta algumas possibilidades de interconexão encontradas atualmente no mercado para a confecção de máquinas paralelas baseadas em agregados e propõe diferentes configurações destacando seu custo e possíveis áreas de aplicação.

Palavras-chave: Arquiteturas de Computadores, Processamento Paralelo e Distribuído, Máquinas Paralelas Baseadas em Agregados

Abstract

Cluster based architectures are standing out in the last years as an alternative for the construction of versatile, low cost parallel machines. This versatility permits their use as much as a teaching as a research environment in the field of parallel and distributed processing. This work describes some of the possibilities found today on the market for the construction of cluster based parallel machines and proposes different configurations based on cost and application areas.

Keywords: Computer Architectures, Parallel and Distributed Processing, Cluster Based Parallel Machines

*Professor da Faculdade de Informática da PUCRS; Mestre em Ciência da Computação (CPGCC/UFRGS, 1998); Áreas de Interesse: Arquitetura de Computadores, Segurança em Sistemas de Computação, Processamento Paralelo e Distribuído; E-mail: mai@inf.pucrs.br

†Professor da Faculdade de Informática da PUCRS; Doutor em Informática (Universidade Fridericiana de Karlsruhe, 1998); Áreas de Interesse: Arquitetura de Computadores, Sistemas Operacionais, Processamento Paralelo e Distribuído; E-mail: derose@inf.pucrs.br

1 Introdução

Sistemas de processamento paralelo vem tornando-se mais populares em função da demanda sempre crescente por poder computacional . Infelizmente, os sistemas que oferecem a capacidade de processamento para satisfazer à demanda ou tem custo muito elevado, ou são difíceis de programar, ou ambos.

Neste contexto tem se investido muito nos últimos anos na pesquisa de máquinas paralelas baseadas em agregados, que aliam as vantagens das outras três classes de máquinas, resultando em uma máquina de custo mais baixo e mais flexível podendo ser configurada dependendo da área de atuação.

Neste artigo serão apresentadas as principais características das máquinas paralelas baseadas em agregadas juntamente com um panorama das principais tecnologias de interconexão para esta classe de máquinas disponíveis hoje no mercado.

Tendo estas tecnologias como base, são propostas diferentes arquiteturas para a confecção de um laboratório de processamento de alto desempenho para o ambiente acadêmico, onde pode ser desenvolvido tanto o ensino como a pesquisa na área de processamento paralelo e distribuído. Para cada arquitetura apresentada são destacados o fator custo e possíveis atividades no ensino e na pesquisa.

As configurações e conclusões contidas neste artigo fazem parte de um estudo que esta sendo realizado no Instituto de Informática da Pontifícia Universidade Católica do Rio Grande do Sul para a implantação de um laboratório de processamento de alto desempenho.

Na seção 2 é apresentado o estado da arte na área de sistemas paralelos para processamento de alto desempenho. Na seção 3 são apresentadas as principais características das máquinas paralelas baseadas em agregados e descritas as principais tecnologias de interconexão para esta classe de máquinas, disponíveis hoje no mercado. Na seção 4 são propostas diferentes configurações para a confecção de máquinas baseadas em agregados. As conclusões do trabalho são apresentadas na seção 5.

2 O Estado da Arte

O processamento de alto desempenho é considerado uma ferramenta fundamental para as áreas de ciências e tecnologia, pois diversas áreas da ciência exigem alto desempenho (e.g. biologia molecular, química, meteorologia). Sua importância é demonstrada pelas iniciativas de governos do mundo todo em financiar pesquisas e desenvolvimentos nesta área. O processamento de alto desempenho, contudo, depende por sua vez fundamentalmente de técnicas do processamento paralelo, capaz de prover o desempenho necessário para aquelas aplicações [13].

Desta forma, sistemas de processamento paralelo tem tornado-se mais populares em função da necessidade crescente por poder computacional. Na maioria dos casos, estes sistemas ou tem custo muito elevado, ou são difíceis de programar, ou ainda ambos. Os sistemas paralelos disponíveis correntemente podem ser divididos em três classes [5] [17].

- Multiprocessadores Simétricos com memória compartilhada (e.g. SGI Power Challenge);

- Sistemas Massivamente Paralelos com memória distribuída baseados em uma rede de alta velocidade (e.g. Intel Paragon, IBM RS6000/SP, Cray T3E, Thinking Machines CM-5);
- Redes de Estações de Trabalho (NOWs - Network Of Workstations) (e.g. Estações Sun interligadas por rede Ethernet).

As três classes de sistemas têm suas vantagens e desvantagens. A programação dos Multiprocessadores Simétricos (SMPs) é simples mas o tamanho destes sistemas é limitado a uns poucos processadores devido à baixa escalabilidade destas máquinas, pois o barramento que interliga os processadores se torna rapidamente o gargalo do sistema com o aumento do número de processadores. Os sistemas Massivamente Paralelos (MPPs) têm boa escalabilidade mas são caros (necessitam de redes de comunicação de alta velocidade) e de difícil programação pois não existe memória compartilhada e a comunicação é feita por troca de mensagens. Uma das principais vantagens para as NOW é o baixo custo, tanto do hardware como do software uma vez que estão disponíveis softwares como o PVM (Parallel Virtual Machine [10]) que permite explorar a força da computação distribuída para aplicações paralelas praticamente sem custos adicionais. Entretanto, as redes de estações possuem uma baixa performance de comunicação, uma vez que os mecanismos de comunicação sobre redes, por troca de mensagens, são usualmente baseados em protocolos pesados para transferência segura de dados em redes (e.g. TCP/IP). O resultado é uma alta latência (o tempo necessário para o envio de uma mensagem de tamanho zero), de uma ordem de magnitude cerca de três vezes pioir que nos multiprocessadores. É interessante observar que mesmo em redes de alta velocidade (baseadas em ATM) a comunicação utilizando protocolos TCP/IP possui alta latência [9].

Uma abordagem que tem sido utilizada para proliferar ainda mais a utilização da computação paralela é a adoção de redes de PCs para trabalhar em paralelo uma vez que os PCs possuem excelentes taxas de custo/perfomance. Além disso, os sistemas operacionais como Linux e Windows NT provém suporte confiável para comunicação sobre LANs (Local Area Network). Uma vez que o PVM encontra-se disponível para Linux e, mais recentemente, para Windows [12], as redes de PCs estão se tornando plataformas atraentes para processamento paralelo.

3 Máquinas Baseadas em Agregados

Apesar dos problemas apresentados, as Redes de Estações têm tornado-se mais atrativas por causa das novas tecnologias de redes locais de alta velocidade. Com esta combinação, conhecida por Cluster Computing (Máquinas baseadas em Agregados), procura-se aliar as vantagens das três classes apresentadas, construindo-se máquinas paralelas com as seguintes características:

- Comportam-se como redes de estações porque os nós da rede são estações ou PCs normais e podem ser usados em aplicações convencionais;
- Com o uso de placas de comunicação de alta velocidade o sistema de comunicação tem um desempenho que se aproxima dos MPPs (vazão da ordem de centenas de Mbytes/s e latências de uns poucos microssegundos);
- Estações e placas são produzidas em grande escala o que resulta em um custo total bem menor que um MPP, tanto na compra quanto na manutenção;
- Algumas placas de interconexão oferecidas no mercado suportam memória logicamente compartilhada e caches coerentes, o que implica na utilização do modelo de programação de memória compartilhada, como nas SMP's, resultando em uma maior facilidade de programação.

Nos últimos anos muita pesquisa vem sendo desenvolvida em mecanismos mais eficientes de comunicação para máquinas baseadas em agregados. A base desta pesquisa é a criação de um ambiente de interconexão dedicada implementada em hardware para interface de rede que remova a ação do sistema operacional e de processamento de software em geral da linha crítica de comunicação (como mostra a figura 1). O estado da arte nas arquiteturas baseadas em agregados tem obtido na sua comunicação latências da ordem de poucos microssegundos (μs) [6] [11].

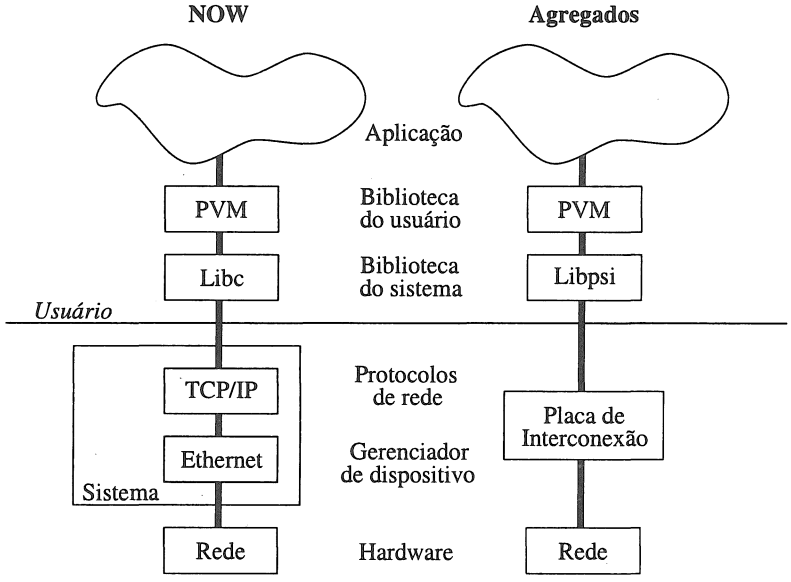

Figura 1: Estrutura de comunicação

Para garantir um melhor desempenho na comunicação, vários padrões de interconexão estão sendo desenvolvidos para conectar os nós destas máquinas, os mais citados na literatura são: *Myrinet* [15], SCI [3] [4], e a *ParaStation* [8].

3.1 *Myrinet*

É um novo tipo de rede que utiliza uma tecnologia baseada em comunicação através de pacotes. As características que tornam a *Myrinet* uma rede de alto desempenho, incluem o desenvolvimento de canais robustos de comunicação com controle de fluxo, pacotes e controle de erro, baixa latência, interfaces que podem mapear a rede, rotas selecionadas e tradução de endereços da rede para estas rotas, bem como manipulação do tráfego de pacotes e software que permite comunicação direta entre os processos a nível de usuário e a rede.

A *Myrinet* foi originalmente desenvolvida para ser utilizada em sistemas multicomputadores (MPP's e NOW's), que consistem de uma coleção de nós de computação, cada um com sua própria memória, conectados por uma rede de troca de mensagens. Atualmente a *Myrinet* vem sendo utilizada em máquinas baseadas em agregados. Em comum com as LANs, os nós de uma máquina baseada em agregados que utilizam uma rede *Myrinet* enviam e recebem dados na forma de pacotes. Qualquer nodo pode enviar um pacote para qualquer outro nodo. Um pacote consiste de uma sequência de bytes iniciando com uma cabeçalho que é examinado pelos circuitos de roteamento para encaminhar o pacote através da rede. Em contraste com as LANs comuns, porém, esta rede baseada em *Myrinet* possui altas taxas de transferência. Uma ligação *Myrinet* é composta por um par de canais full-duplex que permite uma taxa de transferência de cerca de 1.28 Gbit/s cada um.

Uma rede *Myrinet* utiliza normalmente topologias regulares, tipicamente malhas de duas dimensões, embora ela permita a utilização de uma topologia arbitrária uma vez que um cabo *Myrinet*

pode conectar hosts entre si, ou ainda ligar uma placa a um switch ou ainda dois switches entre si. Ao contrário de uma LAN típica onde todo o tráfego de pacotes compartilha um mesmo canal físico, uma rede *Myrinet* com uma malha bidimensional pode ser considerada escalável, pois a capacidade dos agregados cresce com o número de nós devido ao fato de que muitos pacotes podem trafegar de forma concorrente por diferentes caminhos da rede. Uma rede *Myrinet* é composta de ligações full-duplex ponto-a-ponto que conectam hosts e switches. Os switches com múltiplas portas podem ser conectados por ligações para outros switches e para outros hosts em topologias variadas.

A *Myrinet* é uma tecnologia de chaveamento e comunicação de pacotes de alta performance (ela permite uma latência de cerca de 5 μs) e um custo relativamente baixo que está sendo amplamente utilizada para interconectar máquinas baseadas em agregados.

3.2 *ParaStation*

A interface de programação apresentada pela *ParaStation* consiste de uma emulação de sockets UNIX e de ambientes amplamente utilizados para programação paralela, como PVM [10]. Isto permite portar uma grande quantidade de aplicações paralelas e cliente/servidor para a *ParaStation*. Assim como a *Myrinet* a *ParaStation* remove o kernel e os protocolos comuns de rede da linha de comunicação. Enquanto isto, a *ParaStation* mantém a proteção em um ambiente multiusuário, através da implementação de semáforos com o uso de uma biblioteca do sistema, ao nível do usuário. Algumas implementações inicias da *ParaStation* atingiram uma latência em torno de 2 μs e uma largura de banda de 15 Mbyte/s por canal de comunicação. Uma rede *ParaStation* utiliza uma topologia baseada em uma malha toroidal de duas dimensões, mas para sistemas pequenos uma topologia em anel é suficiente.

O objetivo da *ParaStation* é prover uma padronizada e eficiente interface de programação no topo da rede. A rede é dedicada a aplicações paralelas e não pretende substituir LANs comuns , desta forma os protocolos padrão de LANs podem ser eliminados. Isto permite utilizar propriedades mais especializadas na rede, como protocolos especializados ponto-a-ponto e controle da rede ao nível do usuário sem interação com o sistema operacional. O protocolo *ParaStation* implementa múltiplos canais lógicos de comunicação em uma ligação física. Em contraste com outras redes de alta velocidade, como a *Myrinet* por exemplo, na *ParaStation* não há custo adicional para componentes de switch central.

3.3 SCI (*Scalable Coherent Interface*)

SCI é um padrão recente que especifica um inovador hardware e protocolo para conexão de até 64K nós em uma rede de alta velocidade com características de comunicação de alta performance[9][10]. O SCI define serviços de barramento oferecendo soluções distribuídas para a sua realização. O mais notável destes serviços é um espaço de endereçamento físico de 64 bits entre os nós SCI que permite transações de escrita, leitura e a criação de áreas de memória compartilhada entre os nós. Dos 64 bits de endereçamento para a DSM (Distributed Shared Memory), 16 bits são utilizados para endereçar os 64 nós possíveis (2 16 = 64K) e os restantes 48 bits para endereçamento em cada nodo. A placa SCI permite construir máquinas com características NUMA (Non Uniform Memory Access), uma vez que estas placas permitem acessos à memória remota (DSM) realizados pelo hardware, mas que são mais lentos que os acessos locais, o que caracteriza acessos não uniformes à memória [5]. Protocolos para coerência de cache em memória compartilhada distribuída podem ser desenvolvidos para estes sistemas baseados em NUMA.

O SCI evita a limitação física dos barramentos pelo emprego de ligação unidirecional ponto a ponto. Deste modo, não há maiores dificuldades para a escalabilidade. As ligações podem ser

rápidas e sua performance pode aumentar com a utilização de tecnologia de ponta. Tais ligações podem ser implementadas com linhas de transmissão paralela ou serial baseadas em diferentes mídias (p.ex: fibra ótica). O SCI especifica uma largura de banda inicial de 1 Gbit/s para ligação serial e 1 Gbyte/s usando uma canal paralelo, ambos sobre curtas distâncias.

A construção básica de blocos SCI é através de pequenos anéis. Sistemas maiores podem ser obtidos através da criação de anéis de anéis, interconectados via SCI switches. Desta forma, além de permitir a troca de mensagens utilizando um hardware especial o SCI ainda possui a capacidade de implementar via hardware uma memória compartilhada distribuída (DSM), através de operações de escrita e leitura em regiões de memória mapeadas em memórias remotas. Isto se traduz em baixa latência (taxa na ordem de poucos μs) num ambiente baseado em agregados.

4 Configurações Propostas

Nesta seção são propostas três configurações para a confecção de uma máquina paralela baseada em agregados para a implementação um laboratório de alto desempenho, onde podem ser desenvolvidos tanto o ensino como a pesquisa na área de processamento paralelo e distribuído.

A idéia aqui é aproveitar uma grande vantagem das máquinas baseadas em agregados que é a possibilidade de usar os recursos já disponíveis como nós da máquina (PC´s ou estações de trabalho), investindo apenas na sua interconexão. Neste ponto foram utilizadas as tecnologias que mais se destacam no mercado para a interconexão de nós nesta classe de máquinas, apresentadas na seção 3. É claro que se houver recursos disponíveis pode se investir no poder de processamento dos nós, utilizando PC´s ou estações de trabalho de última geração com vários processadores. Como sistema operacional é recomendado o uso do sistema Linux, que alia as vantagens de permitir a alteração do código para adaptação de partes do sistema caso desejado com o fato de ser gratuito.

Nas três configurações serão analisados aspectos como programação das máquinas, possíveis usos no ensino e na pesquisa e principalmente o fator custo, indicado sempre em dólares americanos (US$) por motivos de variação cambial. A questão do número de nós foi deixada propositadamente de lado, já que a intenção é o uso acadêmico destas máquinas na pesquisa e no ensino e não a obtenção pura e simples de desempenho. As configurações apresentadas são sempre de 4 nós e são indicadas formas de expansão caso existam recursos disponíveis.

A estação utilizada para conectar a máquina paralela ao mundo exterior sempre recebe um tratamento especial nas configurações. Esta máquina é denominada hospedeira e não é contada como nó processador da máquina paralela. Como esta máquina é responsável por toda a parte de E/S da máquina paralela e ainda tem funções de carga de programas e de monitoração, ela já sofre uma carga considerável. Isto naturalmente não impede que ela seja usada também no processamento de aplicações paralelas como as outras, mas para fins de análise de desempenho esta sobrecarga tem que ser considerada.

4.1 Configuração Mínima

A figura 2 apresenta o que pode ser chamado de uma configuração mínima para esta classe de máquinas. Neste caso é utilizada uma *Switch Fast-Ethernet* para a interconexão dos nós da máquina.

É importante ressaltar que apesar da diferença para uma rede local normal parecer pequena, esta *Switch* garante uma latência muito menor na comunicação entre as máquinas, através da emulação de uma conexão ponto-a-ponto entre todas as máquinas (é feito um chamamento em hardware ligando os parceiros a cada comunicação). Este é o ponto determinante que faz com que

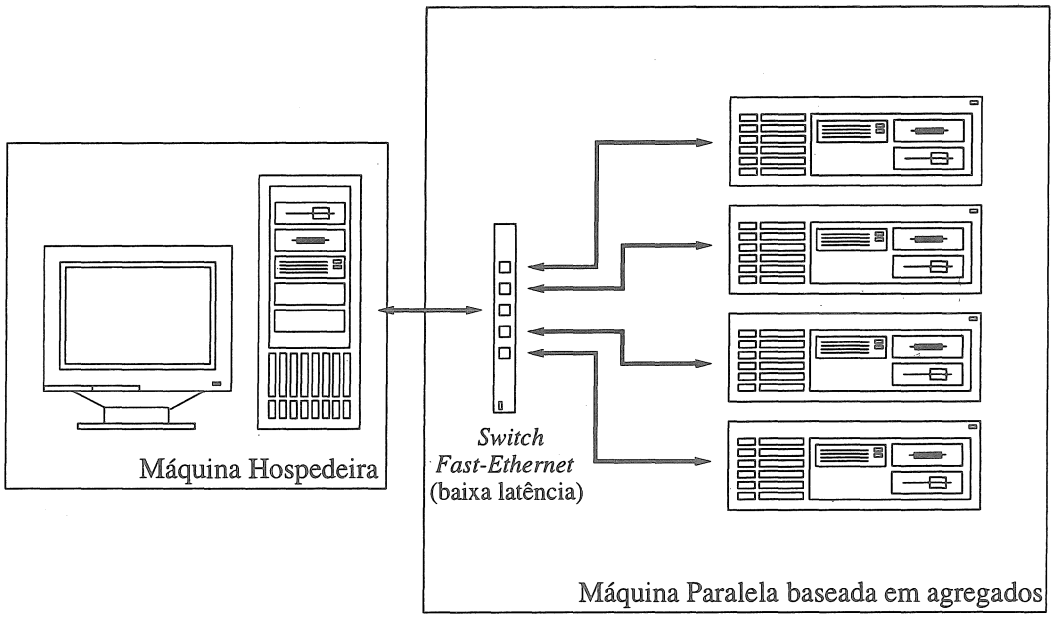

Figura 2: Configuração mínima

esta máquina pertença a classe de máquinas baseadas em agregados e não a classe de redes de estações (NOW´s). Como a interconexão usa placas convencionais *Fast-Ethernet*, a vazão nominal é de 100 Mb/s.

Esta configuração é denominada mínima porque o uso de placas convencionais implica na implementação das camadas de rede em software (figura 1) o que compremete a latência de forma significativa. Nas outras configurações propostas neste trabalhoe estas camadas são implementadas em hardware, o que melhora a latência da comunicação.

A tabela 1 mostra o custo para a confecção de uma máquina baseada em agregados com 4 nós na configuração mínima. A *Switch* utilizada tem 8 portas o que permite a expansão desta máquina até 8 nós apenas com a compra de mais placas de rede. *Switches* com mais portas se encontram disponíveis no mercado e podem ser usadas para confecção de máquinas com mais nós.

Tabela 1: Custo da configuração mínima com 4 nós

Descrição	**Custo** (unid.)	**Quantidade**	**Custo**
Cabo (par-trançado, nível 5, 2m)	4 $	5	20 $
Placas de rede *Fast-Ethernet*	80 $	5	400 $
Switch Fast-Ethernet com 8 portas	2500 $	1	2500 $
		Custo Total	**2920 $**

A programação destas máquinas pode ser feita com bibliotecas padrão para a programação paralela como PVM que se encontram disponíveis para o sistema operacional Linux e são gratuitas. Ambas trabalham com o modelo de comunicação de troca de mensagens que se adapta bem a memória distribuída desta configuração. Outra possibilidade é a programação utilizando o mecanismo de *Sockets* [16] disponíveis no Linux, o que também se enquadraria no paradigma de troca de mensagens.

Esta máquina, apesar de simples, já permite que sejam explorados os paradigmas da progra-

mação paralela e distribuída em atividades de ensino. Com o uso de um monitor de carga na máquina hospedeira, que mostre constantemente a carga de processamento dos nós e o tráfego de mensagens na máquina, podem ser também exploradas de forma mais didática questões de modelagem, desempenho e balanceamento de carga de aplicações paralelas.

Como esta configuração não implementa uma memória global por hardware, como a configuração avançada que veremos a seguir, a implementação de uma memória global distribuída em software e suas implicações é uma possível área de pesquisa. Outra área interessante é a implementação do monitor de carga mencionado anteriormente e suas implicações no funcionamento da máquina como um todo, já que monitor e aplicações compartilham os mesmos canais de comunicação.

4.2 Configuração Básica

A figura 3 apresenta o que pode ser chamado de uma configuração básica para esta classe de máquinas. Neste caso é utilizada uma rede de baixa latência para a interconexão dos nós. Esta denominação foi utilizada para representar a interconexão dos nós por placas de baixa latência e não por placas de rede convencionais.

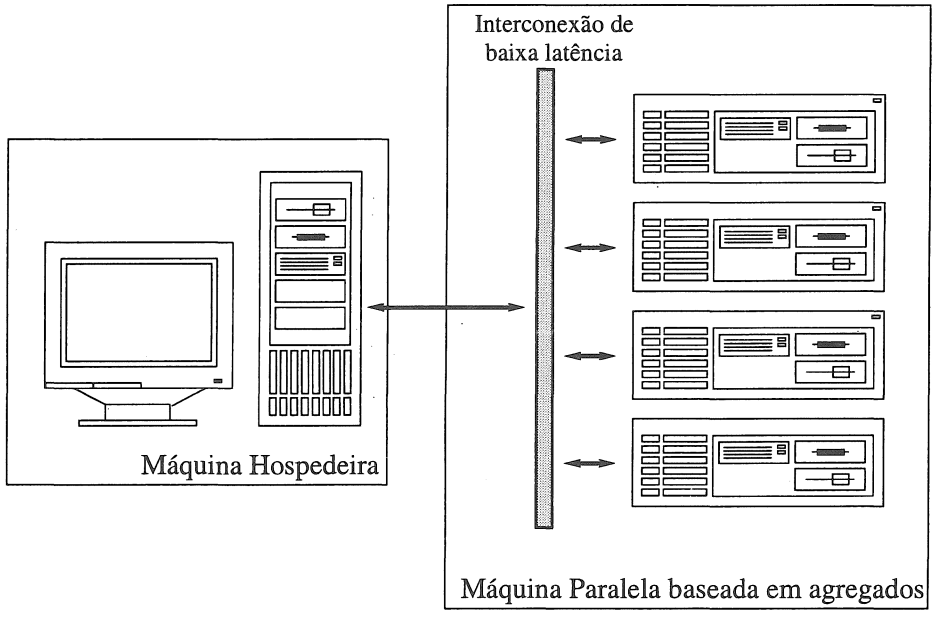

Figura 3: Configuração básica

Nesta configuração é sugerido o uso das placas *Myrinet* ou *Parastation*. A figura 4 mostra a diferença na forma de ligação entre os nós dependendo da placa escolhida. A placa *Myrinet* precisa de uma *Switch* e todas as placas são ligadas a ela por ligações ponto-a-ponto (figura 4a). As placas *Parastation* por sua vez podem ser ligadas entre si por conexões ponto-a-ponto, e para um pequeno número de nós (2-10) se recomenda a ligação em anel (figura 4b).

A principal diferença para a configuração mínima é que as camadas de rede são implementadas em hardware nas placas, e não em software como na configuração anterior, o que melhora a latência na comunicação. Com a interconexão sendo realizada por placas especiais a latência fica em torno de uns poucos μsegundos, consideravelmente menor que na configuração anterior (que depende da *Switch* utilizada, mas varia em torno de algumas dezenas de μsegundos).

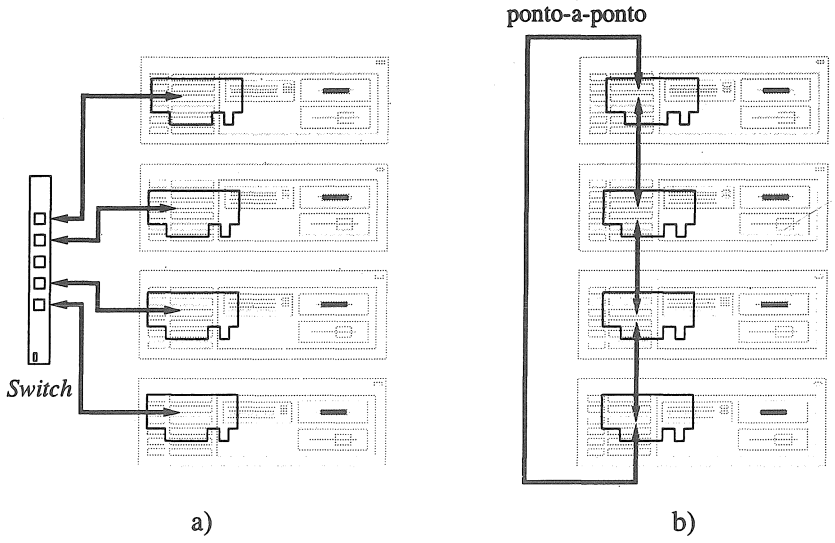

Figura 4: Posseis formas de interconexão dos nós

As tabelas 2 e 3 mostram o custo para a confecção de uma máquina baseada em agregados com 4 nós na configuração básica, dependendo do tipo de placa utilizada na interconexão dos nós. A *Switch Myrinet* utilizada tem 8 portas o que permite a expansão desta máquina até 7 nós (descontando o hospedeiro), apenas com a compra de mais placas de rede e cabos. *Switches* com mais portas se encontram disponíveis no mercado e podem ser usadas para confecção de máquinas com mais nós.

Tabela 2: Custo da configuração básica com 4 nós (Myrinet)

Descrição	Custo (unid.)	Quantidade	Custo
Cabo *Myrinet*	20 $	5	100 $
Placa de Interconexão *Myrinet*	1000 $	5	5000 $
Switch Myrinet com 8 portas	3000 $	1	3000 $
		Custo Total	**8100 $**

Tabela 3: Custo da configuração básica com 4 nós (Parastation)

Descrição	Custo (unid.)	Quantidade	Custo
Placa de Interconexão *Parastation*	1400 $	5	7000 $
		Custo Total	**7000 $**

Como na configuração mínima, a programação destas máquinas pode ser feita com bibliotecas padrão para a programação paralela como PVM e através do mecanismo de *Sockets*. Como na configuração anterior, se trabalha com o modelo de comunicação de troca de mensagens já que aqui também não existe uma memória global entre as máquinas. Desta forma, a implementação de uma memória global distribuída em software e suas implicações é, também nesta configuração, uma possível área de pesquisa. Na prática todas as áreas de aplicação em ensino e pesquisa da configuração anterior são possíveis aqui, tendo que ser apenas levada em consideração a diferença na latência das comunicações. Como este valor se aproximou consideravelmente das máquinas MPP, já se torna possível neste caso comparar as duas arquiteturas (agregados e MPP) em nível de desempenho.

4.3 Configuração Avançada

A figura 5 apresenta o que pode ser chamado de uma configuração avançada para esta classe de máquinas. Neste caso são utilizadas duas redes de interconexão distintas, uma que se utiliza de uma *Switch Fast-Ethernet* (equivalente a configuração mínima) e outra que utiliza placas de interconexão especiais do padrão SCI. A idéia aqui é utilizar a rede *Fast-Ethernet* para a tráfego de E/S, monitoração e gerência de recursos do sistema (lado esquerdo da figura, incluindo o hospedeiro), liberando a rede de menor latência para o tráfego exclusivo de mensagens das aplicações paralelas.

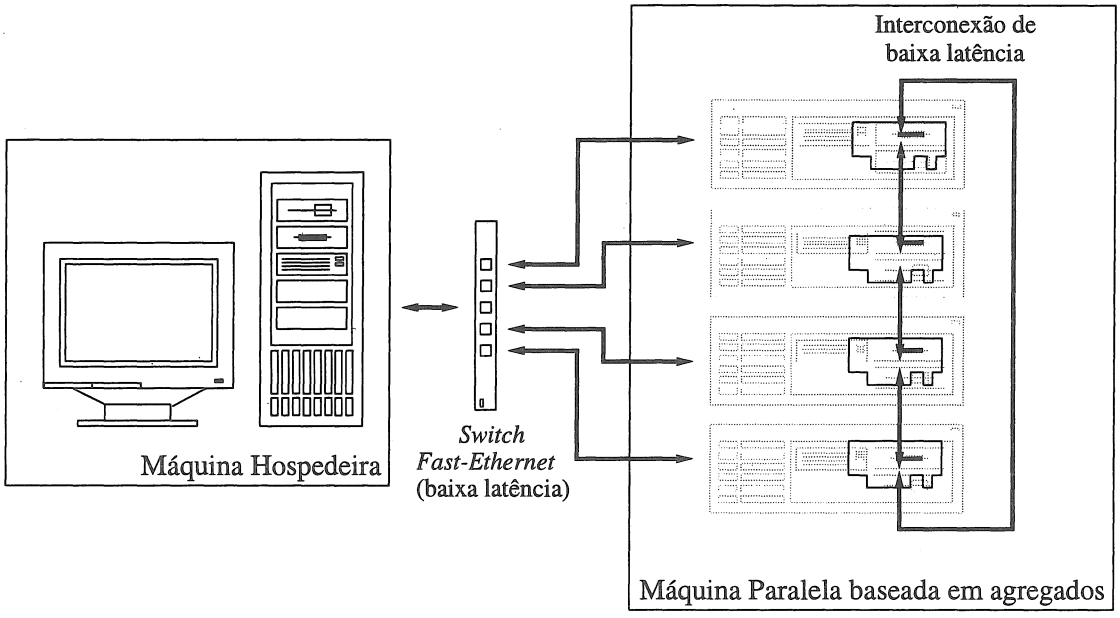

Figura 5: Configuração avançada

As placas SCI são ligadas entre si por conexões ponto-a-ponto e para um pequeno número de nós (2-10) se recomenda a ligação em anel, como pode ser visto na figura 4.3. É importante ressaltar que a principal diferença em nível de arquitetura da máquina é que a placa SCI implementa também uma memória global em hardware, dando uma maior versatilidade na programação desta configuração. A latência da placa SCI é equivalente as placas usadas na configuração básica (poucos μsegundos), pois também implementa as camadas de rede em hardware.

A tabela 4.3 mostra o custo para a confecção de uma máquina baseada em agregados com 4 nós na configuração avançada. A *Switch Fast-Ethernet* utilizada tem 8 portas o que permite a expansão desta máquina até 8 nós (incluindo o hospedeiro), com a compra de mais placas de rede e placas SCI. *Switches* com mais portas se encontram disponíveis no mercado e podem ser usadas para confecção de máquinas com mais nós.

Tabela 4: Custo da configuração avançada com 4 nós (SCI)

Descrição	**Custo** (unid.)	**Quantidade**	**Custo**
Placa de Interconexão SCI (com 2 cabos)	1800 $	4	7200 $
Cabo (par-trançado, nível 5, 2m)	4 $	5	20 $
Placas de rede *Fast-Ethernet*	80 $	5	400 $
Switch Fast-Ethernet com 8 portas	2500 $	1	2500 $
		Custo Total	**10120 $**

Como na configuração básica, a programação destas máquinas pode ser feita com bibliotecas

padrão para a programação paralela como PVM e através do mecanismo de *Sockets*. Como aqui também existe uma memória global entre as máquinas podem ser usadas bibliotecas que trabalhem com memória global como a biblioteca *TreadMarks* [14]. Na prática todas as áreas de aplicação em ensino e pesquisa da configuração anterior são possíveis aqui, incluindo áreas de aplicação que trabalhem com a questão da memória global, como por exemplo o uso de memórias cache para acelerar os acessos a esta memória.

Uma área de pesquisa muito interessante nesta configuração é a utilização da rede de interconexão de E/S na gerência de recursos, o que possibilita o uso de estratégias mais complexas sem que o trafego das aplicações (que se utiliza da rede de interconexão SCI) seja afetado. Assim, operações complexas como a alocação dinâmica de processadores, particionamento da máquina paralela, balanceamento de carga e até migração de processos, podem ocorrer paralelamente a execução das aplicações, sem que ocorra interferência do tráfego gerado em ambas as redes [2].

5 Conclusões

Neste trabalho foi apresentada primeiramente a arquitetura de máquinas baseadas em agregados como uma alternativa em relação às classes de arquiteturas paralelas existentes. Estas arquiteturas baseadas em agregados procuram unir as principais vantagens das outras três para prover uma máquina versátil e de baixo custo. Foi abordado também o fato de que estas máquinas propõe basicamente uma melhoria acentuada na comunicação entre os nós, pela retirada do software que estaria tornando lento o processo na linha crítica de comunicação, permitindo, desta forma, uma redução considerável na latência associada à comunicação, que ficaria na ordem de poucos μs.

Neste contexto foram apresentadas três placas que permitem uma melhora nesta comunicação e que tem sido bastante citadas na literatura, são elas: Myrinet, ParaStation e SCI.

A principal contribuição do trabalho são as três configurações diferentes apresentadas no capítulo 4, destacando o baixo custo e as possibilidades de ensino e pesquisa a serem exploradas com a construção de um laboratório para processamento paralelo e distribuído. Estas configurações vão desde a configuração mínima, que utiliza uma switch Fast-Ethernet para interconexão dos nós da máquina, o que permite uma latência menor que a de uma rede local normal, até uma configuração avançada, onde seriam utilizadas duas redes de interconexão distintas, uma utilizando Fast-Ethernet apenas para tráfego de E/S em geral e monitoração do sistema e a outra utilizando placas SCI ficaria liberada para as mensagens das aplicações paralelas, que necessitam de menor latência. A placa SCI foi escolhida para esta configuração avançada por ser a mais versátil das três, uma vez que permite tanto a troca de mensagens como implementa uma memória global (DSM) em hardware. Esta configuração avançada permitiria um melhor desempenho, pois poderiam ser realizadas operações de gerência de recursos do sistema e monitoração da carga de mensagens e de processamento, sem que isto interfira no desempenho da máquina paralela como um todo.

É importante ressaltar que a tecnologia de agregados ainda está em evolução, uma vez que vários fabricantes estão produzindo placas de baixa latência e novos padrões para estas placas estão sendo desenvolvidos. Cabe a comunidade científica fazer a sua parte pesquisando e desenvolvendo ambientes de execução para estas placas, traçando assim o caminho a ser seguido por esta nova classe de máquinas paralelas.

Referências

[1] M. Borden et al. **Support for ReSerVation Protokol (RSVP) in ATM Networks.** ATM Forum 96-0039, Fevereiro 1996.

[2] C. A. F. De Rose; P. Navaux. *Um modelo Distribuído para a alocação e Gerência de Processadores em Multicomputadores.* **Proceedings do IX Simpósio Brasileiro de Arquitetura de Computadores e Processamento Paralelo - SBACPAD**. Campos do Jordão, SP, 1997.

[3] IEEE: IEEE Standart for Scalable Coherent Interface (SCI). **IEEE standart 1596-1992**, New York, 1993.

[4] H. Hellwanger, W. Karl, M. Leberecht. **Enabling a PC Cluster for High Performance Computing.** LRR-TUM Muenchen, 1998.

[5] Kai Hwang. **Advanced computer architecture: parallelism, scalability, programmability**. MacGraw-Hill Series in Computer Science. McGraw-Hill, 1993.

[6] Proc. Workshop on Communication and Architectural Support for Network- based Parallel Computing (CANPC´97). **http://www.cis.ohio- state.edu/ panda/canpc97.html**,

[7] C. L. Seitz et. al. *Myrinet - A Gigabit-per-Second Local-Area Network.* **IEEE-Micro**, vol. 15, n. 1, Fevereiro 1995, pp. 29-36.

[8] T. M. Warschko et. al. *The ParaStation Project: Using Workstations as Building Blocks for Parallel Computing.* **In proceedings of the International COnference on Parallel and Distributed Processing, Techniques and Aplications (PDPTA "96)**, Agosto 1996, CA, vol 1, pp. 375-386.

[9] T.von Eicken, A.Basu, V. Buch. **Low-Latency Communication Over ATM Networks Using Active Messages. IEEE Micro, Feb.1995, 46-53**

[10] A. Geist, A. Baguelin, J. Dongarra, W. Jiang, R. Manchek, V. Sunderam. **PVM: Parallel Virtual Machine. A User´s Guide and Tutorial for Networked Parallel Computing**. MIT Press 1994.

[11] Proc. Second NOW/Cluster Workshop: Building Systems of systems. **http://w̃ww.csag.cs.uiuc.edu/ãsplos/now-cluster.html**,

[12] PVM for the WIN32 API. **http://www.netlib.org/pvm3/pvm_win32.zip**,

[13] Ted G. Lewis; H. El-Rewini. **Introduction to Parallel Computing**. Prentice-Hall International, Englewood Cliffs, 1992.

[14] Cristiana Amza et. al. *TreadMarks: Shared Memory Computing on Networks of Workstations.* **ACM Computer**, 1995.

[15] C. L. Seitz et. al. *Myrinet - A Gigabit-per-Second Local-Area Network.* **IEEE-Micro**, vol. 15, n. 1, Fevereiro 1995, pp. 29-36.

[16] Arthur Dumas. **Programming WinSock.** Sams Publishing, 1995.

[17] Albert Y. Zomaya, editor. **Parallel and Distributed Computing Handbook**. McGraw-Hill, New York, 1996.