Evolução das Redes de Interconexão

Marino Hilário Catarino
Programa de Ciências da Computação - IME-USP
MAC5742 - Computação Paralela e Distribuída
12 de junho de 2015

Resumo

A interconexão direta de todos os processadores, entre si, não é viável quando o número dos mesmos aumenta. Como regra geral é utilizado um padrão para definir as ligações. Este padrão é denominado de topologia da rede de interconexões. Neste artigo serão vistos a evolução das redes de interconexão, as suas classificações, as topologias utilizadas nas maquínas Top 500.

1 – Introdução

Com a evolução das arquiteturas de computadores, as redes de interconexão evoluiram juntamente. Redes de interconexão podem ser usadas tanto para conexão interna entre processadores, módulos de memória, e dispositivos de entrada e saída quanto para formar uma rede distribuída de nós em um sistema multicomputador. Os centros de dados e redes de alta velocidade são compostos por um conjunto de máquinas que necessitam se comunicar para trocar dados, o que é realizado através da rede de interconexão. Em grande parte das vezes o tempo necessário para que se realize esta comunicação é muito grande e, assim, acaba comprometendo o tempo final das aplicações que executam sobre este tipo de ambiente. Com isso, a evolução na topologia das redes de interconexão avançaramde modo a melhorar o desempenho e serem possíveis de proporcionar um tráfego de dados de alta velocidade, um aumento da produtividade e uma melhor administração das informações e dos equipamentos da rede.

A arquitetura dos multiprocessadores é fortemente acoplada, sendo que os processadores e memória estão fortemente interligados através de seu sistema local de interconexão. A interconexão local de processadores e memória, quando efetuada por intermédio de uma barra, garante a facilidade de configuração compartilhada. Contudo a interligação de processadores e memórias através de um equipamento de comutação estabelece uma configuração comutada simples, podendo também se estenderaté a múltiplos níveis.

Deve-se observar que independente do hardware de interconexão, independente de ser via barra ou um elemento comutador (switch), a arquitetura de um multiprocessador é caracterizada pelo compartilhamento global de uma memória pelos diversos processadores do ambiente.

Independnete do tipo da arquitetura, todo computador paralelo necessita de uma rede de interconexão para criar canais de comunicação entre os seus diversos recursos de processamento, armazenamento e entrada/saída. Considerando a diversidade das alternativas tecnológicas, esta seção vai explorar aspectos centrais pertinentes ao tema, a partir dos quais, podem ser entendidas as várias alternativas possíveis para as redes de interconexão. Na figura 1 é visto a arquitetura de um multiprocessador do tipo UMA.

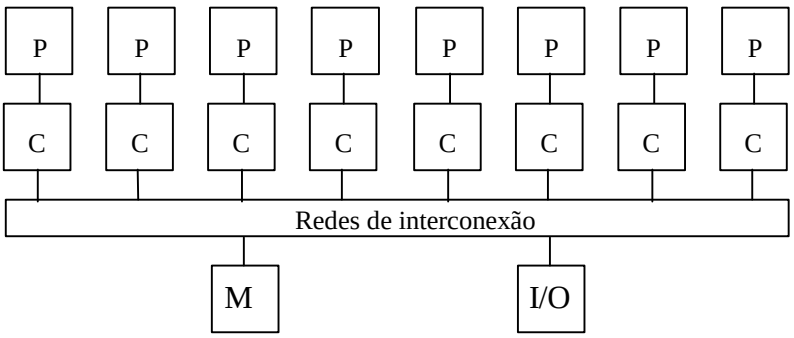


Figura 1 - Arquitetura de um multiprocessador do tipo UMA

2 - Características

Além da topologia de rede de interconexão, outras características são importantes para a classificação quanto o desempenho de cada uma:

Escalabilidade: possibilidade de acréscimo de dispositivos sem a necessidade de alteração das características da rede;

Desempenho: está relacionado com as distâncias envolvidas e om o número de operações simultâneas. O desempenho tem como métricas a latência e a taxa de transferência. A primeira corresponde ao tempo necessário para a transferência dos dados e a segunda representa a quantidade de dados que podem ser comunicados por unidade de tempo. Uma outra questão com impacto no desempenho é se a rede é unidirecional ou bidirecional;

Custo: basicamente é proporcional ao desempenho desejado e ao número de ligações existentes;

Confiabilidade: especifica a capacidade de comunicação da rede mediante falha em alguma ligação. Está associada com a existência de caminhos redundantes entre os componentes;

Funcionalidade: indica os serviços extras oferecidos pela rede como armazenamento temporário, ordenação e roteamento automático por hardware.

Também devem ser levados em conta a largura de banda do canal. Que é o número de bytes por segundo que pode fluir entre dois nós com conexão direta. A largura de banda é dependente do número de pulsos por segundo da arquitetura (clock) e do número de bits possíveis de serem enviados por pulso. A latência de comutação, que é o tempo inerente à operação da chave de comutação. Se dois processadores precisam trocar dados, e não existe um canal interligando os dois diretamente, as chaves de comutação intermediárias precisam propagar a mensagem através da rede de interconexão. As latências elevadas trazem prejuízo ao desempenho da arquitetura paralela, sobretudo quando a mensagem necessita passar por diversas chaves.

A independência de processador, que se caracteriza a necessidade de o processador ser ou não interrompido, para auxiliar na atividade de comunicação. Muitas das atuais implementações de redes de interconexão permitem que o processador continue sua computação enquanto uma mensagem está sendo transmitida, recebida ou roteada. Isto minimiza o custo introduzido pela necessidade de comunicação entre processadores. E por fim a contenção, que pode ocorrer a recepção praticamente simultânea de duas mensagens por uma determinada chave, e ambas podem necessitar usar o mesmo canal de saída. Uma obrigatoriamente terá de aguardar. O atraso na computação do processador que aguarda a mensagem retida pode resultar em perda de desempenho. Uma possibilidade é o hardware de comutação prever uma política de tempo compartilhado para as portas das chaves; isto dividiria o custo de espera entre os dois processadores destinatários, porém introduziria custos de comutação .

3 - Redes estáticas

São as que especificam uma ligação direta dedicada entre dois componentes quaisquer. Muito utilizada em multicomputadores. O número de ligações diretas de cada componente define o grau do nó. A maior distância (em número de ligações) entre dois componentes quaisquer é chamada de diâmetro da rede.

No exemplo da figura 2 temos:

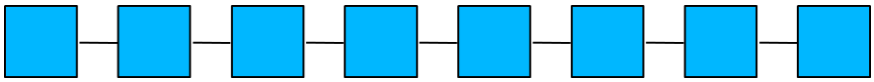


Figura 2 – topologia em array linear

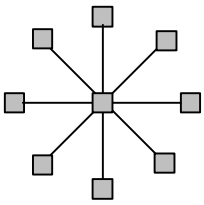
Ligações (L) entre os N nós/componentes: L = N -1 = 7.
Grau (g) do nó: número de ligações diretas de cada componente: g = máximo, que é 2.
Diâmetro (D): A maior distância (em número de ligações) entre dois componentes quaisquer é chamada de diâmetro da rede:
D = N – 1 = 7.

As topologias de redes estáticas são:

Array linear:

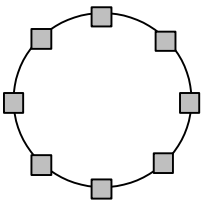
- sem caminhos alternativos

Estrela:



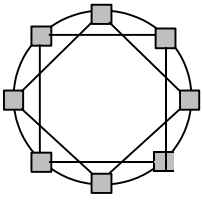
- tráfego intenso no nó central
- problemas no nó central bloqueiam a rede

Anel:



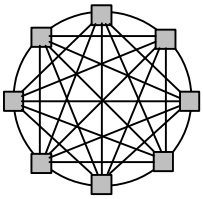
- baixo custo
- diâmetro cresce de forma linear com os nós
- sem caminhos alternativos
- tráfego intenso

Anel chordal:



- menos tráfego no anel central
- caminho alternativos

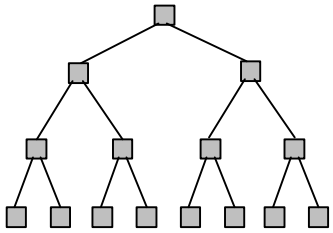
Totalmente conectada:



- alto custo
- grau de nó = número de nós - 1
- diâmetro 1 (ideal)

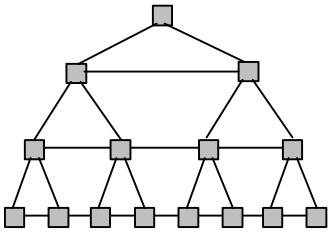
Outro critério para a avaliação de uma rede é a sua adequação a um algoritmo específico. Uma rede do tipo árvore binária, por exemplo, é ideal para a execução de algoritmos do tipo divisão e conquista (*divide and conquer*). A topologia Hipercubos surgiu em 1983, pela Caltech, e a topologia Torus surgiu em 1985, também pela Caltech.

Árvore binária:



- diâmetro cresce de forma linear com a altura *h*
- grau de nó máximo 3
- sem caminhos alternativos
- nó raiz é um gargalo

X-Tree:

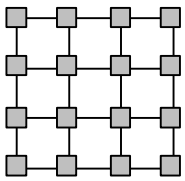


- caminhos alternativos
- grau de nó máximo 5

A clássica árvore binária, com processadores nas suas folhas, tem se mostrado uma boa opção de topologia para arquiteturas paralelas. O diâmetro de uma árvore completa é 2log2((N+1)/2), bastante similar ao do hipercubo (N é o número de processadores). A largura da bisseção, por sua vez, é somente 1, o que pode introduzir um severo gargalo quando processadores de uma metade da árvore precisarem se comunicar com os da outra metade.

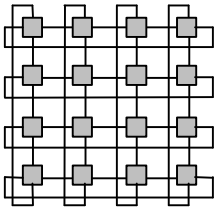
A solução para pequeníssima largura da bisseção da árvore é utilizar uma variante denominada árvore larga. Em uma árvore larga (vide Figura 3.4), a largura dos ramos (canal) cresce a medida em que se sobe das folhas em direção à raiz. A largura da bisseção da árvore larga plena é N e o seu diâmetro proporcional a 2(logN). A arquitetura da CM-5 da Thinking Machines utiliza uma versão modificada da árvore larga.

Malha bidimensional:



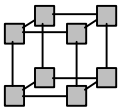
- grau de nó máximo 4
- facilidade de incremento de elementos

Torus:



- grau de nó 4
- diâmetro reduzido em relação ao número de nós

Hipercubos:

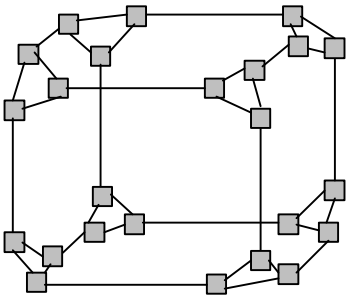


- escalabilidade restrita a potências de 2
- diâmetro = grau de nó
- diâmetro cresce logaritmicamente
- grau de nó = dimensão do cubo (exemplo = 3)

Toda rede de interconexão hipercúbica está alicerçada sobre uma estrutura multidimensional baseada em endereços binários. Os tamanhos do hipercubo são definidos por potências de 2; N=2D onde D é a dimensão do hipercubo e N o número de processadores. Em função disto, todos os nós podem ser identificados por um número binário. Cada nó é conectado a todos os seus vizinhos; isto faz com que o hipercubo tenha grau variável e de valor D. A topologia hipercúbica confere boas propriedades à rede de interconexão; a largura da bisseção é N/2, e o diâmetro é log2. Apesar de apresentar bom desempenho para muitos padrões de comunicação, sua eficiência se mostra bastante dependente do algoritmo de roteamento a ser empregado.

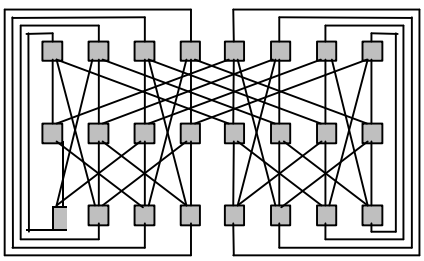
Um aspecto inconveniente do hipercubo é sua escalabilidade, o número de processadores sempre cresce em potência de 2. Além disso, como o grau de cada nó é em função do tamanho do cubo, toda expansão no número de processadores implica em adicionar mais um canal de comunicação a todos os nós. Para cubos maiores, estas características podem trazer inconvenientes para a administração do custo/benefício quando da expansão da arquitetura. Um equipamento que emprega esta topologia é o Ncube 2.

Cubo CCC (*Cube Connected Cycles*):



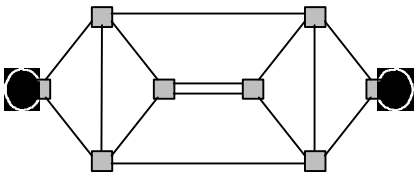
- hipercubo em que cada nó é um anel
- hipercubo de dimensão *d* = anel com *d* nós
- diâmetro cresce logaritmicamente
- grau de nó 3 para qualquer diâmetro

Butterfly:



- grau de nó 4
- diâmetro menor que um cubo CCC
- diâmetro cresce logaritmicamente
- o exemplo é de dimensão 3

Grafo de DeBrujn:



- grau de nó 4
- grafo de dimensão d = 2 x d nós
- diâmetro cresce logaritmicamente
- o exemplo é de um grafo de dimensão 3

4 - Redes dinâmicas

Corresponde às redes em que as conexões são feitas sob demanda. Não existem ligações fixas entre os componentes. São as mais utilizadas em multiprocessadores. Uma questão importante nas redes dinâmicas é a possibilidade de serem bloqueantes ou não, ou seja, é a possibilidade de conexão entre dois elementos da rede impedir a conexão entre dois outros.

As redes dinâmicas podem ser de três tipos:

- Barramento
- Matriz de chaveamento
- Rede multinível

O barramento é a forma mais simples de conexão mas tem como grande desvantagem o compartilhamento do mesmo meio físico por parte de todos os elementos do sistema. Este fato caracteriza o barramento como altamente bloqueante e de baixa confiabilidade. O fato de ser bloqueante limita a escalabilidade do barramento que é indicado para sistemas com menos de 50 processadores. Uma forma de minimizar as desvantagens deste tipo de conexão é a utilização de mais de um barramento na conexão dos elementos.

Nesta topologia, todos os processadores estão conectados em um único barramento compartilhado. Quando um processador necessita comunicar-se com outro, ele aguarda que o barramento esteja livre e propaga no mesmo a mensagem; o destinatário, por sua vez, identifica que a mensagem é para si e a recebe.

No caso de duas transmissões simultâneas, o software detector de colisões interrompe as transmissões e os processadores voltam a tentar novamente após um período de tempo determinado aleatoriamente. Assim sendo, a sua largura da bisseção é 1. Isto significa que esta topologia não permite mais do que um par de processadores em comunicação simultaneamente. Do ponto de vista do desempenho, esta topologia somente é viável para um pequeno número de processadores e/ou classes de problemas cujos algoritmos implementem pouca comunicação. Esta topologia é bastante usual em pequenos agrupamentos (clusters) de estações de trabalho interligadas por redes locais.

Do ponto de vista do desempenho, esta topologia somente é viável para um pequeno número de processadores e/ou classes de problemas cujos algoritmos implementem pouca comunicação. Esta topologia é bastante usual em pequenos agrupamentos (clusters) de estações de trabalho interligadas por redes locais.

A matriz de chaveamento, ou crossbar switch, é uma alternativa não bloqueante de interconexão. Quaisquer elementos podem ser interligados dinamicamente mas esta característica é originada na grande disponibilidade de hardware, o que se reflete em um alto custo do sistema para um grande número de processadores. A escalabilidade fica limitada apenas pelos aspectos econômicos. Uma possível solução para a questão dos custos é a conexão de várias matrizes menores para a implementação de redes com grande número de processadores, mas esta é uma alternativa que torna a rede bloqueante.

Os processadores nesta topologia tem um canal de comunicação direto com o seu vizinho (a). Uma variação que é utilizada consiste em interligar as extremidades da grade, criando uma configuração denominada malha toroidal (b), a qual reduz o diâmetro da malha por um fator de 2. A largura da bisseção de uma malha é $\sqrt{N}$ onde N é número de

processadores. A largura da bisseção dobra para a malha toroidal. O diâmetro da topologia em malha é $2(\sqrt{N}-1)$, e o seu grau é fixo e de valor 4.

O hardware para este tipo de tecnologia é de simples construção e expansão. A malha se adapta bem a algoritmos utilizados em cálculos científicos, onde se destaca a manipulação de matrizes.

As redes multinível são uma variação das matrizes de chaveamento, de maneira que se possa utilizar matrizes de chaveamento padrão (2 x 2, por exemplo) para a interconexão dos elementos. Tais matrizes são organizadas em diversos níveis de conexões, de forma a minimizar a possibilidade de conflito na ligação entre dois componentes quaisquer.

5 - Dispositivos de interconexão

Já estão disponíveis comercialmente dispositivos de interconexão que propiciam a criação de ambientes similares a multicomputadores ou multiprocessadores, utilizando computadores convencionais.

Existem atualmente duas grandes classes de dispositivos de interconexão para alto desempenho. Uma primeira classe é formada por dispositivos cuja solução é baseada em programação por troca de mensagens entre processadores no nível de placa de rede, esta solução permite a criação de multicomputadores. Exemplos de equipamentos desta classe são: Myrinet, Gigabyte System Network e Giganet, sistemas que utilizam rede Gigabit ethernet também são encontrados, mas com desempenho de rede mais baixo. Não se pode confundir as tecnologias Gigabit ethernet, Gigabyte System Network e Giganet. A Gigabit ethernet é a mais conhecida, utilizada e de menor custo, todavia o seu desempenho é muito menor comparado com as outras soluções.

A segunda classe é formada por interconexões e tem como peculiaridade uma solução que cria a abstração de uma memória virtual única (multiprocessador) entre todos os computadores interligados no dispositivo. Exemplo desta são o Quadrics Network (QSNET) e Dolphin SCI.

6 – Top 500

Segundo a lista de 2013 dos supercomputadores Top 500, a topologia Torus é a mais utilizada, alterando-se a quantidade de dimensões em cada configuração de computador:

- IBM Blue Gene/L e Blue Gene/P, Cray XT3: 3D torus
- IBM Blue Gene/Q: 5D torus
- Fujitsu K: 6 D torus chamado Tofu

O k Computer foi concebido por meio de uma parceria da Fujitsu e a RIKEN. Este supercomputador em 2011 ocupava a segunda posição da lista Top500. Esse sistema utiliza uma conexão direta e foi projetado para suportar mais de 80 mil nós. Com este tipo de conexão, juntamente com algoritmos especializados, é possível alocar um grupo de k nós a uma aplicação ou usuário específico, por isto o nome K Computer.

7 - Conclusão

A evolução de tecnologias para redes leva a utilização de arquiteturas de interconexão mais eficientes que é capaz de porporcionar várias funcionalidades, além da alta taxa de transferência de dados, que são requeridas por gerentes de tecnologia da informação de grandes organizações. A tendência da arquitetura de redes é ser utilizada em conjunto com as tecnologias padrão existentes atualmente e, dessa forma, construir ambientes que melhorem a vida dos usuários e dos administradores de ambientes computacionais. Redes de interconexão são importaantes pois acabam se tornando gargálos do tempo de execução de processamente.

Esta evolução é visível no ranqueamento dos top 500, no qual nota-se uma padronização recente.

8 - Referência bibliográfica

- Hwang, K.; Xu, Z. Scalable Parallel Computing: technology, architecture, programming. McGraw-Hill, 1998. (Capítulo 6)
- de Rose, C. A. F. Fundamentos de Processamento de Alto Desempenho. Curso Permanente. Escola Regional de Alto Desempenho, 2006.
- A 39ª lista Top500. Os computadores mais rápidos do mundo. LEONARDO GARCIA TAMPELINI
- Interconnection Network Architectures for High-Performance Computing, Cyriel Minkenberg IBM Research ─ Zurich, 2013.
- Nedialkov, N., Interconnection Networks, CS/SE 2015
- Dally, W. J. From Hypercubes to Dragonflies, a short history of interconnect, Stanford University, 2008.
- Minkenberg, C.,Interconnection Network Architectures for High-Performance Computing, IBM, 2013.
- Silva, J. M. O., Estudo e construção de um Ambiente de Alto Desempenho utilizando Cluster Computacional, UFPI, 2009.