O Subsistema de Interconexão do Multiprocessador Multiplus

Gerson Bronstein [1]
NCE/UFRJ

Resumo

Neste artigo apresentamos o subsistema de interconexão do multiprocessador Multiplus. Este subsistema é constituído de dois módulos: a Rede de Interconexão Multi-estágio e a Interface de Rede. A rede de interconexão tem topologia n-cubo invertido e é composta por elementos comutadores 2 × 2. A interface de rede controla a transmissão/recepção de mensagens e faz a interface da rede com o sistema de barramentos. São relacionadas as características gerais destes dois módulos bem como aspectos relacionados a implementação dos mesmos. Ao final, apresentamos algumas sugestões para o aumento do desempenho global do Multiplus.

Abstract

In this paper we present the interconnection subsystem of the Multiplus multiprocessor. This subsystem has two main modules: the Multistage Interconnection Network and the Network Interface. The interconnection network is an inverted n-cube and is made of 2 × 2 switching elements. The network interface controls message transmission/reception and interfaces the network with the bus system. General characteristics of both modules are related and implementation-dependent features are discussed. At the end, we present some suggestions to improve general performance of the Multiplus.

[1] M.Sc. 1991 pela COPPE/UFRJ. Pesquisador do Núcleo de Computação Eletrônica - UFRJ. Áreas de interesse: Arquitetura de Computadores, Redes de Interconexão, Dispositivos Lógicos Programáveis. Email: gerson@barra.nce.ufrj.br.

1 Introdução

A demanda por computadores de alto desempenho tem aumentado a cada dia e hoje estas máquinas são utilizadas nas mais diversas áreas. Se levarmos em conta que a velocidade dos dispositivos e componentes possui um limite físico, os esforços para o aumento do desempenho dos sistemas de computação estarão voltados para a utilização de arquiteturas paralelas e a exploração do paralelismo das aplicações. Dentro deste contexto, o subsistema de interconexão dos processadores desempenha um importante papel, fornecendo os meios para a exploração deste paralelismo.

Neste trabalho apresentaremos o subsistema de interconexão do multiprocessador Multiplus, em desenvolvimento no Núcleo de Computação Eletrônica da Universidade Federal do Rio de Janeiro. Inicialmente daremos uma descrição sucinta da arquitetura do Multiplus. Nas Seções 3 e 4 apresentaremos os dois módulos que compõem o subsistema de interconexão: a rede de interconexão e a interface de rede ou serializador de mensagens. Finalmente, na Seção 5, apresentaremos algumas sugestões para o aumento do desempenho do Multiplus em geral e do subsistema de interconexão em particular.

2 A Arquitetura do Multiplus

O Multiplus [1] é um multiprocessador com uma arquitetura modular capaz de suportar até 1024 nós de processamento (NP) (Fig. 1). Cada NP é constituído de um processador RISC com arquitetura SPARC [2], 64 M*bytes* de memória local que representam parte do espaço de endereçamento global, um *cache* de dados e um de instruções com 64 K*bytes* cada, um co-processador de ponto flutuante e *hardware* de suporte à gerência de memória. Até oito nós de processamento e dois processadores de E/S (PES) [4] podem ser conectados através de um sistema de barramentos formando um *cluster* de processamento.

Os barramentos dentro de um *cluster* são especializados: um para leitura de instruções e transferências de DMA e outro para leitura/escrita de dados. Ambos tem largura de 64 *bits*. Optamos por realizar as transferências de DMA pelo barramento de instruções pois este apresenta uma taxa média de ocupação bem menor que a do barramento de dados [3]

Como a memória local de um NP faz parte do espaço global de endereçamento, podemos ter 3 classes de acessos: acesso local ao NP, acesso local ao *cluster* e acesso remoto. Apenas esta última classe de acessos utiliza o subsistema de interconexão

O Protocolo utilizado nos barramentos [5] é uma versão assíncrona do SPARC MBus [2] e é composto de duas fases: fase de endereçamento e fase de dados. Na fase de endereçamento, o *master* (módulo origem do acesso) fornece o endereço da transação e aguarda a decodificação dos demais membros do barramento. Na fase de dados podem ocorrer uma ou mais transferências entre *master* e *slave* (módulo destino do acesso). A cada transferência, o*master* sinaliza autorizando o início da mesma e o *slave* retorna um código informando o *status* da mesma.

Para que seja garantida a coerência dos acessos remotos, mesmo durante um acesso remoto de escrita o *master* aguarda a chegada da confirmação da realização do acesso. Portanto, cada membro de um cluster só pode ter um acesso remoto pendente.

3 A Rede de Interconexão do Multiplus

A rede de interconexão do Multiplus [6] é o meio pelo qual os *clusters* de processamento do Multiplus são interligados. Ela é do tipo multi-estágio [7] e é composta por elementos comutadores 2 × 2 com *buffers* do tipo FIFO em cada uma das saídas [8]. Ela utiliza comutação por pacotes com roteamento do tipo *wormhole* [7]. A topologia adotada provê a arquitetura do Multiplus com duas importantes características: modularidade e partibilidade [7].

A modularidade permite que se aumente o número de *clusters* de processamento pela simples adição de estágios extras na rede, enquanto que a partibilidade possibilita a divisão da rede em sub-redes de forma que o tráfego em uma sub-rede não interfira no tráfego das demais sub-redes.

As mensagens tem tamanho variável e são compostas por pacotes de 8 bits mais 1 bit de paridade. Ficou estabelecido, por questões de confiabilidade do projeto, que cada mensagem gerada (mensagem-transação) deve ter sempre uma resposta (mensagem-resposta). Foram definidos 3 tipos de mensagem-transação e suas respectivas mensagens-resposta: leitura (resposta-leitura), escrita (resposta-escrita) e DMA (resposta-dma).

As mensagens podem ter até 3 campos distintos: um cabeçalho, composto de 4 pacotes; um preâmbulo, composto de 8 pacotes; um campo de dados, composto de 1 a 128 pacotes.

O cabeçalho, que é o único campo presente em todas as mensagens, contém informações de roteamento e controle. O primeiro pacote possui o endereço de destino. Cada elemento comutador verifica o *bit* menos significativo deste pacote e decide para qual das duas saídas rotear a mensagem. Para que todos os elementos comutadores verifiquem sempre o mesmo bit do endereço de destino, este pacote é sempre deslocado de um *bit* para a direita ao ser transmitido para o estágio seguinte. O segundo e terceiro pacotes são idênticos. Eles possuem o tamanho da mensagem (4 *bits*) replicado 4 vezes. Isto foi feito como uma medida extra de segurança, pois atualmente o tamanho da mensagem é a única forma que o elemento comutador tem de saber quando uma mensagem termina. O quarto e último pacote possui o tipo da mensagem e identificação do módulo (NP ou PES) que a originou.

O preâmbulo contém as informações de controle do barramento geradas durante o acesso no *cluster*-origem. Apenas as mensagem-transação possuem este campo.

O campo de dados possui os dados correspondentes a um determinado acesso. Apenas as mensagens-transação de escrita e DMA e as mensagens-resposta de leitura possuem este campo.

Outro fator importante para a definição da arquitetura da rede de interconexão foi a

tecnologia a ser utilizada na sua implementação. Foram analisadas diversas alternativas tais como lógica discreta, *gate-arrays*, implementação VLSI dedicada e dispositivos lógicos programáveis. Após uma cuidadosa análise das opções acima, decidimos pela utilização de dispositivos lógicos programáveis, em particular a família EPM5000 da Altera [9]. Os fatores que motivaram esta escolha foram a alta densidade de portas lógicas desta família, flexibilidade na correção/alteração do projeto e a existência de um pacote integrado de programas de suporte à projetos [10].

4 A Interface de Rede do Multiplus (IR)

A interface de rede é a responsável pela serialização e transmissão/recepção de mensagens (Fig. 2). Cada um dos módulos é composto por uma ou mais máquinas de estado independentes. A filosofia básica adotada durante todo o projeto da IR foi a de minimizar ao máximo o impacto causado pelos acessos remotos, lentos se comparados aos locais, na utilização dos barramentos.

O protocolo de barramento utilizado é uma versão assíncrona do SPARC MBus. Durante a fase de endereçamento, a IR identifica se o acesso é local ou remoto. No caso de um acesso remoto, a IR atua como *slave* e existem 2 possibilidades: acesso remoto de leitura ou de escrita.

No acesso remoto de escrita, as informações de controle do acesso são copiadas para a memória de transmissão de mensagens e posteriormente processadas pelo módulo de transmissão de mensagens. Para que o barramento não permaneça ocupado inutilmente enquanto aguarda a resposta, a IR retorna para o *master* o código **Relinquish&Retry** [5]. Este código informa ao *master* que ele deve liberar o barramento, aguardar um determinado tempo e depois repetir o acesso. Este procedimento é repetido pelo *master* até que os dados solicitados cheguem através da rede.

No acesso remoto de escrita, todas as informações (controle do acesso + dados) são copiadas para a memória de transmissão de mensagens e a IR retorna ao *master* o código **Acesso-pela-Rede** [5] ao final da fase de dados. O *master* permanece inativo até a resposta do acesso de escrita retorne.

As mensagens que chegam através da rede são tratadas pelo módulo de recepção de mensagens e ficam armazenadas na memória de recepção de mensagens. Estas mensagens são posteriormente retiradas e tratadas pelo módulo de interface com o barramento.

Erros ocorridos durante as transações ou durante a transmissão/recepção de mensagens são informados através de códigos próprios. Além disto, para cada mensagem transmitida é disparado um temporizador. Caso a mensagem-resposta correspondente não retorne em um intervalo de tempo determinado, o*master* é informado. Isto evita que o *master* permaneça em estado de espera indefinidamente.

Cada IR possui um controlador de DMA. Este controlador só pode ser utilizado pelos membros do cluster ao qual ele pertence. Sua programação é feita através do barramento de

dados, porém as transferências são realizadas pelo barramento de instruções.

Visando aumentar o desempenho do subsistema de interconexão e pelo fato de não haver troca de informações entre os barramentos de dados e de instruções, optamos pela utilização de duas redes distintas: uma para dados e outra para instruções e DMA.

A tecnologia escolhida para a implementação foi a mesma utilizada na rede de interconexão: dispositivos lógicos programáveis.

5 Sugestões para Versões Futuras

No decorrer deste trabalho, algumas decisões foram tomadas a fim de simplificar o projeto da primeira versão do subsistema de interconexão e do Multiplus em geral. Apresentamos aqui algumas sugestões que visam aumentar o desempebho global da máquina.

Devido às dimensões do *backplane* dos barramentos (cerca de 80cm) e da frequência de operação das interfaces (25 MHz), decidimos utilizar um protocolo assíncrono nesta primeira versão. Porém, se tormarmos o devido cuidado com os atrasos na propagação dos sinais, poderemos utilizar em versões futuras um protocolo síncrono. Isto diminuiria significativamente o tempo necessário para a realização de uma transação no barramento.

O fator limitante no desempenho dos elementos comutadores da rede de interconexão é a velocidade dos *buffers* tipo FIFO. Com a utilização de dispositivos lógicos programáveis mais densos (a família EPM7000 da Altera, por exemplo [11]), poderíamos embutir estes *buffers* no elemento comutador. Isto certamente aumentaria o desempenho da rede.

Na interface de rede, o paralelismo das transações poderia ser aproveitado de forma mais efetiva.

6 Agradecimentos

Gostaríamos de agradecer ao Prof. Júlio Salek Aude pelas inúmeras sugestões dadas durante a realização deste trabalho e à Finep pelo apoio dado ao Projeto Multiplus.

Referências

[1] J. S. Aude et ali., "Multiplus: A Modular High-Performance Multiprocessor", *Proceedings of Euromicro 91*, pp. 45-52, Viena, 1991.

[2] *The SPARC Architecture Manual*, Sun Microsystems Inc., 1987.

[3] A. M. Meslin et ali., "A Comparative Study of Cache Memory Architectures for the Multiplus Multiprocessor", *Proceedings of Euromicro 92*, pp. 555-562, Paris, 1992.

[4] S. C. Oliveira e J. S. Aude, "Uma Avaliação do Impacto das Operações de E/S do Multiprocessador Multiplus", *Anais do IV SBAC-PAD*, pp. 379-394, São Paulo, 1992.

[5] *Manual do NCEBus*, Relatório Interno do Projeto Multiplus, 1991.

[6] G. Bronstein, *Projeto de uma Rede de Interconexão para uma Máquina Paralela de Alto Desempenho*, Tese de Mestrado, COPPE-UFRJ, 1991.

[7] H. J. Siegel et ali., "Using the Multistage Cube Network Topology in Parallel Supercomputers", *Proceedings of the IEEE*, pp. 1932-1953, no. 12, Dez/89.

[8] Y. Tamir e G. L. Frazier, "High-Performance Multi-Queue Buffers for VLSI Communication Switches", *ACM Computer Architecture News*, v. 16, Mai/88.

[9] *The Maximalist Handbook*, Altera Corporation, Jan/1990.

[10] *MaxPlus User's Guide*, Altera Corporation, 1989.

[11] *MAX 7000 Data Book*, Altera Corporation, 1993.

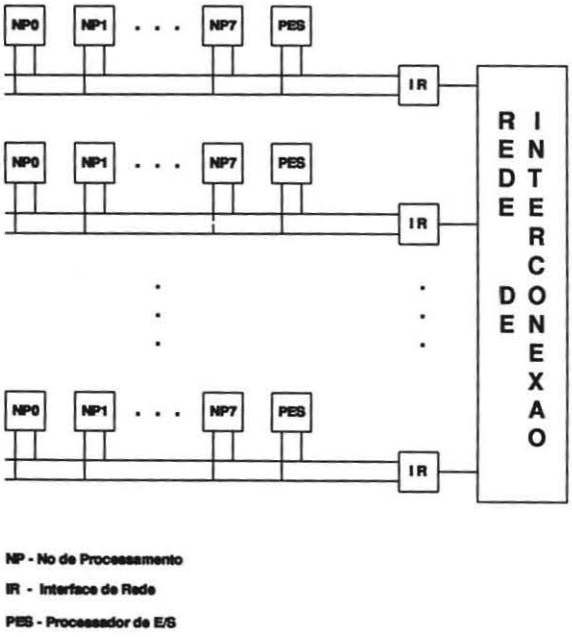

Figura 1: Arquitetura do Multiplus

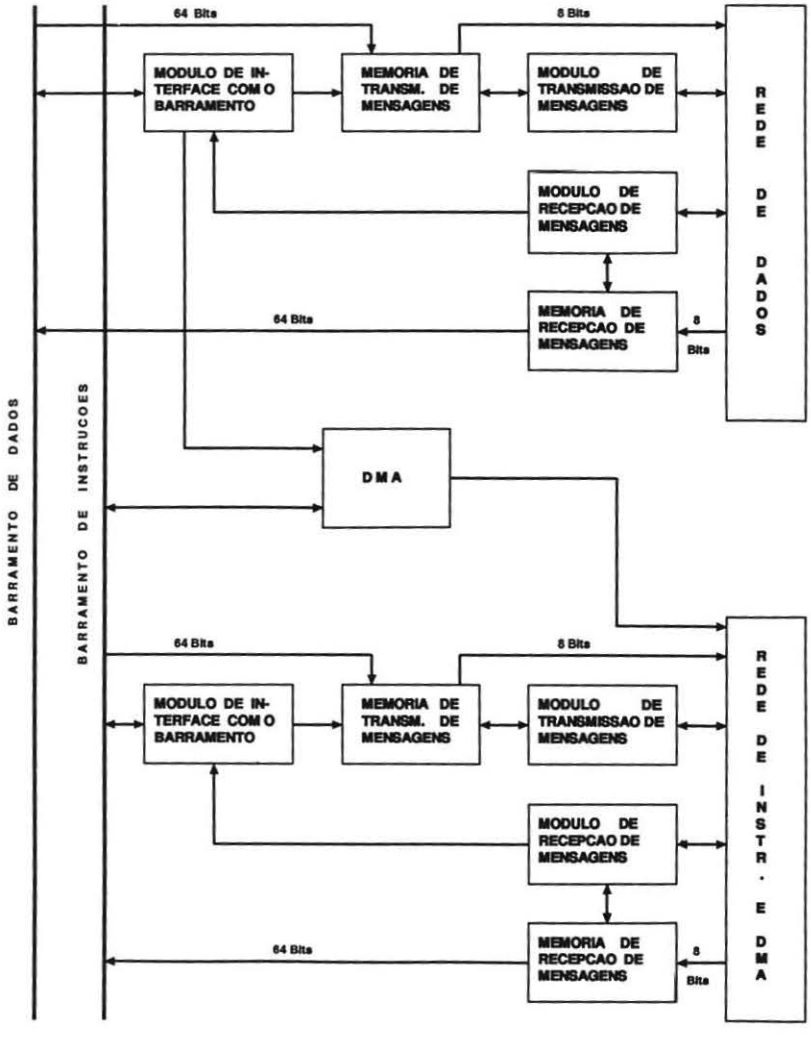

Figura 2: Arquitetura da Interface de Rede