# COMPUTADORES PARALELOS COM ARQUITETURA DE DUTOS

Eng. Carlos Alberto Sangiorgio.
Laboratório de Sistemas Integráveis.
Escola Politécnica da Universidade de São Paulo
Av. Prof. Luciano Gualberto, 158 - Trav. 3
CEP 05508 - São Paulo - SP.

RESUMO

Computadores paralelos com arquitetura de dutos vem sendo cada vez mais utilizados pois representam uma opção atraente economicamente e de implementação rápida tanto a nível de hardware como de software. Partes do projeto do computador multiprocessador MS8701 (LSI-USP) são aqui apresentadas, sendo analisadas diferentes topologias de duto que foram objeto de estudo durante a definição da arquitetura do computador.

ABSTRACT

Parallel computers based on buses have been increasingly used because they are a cheaper and easily implementable option, both at the hardware and software levels. Some parts of the design of the multiprocessor computer MS8701 (LSI-USP) is presented, and several bus topologies, that were studied during the architectural definition phase, are analised.

## 1. INTRODUÇÃO

Diversos fatores tem contribuido para o sucesso de computadores baseados em multiprocessamento. Entre eles destacam-se :

- O desenvolvimento cada vez maior de circuitos VLSI fazendo emergir uma nova geração de microprocessadores de elevado desempenho, permitindo assim que sistemas baseados em multi-microprocessadores alcancem uma relação custo/desempenho extremamente competitiva;

- A maior confiabilidade que se consegue atingir nos sistemas multiprocessadores em relação aos sistemas monoprocessador. Esta maior confiabilidade se deve à possibilidade de implementação de algoritmos de reconfiguração que exploram a redundância intrinseca dos sistemas multiprocessadores.

A utilização de computadores paralelos torna-se ainda mais viável para classes de aplicações que sejam intrinsecamente paralelas, como por exemplo processamento de imagem e sinais, cálculos sobre vetores e matrizes e sistemas "time-sharing" .

Este artigo tem como objetivo principal mostrar algumas topologias de conexão que foram objeto de análise durante a fase de definição de parte da arquitetura de um minissupercomputador (referenciado de agora em diante como MS8701) que está em fase de implementação no Laboratório de Sistemas Integráveis da USP. São apresentadas as topologias e alguns cálculos associados, principalmente no que diz respeito as bandas dos dutos de conexão e dos bancos de memória. Com os resultados destes cálculos procura-se justificar algumas opções seguidas e a topologia de conexão finalmente adotada.

Para que haja um melhor acompanhamento desta descrição (dada basicamente no ítem 3) é apresentada no ítem 2 uma visão resumida do hardware do MS8701, procurando mostrar as características e capacidades de cada módulo que o compõe, bem como a topologia na qual estes módulos estão conectados. No ítem 4 procura-se ressaltar a importância da

perfeita integração software/hardware para o bom desempenho de computadores paralelos. No ítem 5 são apresentadas algumas conclusões e finalmente no ítem "REFERÊNCIAS" é apresentada parte da bibliografia que foi utilizada como referência pelos projetistas do MS8701 durante a fase de definição de sua arquitetura.

## 2. DESCRIÇÃO DO MS8701 :

O MS8701 é um computador multiprocessador de propósito geral (pois deverá atender tanto aplicações intensivas em entrada e saída como intensivas em processamento) onde será instalado um sistema operacional compatível com o UNIX V Rel. 3, com características de multiprogramação, multiusuário e permitindo aplicações que envolvam processamento em tempo real.

Na versão 1.0 o sistema deverá apresentar as seguintes características:

-capacidade de processamento: até 100Mips e 12Mflops;

-capacidade de memória física: até 512MBytes;

-capacidade de armazenamento em memória de massa: até 100GBytes;

-número de terminais: até 256;

-suporte a redes: X.25 e ETHERNET.

O sistema foi projetado de forma modular, procurando-se assim torná-lo flexível o suficiente para seu aproveitamento em diversos ambientes. Os principais módulos que o compõe são descritos abaixo e estão mostrados na figura 1 :

I) Placa de Processamento Geral (PPG): Representam o núcleo de processamento do MS8701. Esta placa contém quatro processadores (com desempenho aproximado de 2,5Mips [1] e 300kflops [2] cada). Cada processador possue um sistema de memória cache (com 64kbytes) e uma memória rápida (com 256kbytes), formando um primeiro "núcleo de processamento". Cada placa possue portanto 4 destes "núcleos de processamento" que se comunicam através de uma memória compartilhada (com até 32Mbytes), formando assim um "cluster de processamento" (ver figura 1).

II) Computador de Memória de Massa (CMM): Este sistema é o responsável por todo o gerenciamento das operações de memória de massa. A opção pela centralização deste sistema não será aqui abordada e poderá ser encontrada em [12]. O CMM também possui um projeto modular permitindo diversas configurações. Estas configurações deverão ser escolhidas levando-se em conta principalmente os seguintes fatores:

- Número de placas de processamento geral presentes no sistema;

- Tipo de processamento que irá ser executado na máquina (intensivo em entrada e saída ou em processamento).

Em sua configuração máxima o CMM deverá ser capaz de fornecer 16Mbytes/s (de entrada e saída de memória de massa) e até oito dutos de conexão com dispositivos padrão SCSI.

III) Placa de Processamento de Comunicação (PPC): Responsável pelo gerenciamento do sistema de comunicação com os terminais e gerenciamento do sistema de redes. Este sistema de redes inclui duas redes locais (padrão ETHERNET) para conexão com terminais a caracter e estações gráficas.

IV) Placa de Gerenciamento de Sistema (PGS): Encarregada do controle do subsistema de interrupções e do armazenamento de algumas estruturas de dados globais.

Observando-se novamente a figura 1 pode-se notar a existência de diversos dutos conectando os subsistemas (arquitetura hierárquica de dutos). A justificativa para as soluções adotadas será apresentada no item 3 deste trabalho. As premissas adotadas para o desenvolvimento do MS8701 foram:

- manter constante a quantidade de dados capaz de ser fornecida aos processadores (banda), para obter o desempenho esperado;

- minimizar a complexidade e os custos de implementação do hardware;

- procurar simplificar o trabalho de instalação do software básico.

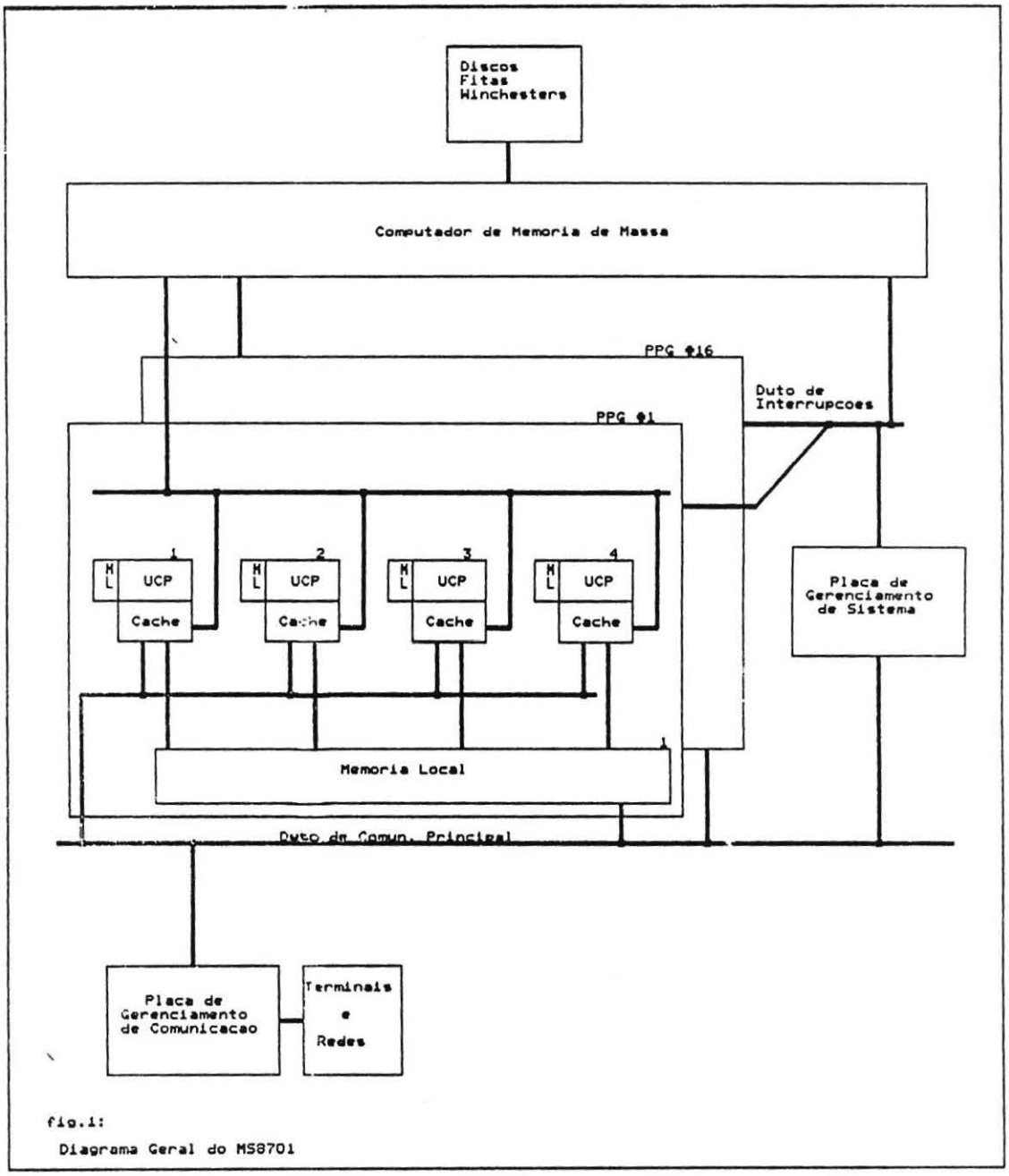


fig.1:
Diagrama Geral do MS8701

## 3. EVOLUÇÃO DA TOPOLOGIA DE HARDWARE

A seguir serão apresentadas quatro topologias de conexão por sistemas de duto que serão denominadas:

I) Sistema com memória global única.

II) Sistema com memória global única e um cache para cada processador.

III) Sistema com memória global única e um cache e uma memória local para cada processador.

IV) Sistema com memória local compartilhada e um cache e uma memória local para cada processador.

Para cada uma destas configurações serão calculadas as bandas necessárias para os diversos dutos e memórias. Nestes cálculos serão assumidos os seguintes dados:

A) Desempenho esperado de cada processador: 2,5Mips [1]. Este dado foi retirado de medidas de desempenho do processador MC68020 em sistemas anteriormente projetados pelo LSI.

B) Número médio de acessos à memória por instrução: 1,7 [1]. Calculado através de um conjunto médio de instruções geradas pelos compiladores para o MC68020 e do número de bytes necessários para cada uma destas instruções. Cada acesso representa a busca de quatro bytes.

C) Tempo de ciclo da memória: 250ns. Este valor foi retirado de medidas sobre o sistema de memória que será utilizado no MS8701. Para uma memória com 32bits de largura (que é a mesma largura dos dutos de dados do MS8701) tem-se a seguinte banda da memória (BWM):

BWM = [1 ciclo/250ns] *
* [4bytes/s] = 16Mbytes/s

D) Sistema de memória cache com 64kbytes e política de escrita "write-through", que irá propiciar uma taxa de acerto ("hit rate") médio nos ciclos de leitura de aproximadamente 90% [3].

E) Porcentagem de acessos de escrita em relação aos acessos totais: 20% [1]. Calculado através de um conjunto médio de instruções geradas pelos compiladores para o MC68020.

F) Todos os cálculos feitos pelo valor médio das grandezas. As grandezas envolvidas nos cálculos são de natureza aleatória e não determinística. Neste trabalho os cálculos são feitos levando-se em conta apenas o valor médio das grandezas (primeiro momento).

### 3.1. Sistema com Memória Global Unica.

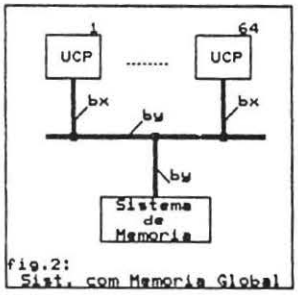

fig.2:
Sist. com Memoria Global

A figura 2 mostra este tipo de conexão. Nos dutos tipo "bx" tem-se a seguinte banda :

BW bx = (número de instruções por segundo) * (número de bytes por acesso) * (número de acessos por instrução)

BW bx = 2,5 * 1,7 * 4 =
= 17Mbytes/s.

Nos dutos tipo "by" tem-se a seguinte banda:

BW by = (número de dutos tipo "bx") * (banda dos dutos tipo "bx")

BW by = 64 * 17 = 1088Mbytes/s.

O número de bancos de memória intercalados ("interleaved") será então:

NM = (banda total do sistema) / (banda de uma memória).

NM = 1088 / 16 = 68 bancos.

Uma análise dos resultados obtidos deve levar em conta que um sistema com "NM" bancos de memória intercalados possue uma banda próxima a "NM" vezes um único banco somente para valores pequenos de "NM" (em geral "NM" menor ou igual a 4). No caso, o valor de "NM" igual a 68 não deverá ser atingido com soluções simples de intercalamento de memória. Uma opção alternativa seria a de aumentar a largura da memória de 32bits para por exemplo 128bits. Isto implicaria numa queda de "NM" para 17 bancos. Contudo os números apresentados ainda são restritivos tanto a nível de número de bancos de memória como a nível de dutos necessários para conexão entre os módulos.

3.2) Sistema com memória global única e um cache para cada processador.

Uma evolução da topologia anterior está em acrescentar um sistema de memória cache para cada processador. Isto irá fazer com que apenas os dados que não estejam presentes no cache sejam carregados da memória principal. A nova configuração é mostrada na figura 3.

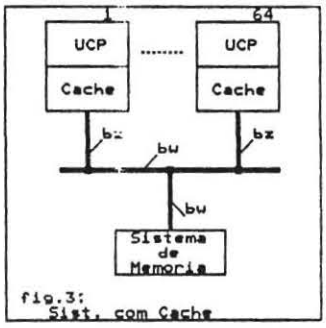

fig.3: Sist. com Cache

Nos dutos tipo "bz" tem-se a seguinte banda:

BW bz = {(% de ciclos de escrita) + [(% de ciclos de leitura) * (1 - "hitrate")]} * * (banda sem cache)

BW bz = {0,2 + [ 0,8 * 0,1 ] } * * 17 = 4,8 Mbytes/s.

Nos dutos tipo "bw" tem-se a seguinte banda :

BW bw = (número de dutos tipo "bz") * (banda dos dutos tipo "bz").

BW bw = 64 * 4,8 = 307 Mbytes/s.

O número de bancos de memória intercalados ("interleaved") será então:

NM = (banda total do sistema) / / (banda de uma memória).

NM = 307 / 16 = 20 bancos.

Uma análise dos resultados mostra que o número de bancos de memória tornou-se bem mais razoável do que no item anterior.

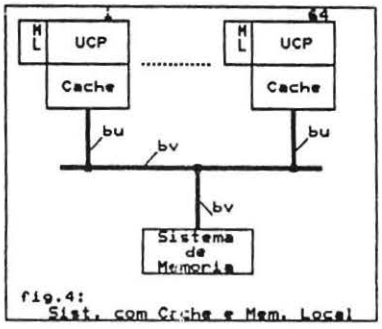

fig.4: Sist. com Cache e Mem. Local

3.3) Sistema com memória global única e um cache e uma memória local para cada processador.

A próxima evolução está baseada no fato de que em média 20% a 30% dos acessos feitos pelos processadores são de leitura de código do sistema operacional. Logo, acrescentando-se uma memória local (uma para cada processador) que armazene o sistema operacional, as bandas totais dos dutos bem como o número de bancos de memória poderão ser reduzidos. A nova configuração é mostrada na figura 4.

Nos dutos tipo "bu" tem-se a seguinte banda :

BW bu = [1 - (% acessos a código do S.O.)] * (banda sem memória local)

BW bu = [1 - 0,2] * 4,8 =
= 3,8Mbytes/s.

Nos dutos tipo "bv" tem-se a seguinte banda:

BW bv = (número de dutos tipo "bu") * (banda dos dutos tipo "bu")

BW bv = 64 * 3,8 = 243 Mbytes/s.

O número de bancos de memória intercalados ("interleaved") será então:

NM = (banda total do sistema) /
/ (banda de uma memória)

NM = 243 / 16 = 16 bancos.

3.4. Sistema com memória local compartilhada e um cache e uma memória local para cada processador.

A topologia final apresentada é também a adotada no MS8701 e está mostrada na figura 5. Para se chegar a esta configuração os seguintes fatores foram levados em consideração :

- O " back plane " de conexão dos módulos deveria ser de no máximo 384 linhas (4 conectores EURO de 96 pinos).

- Número máximo de placas de processamento geral igual a dezesseis com quatro processadores em cada placa. Este número de quatro processadores é indicado como um valor ótimo em diversas simulações e artigos de avaliação de desempenho de sistemas multiprocessadores [4],[5],[13].

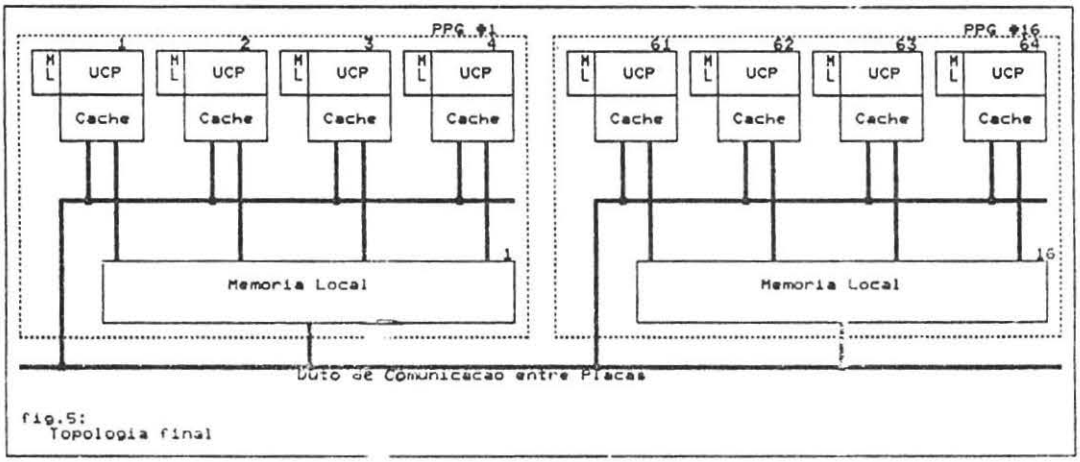


fig.5:
Topologia final

Como pode ser observado pela figura 5, cada uma das placas de processamento geral possue quatro "núcleos de processamento" (ver item 2) que se comunicam através de uma memória local (de até 32Mbytes). Esta memória é também acessível por todas as outras placas do sistema através do duto de comunicação entre placas. Logo, quando os processadores de determinada placa acessarem dados presentes em sua memória local, o duto de comunicação entre placas não sofrerá sobrecarga.

Nesta configuração um ponto importante a ser analisado é o número máximo de acessos que uma placa de processamento geral necessita executar na memória de outra placa.

Seja "P_ext" a taxa relativa de ciclos que cada processador deve realizar através do duto principal para acessar dados presentes na memória de outra placa. Evidentemente se P_ext = 1 implica numa banda do duto principal igual a 243 Mbytes/s (calculada no item anterior). Quando P_ext diminui a banda do duto principal pode também ser diminuida. Em particular, no MS8701 foram adotados dois dutos de 15 Mbytes/s cada totalizando 30 Mbytes/s de banda para comunicação entre as placas de processamento geral. Supondo-se a carga distribuida equitativamente entre as placas pode-se concluir o seguinte:

Banda máxima do duto principal que cada processador pode utilizar:

BW_1 = (banda total dos dutos principais) / (número de processadores)

BW_1 = 30 / 64 = 0,48 Mbytes/s.

-Máximo P_ext para não saturar os dutos principais :

P_ext_max = (banda total dos dutos principais) / (banda total dos processadores)

P_ext_max = 30 / 243 = 12 %

Ou seja, os dutos de comunicação entre as placas de processamento geral saturam para uma taxa de acessos superior a 12% dos acessos totais externos ao cache e à memória local do processador.

## 4. ANALISE DA SOLUÇÃO ADOTADA

A análise feita no item 3 merece alguns comentários e considerações. Em primeiro lugar, foi suposto um fator de utilização para as memórias de 100%, o que não deve ser feito uma vez que os processos (que serão processados no sistema) são de natureza aleatória e não determinística [6] (o que foi feito é um cálculo pelo valor médio). Logo o número de bancos de memória calculado está subestimado como já foi citado anteriormente.

Em segundo lugar, sistemas com vários dutos e memórias com múltiplos portos de acesso requerem cuidados especiais principalmente no projeto dos árbitros [7]. Em particular, no MS8701 foi adotado um sistema de arbitração paralela que realiza este processo durante o ciclo de acesso à memória por parte de outro processador. Como o tempo de arbitração é menor do que o tempo de ciclo da memória, este tempo (de arbitração) pode ser desprezado supondo-se o sistema em regime, com vários pedidos.

Outra consideração que deve ser feita é quanto ao número de dutos necessários para atingir as bandas calculadas. Sistemas de múltiplos dutos foram estudados [8] e não serão abordados neste artigo. Apenas seria bom lembrar que sistemas baseados em comunicação por fibra ótica podem representar um grande avanço nessa área.

Deve também ser notado que, em computadores paralelos, um dos principais fatores que irá determinar seu desempenho é o casamento hardware/software. Portanto, para cada topologia de hardware encontrada tem-se uma determinada complexidade no software envolvido. Assim, por exemplo, portar o S.O. UNIX [9],[10] para sistemas multiprocessador é trabalho que depende da arquitetura do hardware envolvido. Das várias implementações já realizadas sabe-se que arquiteturas do tipo fortemente acopladas são as que oferecem maior facilidade, enquanto arquiteturas do tipo fracamente acopladas (por exemplo hipercubo) merecem implementação mais cuidadosa [8]. Um ponto que mostra bem este relacionamento hardware/software é o sistema de memória cache. A inclusão de sistemas

cache reduzem em muito o número de bancos de memória bem como a banda dos dutos, reduzindo assim o custo do hardware do sistema (como visto nos cálculos anteriores). Em contrapartida, torna-se necessário ao software manter a coerência entre os dados dos diversos caches existentes no sistema. Mecanismos de hardware como "bus watch" e "cache disable" podem ser implementados para facilitar a tarefa do software. No caso do MS8701 um ponto vital na determinação de seu desempenho (e que novamente ressalta a necessidade do bom casamento hardware/software) está em se procurar concentrar os dados e o código dos processos de determinada placa de processamento geral dentro de sua própria memória local, evitando-se assim a saturação do duto de comunicações entre placas (ver 3.4).

Outros problemas que aparecem, e que serão apenas citados, são o balanceamento de cargas no sistema, a distribuição de carga para evitar contenção nos dutos, mecanismos de acesso a dados globais compartilhados, a comunicação com o sistema de armazenamento de massa e entrada e saída [9],[10],[11].

Finalmente deve-se ressaltar que vários sistemas multiprocessados baseados em topologia de dutos foram desenvolvidos (por exemplo o Balance, o Counterpoint System 19 [9], o C.mmp [4], o Convergent Techonologie Megaframe [10] e o MS8701), bem como diversos sistemas operacionais (desenvolvidos ou adaptados para este tipo de máquina), representando assim, uma ótima bibliografia para o início de desenvolvimento de projetos nesta área.

## 5. CONCLUSÕES

A opção por uma arquitetura de dutos na implementação de computadores paralelos pode ser atraente para sistemas com um número de processadores que varie de algumas dezenas até poucas centenas, pois além de permitir uma implementação de hardware com componentes de mercado, é o que traz menores modificações do software em relação ao sistema monoprocessador. Dentro desta opção de dutos, foram apresentadas algumas topologias possíveis. Cabe ao projetista escolher a mais adequada ao seu problema, levando-se em conta diversos fatores, como o número de processadores, desempenho de cada processador, características das memórias, espaço físico e custo.

## AGRADECIMENTOS

O projeto MS8701 está sendo realizado no Laboratório de Sistemas Integráveis da USP e conta com o apoio financeiro da FINEP e do CNPq.

## REFERÊNCIAS

[1] MC68020 32-Bit Microprocessor User's Manual. Prentice - Hall, Inc., Englewood Cliffs NJ, 1984.

[2] MC68881 Floating Point Coprocessor User's Manual. Prentice- Hall, Inc., Englewood Cliffs NJ, 1984.

[3] Clipper 32-Bit Microprocessor, Introduction to the Clipper Architecture, Fairchild Semiconductor Corporation, USA, June 1987.

[4] Computer Architecture and Parallel Processing, Kai Hwang e Fayé A. Briggs. McGraw-Hill International Edition, New York, NY, 1987.

[5] Avaliação de Desempenho de um Sistema Multiprocessador, Zelenovsky, Ricardo, Instituto Militar de Engenharia, Rio de Janeiro, RJ, 1988.

[6] Computer Systems Performance Modeling, Charles H. Sauer e K. Mani Chandy, Prentice-Hall, Inc., Englewood Cliffs, NJ, 1981.

[7] Asynchronous Arbiter Module, R.C.Pearce, J.A.Field e W.D.Litlle, IEEE Trans. Computers, September 1975.

[8] Multiple Bus Architecture, T.N. Mudge, J.P.Hayes e D.C.Winsor, University of Michigan, Computer, June 1987.

[9] Application Dictates Your Choice of a Multiprocessor Model, J.Kent Peacock, EDN, June 1987.

[10] Punching up UNIX Performance, Edward L.Patriquin Jr e Stephen Ricossa, Computer Design, July 1983.

[11] Synchronizing Multiprocessor Access to Shared Operating System Data Structures, Jason Gait, Computer Systems Science and Engineering, October 1987.

[12] Projeto de um Subsistema de Memória de Massa para um Computador de Arquitetura Paralela, Prado, Claudio Almeida, LSI-DEE-EPUSP, Agosto de 1988.

[13] Simulação da Placa de Processamento Geral do MS8701 , Sangiorgio, Carlos Alberto, LSI-DEE-EPUSP, Maio de 1988.