UMA PROPOSTA DE ARQUITETURA DE EIS PARA O MULTIPROCESSADOR MULTIPLUS

Sidney de Castro Oliveira

TESE SUBMETIDA AO CORPO DOCENTE DA COORDENAÇÃO DOS PROGRAMAS DE PÓS-GRADUAÇÃO EM ENGENHARIA DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA DE SISTEMAS E COMPUTAÇÃO.

Aprovada por:

Júlio Salek Aude, Ph.D.
(Presidente)

Edil Severiano Tavares Fernandes, Ph.D.

Adriano Joaquim de Oliveira Cruz, Ph.D.

RIO DE JANEIRO, RJ - BRASIL

FEVEREIRO DE 1992

OLIVEIRA, SIDNEY DE CASTRO

Uma Proposta de Arquitetura de E/S para o Multiprocessador Multiplus
Rio de Janeiro, 1992.
viii, 84 p., 29,7 cm (UFRJ/COPPE, M.Sc., Engenharia de Sistemas e Computação, 1992)

Tese - Universidade Federal do Rio de Janeiro, COPPE.
1 - Arquitetura de E/S 2 - Arquitetura de Computadores
I. COPPE/UFRJ II. Título (série)

Para minha esposa, Paula,
por toda compreensão e apoio recebidos,
indispensáveis ao sucesso desta jornada.

AGRADECIMENTOS

Ao Professor Júlio Salek Aude, pela valiosa orientação recebida.

À minha Esposa, pela arte final das figuras contidas neste trabalho.

Ao colega Norival Ribeiro Figueira, pelas sugestões dadas na definição da arquitetura do Processador de E/S.

Ao colega Alexandre Malheiros Meslin, pela concessão do simulador utilizado na avaliação da arquitetura proposta.

Ao NCE/UFRJ, pelo suporte humano e material para realização deste trabalho.

Ao CNPq e à FINEP, pelo apoio financeiro ao Projeto Multiplus, tornando possível a elaboração desta tese.

Resumo da Tese apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

UMA PROPOSTA DE ARQUITETURA DE E/S PARA O MULTIPROCESSADOR MULTIPLUS

Sidney de Castro Oliveira

Fevereiro, 1992

Orientados: Júlio Salek Aude
Programa: Engenharia de Sistemas e Computação

RESUMO

Este trabalho apresenta a definição de uma arquitetura de E/S para uma máquina paralela de alto desempenho, o Multiplus. Inicialmente são relacionados os problemas inerentes à entrada e saída de dados, e os fatores que fazem do correto dimensionamento de um subsistema de E/S um ponto fundamental para o alto desempenho de um sistema computacional. Em seguida, é apresentado o Multiplus, juntamente com a proposta de arquitetura para seu subsistema de E/S. Por fim, é feita uma avaliação, através de simulações e análises quantitativas, do desempenho da arquitetura proposta e do impacto das operações de E/S na operacionalidade do Multiplus.

Abstract of the Thesis presented to COPPE/UFRJ as partia1 fulfillment of the requirements for the degree of Master of Science (M.Sc.)

A PROPOSAL OF AN I/O SYSTEM ARCHITECTURE FOR THE MULTIPLUS MULTIPROCESSOR

Sidney de Castro Oliveira

February, 1992

Thesis Supervisor: Júlio Salek Aude
Department: Computes Science and Systems Engineering

ABSTRACT

This thesis presents the definition of an 1/0 System Arquitecture for a high-performance parallel machine, the MULTIPLUS. Initially, intrinsic problems of data 1/0 and the aspects that make the design of the 1/0 Subsystem a fundamental issue for the high performance of a computational system are discussed. Following, the MULTIPLUS System is presented and an architecture for its 1/0 Subsystem is proposed. Finally, the performance of the proposed architecture and the impact of 1/0 operations on the MULTIPLUS computational power are evaluated through simulation and quantitative analysis.

ÍNDICE

CAPÍTULO I

INTRODUÇÃO

A velocidade dos dispositivos eletrônicos, especialmente o microprocessador, tem aumentado acentuadamente nos últimos anos. A utilização destes dispositivos em arquiteturas avançadas resulta em computadores de alto desempenho. Diversas áreas de aplicação desta classe de computadores demandam grandes quantidades de dados, como por exemplo, o processarnento científico em geral: meteorologia, análise estrutural, interpretação de imagens, etc. Estes dados são mantidos em dispositivos de armazenamento e, periodicamente, intercambiados com os elementos processadores, conforme sua necessidade. A rapidez neste intercâmbio de dados é fundamental para o alto desempenho destes computadores.

A parte da arquitetura do computador responsável pelo armazenamento em massa dos dados e seu intercâmbio com os elementos processadores é chamado Subsistema de E/S. Ele engloba os dispositivos de armazenamento, processadores de E/S e canais de comunicação. Para que este subsistema atenda, satisfatoriamente, a demanda de dados dos elementos processadores, é necessário um equilíbrio de desempenho entre os dois, ou seja, que o desenvolvimento tecnológico do subsistema de E/S seja similar ao dos elementos processadores.

Existem basicamente duas estratégias para o aumento do desempenho dos subsistemas de E/S: o aumento da velocidade de seus dispositivos e componentes e a utilização de arquiteturas mais eficientes. A primeira delas esbarra em limitações físicas decorrentes da tecnologia empregada nos dispositivos de armazenamento. Desta forma, a descoberta de arquiteturas de E/S mais avangadas e eficientes tem sido o principal instrumento para o aumento do desempenho dos subsistemas de E/S, sendo alvo de muita pesquisa no projeto de computadores de alto desempenho.

O objetivo deste trabalho é definir um subsistema de E/S que atenda às necessidades do Multiplus. O Multiplus é um sistema do tipo multiprocessador com memória global compartilhada em desenvolvimento no NCE/UFRJ. Sua

proposta é obter, através de uma arquitetura modular, um sistema paralelo de alto desempenho com o uso de microprocessadores RISC.

O capítulo II faz um estudo dos problemas relacionados à entrada e saída de dados em sistemas computacionais e dos fatores que fazem do subsistema de E/S um ponto fundamental para seu alto desempenho. Este capítulo procura identificar e caracterizar estes fatores, iniciando com uma comparação da evolução tecnológica sofrida pelos componentes básicos de um sistema computacional. Em seguida, faz uma avaliação do desempenho dos subsistemas de E/S, tanto em função das próprias características dos dispositivos de armazenamento, quanto em função do tipo de aplicação a que eles se destinam. Por último, são descritas algumas alternativas capazes de melhorar, futuramente, o desempenho dos dispositivos de armazenamento de massa.

O Capítulo III expõe algumas opções para configuração de um subsistema de E/S em função das diversas formas de se associar os processadores de E/S aos nós de processarnento. Dependendo da forma de associação e do tipo de aplicação a que o sistema se destina, determina-se o dirnensionamento ótimo do s~ibsistemade E/S. O mau dimensionamento deste subsistema pode representar um gargalo para todo o sistema computacional.

O capítulo IV inicia a definição do subsistema de E/S do Multiplus. Ele procura definir uma arquitetura para este subsistema de forma satisfatória às diversas configurações que o Multiplus pode assumir. Primeiramente, o capítulo apresenta a arquitetura geral do Multiplus, descrevendo suas características principais, suas configurações possíveis e o ambiente de processamento a que ele se destina. Em seguida, procura-se mostrar a evolução das idéias que resultaram na definição do subsistema de E/S do Multiplus. Por fim, é mostrada a arquitetura proposta para este subsistema.

O capítulo V prossegue na definição do subsistema de E/S do Multiplus, abordando agora o Processador de E/S. Inicialmente são discutidas as principais questões inerentes à arquitetura de um processador de E/S, identificando, em cada uma delas, a melhor alternativa para o Multiplus. Entretanto, face às restrições impostas à implementação do Multiplus, nem sempre a melhor alternativa pode ser adotada. Assim, este capítulo, além de expor a arquitetura proposta para o Processador de E/S do Multiplus, busca avaliar o impacto das restrições impostas sobre seu desempenho.

O capítulo VI apresenta os resultados obtidos através de simulações e análises quantitativas que pudessem retratar a operacionalidade do subsistema de E/S dentro do contexto do Multiplus. Os principais aspectos avaliados foram: o comportamento do subsistema quando submetido a uma carga de E/S e o impacto das operações de E/S no desempenho dos elementos processadores do Multiplus.

Por último, o capítulo VII reúne os pontos mais importantes e as principais conclusões deste trabalho, bem como algumas sugestões para a sua continuidade.

CAPÍTULO II

PROBLEMAS RELACIONADOS A ENTRADA E SAÍDA

O avanço da tecnologia de microeletrônica nos últimos anos tem resultado em dispositivos eletrônicos cada vez mais velozes e poderosos. A área de microprocessadores é uma das que mais tem se beneficiado deste avango. Atualmente, a capacidade de processamento de informação dos microprocessadores é bastante elevada e crescente a cada ano. A exploração do paralelismo em arquiteturas computacionais aumenta muito a demanda por instrugões e dados pelos microprocessadores. Estas instruções e dados estão presentes na memória do sistema, seja na memória principal, que é menor e mais rápida, seja na memória secundária, que é maior e mais lenta. A memória principal é formada pelos *chips* de memória e a secundária, também conhecida como memória de massa, pelos dispositivos de armazenamento, tais como discos e fitas magnéticas. É a memória secundária que fornece dados para a memória principal. Entretanto, enquanto a tecnologia de microeletrônica avança, proporcionando microprocessadoresmais velozes e dispositivos de memória cada vez mais densos e baratos, a tecnologia dos dispositivos de armazenamento de massa não progride em igual proporção, criando um desequilíbrio entre a capacidade de processamento de informação dos sistemas computacionais e sua demanda por entrada e saída de dados. A minimização deste desequilíírio, através de novas arquiteturas, tem sido bastante investigada pelos projetistas de computadores de alto desempenho.

Este capítulo resulta de um estudo dos problemas relacionados à entrada e saída de dados através de dispositivos de armazenamento de massa em arquiteturas computacionais de alto desempenho. Ele está dividido, basicamente, em três partes. Na primeira delas procurou-se indentificar e caracterizar os fatores que fazem do processamento de E/S um ponto crítico para o bom desempenho do sistema. Para isto, comparou-se a evolução tecnológica nos últimos anos dos três componentes básicos de um sistema computacional: a unidade processadora, a memória principal e a memória secundária. A segunda parte avalia o desempenho de um subsistema de E/S tanto em função das próprias características dos dispositivos de armazenamento, quanto em função do tipo de aplicação a que ele se destina. Por último, são descritas algumas alternativas promissoras para os

dispositivos de armazenamento, capazes de torná-los, futuramente, mais eficientes quanto ao desempenho.

II.1 - Caracterização do Gargalo do Processamento de E/S

Para melhor entender a situação atual da tecnologia, que determina, hoje, o desempenho da unidade processadora, da memória principal e dos dispositivos de entrada e saída que compõem a memória secundária, será feita uma breve discussão sobre a evolução tecnológica de cada um deles nos últimos anos. Esta discussão permitirá identificar os fatores que resultaram num desequilíbrio entre a capacidade de processamento de informação das unidades processadoras e a capacidade de fornecimento de dados da memória secundária.

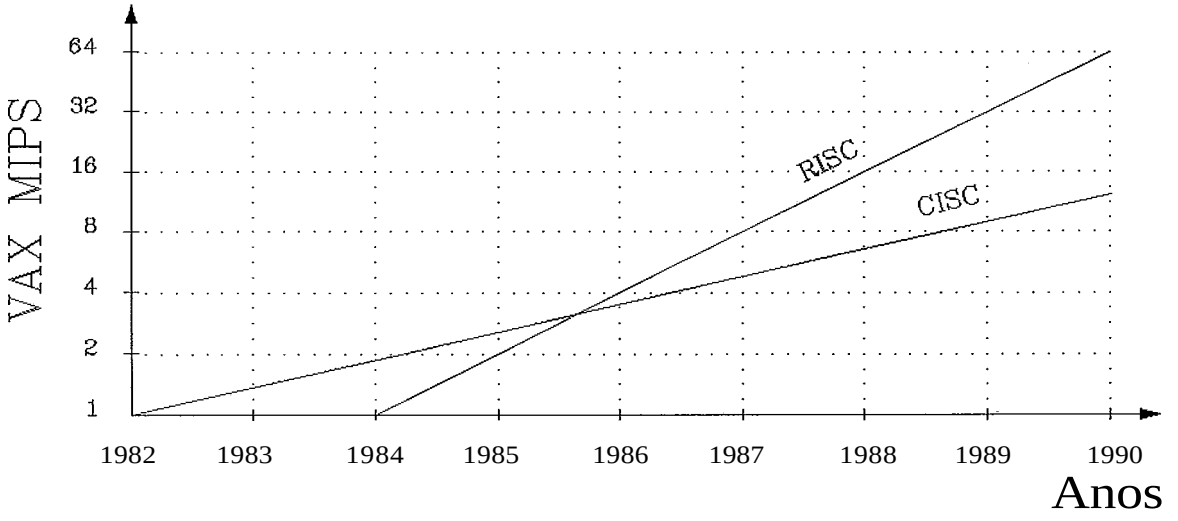

FIGURA 1: Evolução da capacidade de processamento dos microprocessadores Intel ao longo dos anos.

A unidade processadora é composta principalmente por microprocessadores. Estes, tiveram um desenvolvimento acentuado na última década. Sem dúvida, foram os que mais se desenvolveram. Atualmente é possível encontrar microprocessadores, tais como o Intel *i860*, com uma capacidade de processamento similar ao do primeiro computador da linha CRAY, aproximadamente 33 MIPS escalar, e com um custo muito reduzido. Usando a própria família de microprocessadores Intel como exemplo, pode-se observar que a capacidade de processamento vem dobrando a cada 2,25 anos como ilustrado na figura 1. Considerando o surgimento comercial dos microprocessadores RISC em

1984, nota-se uma taxa de crescimento ainda maior, com a capacidade de processamento dobrando a cada ano [KATZ89].

Para suportar este aumento de demanda por instruções e dados ocasionado pelo avanço tecnológico dos microprocessadores, o sistema de memória teve que se tornar maior e mais rápido. De forma empírica, pode-se dizer que cada instrução por segundo do microprocessador requer um byte de memória principal, ou seja, um mega byfe para cada MIPS [KATZ89]. Desta forma, a capacidade dos *chips* de memória deve acompanhar a taxa de crescimento da velocidade de processamento dos microprocessadores. Similarmente, fazendo-se uma analise da evolução da capacidade de armazenamento dos *chips* de memória dinâmica (DRAM) no decorrer dos anos, pode-se observar que ela vem, aproximadamente, quadruplicando a cada 3 anos nos últimos vinte anos. O gráfico da figura 2 ilustra esta tendência. Aliado a isto, houve urna queda muito acentuada no preço destes *chips* de memória. Apesar de estarem mais densos, o preço por mega byte foi muito reduzido. Isto permitiu que, apesar da densidade dos *chips* de memória ter crescido a uma taxa até inferior a da velocidade de processarnento dos microprocessadores, alguns sistemas se configurassem com até três mega bytes para cada MIPS.

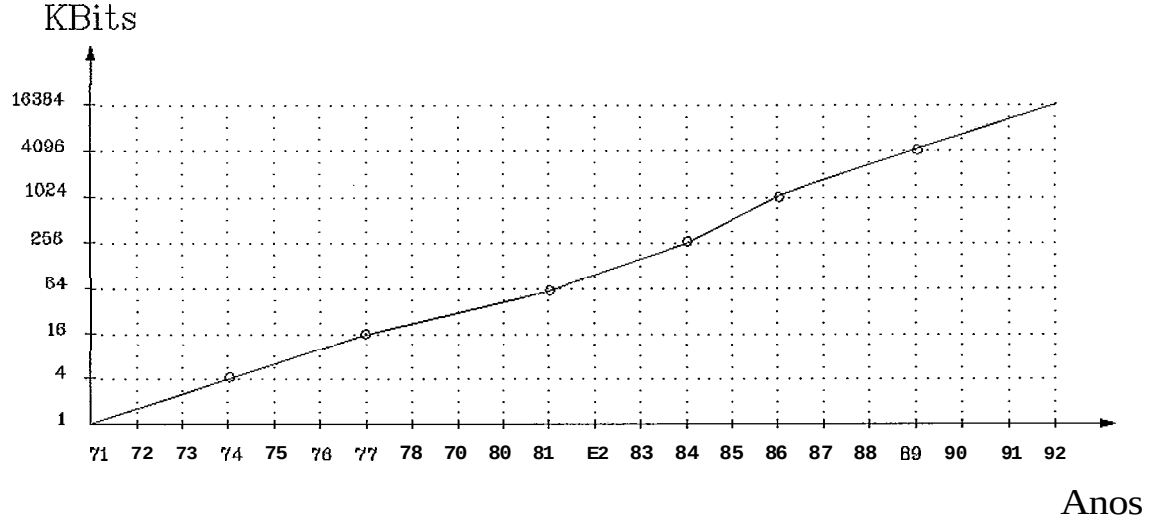

FIGURA 2: Evolução da densidade dos chips de memória ao longo dos anos.

Mas nem todas as instruções e dados requeridos pelo microprocessador estão presentes na memória principal do sistema. Muitas vezes é necessário buscar

estas informações na memória secundária. Para manter todo o sistema em equilíbrio, é fundamental que o desempenho da memória secundária acompanhe o aumento de desempenho das outras partes do sistema. O principal elemento da memória secundária é o disco magnético. Para se medir o avanço tecnológico deste componente, considera-se a quantidade de *bits* armazenados por polegada quadrada, isto é, o número de *bits* por polegada em uma trilha vezes o número de trilhas por polegada. Como exemplo, a figura 3 mostra a evolução da densidade de *bits* dos discos magnéticos que equiparam os computadores IBM nas últimas duas décadas [KATZ89]. Pode-se observar que esta evolução tem permitido dobrar a capacidade de armazenamento somente a cada 3 anos. Com relação ao custo, tanto quanto os *chips* de memória, o preço por mega *byte* dos discos magnéticos reduziu-se acentuadamente neste mesmo período.

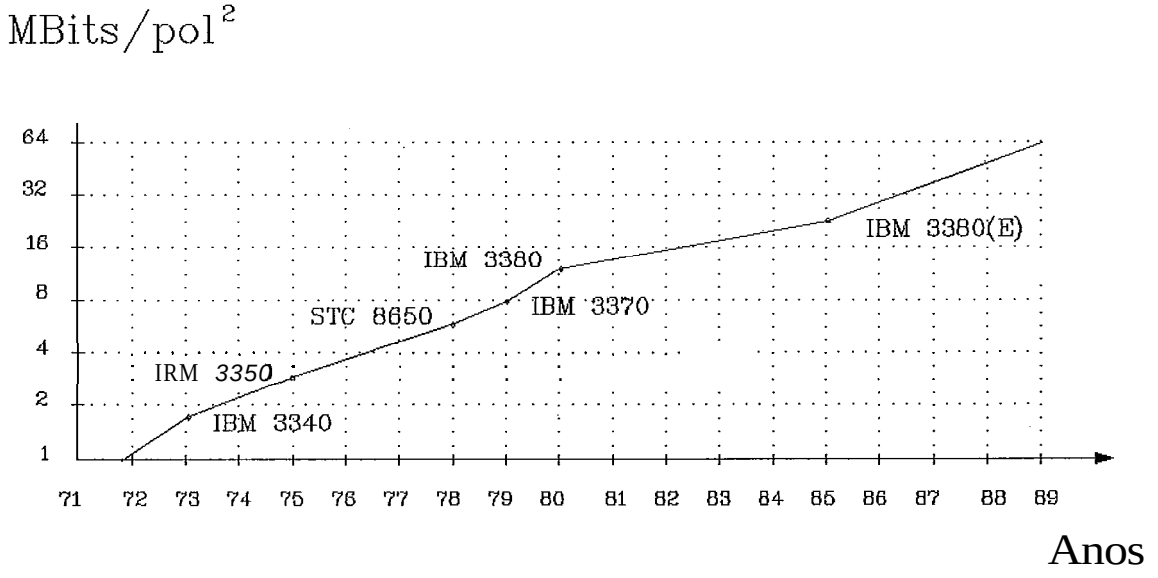

FIGURA 3: Evolução da densidade do meio magnético dos discos IBM ao longo dos anos.

Para comparar o desenvolvimento tecnológico entre os microprocessadores, a memória principal e a memória secundária, pode-se considerar que a velocidade dos microprocessadores vem dobrando a cada ano, que a densidade dos *chips* de memória vem quadruplicando a cada três anos, e que a densidade dos discos magnéticos vem dobrando a cada três anos. A figura 4 ilustra graficamente esta situação. Pode-se notar que os últimos dez anos são suficientes para ocasionar uma grande discrepância entre o desenvolvimento tecnológico dos microprocessadores, tanto em relação à memória principal quanto

à memória secundária, principalmente em relagão a esta última. Pela situação atual da tecnologia, esta discrepância tende a aumentar, não havendo motivos para se acreditar numa reversão desta tendência.

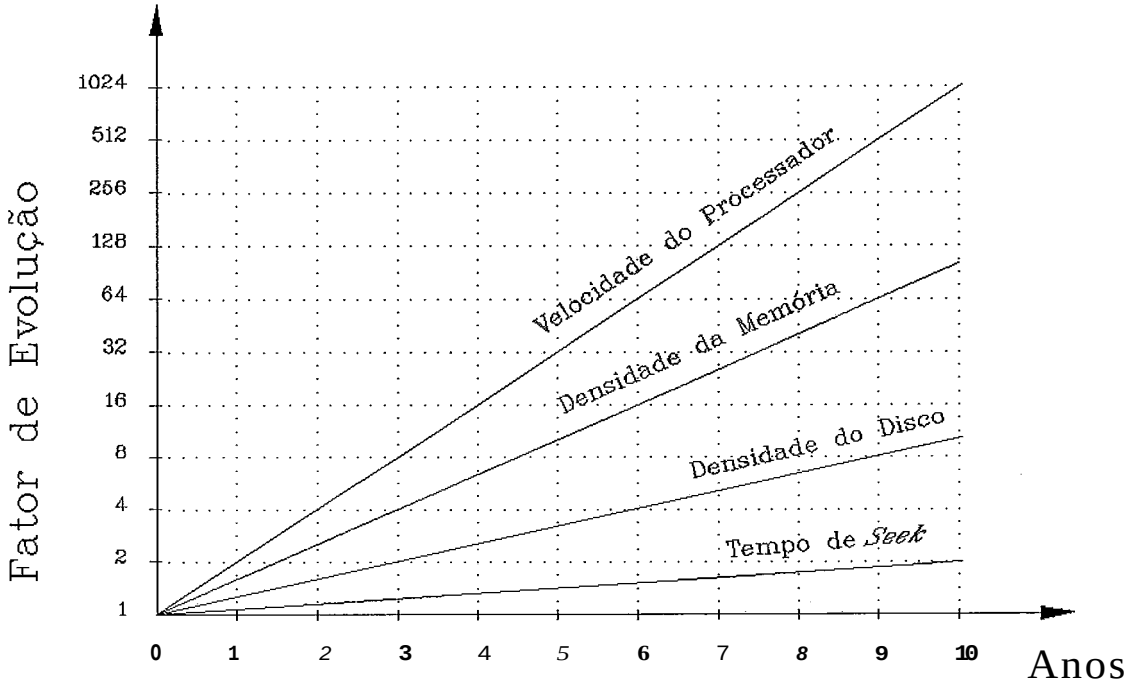

FIGURA 4: Comparagão entre o desenvolvimento tecnológico dos principais elementos de um sistema computacional.

Por outro lado, a densidade ou a capacidade de armazenamento tanto da memória principal quanto da memória secundária não é a ímica de suas características que precisa acompanhar o desenvolvimento tecnológico dos microprocessadores para manter o equilíbrio do sistema. Mesmo porque, quantidade de memória é um fator que sempre pode crescer pela simples interligação de dispositivos adicionais. O desenvolvimento tecnológico dos microprocessadores está relacionado ao número de instruções por segundo que eles são capazes de executar, isto é à sua velocidade de processamento. Desta forma, um outro fator até mais importante para o equilíbrio do sistema é a velocidade com que as instruções e dados requisitados pelos microprocessadores chegam até eles, ou seja, quão rápido os dispositivos de memória conseguem responder a um pedido de informação. O avanço tecnológico desta velocidade de resposta deve ser tanto para a memória principal quanto para a memória secundária.

A memória principal, por ser da ordem de grandeza de algumas dezenas de mega bytes, é geralmente implementada com chips de memória dinâmica (DRAMs). Estes dispositivos, como visto anteriormente, tiveram um progresso muito grande em sua densidade de bits, entretanto, sua velocidade ou tempo de acesso não se desenvolveu em proporção semelhante, evidenciando um desequilíbrio com a velocidade dos microprocessadores. Apesar disto, alguns fatores tem permitido à memória principal contornar satisfatoriamente esta defasagem:

- Interleaving de memória
- Técnica de caching
- Desenvolvimento tecnológico das SRAMs

A técnica de interleaving consiste em dividir a memória principal em vários bancos, sendo que o acesso a cada banco de memória é mutuamente exclusivo e independente dos demais. Assim, pode-se organizar a memória principal de forma a que seus endereços consecutivos se situem em bancos de memória distintos. Isto permite a busca antecipada de dados ou instruções contidos em endereços subsequentes concorrentemente com o acesso vigente do microprocessador. Cabe ressaltar que o bom desempenho desta técnica se baseia na sequencialidade dos acessos do microprocessador.

Na técnica de caching, um *buffer* pequeno e muito rápido é colocado entre o microprocessador e a memória principal. Este *buffer*, também chamado de memória cache, contém cópias de informações contidas na memória principal. Isto permite, pelo princípio da localidade de referência dos programas computacionais, que uma grande parte das informações requeridas pelos microprocessadores durante a execução de uma tarefa já estejam presentes neste *buffer*, possibilitando uma resposta muito mais rápida. Este *buffer* é implementado com chips de memória estática (SRAMs), cuja velocidade de acesso tem dobrado a cada dois anos nos últimos anos. Desta forma, pode-se construir sistemas de memória principal extensos e muito rápidos, capazes de fornecer as informagões requisitadas na velocidade máxima dos microprocessadores, desde que a taxa de acerto no *buffer*, também conhecida como taxa de hit na memória cache, seja alta.

Para a memória secundária a situação é bem diferente. Dos diversos dispositivos que compõem a memória secundária, a maior carga de trabalho recai sobre os discos magnéticos. Por ser um elemento eletromecânico, seu desempenho

está condicionado à evolução de uma tecnologia distinta àquela dos microprocessadores, ou seja, não mais à tecnologia de microeletrônica. Constata-se que o desempenho dos discos magnéticos cresceu apenas modestamente nas últimas duas décadas. Os principais elementos que medem a velocidade dos discos são:

- Tempo de seek
- Latência rotacional

Analisando as características dos discos magnéticos IBM citados na figura 3, pode-se verificar que o tempo de seek, necessário para se mover a cabega do disco, tem diminuído muito lentamente no decorrer dos anos, na ordem de 7% ao ano [KATZ89]. O tempo de latência rotacional é proporcional à velocidade angular do disco, isto é, ao seu número de rotações por minuto. Entretanto, esta velocidade não tem se alterado há muito tempo, mantendo constante a latência rotacional dos discos magnéticos ao longo destes últimos anos. A próxima seção aborda com mais detalhes as características e os elementos que determinam o desempenho dos discos magnéticos.

Assim, pode-se verificar que o desenvolvimento tecnológico no setor de informática ocorreu de forma desigual entre os principais elementos de um computador: o microprocessador, a memória principal e a memória secundária. Basicamente, houve um progresso muito acentuado na área de microeletrônica, possibilitando o surgimento de chips muito mais poderosos, determinando as características atuais dos microprocessadores e da memória principal. Como a evolução da memória secundária dependeu de um progresso eletromecânico de seu principal componente, o disco magnético, ou do surgimento de novos dispositivos de armazenamento, fatos que não ocorreram ou que não corresponderarn às expectativas, surgiu um gap de desempenho entre a memória secundária e os microprocessadores. Este *gap* faz com que a definigão de subsistemas de E/S deva ser bastante cuidadosa, a fim de se evitar que os dispositivos de armazenamento de massa tornem-se um gargalo para o sistema computacional. A busca de soluções para um maior equilíbrio no desempenho dos microprocessadores e das memórias principal e secundária é um dos desafios atuais dos projetistas de arquiteturas de computadores, principalmente as de alto desempenho.

11.2 - Caracterização do Desempenho dos Subsistemas de E/S

O desempenho de um subsistema de E/S está fortemente ligado ao tipo de processamento de entrada e saída a que ele está associado. Este processamento é função do tipo de aplicação a que o sistema, como um todo, se destina. Para se avaliar o impacto do *gap* de desempenho entre os microprocessadores e a memória secundária na *performance* global de um subsistema de E/S, é necessário saber como ele se comporta quando submetido a uma carga de trabalho de E/S. Isto significa identificar as características desta carga de trabalho e avaliar como se distribuem as frações de tempo durante sua execução frente às diversas tarefas que a compreendem. Um subsistema de E/S engloba um conjunto de dispositivos de armazenamento e processadores de E/S. Estes processadores são os responsáveis pelo gerenciamento das operações de E/S e pela ligação entre os dispositivos de armazenamento e os elementos processadores do sistema, também chamados de nós de processamento. Desta forma, e inicialmente, pode-se dividir o desempenho de um subsistema de E/S em duas partes: uma relativa aos dispositivos de armazenamento, outra relativa aos processadores de E/S.

II.2.1 - *O* desempenho dos Dispositivos de Armazenamento

O principal dispositivo de armazenamento é o disco magnético. Antes de se identificar os parâmetros que determinam seu desempenho é necessário esclarecer melhor alguns pontos sobre este dispositivo. A figura 5 esboga uma

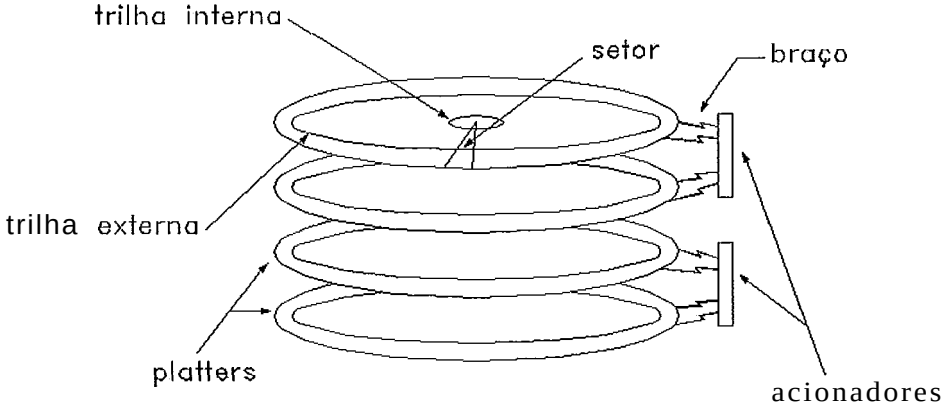

FIGURA 5: Esboço simplificado de uma unidade de disco magnético.

unidade de disco magnético. Ela é composta de um conjunto de *platters,* que consistem de discos metálicos cobertos por um material magnético onde são gravadas as informações. Cada *platter* contém uma quantidade de trilhas circulares. Estas trilhas são divididas em setores. Um setor constitui, fisicamente, a menor quantidade de dados lidos ou escritos na unidade. Entretanto, as informações são armazenadas em blocos no disco, sendo que cada bloco consiste de uma quantidade fixa e definida de setores. Um bloco constitui, logicamente, a menor quantidade de dados efetivamente lidos ou escritos no disco. As informaqões gravadas são recuperadas por uma cabeça de leitura/escrita posicionada em um braço que se move ao longo de cada *platter* através de um acionador.

O desempenho desta unidade está associado ao tempo gasto por ela para fornecer um bloco de informações dada a sua solicitaqão. Este tempo pode ser dividido em três componentes:

- Tempo de seelc
- Latência rotacional
- Tempo de transferência dos dados

O tempo de seek é o tempo necessário para o disco mover a cabeça de leitura/escrita até a trilha apropriada que contém o dado. Este tempo está associado à inércia para se tirar a cabeça da posição de repouso, que é da ordem de alguns mili-segundos, e ao número de trilhas a serem avançadas. Após a aceleração da cabeça, o tempo de avanço de trilha reduz-se bastante, chegando a um terço do inicial. Tipicamente, o tempo médio de seelc, dadas duas trilhas aleatórias, se situa na faixa de 10 a 20 ms dependendo do disco [PATTER90].

O segundo componente é a latência rotacional. Ela representa o tempo gasto para o setor de início de bloco dentro da trilha se posicionar tangencialmente à cabeça de leitura/escrita, permitindo a leitura ou escrita dos dados. Este tempo é função do número de rotações por minuto do disco, que é atualmente de 3600 RPM. Consequentemente, o tempo de uma revolução completa é de aproximadamente 16 ms. O tempo médio de latência é igual a metade do tempo de uma revolução, ou seja, cerca de *8* ms. Pode-se notar que, no pior caso, o tempo de latência rotacional pode ser até superior ao tempo médio de seek.

O último componente é o tempo de transferência dos dados. Ele corresponde ao tempo gasto para os bytes lidos serem transferidos do disco para o processador de E/S ou vice-versa. Em contrapartida com o tempo de seelc e de latência rotacional, que, como visto anteriormente, independem do tamanho lógico do bloco de informação a ser transferido, o tempo de transferência é função direta da quantidade de blocos transferidos. Assim sendo, existe um compromisso entre a definição do tamanho do bloco e a quantidade média de blocos requisitados de uma só vez. Se as transferências predominantes são de arquivos extensos, é conveniente escolher um tamanho de bloco grande. Isto faz com que o tempo de seek e de latência sejam atenuados frente ao tempo de transferência dos dados, já que, dentro do bloco, as informações são gravadas sequencialmente, e dentro do disco, os blocos são alocados conforme a disponibilidade. Caso contrário, se as transferências predominantes são de arquivos curtos, é conveniente escolher um tamanho de bloco pequeno, evitando transferir informações desnecessárias e otimizando a ocupação do disco.

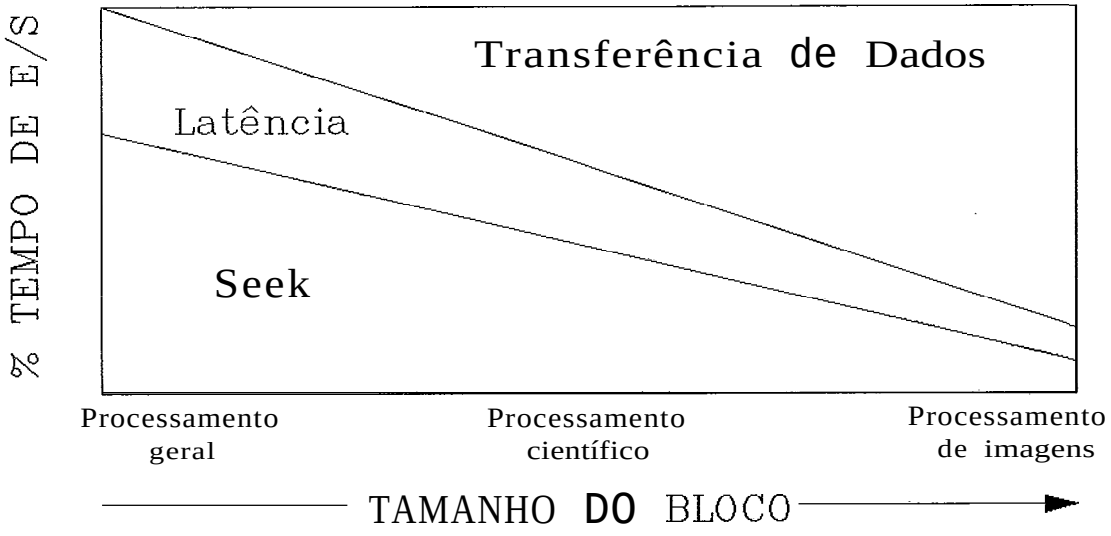

FIGURA 6: Distribuição do tempo de E/S referente ao disco magnético em função do tipo de aplicação a que o sistema computacional está submetido.

O predomínio de transferências de arquivos extensos ou curtos está relacionado com a característica do processamento de E/S referente ao tipo de aplicação a que o sistema computacional se destina. Computadores de uso geral processam simultaneamente um grande número de pequenas tarefas, que manipulam, cada uma, pequenas quantidades de dados. Por outro lado, os

computadores para uso científico processam simultaneamente poucas tarefas, mas que demandam grandes transferências de dados. A figura 6 ilustra esta situação, onde é mostrada a composição dos três elementos que determinam o desempenho dos discos magnéticos em função do tipo de aplicação a que o sistema está submetido. Pode-se notar que o processamento geral gasta a maior parte do tempo de E/S referente aos dispositivos de aimazenamento movimentando a cabeça de leitura/escrita (tempo de *seek)* e na latência rotacional. Em consequência, qualquer avanço na taxa de transferência dos dados não traz grandes benefícios. Por outro lado, a distribuição do tempo de E/S nas aplicações científicas é mais equalitária entre o tempo de *seek* e de transferência dos dados. Desta forma, o desempenho dos discos magnéticos nestas aplicações é bastante sensível a qualquer avanço tecnológico.

II.2.2 - *O* desempenho do Processador de E/S

O desempenho de um subsistema de E/S não depende exclusivamente do desempenho do disco magnético, mas sim do desempenho de todo o subsistema. De uma maneira geral, um subsistema de E/S é composto de um processador de E/S, que controla todo o subsistema, um conjunto de dispositivos de armazenamento tais como discos magnéticos, e um canal de comunicação entre eles, permitindo a transferência dos dados. Analogamente aos discos magnéticos, o desempenho do processador de E/S, quando submetido a uma carga de trabalho, pode ser analisado em função de três fatores:

- *Throughput*
- Latência
- Banda passante

Throughput refere-se ao número de pedidos de E/S atendidos pelo processador por unidade de tempo. Latência representa o tempo gasto para um determinado pedido ser atendido, também conhecido como *oveuhead* ou *queueing time*. A banda passante avalia a quantidade de dados por unidade de tempo que flui entre os dispositivos de armazenamento e o processador de E/S.

Baseado nisto, e em analogia com os discos magnéticos, pode-se também definir as características necessárias a um processados de E/S em função do tipo de aplicagão a que ele se destina. A computação científica pode ser caracterizada, quase inteiramente, por operagões de E/S sequenciais. Tipicamente,

os dados são transferidos em grande quantidade do disco para a memória principal, processados, e os resultados são periodicamente reescritos no disco. Este tipo de aplicagão exige uma larga banda passante do canal de comunicação, com o mínimo de latência do controlados. Isto permite que as transferências sejam agilizadas, já que são extensas. Entretanto, é caracterizado por um baixo *throughput* do processador, pois são poucos os pedidos de E/S por unidade de tempo.

Por outro lado, os operações de E/S em aplicagões de uso geral são caracterizadas por um grande número de pequenas tarefas. Estas tarefas solicitam acessos de forma randômica ao disco magnético. Neste tipo de aplicagão, tanto a banda passante do canal de comunicagão quanto a latência do processador de E/S podem ser apenas moderadas, enquanto o *throughput* tem que ser bastante elevado. O maior desafio na definição de um subsistema de E/S é justamente conseguir um desempenho satisfatório tanto para as aplicagões gerais quanto para as científicas, simultaneamente.

11.3 - Alternativas Futuras para os Dispositivos de Armazenamento

Neste capítulo ficou caracterizado um *gap* de desempenho, principalmente em relagão ao tempo de acesso, entre a memória secundária e os demais componentes do sistema computacional. Entretanto, algumas alternativas estão surgindo com a promessa de atenuar este desequilíbrio.

As expectativas com relação ao aumento de desempenho das unidades de discos magnéticos não são muito animadoras. Seguramente ocorrerá uma diminuição significativa no volume e, consequentemente, no consumo de energia, mas espera-se pouco em termos de aumento de velocidade. Apesar da maior densidade do meio magnético resultar em maior taxa de transferência de dados, o tempo de seek não deverá variar muito. A velocidade de rotação tem se mantido em 3600 RPM há mais de uma década. Há, contudo, alguns fabricantes que planejam migrar para 5400 RPM em futuro próximo.

Face a este panorama, várias pesquisas tem sido realizadas com o objetivo de superar as principais deficiências apresentadas pelos discos magnéticos como dispositivos de armazenamento secundário. A discussão a seguir aborda algumas das propostas mais significativas.

II.3.1 - RAM-DISK

Esta proposta baseia-se na utilização de *chips* de memória dinâmica (DRAMs) como dispositivos de armazenamento. Com o auxílio de uma bateria, pode-se criar um *array* de memória não volátil, cujo comportamento seja semelhante ao do disco magnético. Esta técnica é conhecida como RAM-DISK. Um RAM-DISK é dividido em blocos que emulam as trilhas e setores de um disco real. Isto torna seu acesso transparente, para o *software,* quanto ao meio físico acessado.

As vantagens do RAM-DISK são evidentes. Sendo um dispositivo semicondutor com tempo de acesso bastante reduzido, pode-se conseguir substanciais melhorias no tempo seelc e latência, na taxa de transferência e na confiabilidade se comparado aos discos magnéticos. Entretanto, sua principal desvantagem o torna inviável comercialmente para sistemas que requerem uma grande capacidade de armazenamento: o custo por mega *byte* ainda é muito alto, pelo menos dez vezes superior ao do meio magnético.

II.3.2 - Discos óticos

Os discos óticos tem sido apontados como uma alternativa promissora na substituição dos discos magnéticos como principal dispositivo de armazenamento, especialmente considerando duas características essenciais à memória secundária: grande capacidade e alta densidade de armazenamento. Isto é obtido em função da tecnologia mais avançada de recuperação dos dados no disco ótico. Um feixe de luz gerado por um diodo *laser* é refletido na superfície ótica, alterando algumas de suas propriedades físicas conforme as informagiões gravadas nesta superfície.

A maior desvantagem dos discos óticos é a falta de sustentagião da taxa de transferência dos dados. Isto se deve ao alto tempo de acesso às informações no disco e à relativa primitividade das operações de escrita. Entretanto, várias pesquisas estão sendo realizadas com objetivo de solucionar estas dificuldades.

Os discos óticos podem ser classificados em três categorias:

- Read-Only (ROM)
- Write-Once-Read-Many (WORM)
- Erasable/Rewritable

O primeiro tipo, também conhecido como CD-ROM, tem as informagões pré-gravadas pelo fabricante, não permitindo ao usuário alterá-las. Os discos da segunda categoria, como o nome sugere, permitem ao usuário gravar uma única vez as informagões, não podendo apagá-las ou regravá-las posteriormente. Na última categoria se enquadram os discos óticos que permitem, livremente, operagões de leitura e escrita.

O sucesso dos discos óticos está relacionado ao bom resultado das pesquisas que procuram transpor as dificuldades das operagões de escrita sem inviabilização do preço. A técnica mais promissora neste sentido é a combinação simultânea de efeitos óticos e magnéticos, resultando no que se chama magnetooptical disk. Maiores informagões sobre discos óticos podem ser conseguidas em [BERRA89].

II.3.3 - DISK-ARRAY

A alternativa com maiores chances a curto prazo para otimização do armazenamento de dados na memória secundária não é, propriamente, uma inovação tecnológica, mas sim uma nova forma de organização dos discos magnéticos. Uma técnica chamada striping propõe a fragmentação dos blocos de informagões para armazenamento em uma matriz de discos, disk-array, de forma a cada fragmento ser armazenado em um disco diferente. Isto permite um paralelismo nas operagões de leitura e escrita destes blocos, resultando num aumento da taxa de transferência da matriz de dispositivos de armazenamento.

Os argumentos a favor dos *disk-arrays* são simples. A massificagão do comércio de computadores pessoais fez com que surgissem discos magnéticos cada vez menores e baratos. Como o prego por mega byte é praticamente independente do tamanho do disco, está sendo técnica e economicamente mais fácil alcançar uma alta banda passante na memória secundária, característica essencial à computagão científica, utilizando vários pequenos discos paralelamente do que buscar o desenvolvimento de discos magnéticos que, individualmente, tenham alto desempenho.

O principal inconveniente na fragmentação das informações ao longo de vários discos é a menor tolerância a falhas. O parâmetro que mede a confiabilidade de um disco magnético é fornecido pelo fabricantesob o código MTTF (*mean* time to *failure),* que representa a quantidade de horas garantidas sem a ocorrência de

falhas. Cada um dos discos que compõe um dislc-array tem uma confiabilidade semelhante a do disco magnético de alta capacidade. Entretanto, no conjunto, a confiabilidade do dislc-array é inversamente proporcional ao número de discos no array:

$$\text{MTTF do dislc-array} = \frac{\text{MTTF de um dos discos}}{\text{número de discos no array}}$$

Para melhorar a confiabilidade dos dislc-arrays, são inseridos, no array, discos extras contendo informações redundantes capazes de recuperar a informação original quando ocorre falhas em um dos discos, sacrificando, de urna certa forma, a banda passante e a capacidade de armazenamento de todo o array de discos. Desta forma, o tipo de técnica de correção de falhas utilizado nos dislc-arrays acaba determinando seu desempenho e viabilidade.

Um bom exemplo de utilização da técnica de *striping* em dispositivos de armazenamento pode ser constatado nos computadores da linha CRAY. Uma de suas unidades de disco, a DS-40, se apresenta como um simples dispositivo. Entretanto, ele é implementado, internamente, com quatro discos magnéticos. Desta forma, o DS-40 é capaz de atingir uma taxa de transferência de 9,6 *MBytes*/s. Maiores informações sobre *disk-array* podem ser conseguidas em [PATTER88], [GIBSON89] e [NG88].

CAPÍTULO III

OPÇÕES PARA CONFIGURAÇÃO DE UM SUBSISTEMA DE E/S

Conforme visto anteriormente, os microprocessadores são algumas ordens de grandeza mais rápidos que os dispositivos de armazenamento. Este desequilíbrio fez surgir um problema de adequação da velocidade de processamento dos elementos processadores com suas necessidades de entrada e saída de informações. O mau demensionamento do subsistema de E/S pode representar um gargalo para o sistema computacional. Diferentes soluções para este problema podem ser adotadas dependendo do tipo de aplicação a que a máquina se propõe. Este é o tema da discussão neste capítulo.

Um sistema computacional é constituido de subsistemas. O subsistema de E/S é o responsável pelo acesso à memória secundária. Ele é composto de um ou mais processadores de E/S (PES), que controlam e gerenciam todas as operações com os dispositivos de armazenamento. Existem várias formas de se associar estes PES aos elementos processadores do sistema. Dependendo da forma de associação, consegue-se um maior ou menor desempenho em função do tipo de aplicação a que o sistema computacional, como um todo, está submetido.

A idéia mais simples para definição de um subsistema de E/S é a utilização de um computador comercial específico para controlar todas as operações em memória de massa. Constituiria uma unidade completamente independente, interligada através de um canal ou rede de alta velocidade ao sistema principal. Entretanto, ela esbarra em alguns problemas que a tornam pouco recomendável na prática. Primeiramente, existe a dificuldade de se realizar eventuais modificações numa máquina comercial, onde o *hardware* e o *firmware* não são suficientemente abertos para permitir uma adaptação que proporcione um desempenho ótimo na execução desta nova função. Associado a isto, existe o questionamento do próprio desempenho desta opção frente às necessidades de E/S de um computador de alto desempenho voltado para processamento científico. Uma simultaneidade de pedidos de E/S com grande quantidade de dados pode congestionar o canal de comunicação. Até o advento da tecnologia ótica, os canais de comunicação eram muito lentos para este propósito, mesmo quando implementados em múltiplas vias. Canais como ETHERNET, tipicamente

implementados em uma única via de comunicação, operam a 10 *MBits* por segundo, taxa muito aquém das necessidades. Com o surgimento da tecnologia ótica, bandas passantes de até 100 *MBits* por segundo por via de comunicagão são facilmente alcançadas. Existe uma expectativa de que taxas de até 1*GBits* por segundo sejam alcangadas nos próximos anos. Desta forma, canais ou redes de comunicação ainda são lentos para constituir o único meio para as operações de E/S em computadores de alto desempenho, podendo se tornar uma técnica promissora no futuro.

Em contrapartida, a outra idéia que surge é a configuração do subsistema de E/S acoplado diretamente aos elementos processadores do sistema. Primeiramente, qualquer que seja a forma desta configuração, o fato dela ser interna ao sistema determina um compromisso forte com suas caractesísticas. O subsistema deixa de ser completamente independente para moldar-se a uma arquitetura específica, seguindo padrões elétricos e mecânicos que dificilmente se adaptariam a uma outra situação.

Há basicamente duas formas de se configurar internamente um subsistema de E/S. A forma distribuida, onde pequenos processadores de E/S (PES) são associados de forma exclusiva a cada elemento processados ou a pequenos grupos deles, e a forma concentrada, onde um grupo de PES, acessado de forma simétrica por todos os elementos processadores, controla os dispositivos de armazenamento. Cada urna das formas tem suas vantagens e desvantagens como será mostrado adiante. A tabela da figura 7 resume esta discussão.

III.1 - EIS Distribuida

Neste tipo de configuração, o processamento de E/S do sistema computacional está distribuido pelos seus diversos PES. Cada PES está associado de forma exclusiva a cada elemento processador ou a um conjunto deles (*cluster*). Esta associagão se realiza através de um canal de comunicagão de alta velocidade, permitindo a transferência de informagões entre os elementos processadores e o PES associado de forma satisfatória às necessidades de entrada e saída de um computador de alto desempenho. A exclusividade na associagão não significa que somente os processadores de um mesmo cluster podem acessar o PES associado a eles. Qualquer elemento processador pode enviar e receber informações de qualquer um dos PES do sistema. Entretanto, se o PES e o elemento processador requisitante não pertencerem ao mesmo cluster, as transferências deverão ser

realizadas através de caminhos alternativos, que seguramente terão uma banda passante menor que a do canal de comunicação do PES associado, resultando num custo de tempo mais elevado. Dentre as arquiteturas de E/S que se configuram como distribuídas pode-se citar a do BBN-Butterfly [BBN85] e a do RP3 [PFISTER85].

Vantagens:

- *THROUGHPUT*. A fragmentação do subsistema de E/S pela distribuição de seu processamento nos diversos processadores de E/S permite um paralelismo das operações de entrada e saída. Cada PES pode transferir informações para qualquer elemento processador de seu *cluster* simultaneamente com os demais. Assim, o *throughput* ou a capacidade de processamento do subsistema de E/S é a soma do *throughput* de cada processador de E/S. Isto permite ao subsistema atingir um desempenho muito elevado, difícil de ser conseguido isoladamente por um único PES.

- SIMPLICIDADE. Como cada PES está associado a um pequeno grupo de elementos processadores, a quantidade de pedidos de entrada e saída que partem destes elementos é bastante reduzida se compasada com todo o subsistema. Isto acaba refletindo numa maior simplicidade de cada PES. O *throughput* deve ser o suficiente para atender às necessidades de entrada e saída apenas dos elementos processadores do *cluster*. Analogamente, a banda passante do canal de comunicação entre o PES e o *cluster* associado também pode ser menor, resultando em técnicas de projeto mais simples. Além do mais, um menor número de pedidos de E/S permite ao PES respondê-los mais rapidamente, reduzindo também a latência do subsistema de E/S.

- MODULARIDADE. A configuração distribuída também tem a vantagem de ser modular. Cada módulo corresponde a um processador de E/S. Nesta configuração o subsistema de E/S pode crescer à medida em que cresce o número de elementos processadores do sistema, bastando para isso adicionar PES. Entretanto, não há a obrigatoriedade de existir sempre um PES associado a cada *cluster* de elementos processadores. Esta característica é importante na fase de implementação do sistema, onde apenas parte do subsistema de E/S pode ser implementado, deixando o seu crescimento vinculado ao crescimento progressivo de todo o sistema e à necessidade de reforço da capacidade de processamento de E/S.

Desvantagens:

- ESPECIFICIDADE. A distribuição do processamento de E/S torna o subsistema mais apropriado às aplicações específicas, como processamento científico e de imagens, pois muitas destas aplicações podem ser paralelizáveis. Isto se dá pela própria característica do processamento deste tipo de aplicação. São várias tarefas que podem ser executadas em paralelo em diferentes clusters, buscando, de cada PES associado, as informações necessárias ao processamento e reescrevendo, posteriormente, os resultados. Para aplicações gerais o desempenho deste tipo de configuração é questionável, principalmente quanto à distribuição dos arquivos ao longo dos PES.

- MANUTENÇÃO DA COERÊNCIA DE INFORMAÇÕES. Nesta configuração, os dispositivos de armazenamento estão distribuídos sob o controle independente de cada PES. Desta forma, existe uma dificuldade em se manter a coerência das informações gravadas neles. Ou seja, evitar que ocorra multiplicidade de informações e, principalmente, fazer com que as versões desatualizadas sejam invalidadas em todo o subsistema sempre que houver a atualização de algum arquivo.

- TRÁFEGO. Dependendo do tipo de aplicação a que o subsistema de E/S está submetido, a distribuição do processamento pode resultar em muitas transferências de informações entre clusters de elementos processadores. O custo de tempo destas transferências é bem maior que o das transferências intra cluster, pois devem ser realizadas pela rede de comunicação entre elementos processadores de diferentes clusters. O aumento deste tipo de transferência pode saturar esta rede de comunicação, prejudicando o tráfego natural de troca de informações entre elementos processadores que se realiza através dela.

111.2 - E/S Concentrada

Na configuração concentrada, o subsistema de E/S também é composto por um grupo de um ou mais processadores de E/S. Entretanto, este grupo é coeso, gerenciando e executando os pedidos de entrada e saída de todos os elementos processadores do sistema. A comunicação entre o grupo de PES e os elementos processadores é feita através de canais de comunicação. Também pode ocorrer a formação de clusters de elementos processadores com objetivo de acessar o subsistema de E/S, ou seja, a existência de um único caminho de comunicação

OPÇÕES PARA CONFIGURAÇÃO DE UM SUBSISTEMA DE E/S

	DISTRIBUÍDA	CONCENTRADA
VANTAGENS	- THROUGHPUT - SIMPLICIDADE - MODULARIDADE	- GENERALIDADE - GERÊNCIA OTIMIZADA DE TRANSFERÊNCIAS - SIMPLICIDADE NO CONTROLE
DESVANTAGENS	- ESPECIFICIDADE - DIFICULDADE NA MANUT. DA COERÊNCIA DAS INFORMAÇÕES - AUMENTO DO TRÁFEGO	- COMPLEXIDADE DE HARDWARE - CAPACIDADE DE EXPANSÃO RESTRITA

FIGURA 7: Vantagens e desvantagens das configurações de um subsistema de E/S.

com o grupo de PES para cada *cluster*. Dentre as arquiteturas de E/S que se configuram como concentrada pode-se citar a do ES-8701 [PRADO88].

Vantagens:

- GENERALIDADE. A concentração do processamento de E/S permite uma adaptação melhor do subsistema de E/S a uma gama maior de variedades de aplicagões. Isto ocorre por já haver intrinsecamente um compartilhamento de toda a memória secundária entre os elementos processadores. Toda informação é acessada por qualquer elemento processador com o mesmo custo de tempo. Entretanto, esta é uma característica que deve ser analisada em função do propósito da máquina. Da mesma forma que permite um desempenho satisfatório numa gama maior de aplicagões, pode se tornar ineficiente em muitas aplicações onde há a possibilidade de se explorar o paralelismo nas operagões de E/S.

- GERÊNCIA DE TRANSFERÊNCIAS. Existem algumas políticas de otimização de acesso ao disco com objetivo de minimização do tempo médio de resposta a um

pedido de E/S. Usualmente, se os pedidos forem atendidos obedecendo a ordem de chegada, pode resultar em movimentações extensas da cabeça de leitura/escrita ao longo das trilhas e, consequentemente, num dispendimento grande de tempo. Existem basicamente duas alternativas. Na política *Shortest-Seek-Time-First* os pedidos são atendidos na ordem que minimiza o tempo de movimentação da cabeça a partir de sua posição atual. Uma outra opção é o algoritmo SCAN, onde a cabeça de leitura/escrita é movimentada de uma extremidade do disco à outra, atendendo aos pedidos de E/S à medida em que as informações solicitadas se encontrem progressivamente nas trilhas avançadas. Quando se concentra o processamento, pode-se explorar melhor estas políticas. Tem-se uma visão mais global dos pedidos de E/S e, portanto, é possível ordená-los de modo a gerenciar melhor as transferências.

- SIMPLICIDADE NO CONTROLE. A simplicidade no controle das operações de E/S a nível de Sistema Operacional é uma grande vantagem desta configuração. O agrupamento dos dispositivos de armazenamento torna mais simples tanto a localização dos arquivos, quanto a manutengão da coerência deles.

Desvantagens:

- COMPLEXIDADE DE *HARDWARE*. A complexidade é a maior desvantagem desta configuração. A existência de um único grupo de processadores de E/S, onde se concentra todo o processamento de entrada e saída, pode simplificar o Sistema Operacional mas pode também requerer maior complexidade de hardware. Dependendo do número de elementos processadores, a demanda por processamento de E/S pode ser grande, implicando na necessidade de um alto desempenho do grupo de PES. São várias as dificultades inerentes a isto, a fim de se evitar um gargalo no sistema. Elas vão desde o excesso de canais de comunicação entre o grupo de PES e os elementos processadores até a preocupagão com a latência e o *throughput*, que requer o uso de técnicas de projeto mais sofisticadas.

- CAPACIDADE DE EXPANSÃO RESTRITA. Não sendo uma configuração modular, a expansão do subsistema de E/S, em função do número de elementos processadores, em geral influencia a operacionalidade de todo o subsistema. Além disto, requer um projeto que preveja as características de sua configuração máxima, restringindo as expansões futuras.

CAPÍTULO IV

A ARQUITETURA DO SUBSISTEMA DE E/S DO MULTIPLUS

O Multiplus é um multiprocessador científico de alto desempenho com arquitetura modular e memória global compartilhada. A arquitetura é capaz de suportar até 2048 elementos processadores ou nós de processamento. Eles são baseados em microprocessadores RISC de 32 *bits* com arquitetura SPARC e capacidade de processamento de 25 MIPS VAX a 40 MHz. Além do microprocessador, cada nó de processamento possui um co-processador de ponto flutuante, 32 M*Bytes* de memória pertencentes ao espaço de endereçamento global de 32 G*Bytes, cacke* de instrução e dado separados, com 64 K*Bytes* cada um e *hardware* de suporte à gerência de memória.

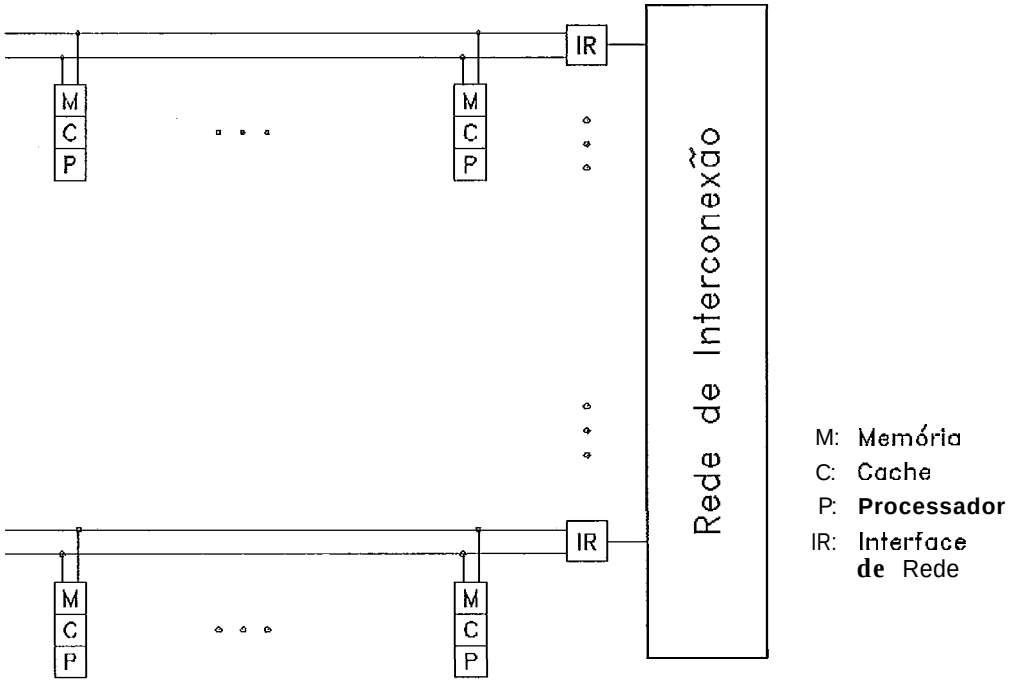

FIGURA 8: Esquema geral da arquitetura do Multiplus.

Até 16 elementos processadores podem ser interconectados através de um barramento duplo de 64 *bits* de largura cada um, formando um *cluster* de processadores. Cada *cluster* possui uma placa multifungão responsável pela arbitragem dos barramentos e geração dos sinais de *ueset.* Os diferentes *clusters* de

processadores se comunicam através de uma rede de interconexão multiestágio do tipo n-cubo invertido. Interfaces inteligentes, com capacidade de armazenamento de mensagens de até 128 *bytes* e de realizagão de operações de DMA nos barramentos, acoplam a rede de interconexão aos *clusters*.

A figura 8 mostra um esquema geral da arquitetura do Multiplus. Com esta arquitetura pode-se criar uma família de computadores cobrindo um espectro que vai desde estações de trabalho de alto desempenho, com configuragões de 1 a 8 nós de processarnento, passando por minisupercomputadores, com 16 a 128 nós, e chegando aos supercomputadores com 256 ou mais nós de píocessamento. Maiores informagõessobre o Multiplus podem ser conseguidas em [AUDE91].

Definir um subsistema de E/S para o Multiplus de forma satisfatória às diversas configuragões que o sistema pode assumir é o tema deste capítulo. Primeiramente cabe distinguir dois tipos de processamento associado a um subsistema de E/S. Um processamento orientado a caracteres, controlando operagões de E/S com terminais, impressoras e redes como ETHERNET, e outro orientado a blocos, controlando operagões de E/S com discos e fitas magnéticas.

Para cada um dos dois tipos de processarnento existe um processados de E/S específico encarregado de controlar as operagões de E/S correspondentes. Este par de processadores é independente em suas funções e, associado a cada *cluster* de elementos processadores, forma o subsistema de E/S do Multiplus. A seguir, será descrita a evolução das idéias que definiram a arquitetura do subsistema de E/S do Multiplus. Devido a esta divisão do processamento associado a um subsistema de E/S, serão analisadas em separado as questões relativas ao processamento orientado a bloco e ao orientado a caracter.

IV.l- O Subsistema de EIS Orientado a Bloco

A parte orientada a bloco do subsistema de E/S processa operagões com discos e fitas. Conforme visto nos capítulos anteriores, as características destas operagões de E/S estão relacionadas à aplicação submetida ao sistema como um todo. O Multiplus é voltado para aplicações científicas. Assim sendo, espera-se que suas operações de E/S sejam esparsas e extensas, devendo ser executadas no menor tempo possível. Isto caracteriza a necessidade de um canal de comunicação rápido e eficiente entre o subsistema de E/S e os elementos processadores. Dentro

deste panorama, foram propostas algumas soluções para o subsistema de E/S do Multiplus.

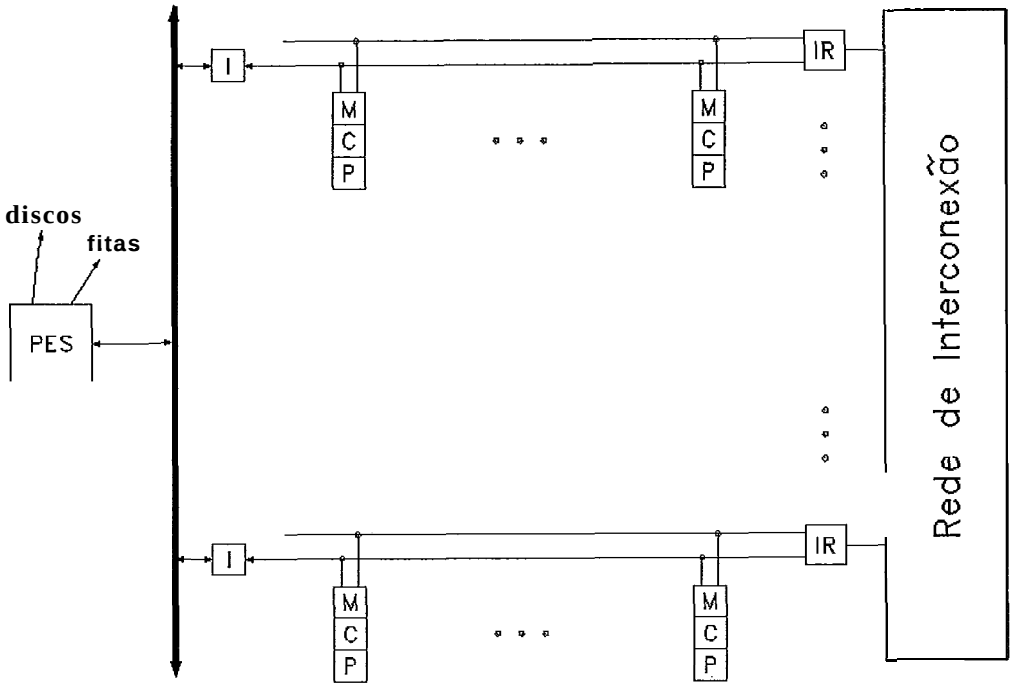

FIGURA 9a: Hipótese concentrada com barramento dedicado para configuragão do subsistema de E/S do Multiplus.

Inicialmente, pensou-se numa configuração concentrada para a arquitetura deste subsistema. A primeira idéia foi interligar os *clusters* através de um barramento de alta velocidade dedicado exclusivamente às operagões de E/S, onde também estariam conectados os processadores de E/S. Esta idéia é espelhada na solução adotada no sistema ES8701 [PRADO88] e é mostrada na figura 9a. Como a quantidade máxima de *clusters* no Multiplus é de 256, existiam duas limitações elétricas significativas:

- Os barramentos de alta velocidade são de comprimento máximo reduzido, implicando numa proximidade física dos *clusters* incompatível com o *lay-out* do Multiplus.

- Existe uma limitação do número máximo de conexões que se pode carregar num barramento em função da impedância das linhas, que é em geral, bem inferior ao número máximo de *clusters*.

Face a estas limitações, procurou-se contornar o problema seccionando o barramento em várias partes, e associando um pequeno grupo de *clusters* a cada uma destas partes, como mostrado na figura 9b. Apesar de se ter resolvido o problema de carregamento do barramento, as limitações de *lay-out* permaneceram. Os barramentos de alta velocidades são implementados em *baclc-planes,* que são placas de circuito impresso fisicamente rígidas. A interconexão destes *back-planes* restringe a disposição dos *clusters.* Além disto, os *back-planes* não são suficientemente compridos para permitirem a conexão de um grupo, ainda que pequeno, de *clusters* do Multiplus (ver figura 16). Por estas razões, esta alternativa se mostrou inviável.

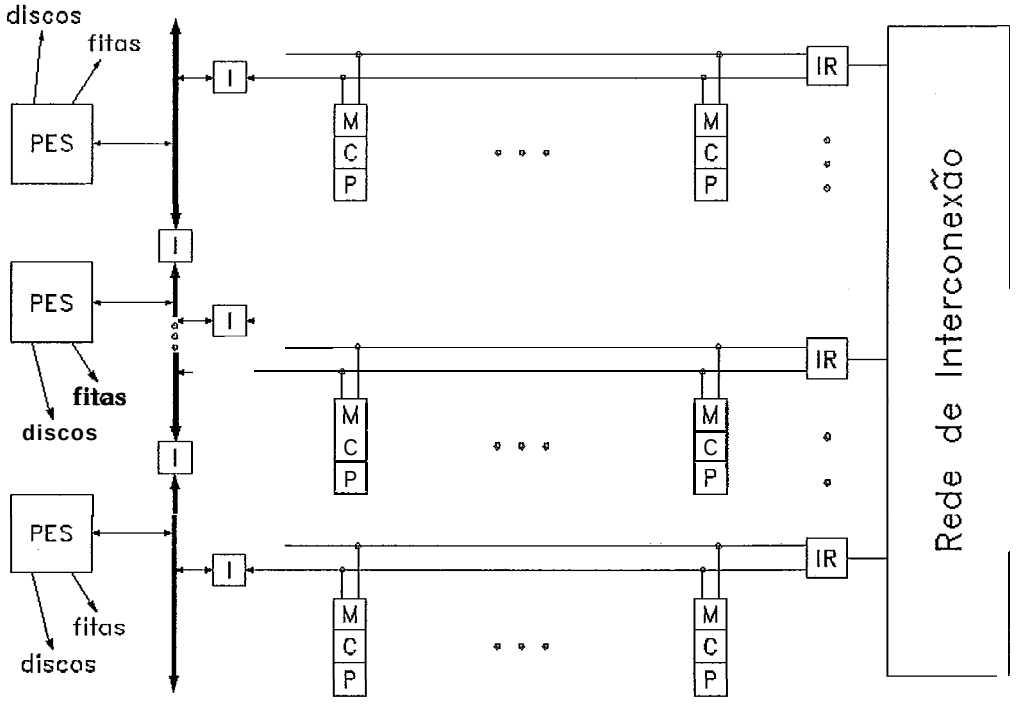

FIGURA 9b: Hipótese concentrada com barramento dedicado e seccionado para configuração do subsistema de E/S do Multiplus.

Uma segunda idéia, ainda na configuração concentrada, foi a interligação dos *clusters* através de uma chave *cross-bar* ao conjunto de PES, como mostrado na figura 10. Entretanto, associar um canal de comunicação direto e de alta velocidade, partindo de cada *cluster* de elementos processadores até a um único grupo de processadores de E/S pode ser bastante custoso e, por conseguinte, inviável na prática. Como o número total de *clusters* de elementos processadores é elevado, haveria uma grande quantidade de canais de comunicação. Além da

dificuldade prática de interligação destes canais, haveria um subaproveitamento devido a baixa periodicidade das transferências. Outro ponto limitante é a própria complexidade da chave cross-bar. Descartou-se, desta forma, uma configuração concentrada para o processamento de E/S do Multiplus.

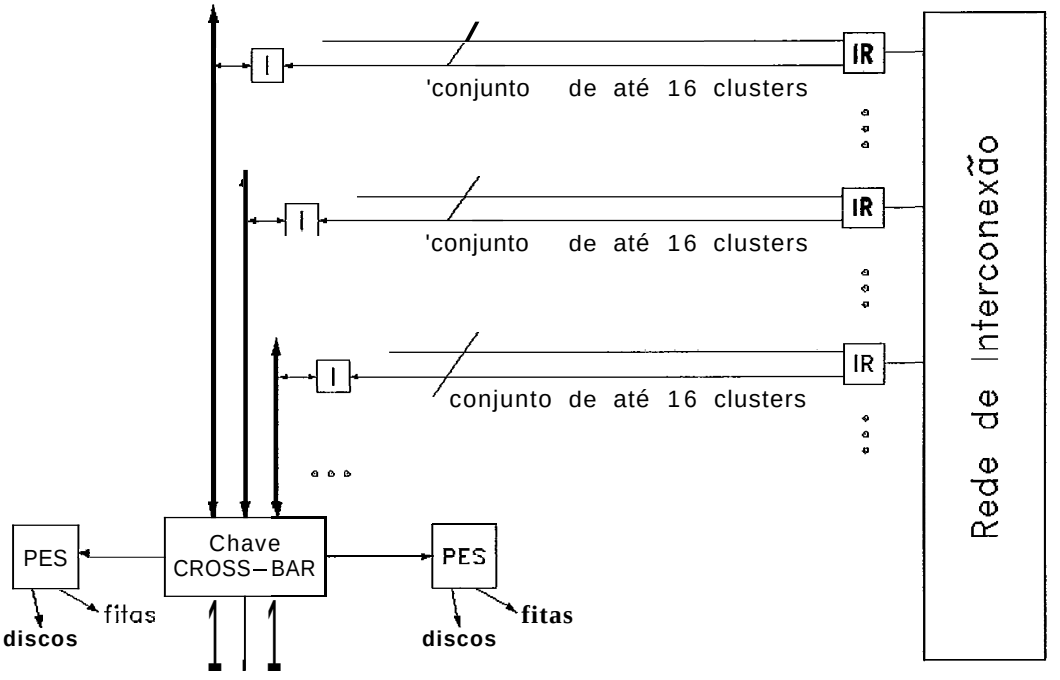

FIGURA 10: Hipótese concentrada com chave cross-bar para configuragão do subsistema de E/S do Multiplus.

Visualizando a arquitetura do Multiplus, pode-se notar que ela própria sugere uma distribuição do processamento de E/S. Como são vários barramentos interligados por uma chave, é sugestivo associar um processador de E/S a cada barramento. Desta forma, o *cluster* de elementos processadores formado por cada barramento possui, através do próprio barramento, um canal de comunicaçao de alta velocidade entre os elementos processadores e o subsistema de E/S. Por outro lado, a própria finalidade a que o Multiplus se destina reforça esta configuração. Uma máquina paralela sugere a realização de várias operações em paralelo. E dentro destas operagões também se encontram as de E/S de dados e instruções para se ter um maior equilíbrio e eficiência no sistema. Este paralelismo é tanto mais explorado quanto mais distribuído for o processamento de E/S. A figura 11 ilustra a idéia inicial surgida para definição do subsistema de E/S do Multiplus dentro da configuragão distribuída. Vale ressaltar que, pela semelhança entre as

arquiteturas do Multiplus e do BBN-Butterfly [BBN85], esta idéia inicial é similar à solução adotada para o subsistema de E/S deste computador de alto desempenho.

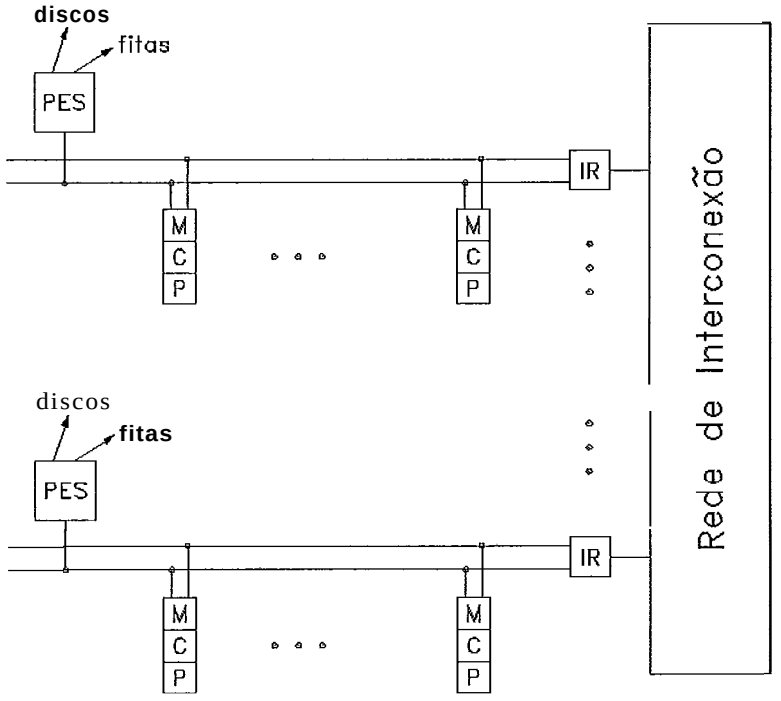

FIGURA 11: Hipótese distribuída para configuração do subsistema de E/S do Multiplus.

Nesta primeira idéia, a comunicação entre os *clusters* de elementos processadores se dava através de uma rede de chaves multiestágio. Entretanto, o custo de tempo desta comunicação é alto frente às necessidades de E/S. Além do mais, sendo as transferências extensas, pode ocorrer a saturação da comunicação pela chave, em detrimento de seu desempenho, caso haja muita troca de dados entre os *clusters.* Assim, buscou-se um caminho alternativo para as transferências de E/S entre os *clusters.* Incorporou-se uma rede de comunicação de alta velocidade ao subsistema de E/S, permitindo aos PES um caminho alternativo de comunicação entre si. Para manter a homogeneidade da arquitetura, esta comunicação é restrita às transferências de entrada e saída. Uma caractesística interessante é que, com esta rede de comunicação, cada elemento processador solicita sempre serviços de E/S ao seu PES associado. Caso a informação solicitada esteja nos dispositivos de armazenamento controlados por este próprio PES, a transferência é realizada normalmente pelo barramento. Caso contrário, o PES solicita a informação ao PES apropriado, recebendo-a através da rede de

comunicação e transferindo-a, posteriormente, ao elemento processador via barramento do cluster. Entretanto, não é essencial que todo clusfer possua um PES associado, podendo existir *clusters* só de elementos processadores. Neste caso, qualquer transferência de E/S é realizada obrigatoriamente pela rede de chaves. A figura 12 ilustra o diagrama da configuração final do subsistema de E/S do Multiplus para as operações com bloco.

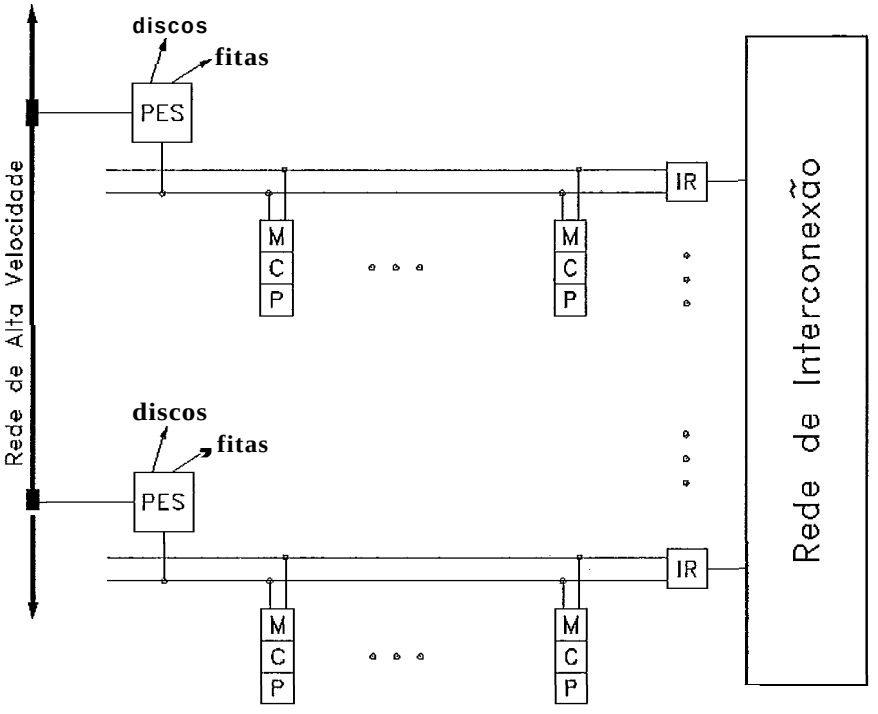

FIGURA 12: Subsistema de E/S do Multiplus orientado a bloco.

Desta forma, o Multiplus se configura com um subsistema de E/S com duas forças de acoplamento aos elementos processadores. Cada clusfer pode ter um PES fortemente acoplado aos seus elementos processadores, permitindo transferências com um custo de tempo bastante reduzido. Caso as informações de E/S não se encontrem no cluster do elemento processador solicitante, o acoplamento se enfraquece, aumentando o custo de tempo das transferências. Consegue-se, assim, explorar, a nível de cluster, as vantagens de uma configuração concentrada, embora o subsistema de E/S como um todo seja distribuído.

IV.2 - O Subsistema de E/S Orientado a Caracter

A parte orientada a caracter do subsistema de E/S processa operações de E/S com terminais e impressoras e interfaceia o Multiplus com redes ETHERNET. Por não representar risco de se configurar como um gargalo para o sistema computacional, não foram abordados, nos capítulos introdutórios, problemas relacionados a entrada e saída de caracteres. Entretanto, para que se possa compreender melhor a definição do subsistema de E/S do Multiplus, algumas considerações serão realizadas.

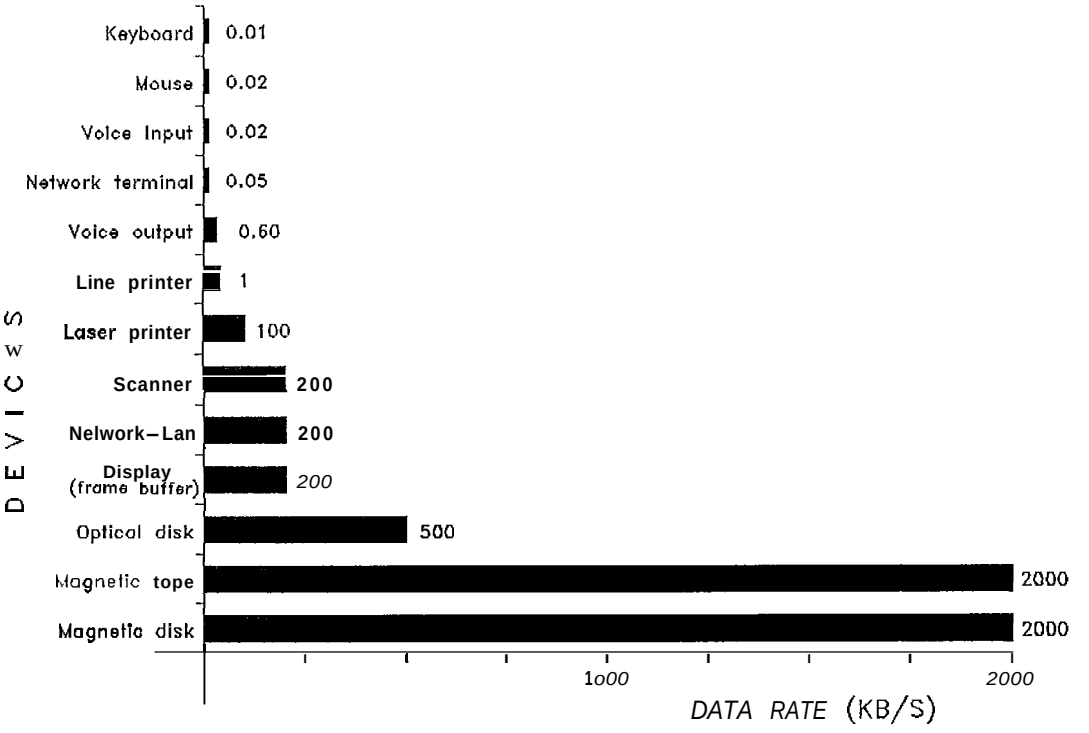

FIGURA 13: Banda passante dos principais dispositivos de E/S.

Os dispositivos de E/S orientados a caracter são, geralmente, utilizados na interação com o sistema computacional. São, por exemplos, teclados, *mouses*, terminais de vídeo, impressoras, que fazem a interface entre o homem e a máquina. Cada um destes dispositivos requer uma taxa de transferência e uma latência associada ao sentido humano envolvido. Desta forma, um subsistema de E/S orientado a caracter tem uma curva de desempenho fixa e conhecida, praticamente independente da complexidade do sistema computacional. A figura 13 mostra alguns dos dispositivos de E/S mais comuns e as respectivas taxas de

transferência necessárias ao bom funcionamento [PATTER90]. Pode-se notar que os dispositivos orientados a caracter requerem uma banda passante muito inferior aos orientados a bloco, como o disco magnético. Assim, o controle destes dispositivos é uma tarefa relativamente fácil e simples, não existindo questões muito preocupantes quanto ao desempenho. Em função disto, e por envolver uma gama de conceitos que fogem ao escopo deste trabalho, não serão discutidos os problemas encontrados na definição do subsitema de E/S do Multiplus orientado a caracter, limitando-se apenas a expor sua arquitetura e as características principais.

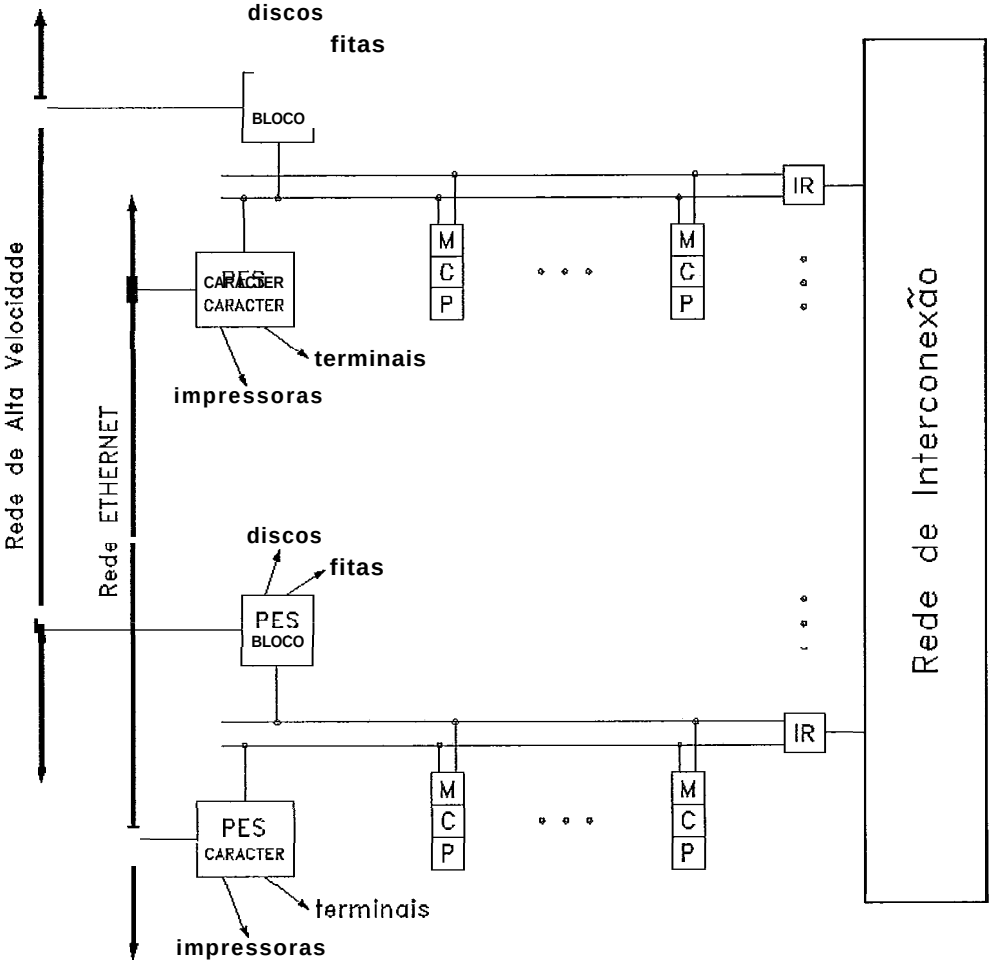

FIGURA 14: Diagrama final do subsistema de E/S do Multiplus.

A figura 14 ilustra a arquitetura final de todo o subsistema de E/S do Multiplus, tanto para as operagões com caracter quanto para as operagões com bloco. A parte do subsistema de E/S responsável pelo interfacearnento com os

dispositivos orientados a caracter também é composta por um conjunto de processadores de E/S (PES), que são associados aos *clusters* e interligados por uma rede ETHERNET. Esta rede permite a comunicagão entre o Multiplus e outros sistemas computacionais.

Os PES orientados a caracter são responsáveis pelo controle dos dispositivos de E/S a eles associados. Suas principais características são as seguintes:

- Processados MOTOROLA MC68020
- Suporte para até 8 terminais de vídeo
- Duas interfaces paralelas padrão CENTRONICS de uso geral
- Duas interfaces seriais padrão RS-232 de uso geral
- Interface com rede ETHERNET
- *Buffer* para recepgão de dados dos elementos processadores.

CAPÍTULO V

O PROCESSADOR DE E/S ORIENTADO A BLOCO

No capítulo anterior definiu-se a arquitetura do subsistema de E/S do Multiplus, e a forma com que os elementos processadores interagem com os dispositivos de armazenamento. Cabe, agora, definir a arquitetura dos processadores de E/S que compõem este subsistema. Este é o tema deste capítulo, onde serão discutidas as questões principais sob o ponto de vista da arquitetura, das limitações impostas para implementação e, por fim, da funcionalidade do PES.

V.1 - Questões Principais na Definição da Arquitetura do PES

O Multiplus é uma máquina que se destina, prioritariamente, às aplicações científicas. Assim sendo, é esperado que suas operações de E/S sejam esparsas e extensas, requisitando um PES com uma altíssima banda passante e um baixo *ovevhead.* Além disto, seguindo a definição do subsistema de E/S, ele deve ter uma interface com o barramento do *cluster* de elementos processadores, uma interface para uma rede de comunicação serial de alta velocidade e, obviamente, uma interface para dispositivos de armazenamento de massa.

A arquitetura proposta para os processadores de E/S do Multiplus é mostrada na figura 15.

V.1.1 - A Divisão dos Barramentos

Pela figura 15, observa-se que o PES é dividido em três barramentos. Dois deles internos, interligando os seus próprios componentes, e um terceiro externo, dedicado a interligação de placas comerciais prontas. Os barsamentos são:

- Barramento da CPU
- Barramento do DMA
- Barramento externo

A divisão do barramento interno teve o objetivo de minimizar o *ovevhead* do PES, fazendo com que, tanto a CPU quanto o DMA tenham o barramento à sua

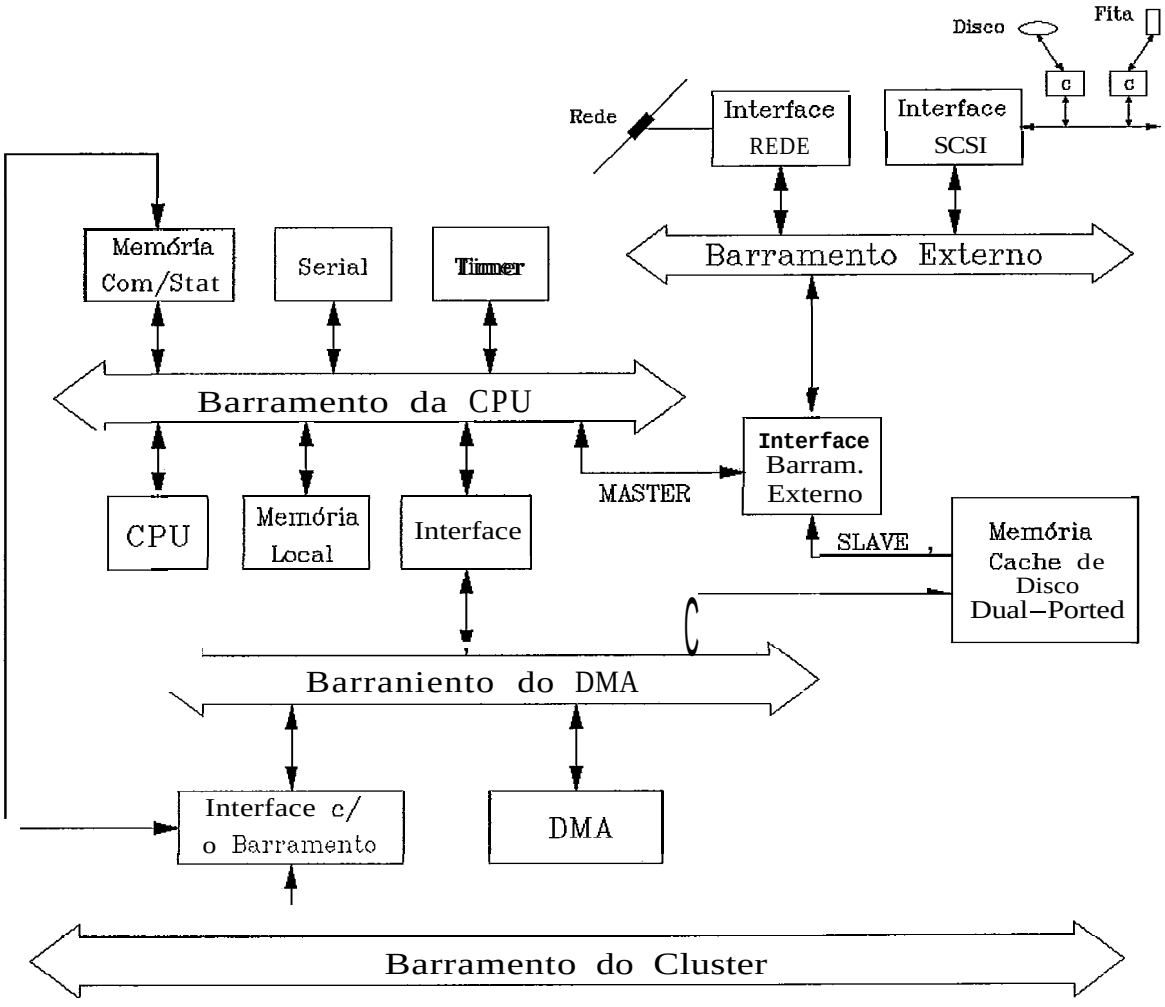

FIGURA 15: Diagrama em bloco da arquitetura do Multiplus.

disposição, sem compartilhamento. Isto permite ao DMA realizar transferências entre o PES e os elementos processadores em modo rajada (*burst*), sem interferir no processamento da CPU. Também permite à CPU, que controla e gerencia todo o PES, atender prontamente a qualquer solicitação, principalmente aquelas provenientes de recursos compartilhados. A memória de comandos, conectada ao barramento da CPU, é um recurso compartilhado por todos os elementos processadores do Multiplus, e como tal, deve ser liberada o mais rápido possível, evitanto espera do Sistema Operacional. A garantia da rápida liberação deste recurso é uma forte justificativa para a divisão do barramento interno do PES do Multiplus.

A comunicação e a passagem de comandos entre a CPU e o DMA e entre a CPU e o barramento externo é feita através de interfaces apropriadas incorporadas ao barramento da CPU.

	VMEbus	Futurebus	MultibusII	IPI	SCSI
NUMERO DE PINOS	96	128	96	16	8
DADO/ENDEREÇO	NAO MULTI-PLEXADO	MULTI-PLEXADO	MULTI-PLEXADO	--	--
PALAVRA	32 BITS	32 BITS	32 BITS	16 BITS	8 BITS
SPLIT TRANSACTION	NAO	OPCIONAL	OPCIONAL	OPCIONAL	OPCIONAL
CLOCKING	ASSINCRONO	ASSINCRONO	SINCRONO	ASSINCRONO	AMBOS
BANDA PASSANTE MAXIMA (BURST)	27,9 MB/S	95,2 MB/S	40 MB/S	25 MB/S	5 MB/S
BANDA PASSANTE: ACESSO A MEMORIA EM 150 ns (BURST)	13,6 MB/S	20,8 MB/S	13,3 MB/S	25 MB/S	5 MB/S
NUMERO MAXIMO DE DISPOSITIVOS	21	20	21	8	7
COMPRIMENTO MAXIMO DO BARRAMENTO	50 cm	50 cm	50 cm	50 metros	25 metros
PADRAO	IEEE 1014	IEEE 896.1	ANSI/ IEEE 1296	ANSI X3.129	ANSI X3.131

FIGURA 16: Características principais de alguns barramentos padrões.

O barramento externo permite a comunicação do PES com o ambiente externo. É através dele que se conectam interfaces com dispositivos de armazenamento e rede de comunicação. A opção por definir um barramento externo, com características padrões, busca versatilidade e expansibilidade. Atualmente existe no mercado uma vasta quantidade de interfaces para dispositivos de E/S que obedecem padrões elétricos e mecânicos de interfaceamento bastante difundidos. A tabela da figura 16 mostra as principais características de alguns padrões de barramentos.

Incorporando ao PES um barramento externo padrão, espera-se poder configurá-lo, conectanto placas comerciais, de forma adequada às necessidades do Multiplus. Além disto, torna-o flexível para suportas, tanto limitações de custo que envolvam inferfaces e dispositivos mais baratos, quanto atualizações tecnológicas mais sofisticadas pela simples substituição das placas conectadas ao barramento externo. Cabe ressaltar que, pelas características da arquitetura do PES, estas placas devem ter a capacidade de ser *Master/Slave* no barramento externo. Isto permite que que elas sejam capazes, tanto de receberem comandos da CPU do PES, quanto de realizarem transferências de dados diretamente com a *cache* de disco.

Na escolha do padrão de barramento externo proairou-se satisfazer algumas condições básicas:

- O barramento não poderia representar risco de se tornar um fator limitante do desempenho do PES
- Ter uma alta banda passante, tendo em vista as características das operações de E/S do Multiplus
- Ser um padrão de interligação bastante difundido e utilizado por fabricantes de interfaces para dispositivos de E/S

A opção foi pela utilização de um barramento externo padrão VME. Além de atender a todas as prerrogativas anteriores, o VMEbus tem ótima confiabilidade e interface elétrica simples, associada a uma grande variedade de chips-set de interfaceamento, cuja utilização é bastante fácil.

V.1.2 - A Cache de Disco

Pela arquitetura do Multiplus, pode-se ter até 16 elementos processadores requisitando operações de E/S a um único PES. Como o desempenho dos dispositivos de armazenamento está aquém das necessidades de E/S dos elementos processadores, cabe ao PES compensar esta defasagem buscando um maior equilíbrio do sistema. Cache de disco é uma técnica de projeto cujo objetivo é melhorar o desempenho do subsistema de E/S. Consiste, basicamente, na colocação de um *buffer* entre os dispositivos de asmazenamento e o PES, que mantém uma cópia de partes das informações contidas nos dispositivos. A CPU do PES implementa o algoritmo de armazenamento de

informações neste *buffer*, controlando todas as operações referentes a ele. Para que esta técnica seja eficiente, é necessário que algumas condições se verifiquem:

- O *buffer* deve capturar uma quantidade significativa de acessos aos dispositivos de armazenamento
- Os tempos de acesso e transferência de dados do *buffer* devem ser muito menores que os dos dispositivos de armazenamento
- O *buffer* não deve introduzir sobrecarga excessiva de processamento

A primeira condição pode ser satisfeita se os acessos aos dispositivos de armazenamento apresentarem alguma localidade, ou seja, se forem restritos a uma certa região destes dispositivos. Esta localidade é comum em acessos a banco de dados, diretórios, leitura sequencial de arquivos e na reutilização frequente de arquivos.

Em sistemas multiprogramados como o Multiplus, é comum a realização de operações de E/S de diversos processos independentes. Cada processo pode apresentar uma determinada localidade diferente. Desta forma, é comum que a totalidade dos acessos aos dispositivos de armazenamento sejam representados por alternância de localidades distintas. Esta alternância implica em movimentos extensos da cabeça de leitura/escrita destes dispositivos, resultando num grande dispêndio de tempo, já que o tempo para movimentação da cabeça, ou tempo de seek, é onde os dispositivos de armazenamento tem seu pior desempenho.

As *caches* de disco devem ser suficientemente grandes para manterem uma cópia das informações de várias localidades diferentes dos dispositivos de armazenamento ao mesmo tempo. Assim, os acessos intercalados às localidades presentes na *cache* não resultam em movimentação da cabeça de leitura/escrita, consequentemente, o tempo médio dos acessos de E/S é reduzido.

Para que a *cache* de disco represente vantagem, a segunda condição também deve ser satisfeita. Entretanto, ela pode ser facilmente alcançada com memórias semicondutoras comuns, visto que seus tempos típicos de acesso (120ns) são muito menores que os dos dispositivos de asmazenamento.

A última condição depende do algoritmo empregado no controle da *cache* de disco. Este algoritmo incorpora, basicamente, três critésios:

- Critério de busca
- Critério de carregamento
- Critério de invalidação

O critério de busca verifica a cada acesso à memória secundária se o bloco de informações solicitado está ou não presente na memória *cache.* Vale lembrar que um bloco é a quantidade mínima de dados efetivamente lida ou escrita nos dispositivos de armazenamento. O tempo gasto para se saber se o bloco está ou não na *cache* é chamado tempo de busca, e deve ser minimizado para não sobrecarregar o *overhead* de processamento. Os critérios de busca mais comuns são:

- BUSCA SEQUENCIAL: Consiste em "varrer" sequencialmente todos os blocos contidos na *cache.*

- ENDEREÇAMENTO DIRETO: Consiste em utilizar parte da referência do bloco solicitado (chave) como endereço de entrada em uma tabela que contém os blocos presentes na *cache.* Blocos com a mesma chave não podem estar presentes simultaneamente.

- ENDEREÇAMENTO ASSOCIATIVO: Critério intermediário entre os dois anteriores. Consiste em usar um endereçamento direto para ter acesso a um conjunto de blocos presentes na *cache* (conjunto associativo), no qual é feito uma busca sequencial. Dentro de um conjunto associativo os blocos possuem a mesma chave. A associatividade da tabela representa o número de blocos com a mesma chave possíveis de estarem presentes simultaneamente na *cache.*

O bom desempenho da *cache* de disco está condicionado à quantidade de blocos solicitados que se encontram presentes nela. Baseado nisto, o critério de carragamento procura determinar quais os próximos blocos a serem requisitados e carrega-os antecipadamente para a *cache.* Em geral, não se pode prever com precisão quais serão estes blocos. Sabe-se, no entanto, que os acessos sequenciais a arquivos são muito comuns. Desta forma, pode-se empregar uma técnica chamada leitura em avanço, carregando um certo grupo de blocos imediatamente sequencial ao bloco solicitado e com grande probabilidade de serem requisitados no futuro.

O último critésio é o de invalidação. Quando um determinado bloco solicitado está ausente da cache, ele é lido dos dispositivos de armazenamento e colocado na cache. Em regime normal, geralmente não há lugar vago na cache, obrigando a invalidação de algum grupo de blocos para que se possa alocar o novo grupo. O critério ótimo de invalidação é aquele que invalida o grupo cuja reutilização será a mais tardia. Lamentavelmente, este procedimento é de difícil irnplementação, já que requer o conhecimento das solicitação futuras. Os critérios de invalidação mais comuns são:

- *FIRST-IN-FIRST-OUT* (FIFO): O grupo invalidado é o que está presente a mais tempo na cache.

- LEAST RECENTLY USED (LRU): O grupo invalidado é o que está a mais tempo presente na cache sem reutilização.

- RANDÔMICO: Invalida um grupo randomicamente.

Para as operações de escrita, o comportamento do critério de invalidação é um pouco diferente. Estas operações podem desencadear invalidações na memória cache em função da política de escrita utilizada, que são basicamente duas:

- WRITE THROUGH: Sempre escreve o bloco nos dispositivos de armazenamento. Caso o bloco esteja presente na cache ele é atualizado.

- WRITE BACK: Nunca escreve o bloco diretamente nos dispositivos de armazenamento. Caso o bloco esteja presente na *cache* ele é atualizado e marcado como escrito. Caso não esteja, ele é colocado na cache. Temporariamente, os blocos escritos na cache são atualizados para os dispositivos de armazenamento, numa operação conhecida como *flush*.

Um estudo completo sobre cache de disco, incluindo várias simulações envolvendo as diversas alternativas dentro de cada um dos critérios do algoritmo de controle da cache, pode ser obtido em [SMITH85] e [FIGUEIRA88] e [PETER85].

V.1.3- Aspectos Relevantes no Projeto do DMA

Entre outros fatores, o bom desempenho de um subsistema de E/S está condicionado a um bom desempenho do PES, já que ele faz a ligação entre os

dispositivos de armazenamento e os elementos processadores. Conforme visto anteriormente, o bom desempenho do PES quando submetido a uma carga de trabalho decorrente de uma aplicação científica está fortemente associado a sua banda passante e seu *oveuhead.*

A banda passante diz respeito à taxa de transferência de dados por unidade de tempo. Como é o DMA que realiza as transferências entre o PES e os elementos processadores, sua banda passante não pode ser um fator limitante na taxa total de transferência do PES. Isto significa que sua banda passante deve ser superior, ou à banda passante da *cache* de disco ou à banda passante do barramento dos elementos processadores, o que for menor dos dois valores.

Uma outra condição necessária ao DMA está relacionada ao seu *oveuhead.* Em sistemas multiprocessados como o Multiplus, várias operações de E/S podem ser solicitadas ao mesmo tempo, envolvendo diversos dispositivos de armazenamento. Para não indroduzir *overhead* por salvamento de contexto de transferências, o DMA deve ser multicanal, ou seja, permitir a realização de diversas transferências simultâneas. O número satisfatório de canais depende da quantidade de dispositivos de armazenamento associados ao PES e, logicamente, da carga de trabalho submetida ao subsistema de E/S.

V.2 - *O* Fluxo de Informações no PES

Para melhor compreender a arquitetura proposta para o PES do subsistema de E/S do Multiplus é necessário conhecer o fluxo de informações no seu interior, ou seja, o tratamento dado aos pedidos de E/S e a forma com que eles desencadeiam as transferências dos dados correspondentes.

O fluxograma da figura 17 ilustra, simplificadamente, o fluxo de informações no PES. As linhas duplas representam o fluxo dos dados, as linhas simples o fluxo de pedidos, e as linhas tracejadas as interrupções de aviso de término de tarefa. O fluxograma é composto por quatro algoritmos principais e independentes. O Algoritmo de Controle Geral (ACG) é o mais importantes deles. Ele é executado pela CPU do PES e controla e gerencia todas as suas operações. Este algoritmo pode ser dividido em três módulos básicos:

- Módulo de controle da *cache* de disco

- Módulo de desmembramento dos pedidos de E/S em sequências de tarefas internas

- Módulo gerenciador das filas de pedidos de E/S e das tarefas internas

O Algoritmo de Controle de Transferências (ACT) é responsável pelas transferências de dados entre a memória cache de disco e as memórias locais dos elementos processadores via barramento do *cluster*, e é executado pelo MC68020 utilizado para emular o DMA do PES.

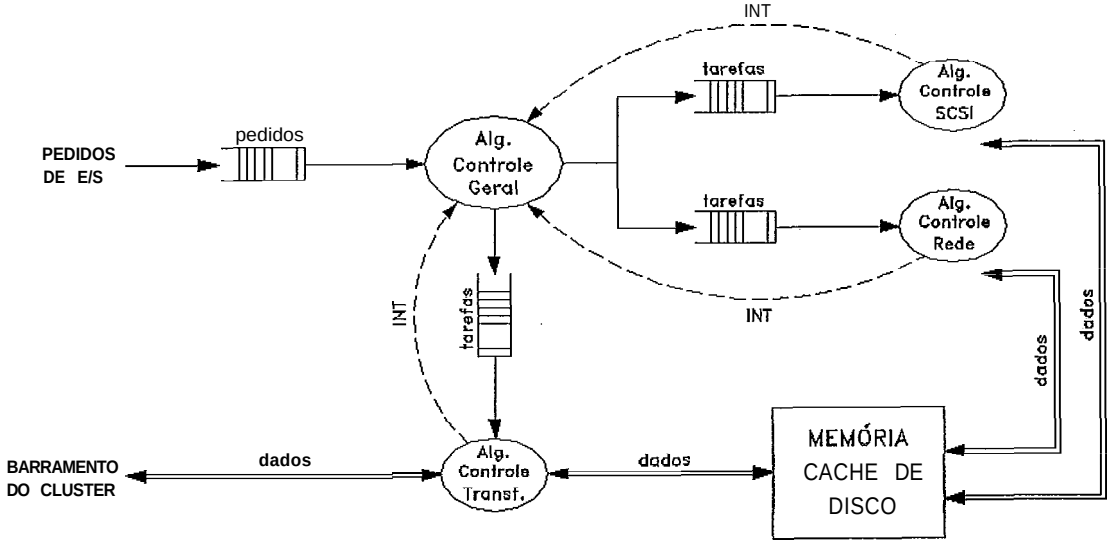

FIGURA 17: Fluxograma ilustrativo do fluxo de informagões no PES.

Por último, os dois algoritmos restantes são os responsáveis pela operagões junto ao barramento externo do PES. O Algoritmo de Controle da Interface SCSI (ACS), executado pela placa controladora SCSI inteligente, é responsável pelo gerenciamento do protocolo de acesso aos dispositivos de armazenamento e pelas transferências de dados entre eles e a cache de disco. Similarmente, o Algoritmo de Controle da Interface da Rede de Comunicação (ACR), executado pela placa controladora inteligente de rede, é responsável pelo gerenciamento do protocolo de acesso à rede de comunicação e pelas transferências de dados entre a rede e a cache de disco.

Os pedidos de E/S são recebidos, no PES, pelo Algoritmo de Controle Geral. Ele faz a sua organização em forma de fila e interpreta-os de forma a desmembrá-los em sequências de tarefas internas ao PES, conforme ilustrado na figura 18. A sequência de tarefas é função do tipo de pedido de E/S e da presenga ou não dos dados solicitados na *cache* de disco.

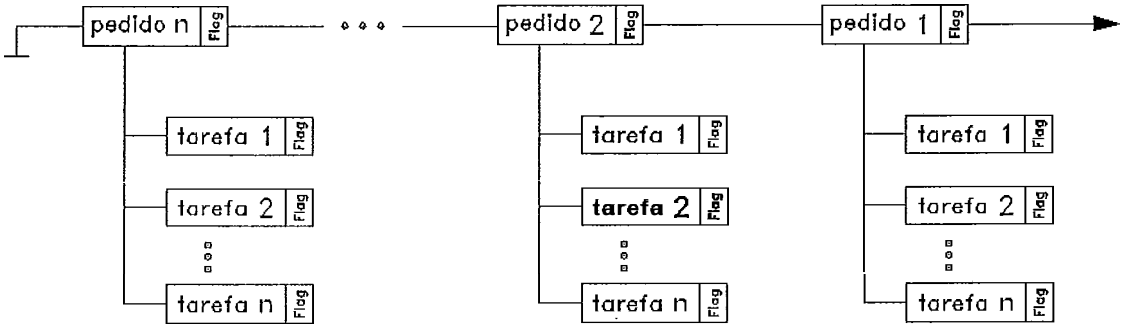

FIGURA 18: Fila de pedidos de E/S no interior do PES.

As tarefas podem ser associadas aos dispositivos de armazenamento, à rede de comunicação ou ao DMA. Para cada um destes recursos, o Algoritmo de Controle Geral organiza uma fila de tarefas, que são ordenadas de modo a otimizar a utilização dos recursos, minimizando o *ovevhead* médio do PES. As tarefas são executadas segundo a disponibilidade dos recursos, porém, obedecendo sua sequencialidade dentro do pedido. A medida em que as tarefas vão se completando, completam-se também os pedidos de E/S. Os pedidos completados são retirados da fila e os elementos processadores solicitantes avisados de seu término.

Quanto ao fluxo dos dados no PES, pode-se perceber, observando o fluxograma, que a memória *cache* de disco é o seu elemento central. Todas as transferências de dados do PES são comandadas por tarefas disparadas pelo Algoritmo de Controle Geral, e envolvem sempre a memória *cache.* Os tipos de transferências são basicamente três:

- Transferência entre o barramento do *cluster* e a memória *cache* de disco

- Transferência entre os dispositivos de armazenamento e a memória *cache* de disco

- Transferência entre a rede de comunicação e a cache de disco

Nem todos os dados envolvidos nestas transferências são *cacheable*. Para simplificar o controle da coerência dos dados na cache de disco, apenas os dados locais ao PES são considerados *cacheable*. Desta forma, os dados remotos ao PES, transferidos através da rede de comunicação, utilizam a memória cache de disco apenas como *buffer*, sendo classificados de not *cacheable* enquanto presentes nela.

Antes de descrever, resumidamente, a sequência de tarefas inerente a cada um dos tipos de pedidos de E/S possíveis de serem enviados ao PES, cabe distinguir duas fontes distintas destes pedidos. Uma através da memória de comunicação com o barramento do cluster, oriunda dos elementos processadores que compõem o cluster a que o PES está associado, e outra através da interface com a rede de comunicação de alta velocidade, oriunda dos demais PES do subsistema de E/S do Multiplus. Ambas as fontes de pedidos de E/S se fazem perceber interrompendo o processamento da CPU do PES. Qualquer que seja a origem do pedido, o tratamento dispensado é homogêneo, isto é, de uma forma ou de outra eles são conduzidos ao Algoritmo de Controle Geral e executados sem distinção.

SEQUÊNCIA DE TAREFAS:

Pedido de Leitura de DADO Local com Hit na Cache de Disco:

TAREFA 1: Executada pelo ACT. Transferir o DADO da cache de disco para a memória local do elemento processador solicitante.

Pedido de Leitura de DADO Local com Miss na Cache de Disco:

TAREFA 1: Executada pelo ACS. Ler o DADO dos dispositivos de armazenamento e colocá-lo na cache de disco.

TAREFA 2: Executada pelo ACT. Transferir o DADO da cache de disco para a memória local do elemento processados solicitante.

Pedido de Escrita de DADO Local:

TAREFA 1: Executada pelo ACT. Transferir o DADO da memória local do elemento processados solicitante para a cache de disco.

TAREFA 2: Executada pelo ACS. Copiar o DADO da *cache* de disco e escrevê-lo nos dispositivos de armazenamento.

Pedido de Leitura de DADO Remoto com Hit na Cache de Disco Remota:

PES ORIGEM	PES DESTINO
TAREFA 1: Executada pelo ACR. Transmitir comando de solicitação de leitura de dado via rede de comunicação, preparando-se para recebê-lo e colocá-lo na *cache* de disco.	
	TAREFA 1: Executada pelo ACR. Ler o DADO da *cache* de disco e transmití-lo via rede de comunicação para o PES origem.
TAREFA 2: Executada pelo ACT. Transferir o DADO da *cache* de disco para a memória local do elemento processador solicitante.	

Pedido de Leitura de DADO Remoto com Miss na Cache de Disco Remota:

PES ORIGEM	PES DESTINO
TAREFA 1: Executada pelo ACR. Transmitir comando de solicitação de leitura de dado via rede de comunicação, preparando-se para recebê-lo e colocá-lo na *cache* de disco.	
	TAREFA 1: Executada pelo ACS. Lei o DADO dos dispositivos de armazemamento e colocá-lo na *cache* de disco.
	TAREFA 2: Executada pelo ACR. Ler o DADO da *cache* de disco e transmití-lo via rede de comunicação para o PES origem.

TAREFA 2: Executada pelo ACT. Transferir o DADO da *cache* de disco para a memória local do elemento processador solicitante.

Pedido de Escrita de DADO Remoto:

PES ORIGEM | PES DESTINO

TAREFA 1: Executada pelo ACT. Transferir o DADO da memória local do elemento processador solicitante para a *cache* de disco.

TAREFA 2: Executada pelo ACR. Transmitir comando de solicitagão de escrita de dado via rede de comunicação.

TAREFA 1: Executada pelo ACR. Transmitir comando de READY via rede de comunicagão, preparando-se para receber o DADO e colocá-lo na *cache* de disco.

TAREFA 3: Executada pelo ACR. Ler o DADO da *cache* e transmiti-lo via rede de comunicação para o PES destino.

TAREFA 2: Executada pelo ACS. Retirar o DADO da *cache* de disco e escrevê-lo nos dispositivos de armazenamento.

V.3 - Implementação do PES

A implementagão dos diversos módulos da arquitetura do PES é discutida nesta segão. Em vários casos, restrigões de ordem prática impuseram limitações na definição da irnplernentagão a ser adotada.

V.3.1 - Interface com os Dispositivos de Armazenamento

Como se sabe, quanto maior a taxa de transferência dos dispositivos de armazenamento, melhor o desempenho do subsistema de E/S. O principal dispositivo de armazenamento é o disco magnético. Apesar de existirem discos, tais como o IBM 3380-AK4 e o FUJITSU M2361A, com taxa de transferência de até 3,0 M*Bytes*/s [PATTER90], a disponibilidade do projeto Multiplus é de discos do tipo *Winchester*, cuja taxa de transferência é de 700 K*Bytes*/s. Em função desta restrição, definiu-se o padrão de interface com os dispositivos de armazenamento como sendo o SCSI, já que é o mais difundido nesta classe de dispositivos.

Mesmo com esta restrição, ainda pode-se conseguir um bom desempenho dos dispositivos de armazenamento. Uma alternativa viável é o agrupamento de discos tipo *Winchester* com o objetivo de se explorar a técnica de *striping*. Já existem, comercialmente, controladoses de *disk-array* para barsamento SCSI que exploram esta técnica, facilitando sua implementação. Desta forma, pode-se assumir a restrição imposta com maior conforto em relação ao desempenho do subsistema de E/S.

V.3.2- CPU

Os elementos processadores do Multiplus são baseados em microprocessadores RISC de arquitetura SPARC. O ideal seria que este tipo de microprocessador fosse a base do processamento de todo o Multiplus, inclusive do PES. Isto tornaria a arquitetura mais homogênea quanto ao tratamento dado ao *software*, permitindo o desenvolvimento de quaisquer novas rotinas usando o próprio Multiplus. Além disto, facilitaria a migração de parte do Sistema Operacional responsável pelas operações de E/S para o PES.

Entretanto, as limitações de custo impuseram restrições quanto a CPU utilizada. Devido ao alto preço do *chip-set* de microprocessadores SPARC e por eles possuirem uma capacidade de processamento muito além das necessidades do PES, optou-se pela utilização de um microprocessador MOTOROLA MC68010 para controlar as operações de E/S do PES. As principais razões para a escolha deste microprocessador foram:

- O Sistema Operacional do Multiplus, o MULPLIX, é uma evolução do PLURIX, Sistema Operacional UNIX-*like* desenvolvido no NCE/UFRJ, e

cuja plataforma original de processamento são os microprocessadores MOTOROLA da família MC68000. Como o domínio sobre este Sistema Operacional é total, e existem no NCE várias máquinas rodando o PLURIX, torna-se fácil, com a utilização deste microprocessador, a tarefa de migração de rotinas do Sistema Operacional para o PES.

- O MC68010 é de fácil interface com o barramento VME, escolhido como barramento externo do PES. Os sinais do VME são, praticamente, uma extensão dos sinais do MC68010, tornando simples o projeto da interface de comunicação entre eles.

O barramento interno do PES que está associado à CPU é, praticamente, uma extensão de seus sinais. Este barramento também é compartilhado com diversos dispositivos. A maioria destes dispositivos tem uma largura de palavra igual ou inferior a 16 bits, que é a largura da palavra do MC68010. Para não haver limitações de desempenho, o barramento deve ter um tamanho de palavra pelo menos igual a dos dispositivos que, nele, realizam operações como *master*. Como o MC68010 é o único *master* deste barramento, sua palavra foi definida com uma largura de 16 *bits*.

V.3.3 - Cache de Disco

A implementação da cache de disco resultou da definição de suas características. Uma das características mais significativas e a primeira a ser definida foi o tamanho da cache. O tamanho que ofereceu melhor custo/desempenho, proporcionando ao PES boa probabilidade de hit na cache nas operações de E/S sem dispor de muito espaço físico na placa, foi o de 32 M*Bytes*. Sua implementação utiliza oito módulos SIMM de memória dinâmica de 4 M*Bytes* cada, compondo um único banco de memória de 64 bits de largura.

Para permitir o acesso à memória cache tanto pelo DMA quanto pelo barramento externo do PES, utiliza-se um controlados *dual-ported* de memória dinâmica, que controla todos os acessos à cache. O tempo necessário à realização de um ciclo completo de acesso à memória cache é de aproximadamente 200 ns, entretanto, nas operações em rajada (*burst*), o tempo de acesso é bem menor, cerca de 125 ns.

O algoritmo de controle da cache de disco é implementado por *software* e executado pela CPU do PES. Na definição das características deste algoritmo procurou-se, sempre, a simplificação de sua implementação. Suas principais características são:

- Política de escrita: WRITE THROUGH
- Critério de busca: Endereçamento associativo com associatividade igual a 4
- Critério de carregamento: Leitura em avanço tanto do bloco solicitado quanto dos subsequentes, até o final da trilha
- Critério de invalidação: LRU

V.3.4 - DMA

A cache de disco é irnplementada com *chips* de memória dinâmica, tem uma palavra de 64 bits e tempo de acesso de 125 ns no modo rajada, resultando numa banda passante de 64 M*Bytes*/s. O barrarnento do clusfer de elementos processadores do Multiplus, implementado numa tecnologia apropriada, o BIL (Backplane Transceiver Logic), possui uma banda passante máxima bem superior, chegando a 100 M*Bytes*/s. Desta forma, o DMA, elemento realizador de transferências entre a memória cache e o barramento do *cluster* deve, para não ser um limitados de desempenho, ser capaz de transferir dados a uma taxa mínima idêntica a da cache, ou seja, de 64 M*Bytes*/s.

Existe um desinteresse por parte dos fabricantes de DMA em desenvolverem versões comerciais que acompanhem a velocidade dos microprocessadores. Salvo algumas exeções, tais como o INTEL 82380, a grande totalidade dos DMAs tem um desempenho aquém do desejado. A MOTOROLA, por exemplo, tem, para a família MC68000, um DMA cuja taxa máxima de transferência é de apenas 5 M*Bytes*/s.

No estudo das arquiteturas de E/S das máquinas de alto desempenho existentes, verifica-se que, geralmente, elas envolvem o projeto de um DMA *custom* adequado às suas necessidades. Como a primeira versão do projeto Multiplus não engloba o desenvolvimento de um DMA *custom*, algumas alte nativas foram estudadas:

- Utilização de um DMA de alto desempenho compatível com outra família de microprocessadores e arcar com o custo de interfaceamento com o MC68010.

- Utilização de um microprocessador especificamente para desempenhar a função de um DMA.

A segunda opção foi a escolhida para implementação, e o microprocessador utilizado foi o MC68020. Isto se deveu, basicamente, a dois fatores. Primeiramente o custo: um MC68020 custa menos de um quinto do valor de um DMA de alto desempenho. Segundo, por ser suficiente às necessidades do PES.

Por possuir internamente uma cache de instruções, a capacidade do MC68020 de transferir dados é bastante agilizada. Um pequeno programa de transferência, executado na cache interna de um MC68020 à 25MHz pode, com auxílio de um hardware bastante simples, realizar transferências de dados na taxa máxima suportada pela cache de disco do PES. Além do mais, esta alternativa simplifica a interface entre os dois barramentos internos do PEÇ. O barramento da CPU não mais necessita ser *master* no barramento do DMA para programá-lo. Utilizando-se um microprocessador como DMA, a comunicagão entre os barramentos pode ser feita através de uma memória *dual-ported*. Maiores informações sobre a utilização de um MC68020 como DMA pode ser conseguida em [MOTOROLA87].

O barramento interno do PES que está associado ao DMA é, praticamente, uma extensão dos sinais do MC68020. Entretanto, para permitir que, através de um artifício de hardware, as transferências envolvendo a cache de disco sejam realizadas em acessos de 64 *bits,* este barramento também foi definido com uma palavra de 64 bits de largura. Somente as operações com a cache explora a largura total do barramento. Para as demais operações, o barramento se comporta como se tivesse 32 bits de largura.

V.3.5 - Demais Componentes do PES

Além da CPU, que controla todo o PES, e do DMA, que realiza as transferência de dados, vários outros componentes integram o processador de E/S do Multiplus.

No barramento da CPU também estão conectados uma interface sesial, um *timer*, uma memória local e uma memória de comandos. A interface serial tem uma função apenas auxiliar, permitindo a conexão de um terminal de vídeo e, consequentemente, que haja uma interação com o PES independentemente dos elementos processadores do Multiplus. Este recurso é muito útil na fase de depuração e manutenção, ou em eventuais testes do PES.

O *timer* é utilizado na monitoração dos acessos ao barramento externo do PES. Com interrupções regulares à CPU, geradas por este componente, o PES é capaz de fornecer, estatisticamente, a taxa de ocupação de cada dispositivo de armazenamento e da rede de comunicação. Estas informações serão úteis no estudo futuro do comportamento do subsistema de E/S, permitindo avaliar o seu desempenho em função das diversas alternativas de distribuição dos processos pelos elementos processadores do Multiplus.

Uma memória de comunicação faz a interação dos elementos processadores com o PES. Através dela, os elementos processadores passam comandos ao PES e recebem o status correspondente. Esta memória tem o tamanho de 2 K*Bytes*, sistema de endereçamento *dual-ported* e é compartilhada por todos os elementos processadores do Multiplus associados a este PES. Por último, uma memória local, composta de 2 M*Bytes* de memória RAM e 64 K*Bytes* de memória ROM é reservada ao *firmware* de controle do PES.

No barramento do DMA, além da memória *cache* de disco, está conectada uma interface com o barramento do *cluster* de elementos processadores do Multiplus.

CAPÍTULO VI

SIMULAÇÃO E ANÁLISE QUANTITATIVA

Nos dois capítulos anteriores definiu-se, respectivamente, a arquitetura do subsistema de E/S do Multiplus e as características dos processadores de E/S (PES) que o compõe. Entretanto, não se abordou o seu comportamento quando submetido a uma carga de E/S. O objetivo deste capítulo é, além de dar uma idéia desse comportamento, avaliar o impacto das operações de E/S no desempenho dos elementos processadores do Multiplus. Para tal, procurou-se gerar dados, através de análises quantitativas e simulações, que pudessem retratar a operacionalidade do subsistema de E/S dentro do contexto do Multiplus.

O capítulo se divide basicamente em duas partes. A primeira delas apresenta as questões avaliadas por dados oriundos de simulações. Descreve o simulador utilizado e as modificações nele introduzidas para incorporar o subsistema de E/S, além dos critérios adotados para simulação e os resultados obtidos. A segunda parte apresenta as questões avaliadas por dados oriundos de análises quantitativas, onde serão discutidas as questões relativas às características internas do PES.

VI.1 - Simulação

Utilizando-se a técnica de simulação pode-se reproduzir, através de um programa computacional chamado simulador, situações próximas às reais de operacionalidade do Multiplus, e extrair informações que só estariam disponíveis após sua implementação.

As principais questões que procurou-se avaliar por meio de simulações foram as relativas à arquitetura do subsistema de E/S do Multiplus. Elas podem ser divididas em dois grupos:

- Questões inerentes aos *clusters* de elementos processadores decorrente das operações de E/S.
- Questões inerentes à rede de comunicação destinada às operações de E/S entre *clusters* de elementos processadores.

A validade dos resultados das simulações está diretamente relacionada à fidelidade do modelamento do comportamento de cada um dos dispositivos que compõem o sistema. Este modelamento é função integrante do simulador, que é descrito a seguir.

VI.1.1- *O* Simulador Original

O simulador utilizado para avaliação das questões mencionadas anteriormente foi uma evolução de um simulador desenvolvido no NCE/UFRJ por [MESLIN91] com objetivo de estudar o desempenho dos elementos processadores do Multiplus frente às diversas opções de política de cache disponíveis. Desta forma, o simulador não constitui parte integrante deste trabalho, apesar de terem sido feitas modificações que resultaram no modelamento do subsistema de E/S, incorporando ao simulador as operações de E/S dos elementos processadores. Ele é considerado apenas como uma ferramenta utilizada neste trabalho.

Para permitir a interpretação dos resultados oriundos da simulação, será descrito, a seguir, o princípio de funcionamento do simulador original e, posteriormente, as modificações nele introduzidas. Informações mais detalhadas à respeito do simulador podem ser conseguidas em [MESLIN91].

Para avaliar as políticas de cache, o simulador retrata a operacionalidade dos elementos processadores do Multiplus. Cada elemento processador possui uma unidade inteira (u.i.), que demanda acessos continuamente, memórias cache de dado e código separadas e uma memória local. Como a memória do Multiplus é global e compartilhada, é permitido a um elemento processador acessar a memória local de outro. Este acesso pode ser interno ao cluster ou via rede de interconexão, caso o elemento processador referenciado pertença a um cluster distinto. A partir de uma distribuição de probabilidade de ocorrência dos diversos tipos de acessos possíveis, o simulador avalia a influência da interdependência destes acessos no desempenho global do Multiplus.

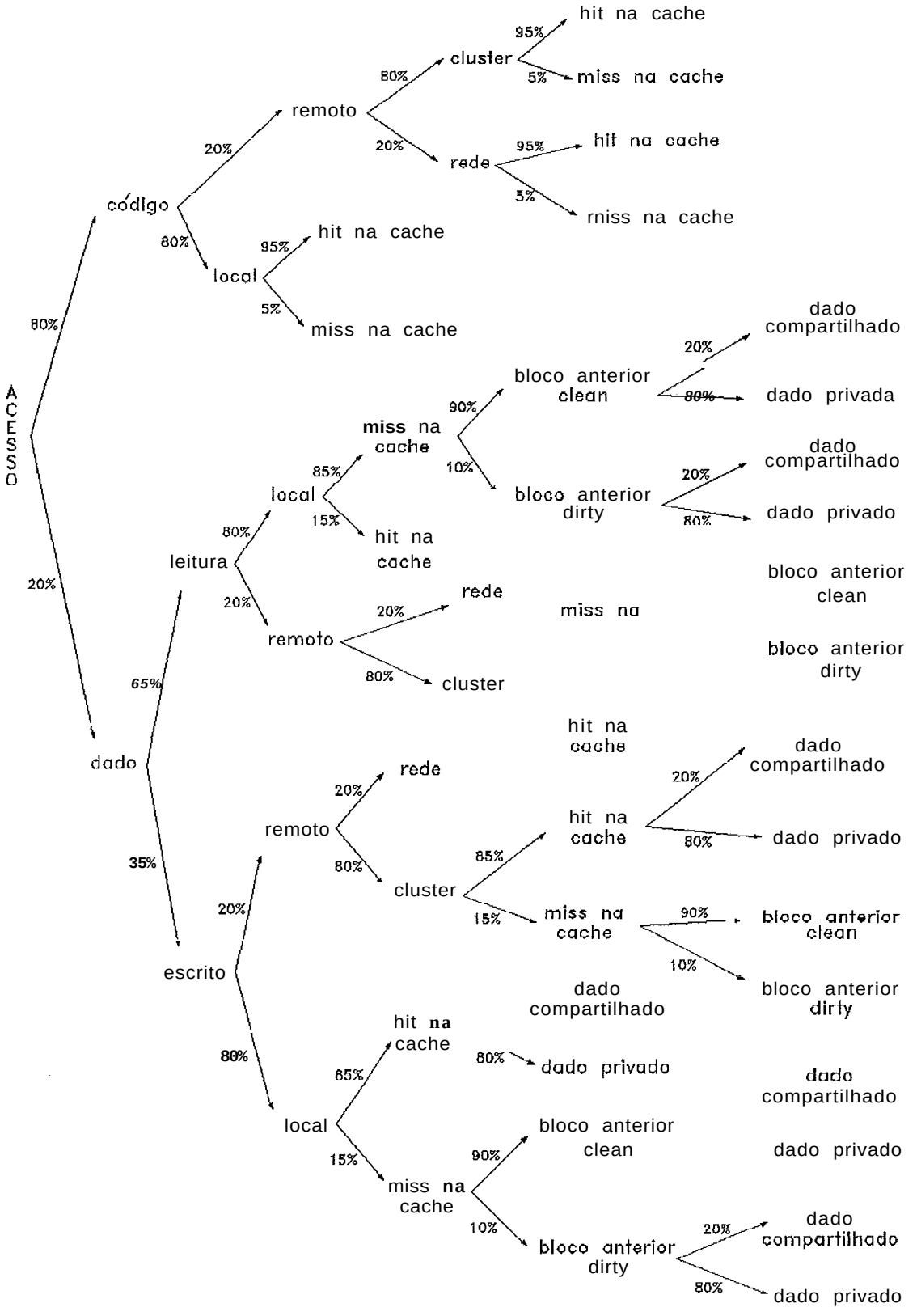

FIGURA 19: Tipos de acessos possíveis às unidades inteiras e suas probabilidades de ocorrência.

O princípio básico de funcionamento do simulador original pode ser representado por um loop de três passos:

```
C PASSO 1: Sorteio do tipo de acesso da u.i.
  PASSO 2: Alocação de recursos referentes ao acesso
  PASSO 3: Gerência dos recursos
```

O primeiro passo determina, através de um sorteio aleatório, o tipo de acesso das unidades inteiras (u.i.) em estado livre de cada um dos elementos processadores. Após o sorteio do acesso, as unidades inteiras passam para o estado espera. Os tipos de acessos possíveis às unidades inteiras, juntamente com suas probabilidades de ocorrência, são mostrados na figura 19.

O segundo passo faz a alocação dos recursos necessários à execução de cada um dos acessos determinados no passo anterior. Caso a alocação seja bem sucedida, ou seja, todos os recursos necessários à execução do acesso estejam disponíveis, eles são ocupados, e a unidade inteira correspondente é colocada no estado execução, caso contrário, a u.i. permanece no estado espera. Cada um dos acessos ocupa os recursos necessários por uma quantidade determinada de ciclos de relógio. Os diversos recursos disponíveis são:

- *Cnche* de dado
- *Cache* de código
- Memória local
- Barramento de dado
- Barramento de código
- Rede de interconexão

O terceiro passo gerencia os recursos e faz a computação dos ciclos, isto é, incrementa o contador total de ciclos e controla os acessos em execução. Quando um acesso se finda, os recursos envolvidos são liberados e a unidade inteira correspondente é colocada no estado livre.

O código de programa a seguir é uma simplificação do núcleo básico de funcionamento do simulador. Ele visa dar uma idéia do contexto no qual cada um destes passos é executado.

```
begin
    inicializa;
    for loop = 1 to número–ciclos do
    begin
        for i.cluster = 1 to número_cluster do
        begin
            for i.ui = 1 to número–LUdo
            begin
                if ui[i.cluster, i.ui].status = livre then
                begin
                    doca–LU;          (PASSO 1)
                end;
            end;
        end;
        case número–política of
            SEMCACHE: aloca_semcache;
            WRITETHROUGH: aloca_writethrough;     (PASSO 2)
            WRITEBACK: aloca_writeback:
        end;
        contabiliza;
        gerencia_recursos;          (PASSO 3)
    end;
    analisa–resultados;
end;
```

Além das rotinas referentes aos três passos citados, algumas outras merecem destaque. A rotina inicializa é responsável pela configuração inicial do simulador. Em função dos parâmetros de simulação recebidos, ela inicializa as variáveis, dando início à simulação. A rotina contabiliza faz a contagem, para todos os recursos, da quantidade de acessos que cada um efetuou e o número de ciclos que eles ficaram livres. Por último, a rotina analisa_resultados organiza os dados em forma de tabelas e salva-os no disco.

VI.1.1.1-As Modificações no Simulador

As modificagões introduzidas no simulador tiveram o objetivo de incorporar o subsistema de E/S. Assim, as operações de E/S passam a compartilhar, juntamente com as unidades inteiras, os diversos recursos do sistema. De forma análoga, o princípio básico de funcionamento do simulador modificado pode ser representado por um *loop* de quatro passos:

PASSO 1: Sorteio do tipo de acesso da u.i.
PASSO 2: Alocação dos recursos referentes ao acesso
PASSO 3: Alocação das operações de E/S
PASSO 4: Gerência dos recursos

Os dois primeiros passos permaneceram inalterados em relação ao simulador original. Já o terceiro passo é inteiramente novo. Ele é o responsável pela alocação das operações de E/S e dos recursos envolvidos nestas operações. Pode-se notar, pela posigão deste passo em relação aos demais, que a prioridade na alocagão dos recursos necessários às operagões de E/S é inferior a das unidades inteiras. O quarto passo sofreu algumas alterações para também gerenciar os PES, que passam a ser mais um recurso disponível no *cluster*.

O código de programa simplificado com as modificações realizadas no simulador é basicamente o mesmo mostrado anteriosmente. A principal diferença é a inclusão da rotina que faz a alocação das operações de E/S. A posição desta rotina dentro do programa é imediatamente acima da rotina contabiliza.

VI.1.2 - Questões Inerentes ao Cluster de Elementos Processadores

A simulação das questões inerentes ao *cluster* de elementos processadores tem o objetivo de avaliar o comportamento dos elementos processadores e dos barramentos do *cluster* em função das operações de E/S. As principais questões são:

- A escolha do barramento onde serão efetuadas as operações de E/S: barramento de código ou de dado.
- A degradação do desempenho dos elementos processadores medido pela duração média dos acessos executados.

- A taxa de transferência média efetivamente obtida nas operações de E/S.

Antes de iniciar a análise dos resultados das simulações é necessário esclarecer alguns dos critérios adotados. Como se sabe, o perfil esperado para as operações de E/S do Multiplus é de transferências esparsas e extensas. Entretanto, estas transferências ocorrem, no barramento, em conjunto com os acessos externos das unidades inteiras dos elementos processadores. Devido a disparidade na periodicidade destes dois eventos, a quantidade de ciclos de relógio simulados necessários para conciliá-los é muito grande, resultando num empecilho prático. A alternativa encontrada para contornar este problema foi simular o comportamento do sistema durante uma fração de tempo na qual as transferências de E/S estão ocorrendo e estender o resultado, analiticamente, para diversas cargas de E/S possíveis de serem impostas pelos elementos processadores.

As transferências de E/S são efetuadas em várias rajadas ininterruptas de 128 bytes cada. Toda rajada é precedida de urna nova arbitração. Desta forma, durante as transferências de E/S, os PES estão sistematicamente tentando alocar os recursos necessários à realização das rajadas. Como a prioridade dos PES na alocação dos recursos é menor que a das unidades inteiras, a quantidade de rajadas realizadas por ciclo de relógio é função da taxa de ocupação dos recursos compartilhados. Quanto mais livres estiverem os recursos, maior a banda passante efetivamente obtida nas transferências de E/S.

Por íiltimo, cabe esclarecer que a avaliação do impacto das operações de E/S no comportamento dos elementos processadores do Multiplus foi restrita às possibilidades reais de implementagão. Desta forma, alguns parâmetros do simulador não foram variados, mantendo-se fixos nos valores correspondentes à definição adotada para implementação do Multiplus:

- Política de *cache*: WRITE THROUGH
- Barramento de código e dado separados
- Largura de cada barramento: 64 *bits*
- Velocidade de processamento das unidades inteiras: 25 MHz

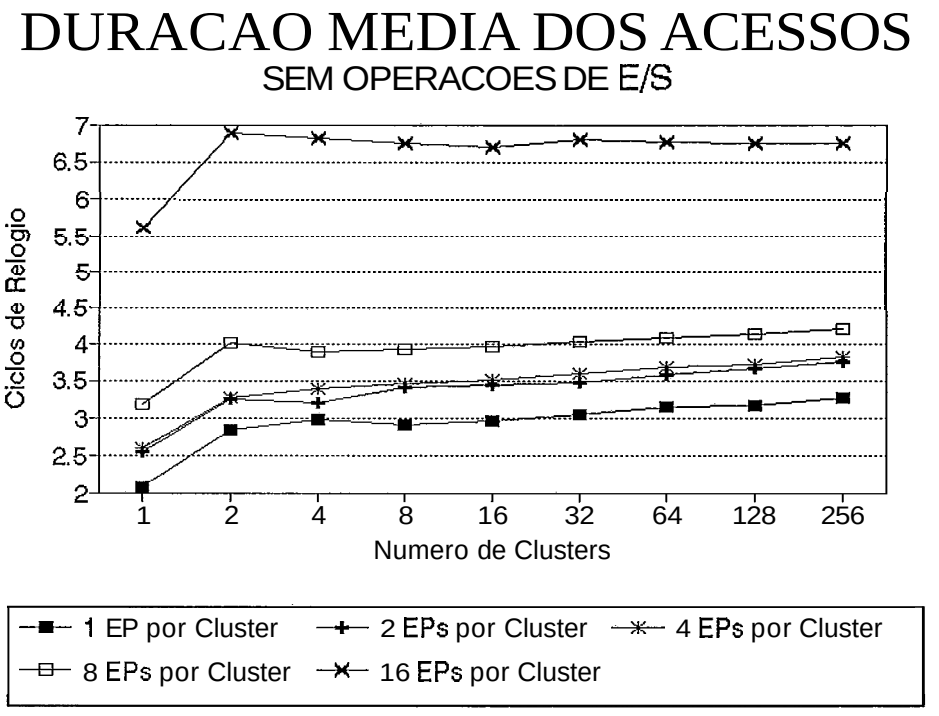

FIGURA 20: Duração média dos acessos externos executados pelas unidades inteiras sem as operações de E/S.

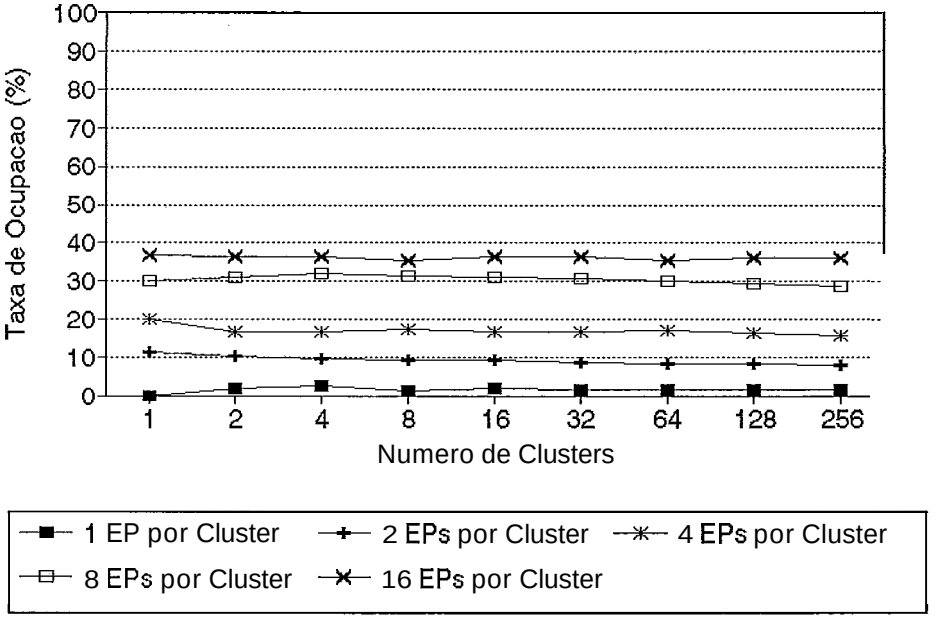

FIGURA 21: Taxa de ocupação do barramento de código do *cluster* de elementos processadores sem operações de E/S.

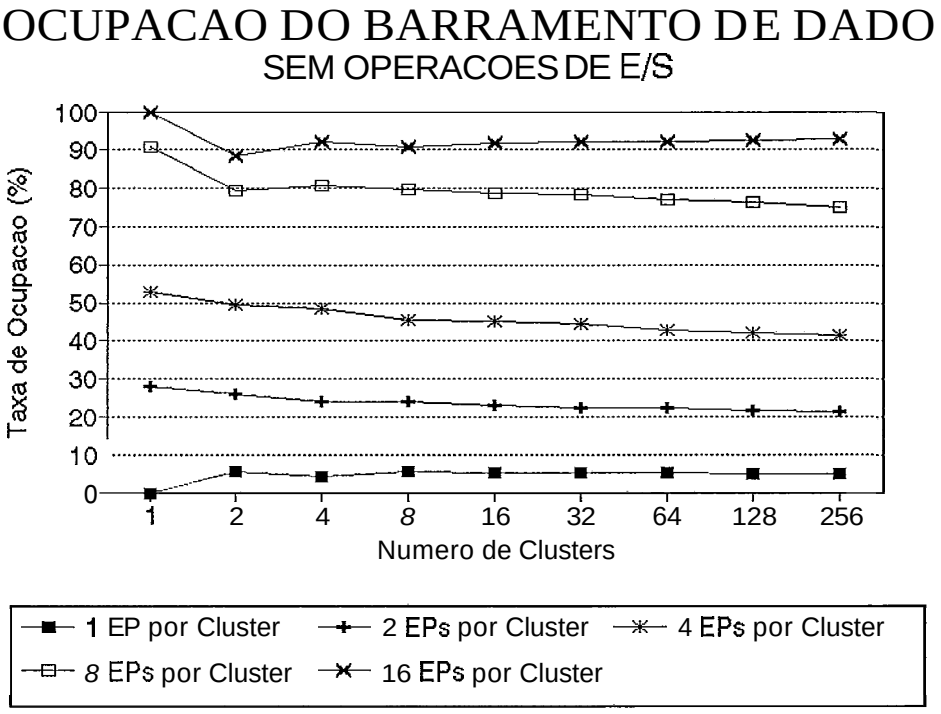

FIGURA 22: Taxa de ocupagão do barramento de dado do *cluster* de elementos processadores sem operagões de E/S.

Para iniciar a análise dos resultados das simulagões cabe, primeiramente, mostrar algumas características do comportamento dos elementos processadores do Multiplus sem o modelamento do subsistema de E/S. O gráfico da figura 20 ilustra a duração média dos acessos externos executados pelas unidades inteiras dos elementos processadores do Multiplus em função das diversas possibilidades de configuração. Correspondentemente, as taxas de ocupação dos barramentos de código e dado são mostradas, respectivamente, nos gráficos das figuras 21 e 22.

A primeira questão avaliada foi a escolha do barramento do *cluster* no qual serão realizadas as operações de E/S. Como as informações contidas nas operações de E/S são tratadas como dados pelos elementos processadores, é sugestivo pensar que, para manter uma homogeneidade, os PES devem utilizar o barramento de dados para realizá-las. Entretanto, mais importante que a homogeneidade está o desempenho do sistema. As figuras 21 e 22 mostram que o barramento de dados, pelas próprias características do Multiplus, é bem mais congestionado que o barramento de código, principalmente nas configurações com muitos elementos processadores por *cluster*. Desta forma, optou-se por realizar as

operações de E/S através do barramento de código, aproveitando sua maior ociosidade. Todas as questões subsequentes serão avaliadas considerando a utilização deste barramento para realização das operações de E/S. Porém, com o objetivo de ratificar esta opção, avaliações semelhantes utilizando o barramento de dados para as operações de E/S são mostradas no Anexo A, permitindo o confronto dos resultados.

A segunda questão avaliada foi a taxa de transferência média efetivamente obtida nas operações de E/S, isto é, a banda passante do barramento do *cluster* para este tipo de operação. O procedimento adotado foi simular o comportamento do barramento de código do Multiplus durante uma transferência de E/S monitorando a quantidade de rajadas efetuadas por unidade de ciclo. Para permitir a interpretação dos resultados obtidos é necessário destacar alguns critérios adotados. Cada acesso ao barramento de código dentro das rajadas tem duração de 2 ciclos. Como as rajadas são de 128 *bytes* e a largura do barramento é de 64 *bits*, são necessários 16 acessos, totalizando 32 ciclos para tranferí-los. Somando-se a estes 32 ciclos outros 8 ciclos referentes ao atraso até a aquisição do barramento, a duração de cada rajada totaliza 40 ciclos de relógio, resultando

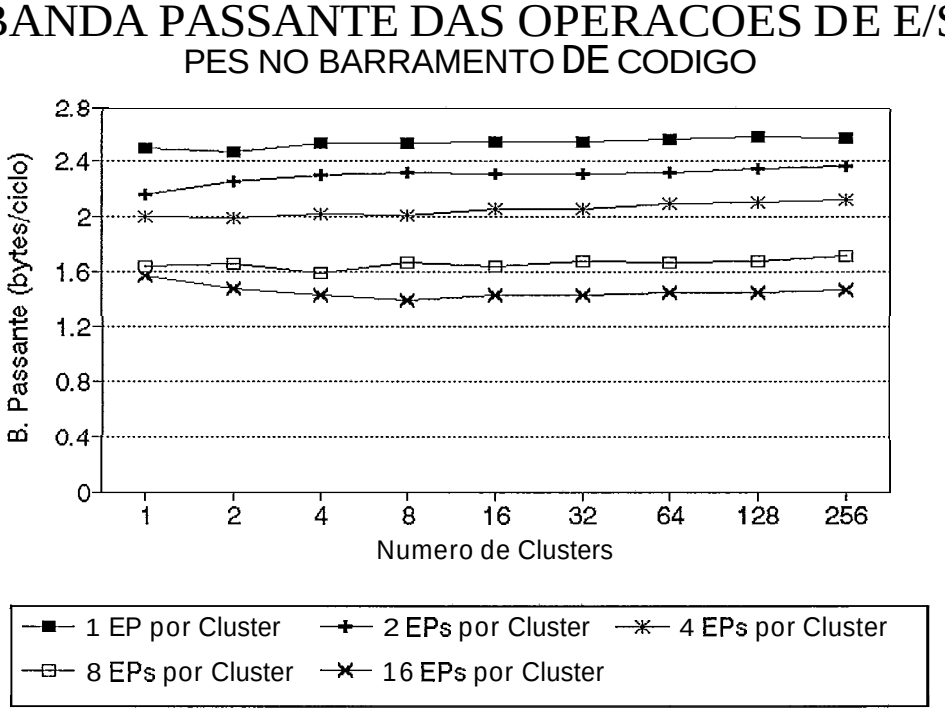

FIGURA 23: Banda passante efetiva do barramento de código para as operações de E/S.

numa taxa de transferência máxima de 3,2 bytes por ciclo nas operações de E/S. Entretanto, devido à menor prioridade do PES para realização de rajadas, elas são espaçadas entre si. Isto resulta numa taxa efetiva de transferência nas operações de E/S menor, que é função da taxa de ocupação do barrarnento de código pelos elementos processadores.

O gráfico da figura 23 ilustra o resultado desta simulagão para diferentes configuragões do Multiplus. Pode-se notar que quanto maior o número de elementos processadores por cluster, menor a taxa efetiva de transferência de E/S obtida no barramento de código.

A íiltima questão que se avaliou foi a influência das operações de E/S no desempenho dos elementos processadores, Esta influência pode ser determinada pela degradação da duração média dos acessos externos executados pelas unidades inteiras. O procedimento adotado foi simular o comportamento das unidades inteiras durante uma sequência de rajadas provenientes de uma transferência de E/S e estender o resultado, analiticamente, pasa diversas cargas de E/S.

O gráfico da figura 24 ilustra os resultados obtidos pela simulação. Comparando-o com o gráfico da figura 20, pode-se obter o fator de degradação da duração média dos acessos externos das unidades inteiras durante uma sequência de rajadas de E/S. Este fator de degradação é mostrado na figura 25. Nota-se que, como a prioridade do PES na alocação das operações de E/S é menor que a das unidades inteiras, o fator de degradação decresce com o aumento da quantidade de elementos processadores por cluster. Isto de deve à maior ocupação do barramento do cluster pelos elementos processadores, resultando numa menor interferência por parte das operações de E/S e, consequentemente, numa menor banda passante do barramento do cluster para estas operações.

É interessante notar, também pelo gráfico da figura 24, que esta degradação decrescente resultou num ponto de mínimo na duração média dos acessos, percebido quando se fixa o número de clusters e varia-se o número de elementos processadores por cluster. Isto de deve à acentuada elevação da duração média dos acessos nos extremos da faixa. Para 1 EP por *cluster,* a duração média dos acessos é alta em decorrência das operações de E/S. Já para 16 EP por cluster, o congestionamento do barramento do cluster é, pelas próprias operações das unidades inteiras, o responsável pela elevação da duração média dos acessos.

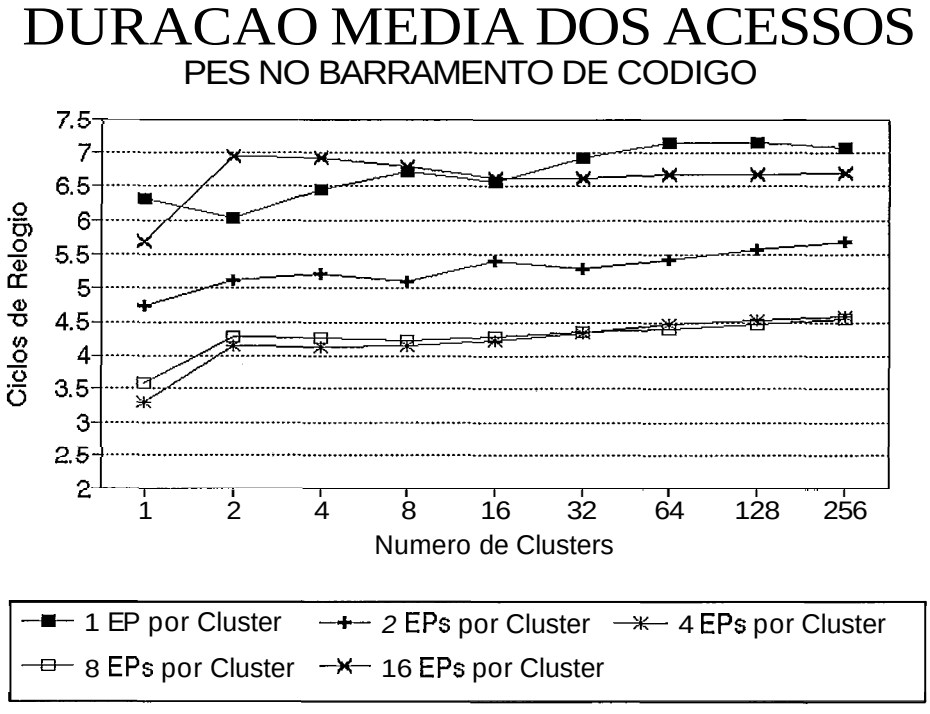

FIGURA 24: Duração média dos acessos externos das unidades inteiras durante um sequência de rajadas de E/S.

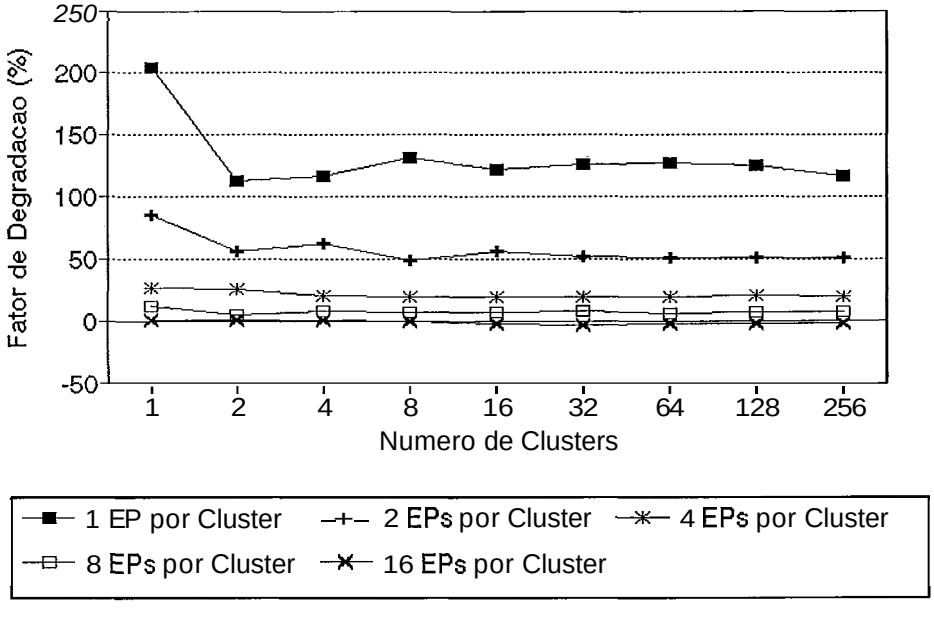

FIGURA 25: Degradação da duração média dos acessos externos das unidades inteiras durante as transferências de E/S.

Para estender o resultado analiticamente, supôs-se quatro cargas de E/S possíveis de serem impostas pelos elementos processadores: 0,5 M*Bytes*/s, 2 M*Bytes*/s, 4 M*Bytes*/s e 16 M*Bytes*/s. A escolha destes valores teve como base a suposição de que cada Mips do processador demanda 1 M*Bit*/s de E/S [SIEWIOR83]. Como cada processador SPARC tem a capacidade de processar 16 Mips a 25 MHz, a expectativa é de que cada um deles imponha uma carga de 2 M*Bytes*/s ao subsistema de E/S. A partir deste valor, supôs-se as demais cargas de E/S. Procurou-se cobrir uma faixa de valores que pudesse, tanto conter as cargas de E/S esperadas para o Multiplus, quanto evidenciar seu impacto no desempenho dos elementos processadores.

Cada uma das cargas de E/S supostas reflete, nas unidades inteiras, um comportamento distinto. Durante as transferências de E/S provenientes destas cargas, a duração média dos acessos executados cresce (ver figura 24). Quanto maior a carga de E/S, maior o tempo em que os acessos das unidades inteiras permanecem nesse patamar mais elevado de duração. Calculando-se, para cada uma destas cargas, a média, ponderada em relação ao tempo, da duragão média dos acessos durante as transferências de E/S (figura 24) e fora delas (figura 20), pode-se obter os gráficos das figuras 26, 27, 28 e 29. Cabe lembrar que, para elaboração destes gráficos, considerou-se que a fsequência dos ciclos de relógio, responsável pela temporização dos acessos das unidades inteiras, é de 25 MHz, frequência que será utilizada na implementação do Multiplus.

Nota-se que a duração média dos acessos externos executados pelas unidades inteiras é pouco influenciada pelas operagões de E/S, desde que estas operações sejam provenientes de cargas de E/S pouco expressivas. Para cargas de 0,5 M*Bytes*/s, 2 M*Bytes*/s e 4 M*Bytes*/s a degradagão foi da ordem de 1%, 4% e 8%, respectivamente. Este resultado pode ser explicado pela própria origem dos dados. Primeiramente, a degradagão da duração média dos acessos durante as rajadas de E/S decresce significativamente com o aumento do número de elementos processadores por cluster (ver figura 25). Segundo, a carga de E/S total é proporcional ao níunero de elementos processadores por cluster. Desta forma, ao se calcular a média ponderada, verifica-se que quando a degradação é mais significativa, a carga de E/S é pequena, influenciando pouco o resultado. À medida em que cresce o número de elementos processadores por cluster, o peso das operações de E/S no cálculo da média ponderada é maior. Entretanto, a degradação da duração dos acessos é pouco expressiva, o que, novamente, ocasiona pouca influência no resultado.

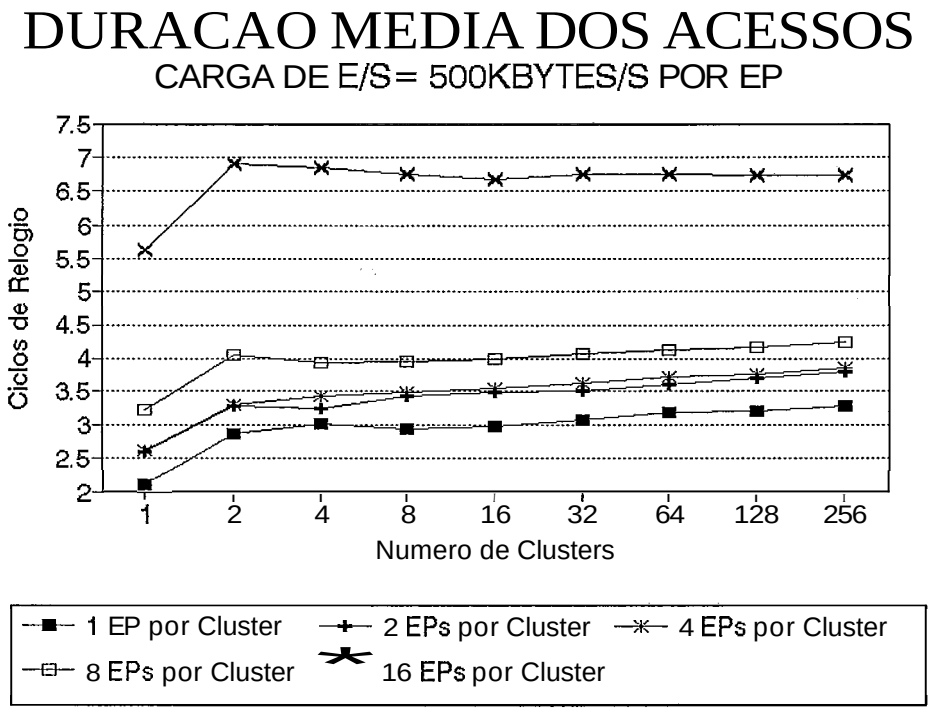

FIGURA 26: Duração média dos acessos externos das unidades inteiras. Carga de E/S de 0,5 M*Bytes*/s por elemento processador.

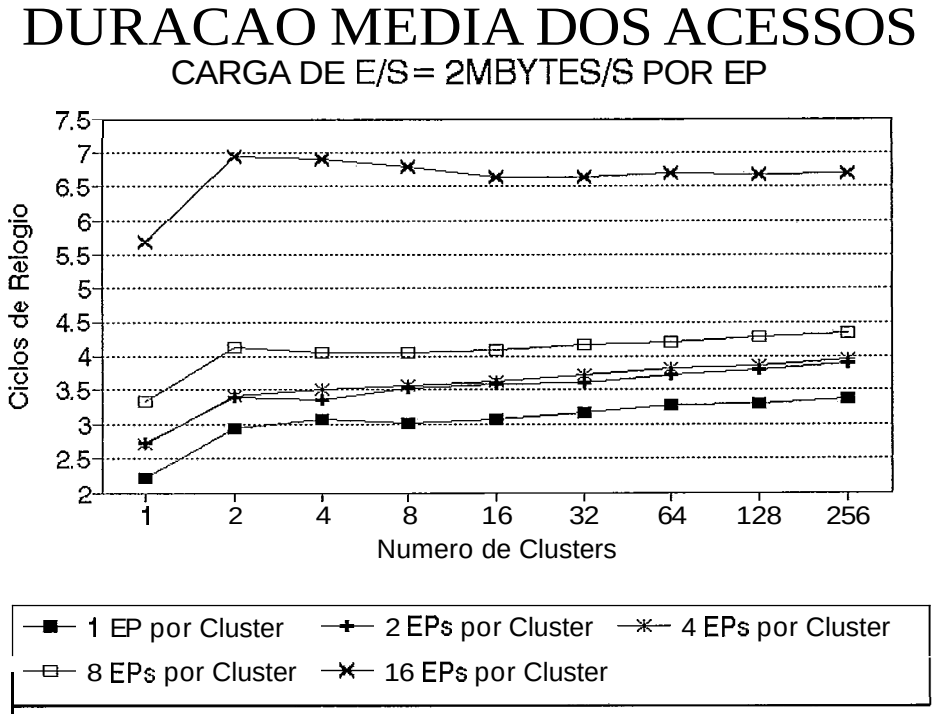

FIGURA 27: Duração média dos acessos externos das unidades inteiras. Carga de E/S de 2 M*Bytes*/s por elemento processador.

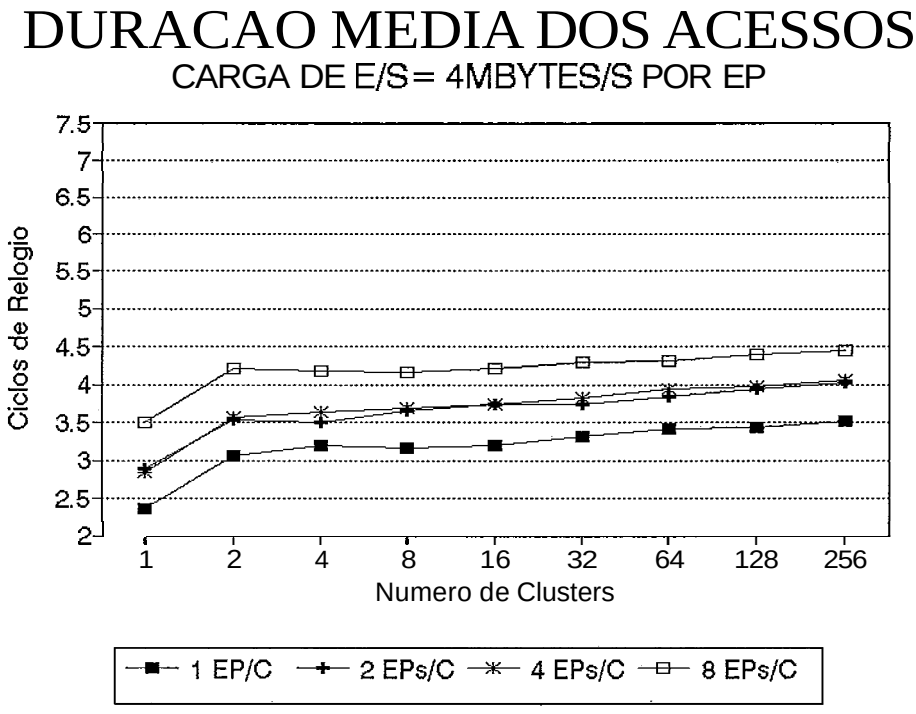

FIGURA 28: Duragão média dos acessos externos das unidades inteiras. Carga de E/S de 4 M*Bytes*/s por elemento processados.

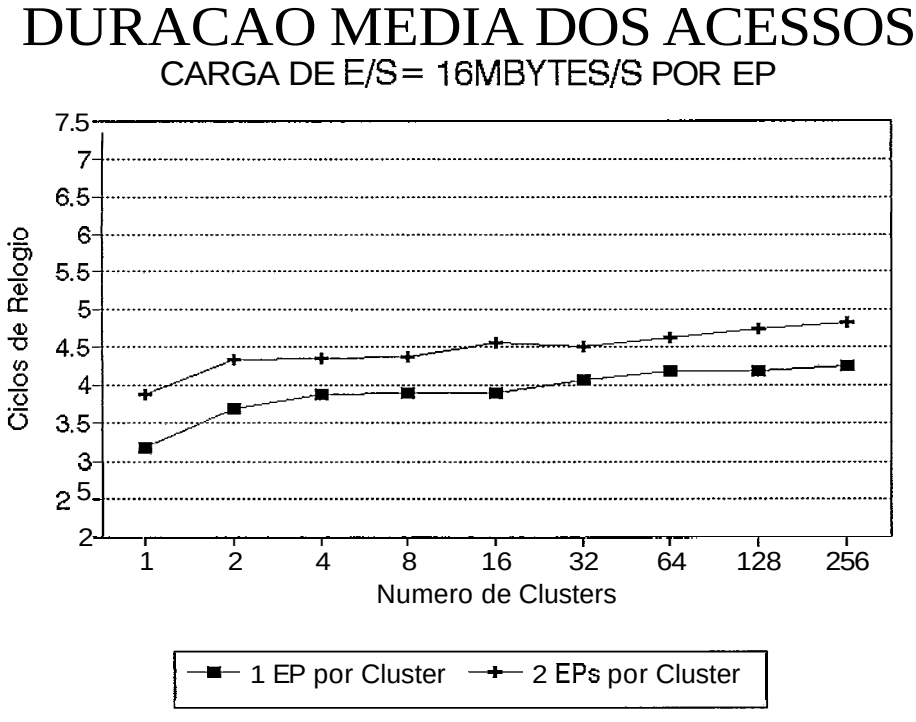

FIGURA 29: Duragão média dos acessos externos das unidades inteiras. Carga de E/S de 16 M*Bytes*/s por elemento processador.

Para que as operações de E/S degradem significativamente a duração média dos acessos externos executados pelas unidades inteiras é necessário que se tenha uma alta carga de E/S e poucos elementos processadores por cluster. O gráfico da figura 29 ilustra esta situação, onde cada elemento processador impõe uma carga de E/S de 16 *MBytes*/s. A degradação resultante é da ordem de 30%.

Nas figuras 28 e 29, os gráficos não foram traçados para algumas configurações. Isto se deve ao fato de, nestas configuragões, a carga de E/S imposta ser superior à banda passante do barramento de código para as operações de E/S, impossibilitando a análise.

Cabe lembrar que estas conclusões são referenciadas no critério de que a prioridade do PES para realização de operações de E/S é menor que a dos elementos processadores. Se a prioridade fosse invertida, chegaria-se a resultados mais favoráveis às operações de E/S, entretanto, a degradação do desempenho dos elementos processadores seria maior.

VI.1.3 - Questões Inerentes à Rede de Comunicação

Parte dos pedidos de E/S que chegam a um PES, proveniente dos elementos processadores, solicitam dados que se encontram em dispositivos de armazenamento associados a outros *clusters* de elementos processadores. Estes dados são transmitidos através de uma rede de comunicação que interliga todos os PES do Multiplus. Com o objetivo de dar uma idéia aproximada do comportamento desta rede de comunicação frente às operações de E/S do Multiplus, procurou-se adaptar o mesmo simulador usado nas questões anteriores para simular sua operacionalidade.

A adaptação consiste em restringir os tipos de acessos possíveis das unidades inteiras, mostrado na figura 19, de modo ao elemento processados se comportar de forma similar a um PES. Assim, definindo uma configuração com apenas um *cluster* e variando o número de elementos processadores por *cluster* pode-se, monitorando o barramento do *cluster,* simular o comportamento da rede de comunicação, onde os elementos processadores representam os PES.

As restrições impostas aos acessos são mostradas na figura 30. A primeira restrição foi quanto ao tipo de acesso. Como todas as operações de E/S na rede de comunicação são de um mesmo tipo, restringiu-se os acessos das unidades

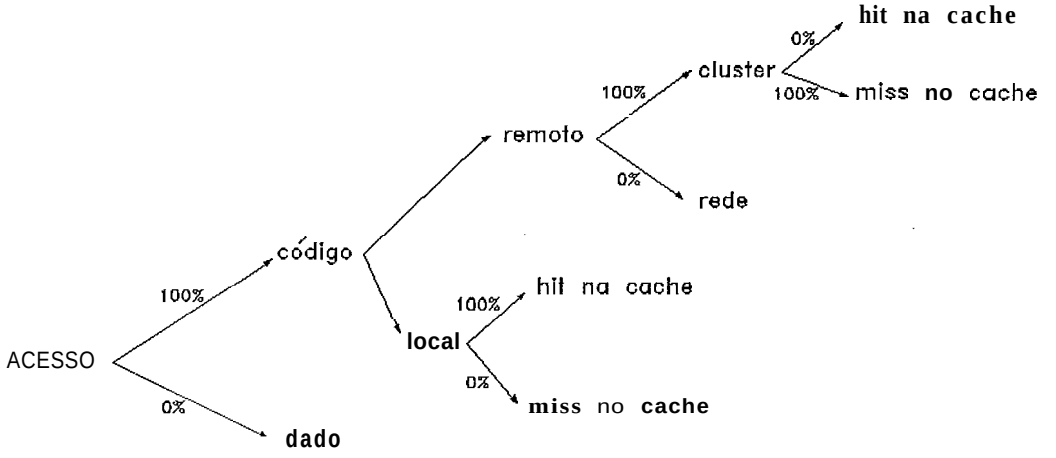

FIGURA 30: Restrições impostas aos tipos de acessos das unidades inteiras para simular o comportamento de um PES.

inteiras como sendo somente de cógido. Os acessos podem ser locais ou remotos ao PES. Os acessos remotos são os que utilizam a rede de comunicação para acessar dados localizados em dispositivos de armazenamento pertencentes a outros PES. A taxa de distribuição dos acessos entre locais e remotos é um dos parâmetros que procurou-se variar nas simulações. Como o objetivo da simulação é avalias o comportamento da rede de comunicação, forçou-se todos os acessos locais a dar hit na cache, já que não importaria como eles se resolveriam internamente ao PES, simplificando a simulação. Em função disto, foram feitos ajustes na taxa de distribuição dos acessos entre locais e remotos, procurando compensar a diferença existente entre as taxas de transferência da rede de comunicação e da cache de disco. Quanto aos acessos remotos, 100% deles devem ser no cluster, pois é o barramento do cluster que se comporta como a rede de comunicação. Por último, todos os acessos ao cluster são seguidos de *miss* na cache, evitando que a unidade inteira deixe de acessas o barramento do cluster.

Alguns parâmetros internos do simulador, referentes ao barramento do cluster, foram alterados para aproximá-lo do modelo de uma rede de comunicação. A rede modelada tem uma banda passante de 100 M*Bits*/s, que pode ser obtida utilizando-se um meio físico ótico. O padrão mais difundido para este tipo de rede é o FDDI.

Certos critérios adotados na simulação merecem destaque. Primeiramente a taxa de aproveitamento da banda passante da rede para transmissão de dados. Considerou-se que 90% das informações contidas numa transmissão são aproveitáveis, ou seja, dados, e 10% são *bits* de controle referentes à implementagão do protocolo de transmissão. Um segundo critério diz respeito ao modo de transmissão. Diferentemente das operagões de E/S realizadas no barramento do *cluster*, as operações na rede de comunicação são efetuadas sem entrelaçamento de tarefas. Todos os dados solicitados são transmitidos de uma única vez, mediante um *lock* na rede. Por último, considerou-se que todo *cluster* de elementos processadores tem um PES associado.

Os resultados da simulação são mostrados nos gráficos das figuras 31, 32 e 33. Procurou-se avaliar a taxa de ocupagão da rede para diversas configurações do Multiplus em função da carga de E/S total submetida ao PES e da porcentagem desta carga que realizam acessos remotos ao PES.

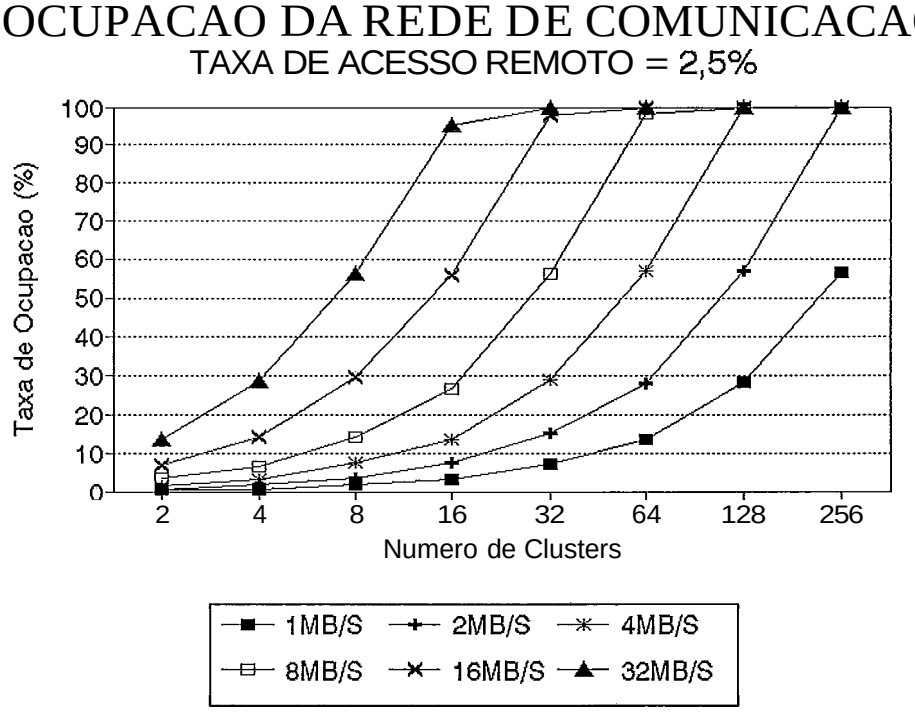

FIGURA 31: Taxa de ocupação da rede de comunicação para diferentes cargas de E/S no PES. Taxa de acesso remoto de 2,5%.

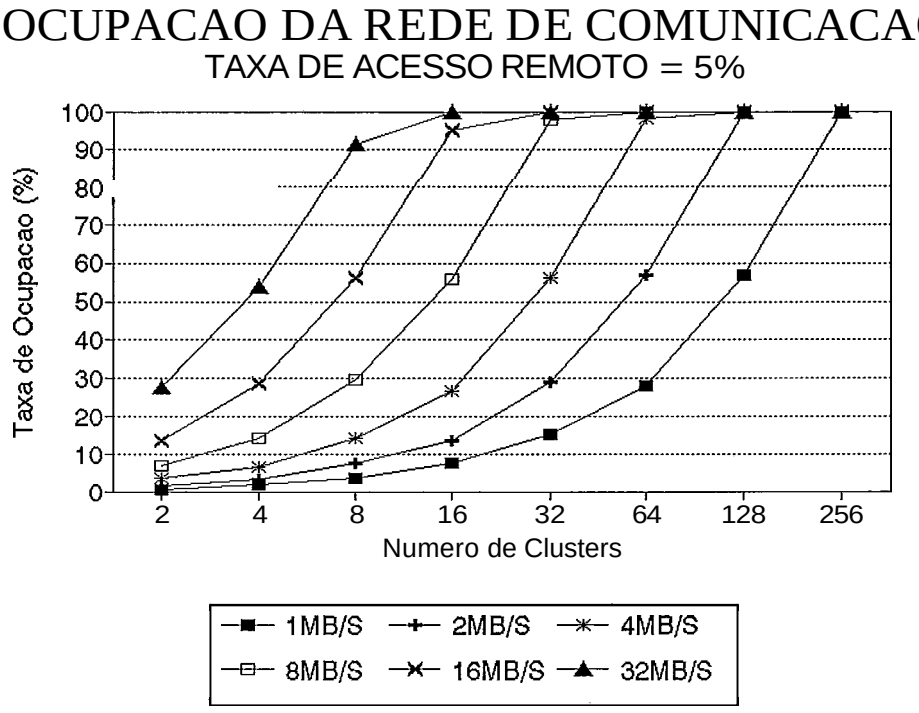

FIGURA 32: Taxa de ocupação da rede de comunicação para diferentes cargas de E/S no PES. Taxa de acesso remoto de 5%.

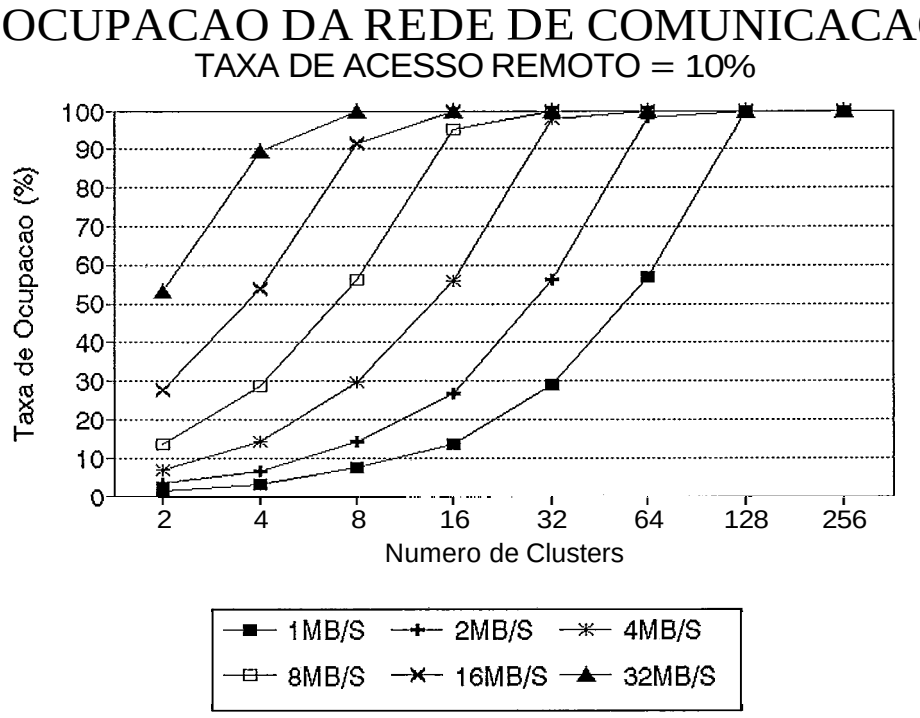

FIGURA 33: Taxa de ocupação da rede de comunicação para diferentes cargas de E/S no PES. Taxa de acesso remoto de 10%.

Pode-se notar que, apesar de se ter modelado uma das redes de maior banda passante disponível no mercado, sua utilização no subsistema de E/S do Multiplus impõe restrições. A dispersão, em diversos PES, dos dados necessários à execução das tarefas dos elementos processadores de um cluster implica numa quantidade de acessos à rede de comunicação capaz de saturá-la em configurações com, até mesmo, poucos clusters. Por outro lado, se a distribuição dos dados for otimizada, a rede pode ser operacional em quase todas as configurações do Multiplus sob diferentes cargas de E/S.

VI.2 - Análise Quantitativa

O subsistema de E/S do Multiplus é composto por um conjunto de PES. Para que ele seja capaz de operar nas cargas de E/S simuladas anteriormente, os PES também têm que ser capazes de suportá-las. Esta análise tem o objetivo de avaliar, quantitativamente, algumas questões relativas à banda passante do PES, face às suas características internas e às limitações impostas para implementação.

Visualizando a arquitetura interna do PES e o fluxo de dados no seu interior, pode-se perceber que as transferências de E/S realizadas no barramento do cluster são sempre provenientes de dados presentes na memória cache de disco. Caso algum dado solicitado não esteja na cache, ele é primeiramente lido dos dispositivos de armazenamento para a cache e transferido para o barramento. Desta forma, a banda passante do PES é diretamente proporcional à taxa de acerto na cache de disco:

$$\frac{1}{\text{B.P. PES}} = \frac{1}{\text{B.P. Cache}} \times (\text{taxa de hit na cache}) + \left(\frac{1}{\text{B.P. Cache}} + \frac{1}{\text{B.P. Disp.}}\right) \times (1 - \text{taxa de hit na cache})$$

$$\frac{1}{\text{B.P. PES}} = \frac{1}{\text{B.P. Cache}} + \frac{1}{\text{B.P. Disp.}} \times (\text{taxa de miss na cache}) \qquad (1)$$

A banda passante da memória cache de disco do PES, supondo um tempo de acesso de 125ns, é:

$$\text{B.P. Cache} = \frac{1}{\text{tempo de acesso}} \times (\#\ \text{bytes/palavra}) \times (\#\ \text{bancos})$$

$$\text{B.P. Cache} = \frac{1}{125\ \text{ns}} \times 8 \times 1$$

B. P. **Cache** = 64 *MBytes*/s.

Os dispositivos de armazenamento utilizados no Multiplus são discos do tipo *Winchester*, cuja banda bassante é de 0,7 *MBytes*/s. Assim, usando a equação (1), pode-se construir o gráfico da figura 34, que mostra a banda passante do PES para diversos valores de taxa de miss na *cache* de disco.

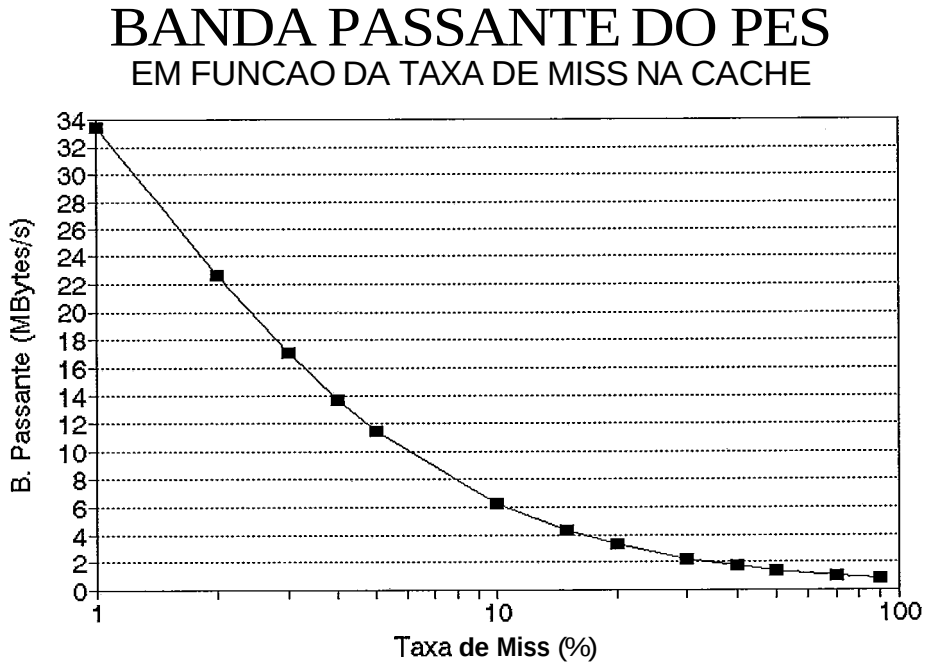

FIGURA 34: Banda passante do PES em função da taxa de miss na cache de disco.

Para avaliar se o PES é capaz de atender às necessidades de E/S do Multiplus pode-se partir de uma regra empírica bastante difundida na literatura: cada instrução por segundo do processador demanda um *bit* de E/S por segundo. Cada PES deve suportar, no mínimo, a carga de E/S dos elementos processadores de seu *cluster*. Um processador SPARC do Multiplus é capaz de processar,

efetivamente, 16 Mips a 25 MHz. Como são, no máximo, 16 SPARC por cluster, o PES deve ser capaz de fornecer dados a uma taxa entre 2 e 32 M*Bytes*/s, dependendo do número de elementos processadores por cluster da configuragão.

Pode-se notar que estas taxas só são satisfeitas mediante uma taxa de *miss* na cache de disco extremamente pequena, incompatível com a política WRITE THROUGH e com o tamanho da memória cache utilizada. Entretanto, partiu-se de uma premissa de que 1 Mips demanda 1 M*Bit*/s de E/S. Analisando as arquiteturas de computadores de alto desempenho existentes, pode-se concluir que esta suposição é raramente cumprida [AKELLA91]. Mesmo porque, é uma suposigão resultante de observagões em máquinas de propósito geral. Sendo o Multiplus uma máquina de propósito científico, espera-se que a relagão entre a capacidade de processamento (Mips) e a demanda de E/S seja mais favorável, permitindo que o PES, apesar das limitagões impostas, satisfaça razoavelmente às necessidades de E/S do Multiplus.

Outro ponto favorável ao desempenho do PES é a redução do número de elementos processadores por cluster. Pelos resultados apresentados neste capítulo, pode-se concluir que o número máximo de elementos processadores por cluster, para que se usufrua das facilidades do barramento do cluster sem congestioná-lo, deve ser limitado a 8, independente das operagões de E/S. Desta forma, a carga de E/S máxima que o PES deve suportar cai a metade, permitindo que sua banda passante máxima também se reduza em igual proporgão.

CAPÍTULO VII

CONCLUSÃO

A definição de um subsistema de E/S exige um cuidado especial quanto ao seu dimensionamento, principalmente quando associado a sistemas computacionais de alto desempenho. Dentre os tipos de processamento inerentes a um subsistema de E/S, o processamento orientado a bloco, por manipular um grande volume de dados, é o que requer maior atenção.

O elemento final de um subsistema de E/S orientado a bloco é o dispositivo de asmazenamento. Quanto maior o desempenho deste dispositivo, mais simples é a arquitetura de E/S capaz de gerenciar um conjunto deles de forma satisfatória às necessidades dos elementos processadores. A existência de um descompasso entre a demanda de dados dos elementos processadores e a capacidade dos dispositivos de armazenarnento de fornecê-los, faz com que recaia sobre o restante do subsistema de E/S a responsabilidade de superação deste desequilíbrio. Excluindo os dispositivos de asmazenamento, os pontos de um subsistema de E/S que mais influenciam no seu desempenho são:

- A forma de associação dos processadores de E/S aos elementos processadores.
- A arquitetura interna do processador de E/S.
- A velocidade do canal de comunicação entre os processadores de E/S e os elementos processadores.

A forma de associação dos processadores de E/S (PES) aos elementos processadoses é função da arquitetura geral do sistema computacional e do tipo de aplicação a que ele se destina. Como o Muliplus é um multiprocessador destinado a aplicações científicas e possui arquitetura modular, onde configurações com até 2048 elementos processadores são permitidas, a distribuição de seu processamento de E/S é condição fundamental à expectativa de alto desempenho. A melhor forma encontrada para distribuir o processamento de E/S do Multiplus foi associar um PES à cada *cluster* de elementos processadores. Cada PES é responsável pelas operações de E/S de seu *cluster*. Para agilizar as operações de E/S entre *clusters* de

elementos processadores, os PES são interligados por uma rede de comunicaqão de alta velocidade, exclusivamente dedicada à estas operações.

Quanto à arquitetura do PES, pode-se obter um maior desempenho explorando o paralelismo de suas tarefas internas. Em função disto, dividiu-se a arquitetura interna do PES do Multiplus em três barramentos: barramento da CPU, barramento do DMA e barramento externo. Isto permite realizar, paralelamente, tarefas relativas ao gerenciamento do PES e à transferência de dados, tanto entre os dispositivos de armazenamento e o PES, quanto entre o PES e os elementos processadores, aumentando o *throughput* do PES. *Além* disto, é fundamental para o desempenho do PES, que ele possua uma *cache* de disco bem dimensionada, bem como um DMA com alta taxa de transferência de dados.

O canal de comunicação entre os PES e os elementos processadores do Multiplus é o mesmo utilizado para acesso à memória global compartilhada, isto é, o barramento do *cluster*. É fundamental para as operaqões de E/S que a taxa de ocupação deste barramento não caracterize um congestionamento, e que sua banda passante seja bastante alta. Desta forma, é possível que as operaqões de E/S compartilhem, satisfatoriamente, o barramento do *cluster* sem interferência significativa na operacionalidade dos elementos processadores do Multiplus.

Como perspectiva de evolução deste trabalho, existe a intenção de expandir o simulador para avaliar questões internas ao processador de E/S proposto para o Multiplus. Isto permitirá conhecer outras características de sua operacionalidade, como por exemplo, *throughput* e *overhead.*

BIBLIOGRAFIA

[AKELLA91] AKELLA, J *et alii.* "Modeling and Measurement of the Irnpact of Input/Output on System Performance", Proceeding of the 18th Annual International Symposium on Computes Architecture, Toronto, Canada, pp 390-399, May 1991.

[AUDE91] AUDE, J. Salek *et alii.* "MULTIPLUS: A Modular High-Performance Multiprocessor", Proceeding of the EUROMICRO, Microprocessing and Microprogramming, North-Holland, vol 32, pp 45-52/1991.

[BBN85] BBN Laboratories Incorporated. "Butterffly (TM) Parallel Processos Overview". BBN Laboratories Incorporated, Jun 1985, 43 pp.

[BERRA89] BERRA, P. Bruce *et alii.* "Optical and Supercomputing". Proceedings of the IEEE, Special Issue on Supercomputer Technology, pp 1797-1815, Dec 1989.

[BORRILL89] BORRILL, Paul L. "High-speed 32-bit Buses for Forward-looking Computers". IEEE Spectrum, pp 34-37, Jul 1989.

[CONNOR87] CONNOR, Gary *et alii.* "Serial Data Races at Parallel Rates for the Best of both Worlds". Eletronic Design, Hayden Publ, Yorktown Heights, NY, pp 79-83, vol 35, no 2, Jan 1987.

[FIGUEIRA88] FIGUEIRA, Norival Ribeiro. "Cache de Disco: Arquiteturas e Algoritmo". Anais do XXI-Congresso Nacional de Informática, Rio de Janeiro, pp 863-869, vol 2, Agosto de 1988.

[FIGUEIRA90] FIGUEIRA, Norival Ribeiro. "Avaliação de Algoritmos Para Cache de Disco". Anais do XXIII-Congresso Nacional de Informática, Rio de Janeiro, Setembro de 1990.

[GIBSON89] GIBSON, Garth A. *et alii.* "Failure Correction Techniques for Large Disk Arrays". Proc. Third Int. Conf. on Architectural

Support for Programming Languages and Op. Sys., Boston, MA, Apr 1989.

[GOTTLI83] GOTTLIEB, Allan et *alii*. "The NYU Ultracomputer-Designing an MIMD Share Memory Parallel Computes". IEEE Transactions on Computers, vol c-32, no 2, Feb 1983.

[HWANG87] HWANG, Kai. "Advanced Parallel Processing with Supercomputer Architectures". Proceeding of IEEE, pp 1348-1379, vol 75, no 10, Oct 1987.

[IVERSEN89] IVERSEN, Wesley R. "Coming Soon: High-performance Son of SCSI". Eletronics, Penton Publishing, Cleveland, OH, pp 104-105, Feb 1989.

[KATZ89] KATZ, Randy H. *et alii*. "Disk System Architectures for High Performance Computing", Proceedings of the IEEE, pp 1842-1858, vol 77, no 12, Dec 1989.

[MESLIN91] MESLIN, Alexandre M. "Estudos de Arquiteturas de Memórias Cache para o Multiprocessador Multiplus". Tese M.Sc. COPPE/UFRJ, Agosto de 1991,114 pp.

[MOKHOF87] MOKHOFF, Nicolas. "Five-chip Token-passing Set Operates LANS at 100 MBits/s". Eletronic Design, Hayden Publ, Yorktown Heights, NY, pp 45-50, vol 35, no 21, Sep 1987.

[MOTOROLA87] MOTOROLA. "Utilizing the MC68020 as a Dedicated DMA Controller". MOTOROLA Semiconductor Products Inc, Design Concept, 1987, 20 pp.

[NG88] NG, Spencer W. "Some Design of Disk Arrays". IBM Research Report, IBM Almaden Research Center, San Jose, CA, Jun 1988, 16 pp.

[PATTER88] PATTERSON, David A. et *alii*. "A Case for Redundant Arrays of Inexpensive Disks (RAID)". A.C.M. SIGMOD Conference, Chicago, IL, pp 109-116, May 1988.

[PATTER90] PATTERSON, David A. *et alii.* "Computer Architecture: A Quantitative Approach". Morgan Kaufmann Publishers, 1990.

[PETER85] PETERSON, J. L. *et alii.* "Operating System Concepts". Addison Wesley Publishing Company, 1985.

[PFISTER85] PFISTER, G. F. "The Architecture of the IBM Research Parallel Processor Prototype (RP3)". IBM Research Report, IBM T. J. Watson Research Laboratoiy, N.Y., Jun 1985.

[PIEPER89] PIEPER, John S. "Parallel **1/O** Systems for Multicomputers". School of Computes Science, Carnegie Mellon University, Pittsburgh, CMU-CS-89-143,1989.

[PRADO88] PRADO, Cláudio Almeida. "Projeto de um Subsistema de Memória de Massa Para Um Computador de Arquitetura Paralela". II-Simpósio Brasileiro de Arquitetura de Computadores, Águas de Lindóia, SP, pp 5.A.4.1-5.A.4.6, Setembro de 1988.

[SIEWIOR83] SIEWIOREK, D. P. *et alii.* "Computes Structures: Principles and Examples", McGraw-Hill Book Company, New York, NY, 1983.

[SMITH82] SMITH, A. J. "Cache Memories". ACM Computing Surveys, pp 473-530, vol 14, no 3, Sep 1982.

[SMITH85] SMITH, A. J. "Disk Cache: Miss Ratio Analysis and Design Considerations". ACM Transactions on Comvuter Systems, New York, pp 161-203, vol 3, no 3, Aug 1985.

[SWAN87] SWAN, R. J. *et alii.* "Cm* - A Modular, Multi-microprocessor". National Computer Conference, Montrale, New Jersey, AFIPS Press, pp 637-667, vol 46,1987.

[WILSON87] WILSON, Ron. "Designers Rescue Supercomputers from 1/0 Bottleneck". Computer Design, pp 61-71, Oct 1987.

ANEXO A

RESULTADOS DA SIMULAÇÃO UTILIZANDO O BARRAMENTO DE DADOS PARA AS OPERAÇÕES DE E/S DO MULTIPLUS

As operações de E/S do Multiplus são realizadas através do barramento de código de seus *clusters* de elementos processadores. Com o objetivo de ratificar esta decisão, este anexo sintetiza os resultados obtidos por simulações, caso se utilizasse o barramento de dados para as operações de E/S.

O gráfico da figura A.1 ilustra a duração média dos acessos externos executados pelas unidades inteiras dos elementos processadores durante uma transferência de E/S no barramento de dados para diferentes configurações do Multiplus. Pode-se notar que não ocorreram alterações significativas na duração média dos acessos se compararmos com os resultados ilustrados na figura 24, onde utilizou-se o barramento de código para as operações de E/S. Observando as taxas de ocupação dos barramentos de código e dado sem as operações de E/S (figuras 21 e 22), verifica-se que elas são bastante semelhantes nas configurações do Multiplus mais suceptíveis à degradação da duração média dos acessos, isto é, poucos elementos processadores por *cluster*. Quando a diferença entre as taxas de ocupação se acentua, a configuração correspondente é pouco sensível às operações de E/S. Desta forma, justifica-se o resultado.

Por outro lado, uma maior taxa de ocupação no barramento onde são realizadas as operações de E/S reflete significativamente na banda passante do barramento para estas operações. É o que acontece utilizando o barramento de dados para as operações de E/S: a quantidade de rajadas efetuadas por unidade de ciclo de relógio reduziu-se acentuadamente. O gráfico da figura A.2 ilustra os resultados. Comparando-o com o gráfico da figura 23 pode-se perceber que a banda passante do barramento de dados para as operações de E/S é bastante inferior ao do barramento de código. Desta forma, sua utilização para este fim restringe bastante a carga de E/S suportada pelo subsistema de E/S, em detrimento da versatilidade e desempenho do Multiplus.

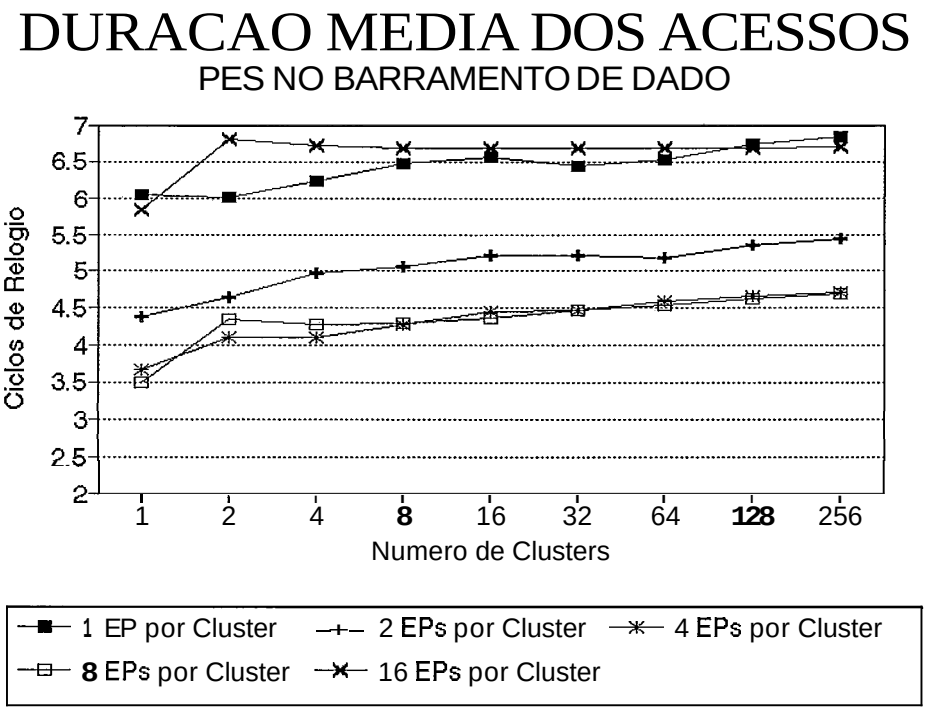

FIGURA A.1: Duração média dos acessos externos das unidades inteiras durante uma sequência de rajadas de E/S no barrarnento de dados.

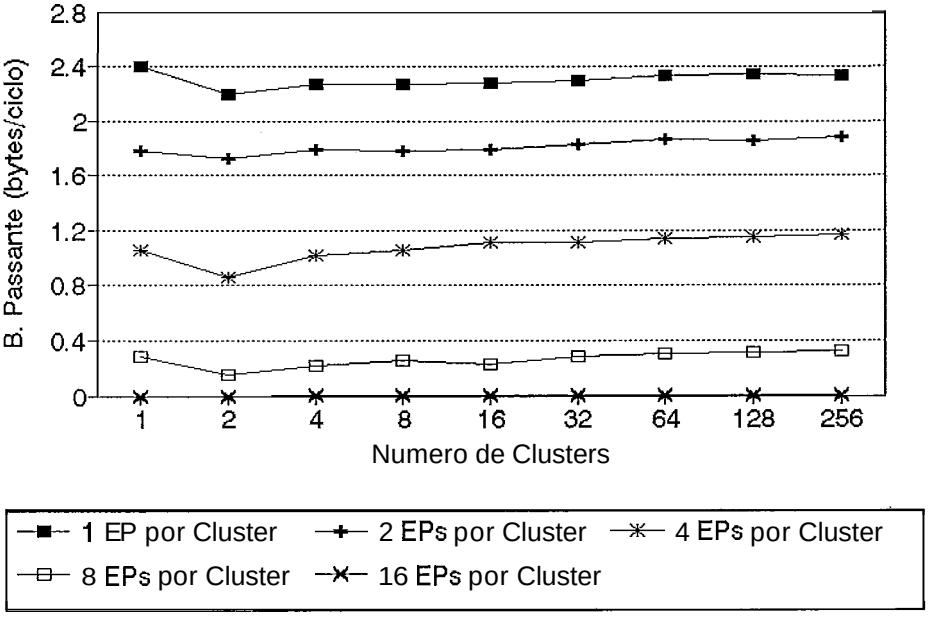

FIGURA A.2: Banda passante efetiva do barramento de dados para as operações de E/S

ANEXO B

TABELAS DE DADOS GERADOS PELO SIMULADOR

SEM OPERAÇÕES DE E/S

#EPs/C	#Clusters	Duracao Acessos	% Barramento Codigo Livre	% Barramento Dado Livre	# Ciclos Simulados
1	1	2.0756	100.000	100.000	10000
1	2	2.8321	98.140	94.330	10000
1	4	2.9780	97.520	95.595	10000
1	8	2.9074	98.682	94.421	10000
1	16	2.9581	98.237	94.889	10000
1	32	3.0593	98.450	94.908	10000
1	64	3.1521	98.513	94.932	10000
1	128	3.1780	98.552	95.149	10000
1	256	3.2693	98.557	95.214	10000
2	1	2.5579	88.800	71.810	10000
2	2	3.2645	89.940	73.925	10000
2	4	3.2099	90.477	76.230	10000
2	8	3.4112	90.855	76.069	10000
2	16	3.4524	90.991	76.974	10000
2	32	3.4754	91.461	77.630	10000
2	64	3.5827	91.612	77.790	10000
2	128	3.6696	91.806	78.378	10000
2	256	3.7663	91.927	78.810	10000
4	1	2.5988	80.000	47.050	10000
4	2	3.2786	83.575	50.405	10000
4	4	3.3981	83.533	51.700	10000
4	8	3.4609	82.817	54.361	10000
4	16	3.5174	83.149	55.024	10000
4	32	3.5993	83.420	55.637	10000
4	64	3.6978	83.086	57.112	10000
4	128	3.7297	83.686	57.847	10000
4	256	3.8322	84.275	58.675	10000
8	1	3.1853	70.110	9.240	10000
8	2	4.0165	69.010	20.720	10000
8	4	3.8960	68.073	19.288	10000
8	8	3.9344	68.565	20.364	10000
8	16	3.9676	69.053	21.116	10000
8	32	4.0333	69.382	21.713	10000
8	64	4.0969	70.178	23.046	10000
8	128	4.1368	70.651	23.645	10000
8	256	4.2137	71.287	24.808	10000
16	1	5.6150	63.200	0.020	10000
16	2	6.8922	63.480	11.375	10000
16	4	6.8232	63.462	8.098	10000
16	8	6.7465	64.441	9.490	10000
16	16	6.7009	63.628	8.436	10000
16	32	6.7972	63.559	8.033	10000
16	64	6.7753	64.434	7.655	10000
16	128	6.7644	63.974	7.611	10000
16	256	6.7606	64.040	7.359	10000

OPERAÇÕES DE E/S NO BARRAMENTO DE CÓDIGO

#EPs/C	#Clusters	Duracao Acessos	%Barramento Codigo Livre	%Barramento Dado Livre	#Ciclos Simulados	Total de Rajadas
1	1	6.3211	21.600	100.000	10000	195
1	2	6.0259	20.565	97.680	10000	385
1	4	6.4443	18.970	97.403	10000	790
1	8	6.7080	19.559	98.054	10000	1580
1	16	6.5665	19.198	97.662	10000	3172
1	32	6.9177	19.085	97.640	10000	6355
1	64	7.1502	18.334	97.767	10000	12817
1	128	7.1501	18.179	97.696	10000	25715
1	256	7.0791	18.228	97.864	10000	51358
2	1	4.7326	26.000	85.220	10000	168
2	2	5.1066	21.270	85.210	10000	351
2	4	5.2213	20.235	84.162	10000	720
2	8	5.0925	21.053	84.645	10000	1447
2	16	5.3959	20.146	85.054	10000	2887
2	32	5.2959	20.163	85.331	10000	5775
2	64	5.4250	19.892	85.824	10000	11591
2	128	5.5761	19.587	85.808	10000	23421
2	256	5.6955	19.077	86.019	10000	47274
4	1	3.3173	19.140	55.420	10000	156
4	2	4.1517	18.950	60.325	10000	310
4	4	4.1279	17.503	62.525	10000	630
4	8	4.1483	17.726	62.913	10000	1255
4	16	4.2217	17.447	64.021	10000	2563
4	32	4.3250	17.393	63.438	10000	5139
4	64	4.4666	16.916	64.109	10000	10465
4	128	4.5290	16.960	64.738	10000	20999
4	256	4.5915	16.684	65.670	10000	42381
8	1	3.5892	11.030	15.530	10000	128
8	2	4.2655	14.500	23.855	10000	259
8	4	4.2531	14.060	23.940	10000	498
8	8	4.2241	13.299	25.470	10000	1040
8	16	4.2810	13.820	26.523	10000	2048
8	32	4.3675	13.877	27.322	10000	4178
8	64	4.3862	13.860	28.385	10000	8323
8	128	4.4692	14.007	29.611	10000	16814
8	256	4.5340	13.468	30.524	10000	34141
16	1	5.6937	3.350	0.250	10000	123
16	2	6.9576	10.140	10.690	10000	232
16	4	6.9234	11.495	8.745	10000	447
16	8	6.7901	10.670	7.180	10000	873
16	16	6.6289	10.570	6.377	10000	1800
16	32	6.6274	10.894	6.580	10000	3590
16	64	6.6907	10.945	6.417	10000	7257
16	128	6.6795	10.611	6.575	10000	14550
16	256	6.6945	10.609	6.594	10000	29281

OPERAÇÕES DE E/S NO BARRAMENTO DE DADO

#EPs/C	#Clusters	Duracao Acessos	%Barramento Codigo Livre	%Barramento Dado Livre	#Ciclos Simulados	Total de Rajadas
1	1	6.0569	100.000	22.120	10000	188
1	2	6.0096	99.535	24.535	10000	343
1	4	6.2422	99.535	22.105	10000	709
1	8	6.4725	99.380	22.161	10000	1417
1	16	6.5695	99.341	21.902	10000	2852
1	32	6.4456	99.283	21.851	10000	5760
1	64	6.5287	99.288	21.192	10000	11642
1	128	6.7394	99.341	20.881	10000	23438
1	256	6.8495	99.267	20.712	10000	46826
2	1	4.3860	93.630	18.920	10000	139
2	2	4.6566	94.220	22.405	10000	269
2	4	4.9717	93.268	20.125	10000	559
2	8	5.0577	93.943	19.639	10000	1111
2	16	5.2088	93.449	19.576	10000	2240
2	32	5.2274	94.048	19.029	10000	4576
2	64	5.1849	94.123	19.272	10000	9315
2	128	5.3571	94.300	18.934	10000	18617
2	256	5.4431	94.468	18.751	10000	37559
4	1	3.6640	85.280	8.660	10000	83
4	2	4.1003	86.765	22.660	10000	135
4	4	4.1035	86.647	18.212	10000	318
4	8	4.2785	85.621	18.210	10000	661
4	16	4.4440	85.811	17.417	10000	1388
4	32	4.4694	86.497	17.646	10000	2780
4	64	4.5915	86.407	17.638	10000	5718
4	128	4.6568	86.975	17.847	10000	11518
4	256	4.7244	87.095	17.866	10000	23425
8	1	3.4868	74.130	2.760	10000	22
8	2	4.3502	72.810	14.875	10000	25
8	4	4.2773	73.352	12.680	10000	70
8	8	4.3016	70.547	12.229	10000	160
8	16	4.3655	73.203	12.638	10000	295
8	32	4.4708	73.225	12.824	10000	711
8	64	4.5364	73.142	13.027	10000	1536
8	128	4.6215	73.584	13.093	10000	3152
8	256	4.6905	73.958	13.455	10000	6553
16	1	5.8286	67.360	0.000	10000	0
16	2	6.7961	67.105	10.785	10000	0
16	4	6.7244	63.645	7.440	10000	2
16	8	6.6762	61.892	7.074	10000	2
16	16	6.6967	63.806	6.193	10000	1
16	32	6.6852	64.212	6.279	10000	4
16	64	6.6975	64.068	6.284	10000	9
16	128	6.6848	63.885	6.265	10000	20
	256	6.7015	63.626	6.322	10000	34