PROJETO MINISSUPERCOMPUTADOR : CARACTERISTICAS GERAIS DO SISTEMA MS - 8701

João Hajime Takeda, Celso Gonzalez Hummel, Fábio Grossmann,
Prof. Sergio Takeo Kofuji,Prof. Dr. João Antonio Zuffo

Universidade de São Paulo - Escola Politécnica - Departamento de Engenharia de Eletricidade - Laboratório de Sistemas Integráveis.
Av. Prof. Luciano Gualberto, trv. 3, no. 158 - CEP 05508

SUMARIO

Descreveremos em termos gerais o Minissupercomputador (Sistema MS 8701). Este computador de alto desempenho é constituído de múltiplos módulos de processamento (até 64) baseados em microprocessadores CISC monolíticos de 32 bits [1]. Discorreremos também sobre a evolução da arquitetura e algumas das características do Sistema Operacional Multiprocessador (LSI-SO.01), ora em implementação no Laboratório de Sistemas Integráveis - LSI DEE EPUSP.

ABSTRACT

In a general way we are going to describe the Mini-supercomputer (MS 8701 System). This high performance computer is made up of manifold processing modules (up to 64), that are based on CISC monolithic microprocessors of 32 bits. We are going to talk over the evolution of the architecture and some characteristcs of the Multiprocessor Operating System (LSI - SO.01), that is at present being fulfilled in the Laboratorio de Sistemas Integráveis - LSI - DEE-EPUSP.

1. INTRODUÇAO

Os avanços tecnológicos em microeletrônica têm proporcionado o desenvolvimento de processadores cada vez mais poderosos, tendo possibilitado, consequentemente, o surgimento de computadores com capacidade cada vez maiores.Além disto, as aplicações que utilizam computadores têm se tornado cada vez mais sofisticadas e exigentes no tocante à capacidade de processamento [2]. Contudo, apesar de todos os avanços, a capacidade individual dos processadores não têm atendido satisfatoriamente as necessidades crescentes de processamento,tornando necessária a busca de soluções alternativas às atualmente utilizadas [3].

Uma das alternativas mais promissoras são os sistemas de computador que utilizam microprocessadores trabalhando de forma cooperativa em paralelo. Estes microprocessadores ampliam o desempenho do sistema e proporcionam uma alternativa de solução para as necessidades de maior potência de computação.

Concorde a esta realidade, procurou-se obter um sistema de alto desempenho capaz de resolver problemas que necessitem de grande capacidade computacional, associada a uma vasta capacidade de armazenamento secundário, operando diversos canais de comunicação e terminais.O projeto do Minissupercomputador MS8701 do Laboratório de Sistemas Integráveis da Escola Politécnica da USP, é um sistema de processamento de dados que integra até 64 processadores de uso geral conjuminados com processadores dedicados a determinadas tarefas. Com um custo efetivo relativamente baixo esta máquina resulta numa relação desempenho-custo extremamente favorável,comparativamente a máquinas da classe "super".

Sinergicamente, o sistema operacional multiprocessador LSI-SO.01, em desenvolvimento, cria para o sistema um ambiente multi-tarefas, multi-usuário e com capacidade de manipulação de mais de um processador. Para este fim, possui extensões que visam o aproveitamento das características especiais da arquitetura do Minissuper, possibilitando e facilitando a programação paralela. O Sistema Operacional instalado no Minissupercomputador MS8701 proporciona melhor aproveitamento das características de divisão de tarefas, aumentando, em geral, o desempenho das aplicações. Este sistema operacional, sendo compatível com o sistema UNIX System V versão 3 permite que seja transportada uma variada gama de aplicações e programas. Há de se observar que a maior parte dos minissupers e supercomputadores vêm adotando o sistema UNIX como um de seus sistemas operacionais como, por exemplo,o Sequent Balance (Sequent Computer System) [4] e a linha de computadores da própria CRAY [5].

AS PLACAS DE PROCESSAMENTO GERAL - PPG

A conexão de 64 processadores - com acesso ao barramento principal (VME) - para compartilhamento dos recursos comuns como: memória geral, placa controladora de disco e placa controladora de terminais, resulta normalmente em elevado tráfego no barramento, e consequente em decaimento do desempenho do sistema. Por exemplo, para o caso do sistema ser construído com processadores MC68020 [6],a taxa necessária para atendimento da demanda seria de 1088 Mbytes/s (considerando-se o desempenho de um processador 68020 em 2,5 Mips e 1,7 palavras por instrução, o total para 64 processadores resulta em 1088 Mbyte/s), não considerando transferências de E/S de blocos e de comunicações, no entanto, a faixa prática do barramento VME [7] é de 15 Mbytes/s.

Para contornar esta limitação, adotou-se uma série de providências que, somadas, permitiram uma solução de compromisso razoável: a) utilizou-se dois dutos padrão VME de 32 bits; b) memória de acesso exclusivo e cache em cada processador; c) os processadores foram dispostos em grupos de 4 e colocados em módulos de processamento,Placa de Processamento Geral (PPG), d) parte da memória principal foi particionada entre os módulos de processamento.Estas providências possibilitaram uma diminuição significativa do tempo médio de acesso à memória e reduziram drasticamente a utilização do barramento principal. Com relação á utilização de memória cache em cada processador, se por um lado isto diminui o tempo médio de acesso ao barramento, por outro introduz novos problemas, associados à manutenção da consistência destas memórias [8].No projeto adotamos a política de "escrita simultânea" ("write throught"), que protege as regiões de memória compartilhada na PPG através de programação da circuitaria ("hardware") de gerenciamento de memória.A implementação de uma memória "cache" [9] sem estados de espera (zero "wait states"), exigiu adoção de endereçamento lógico para esta memória, o que implica em invalidação total do "cache" a cada troca de contexto do processador.

O banco de memória de acesso exclusivo a cada processador é reservado para o acesso em modo supervisor. Este banco contém parte do código do sistema operacional, referente a serviços individualizáveis por processo, ou seja, a parte do sistema operacional que executa pedidos diretamente ligados ao programa de usuário. Esta subdivisão possibilita a execução de grande parte do código do sistema operacional (código comum, executável por todos os processadores, uma vez que o sistema operacional não prevê a existência de um processador mestre) sem a necessidade de acesso e de utilização do barramento, diminuindo,portanto, a taxa de comunicação [10]. Naturalmente, nem sempre existe esta possibilidade quanto aos dados do sistema operacional.

O COMPUTADOR DE MEMÓRIA DE MASSA - CMM

Foi colocado no ítem anterior a forma de processamento paralelo do sistema e sua distribuição. Resta discutir o problema de como fazer chegar rapidamente as PPG's os dados necessários para o processamento, ou seja, a implementação de um subsistema eficiente de acesso a disco. A estrutura de comunicação com o subsistema de controle de disco, convencionalmente utilizada em micros e supermicros, onde uma única placa é acessada através do barramento principal, apresenta as seguintes restrições:

- sobrecarga do barramento principal para transferência de dados;

- existência de grande limitação para que o processador deste controlador gerencie toda a estrutura de entrada/saída do UNIX [11].

- há a possibilidade de duas PPG's fazerem solicitações simultaneamente.

Numa primeira solução, cada placa (PPG) estaria responsável pela montagem do comando na memória compartilhada da placa de disco através do barramento VME e, os dados enviados por dois outros dutos de acesso direto a memória das PPG's (dutos de ADM). Desta forma,a placa de disco fica excessivamente sobrecarregada e os dutos de ADM não suportam a demanda exigida. Para manter compatibilidade com a implementação do sistema operacional UNIX, foi imposto que cada processador - dos módulos das PPG's - execute parte do gerenciamento de bloco, aliviando, assim, a carga do subsistema gerenciador de disco. A adição de um duto particular de ADM, interligando cada PPG ao CMM, veio elevar a capacidade de entrada/saída de disco e aliviar a carga de comunicação do barramento principal.

Para reduzir o tempo médio de acesso aos blocos de disco, o CMM passou a conter uma memória de acoplamento (Buffers) de blocos [12].Ao CMM coube, então, as tarefas de acesso a disco, atendimento de comandos e gerenciamento de interfaces.

A PLACA DE GERENCIAMENTO DE SISTEMA-PGS

Em um sistema multiprocessador de arquitetura paralela é desejável que os processadores de uso geral passem a maior parte do seu tempo executando as tarefas pertinentes aos processos de usuário. Os eventos assíncronos relacionados a processos em estado suspenso (não sendo executado em nenhum processador do sistema) não devem interromper os processos em execução. Desta forma, o atendimento de eventos como: relógio de particionamento de tempo e tempo real, discos e terminais, foi destinado a um único módulo com um processador dedicado a tarefas administrativas, denominado de Placa de Gerenciamento de Sistemas, PGS.

A configuração da PGS é a seguinte :

-um microprocessador MC68020 (16,67 MHz);

-uma memória global de 1 Mbyte onde devem residir as tabelas do sistema e estruturas globais do S.O.,como tabelas de processos,semáforos de sistema;

-relógio de particionamento de tempo;

-relógio de tempo real,com calendário mantido por bateria;

-relógio incremental, para os processos disporem de informação de tempo decorrido;

- interface com os dutos principais;

- interface com o CMM;

- duto de sinalização de interrupções para os módulos das PPG's;

- recursos de diagnóstico remoto do sistema.

Para reduzir o tempo médio de acesso à memória da PGS, esta foi projetada com arquitetura entrelaçada em nivel 2, possibilitando o acesso por duas portas simultaneamente (processador local e barramento principal, por exemplo).

Um duto especial de sinais de interrupção foi criado para agilizar o envio destes sinais da PGS para os processadores gerais [13]. Por este duto a PGS pode enviar sinais de interrupção e mensagens para todos os processadores de cada PPG. Para se enviar um sinal de interrupção para a PGS basta que o processador realize uma operação de escrita em determinadas posições da memória da PGS. Deste modo, os processadores podem enviar sinais de interrupção e mensagens com apenas uma operação de escrita.

O SUBSISTEMA DE COMUNICAÇÕES - SCC

O elo de ligação entre o sistema e o meio externo é o subsistema de comunicações. O usuário do sistema pode estar localizado em um terminal local, remoto ou rede local.Possibilita também a comunicação entre programas em máquinas diferentes interconectadas através deste subsistema.

Este subsistema especifica o controle direto de até 256 terminais assincronos, conectados em portas de E/S utilizando diferentes protocolos definidos no sistema. Suporta rede local (ETHERNET) e rede pública (RENPAC - X.25).

A capacidade suportada diretamente pela configuração do subsistema (PPC, Canal de E/S (duto VSB) e Placas Especializadas de E/S,PES, conectadas nesse duto através de um bastidor) restringiu-se a 64 terminais. A expansão para 256 terminais foi feita por módulos controladores de terminais distribuidos, módulos estes, que conectam-se por "rede local" à Placa de Processamento de Comunicações [14].

A Placa de Processamento de Comunicações (PPC) atua, pois, como gerenciadora do subsistema tendo funções de:

- processadora de comunicações: realiza a tradução de mensagens recebidas do sistema,para as diversas Placas Especializadas de Expansao de E/S PES, e destas novamente para o sistema;

- gerenciadora de controladores de terminais, conectados por uma via de rede;

- gerenciadora de comunicação com rede ETHERNET e X.25;

- gerenciadora das placas especializadas de expansão de E/S.

As Placas Especializadas de Entrada/Saida PES contêm as portas de entrada/saida e são responsáveis pela comunicação efetiva de dados com o terminal, realizando o protocolo de comunicação.

São previstas PES dos tipos básicos:

- placa de terminais de comunicação serial assincrona RS-232/422 (até 38400 bps);
- placa de terminais de comunicação serial sincrona (até 38400 bps); e

- placa de comunicação paralela de alta velocidade.

Os controladores de terminais farão a expansão local destes terminais através de placas de entrada/saida conectadas em um canal de E/S do tipo proposto originalmente pela Motorola [15].

ARBITRAÇAO

Definida a arquitetura, a implementação possibilitou algumas soluções interessantes, dentre as quais podemos destacar a Arbitração do Barramento Principal.

E desejado que cada módulo, quando necessite do barramento principal - constituido por dois dutos padrão VME independentes - adquira o primeiro que esteja disponivel, sendo que todos os módulos têm a mesma prioridade.

A escolha do barramento a ser utilizado é determinada pelo próprio módulo: se este já possui um barramento, o utiliza; caso contrário, é determinado dinamicamente através do sinal de endereço A2.

Foi adotada uma solução que implementa fisicamente o mecanismo varredura cíclica ("round-robin") na priorização dos mestres de barramento, aproveitando a cadeia em cascata ("daisy-chain") do VME para a transmissão dos sinais de comunicação de posse, dotando cada módulo com um bloco lógico que denominamos iniciador [16].

Para agilizar ainda mais o tempo de arbitração foi introduzido o conceito de arbitração paralela, ou seja, uma arbitração pode ocorrer enquanto o recurso ainda está sendo usado (neste caso o barramento VME). Com isto obtivemos um tempo médio de arbitração próximo ao mínimo estipulado pelo barramento VME.[17]

2.SINOPSE

Estruturalmente, a versão 1.0 do sistema MS-8701 é constituída, portanto, dos seguintes módulos:

- 16 Placas de Processamento Geral (PPG), com 4 módulos básicos de processamento (cada módulo com 1 microprocessador 68020 e pastilhas associadas);
- 1 Subsistema de Comunicações a Caracteres - SCC (gerenciado pela Placa de Processamento de Comunicações - PPC);
- 1 Computador de Memória de Massa (CMM);
- 1 Placa de Gerenciamento do Sistema (PGS).

Nestas condições, há a previsão de que o sistema apresente as seguintes características:

- capacidade de processamento: até de 100 MIPS e 12 MFLOPS;
-capacidade de memória física: 512 Mbytes;
- capacidade de memória de massa: 100 Gbytes;
- número de terminais : até 256;
- suporte a redes: X.25 (RENPAC) e ETHERNET (local).

É apresentado, a seguir, o diagrama de blocos da arquitetura do Minissupercomputador (MS-8701), mostrando os principais módulos e o sistema de dutos já mencionados.

No Minissupercomputador MS-8701 as tarefas de administração do sistema são realizadas por um processador dedicado (o processador Placa de Gerenciamento de Sistema,PGS). Assim como o controle de dispositivos , sendo que os 64 processadores das" Placas de Processamento Geral" (PPG's) são exclusivos para a execução dos processos de usuário e funções de serviço do sistema operacional.

O barramento principal, formado por dois dutos de 32 bits de largura de dados, deve suportar o tráfego devido ao acesso de dados nas áreas de memória compartilhada da PGS, da PPC e das PPG's (e.g., dados que devem permanecer acessíveis a todos os supervisores e dados em áreas de memória global das PPG's, ou memória de módulos de controle e operação de dispositivos).

3.CONCLUSÕES

O Minissupercomputador desenvolvido pelo LSI/EPUSP apresenta características modulares e técnicas que o tornam extremamente competitivo, devendo o mesmo,a médio prazo, vir a ser transformado em um produto industrial.

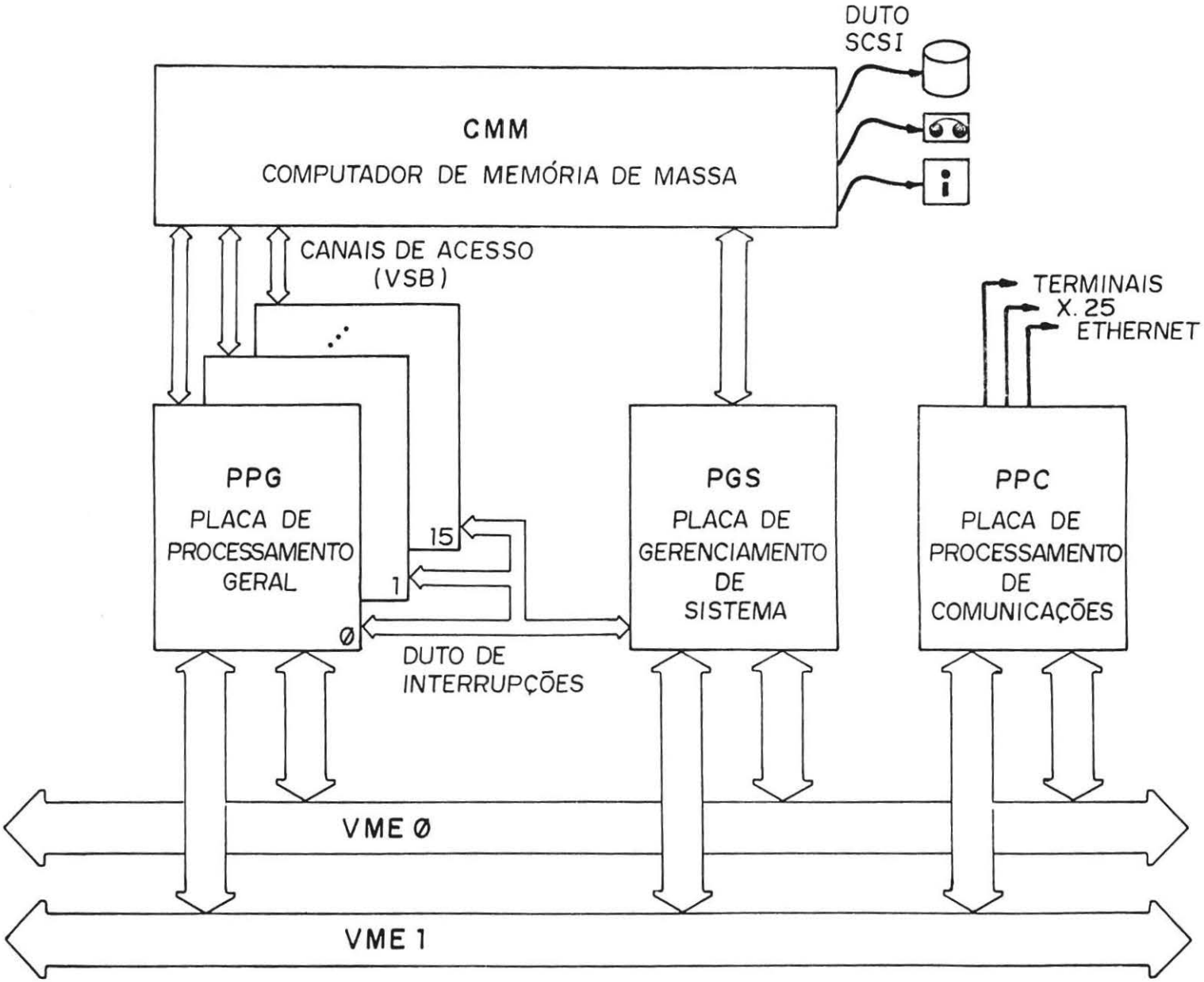
O espírito que norteou seu desenvolvimento, tendo como fundamento software baseado em sistema Unix compatível, tende a dar ao sistema paralelo uma eficiência que torna a relação desempenho/custo desta máquina superior a dos supercomputadores convencionais.

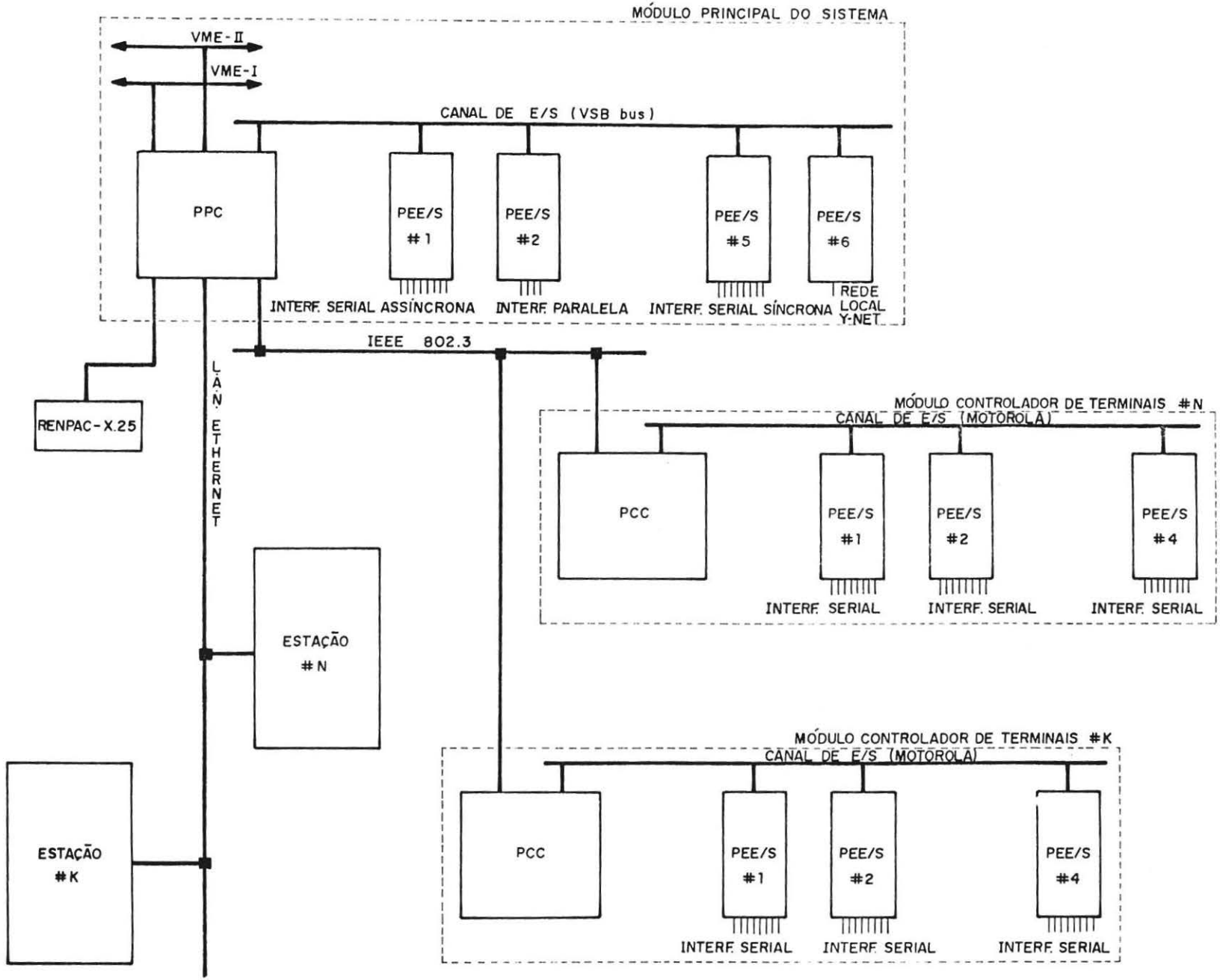
A substituição do microprocessador 68020 por microprocessadores mais avançados poderá levar sua capacidade de processamento acima de 300 mips.

Com relação as operações de ponto flutuante a utilização de unidades de ponto flutuante especiais como as da Weitek possibilitam operações de ponto flutuante que poderão atingir os 150 (cento e cinquenta) MFLOPS.

Atualmente estamos em fase de conclusão do protótipo e de testes operacionais das placas impresso.

MINISSUPER


DUTO
SCSI
CMM
COMPUTADOR DE MEMÓRIA DE MASSA
CANAIS DE ACESSO
(VSB)
TERMINAIS
X.25
ETHERNET
PPG
PLACA DE
PROCESSAMENTO
GERAL
15
1
Ø
DUTO DE
INTERRUPÇÕES
PGS
PLACA DE
GERENCIAMENTO
DE
SISTEMA
PPC
PLACA DE
PROCESSAMENTO
DE
COMUNICAÇÕES
VME Ø
VME 1


MÓDULO PRINCIPAL DO SISTEMA
VME-II
VME-I
CANAL DE E/S (VSB bus)
PPC
PEE/S #1
PEE/S #2
PEE/S #5
PEE/S #6
INTERF. SERIAL ASSÍNCRONA
INTERF. PARALELA
INTERF. SERIAL SÍNCRONA
REDE LOCAL Y-NET
IEEE 802.3
RENPAC-X.25
LAN ETHERNET
MÓDULO CONTROLADOR DE TERMINAIS #N
CANAL DE E/S (MOTOROLA)
PCC
PEE/S #1
PEE/S #2
PEE/S #4
INTERF. SERIAL
INTERF. SERIAL
INTERF. SERIAL
ESTAÇÃO #N
ESTAÇÃO #K
MÓDULO CONTROLADOR DE TERMINAIS #K
CANAL DE E/S (MOTOROLA)
PCC
PEE/S #1
PEE/S #2
PEE/S #4
INTERF. SERIAL
INTERF. SERIAL
INTERF. SERIAL

Obtido sucesso em toda essa integração poderemos ter disponivel, até o fim do corrente ano, uma máquina nacional de grande porte operando em ambiente UNIX com razoável confiabilidade.

A filosofia de projeto do Minissuper, porém, não se limitou apenas a processamento de programas cientificos de grande porte especificos. Esta filosofia procurou adequá-lo também a aplicações de uso geral para que a presente máquina venha concorrer com sucesso em aplicações comerciais convencionais.

Com o desenvolvimento deste projeto visou-se dotar o pais de um computador de grande porte razoavelmente aberto tanto em termos de software, quanto em termos de hardware que poder resolver a baixo custo os problemas computacionais das universidades brasileiras.

4.AGRADECIMENTOS

Gostariamos de expressar nossos sinceros agradecimentos à FINEP (Financiadora de Estudos e Projetos) e ao CNPq (Conselho Nacional de Desenvolvimento Cientifico e Tecnológico), entidades que apoiaram este projeto proporcionando meios para o seu desenvolvimento.

5.REFERENCIAS

[1] JOHNSON, Thomas L.; "The RISC/CISC Melting Port", BYTE, April 1987.

[2] HWANG, K.; "Advanced Parallel Processing with Supercomputer Architetures Proceedings of the IEEE, vol. 75, no.10, October/1987

[3] MOKHOFF, N.; "Parallelism Breeds a New Class of Supercomputers"; Computer Design, vol. 26, no 6, March 15, 1987.

[4] THAKKAR, S.; GIFFORD, P.;FIELLAND, G; "The Balance Multiprocessor System"IEEE Micro, February 1988.

[5] HINDIN, Harve J.;" Minissupercom puter Invade Mainstream Applications"; UNI X/WORLD Magazine, vol.4,no. 12,December 1987.

[6] MOTOROLA INC.; "MC68020 32-BIT Microprocessor User's Manual"; PRENTICE-HALL INC., 1984.

[7] MOTOROLA INC.; "The VMEbus Specification"; VITA, Rev. C.1, October 1985

[8] MAYBERRY, Walter; EFLAND, Gregry;"CacheBoots Multiprocessor Performance", COMPUTER DESIGN, November 1984.

[9] VanAKEN Jerry; "Match Cache Architecture to the Computer System";ELETRONIC DESIGN, March 4, 1982.

[10] SEERY, Jim and LaROCCA; "System level Strategy attacks Key Multiuser Bottlenecks", COMPUTER DESIGN, January, 1988.

[11] WILSON, Ron; "Designers rescue Superminicomputers from I/O Bottle-neck", COMPUTER DESIGN, October, 1987.

[12] GROSSMAN, C.P; "Cache-DASD Storage Design for Improving System Performance", IBM System Journal, 1985.

[13] AGRANAL, D.P and JAIN, R.; "A Pipeline Pseudoparalellism System Architecture for Real-Time Dynamic Scene Analysis", IEEE TRANSACTIONS ON COMPUTERS, October, 1982.

[14] BUCHANAN, Gregory F. and GAULLIER, François; "A Distribuited Terminal Controller for HP Precision Architecture Computers Running the MPE XLOperating System"; HEWLETT-PACKARD Journal, March, 1987.

[15]"I/O Modules - Input/Output Channel Specificacion Manual "MOTOROLA,March, 1982.

[16] CALVO, J. and ACHA, J.I.; "Asynchronous Modular Arbiter", IEEE TRANSACTIONS ON COMPUTER, January,1986.

[17] BEASTON, John and TETRICK, R.Scott; "Designers Confront Metastability in Boards and Buses",COMPUTER DESIGN, March, 1986.