VPP300

- Actualizado (22.12.2005)

Ordenador Vectorial Fujitsu VPP300

O servidor VPP300E deixará de prestar servizo no mes de decembro do 2003. Aqueles usuarios que estean a utilizar este sistema, deben migrar as súas aplicacións a algún dos servidores xa dispoñibles, como o HPC320 ou o cluster SVG. Para máis información, poden dirixirse a .

ORDENADOR	Fujitsu VPP300E
PROCESADOR	Vectorial
CPU´S	6
MEMORIA	12 GB
POTENCIA PICO	14,4 GFLOPS

Arquitectura Vectorial-Paralela

Almacenamento en memoria

Configuración do Sistema:

- Unidade escalar
- Unidade vectorial
- Unidade de almacenamento principal
- Unidade de transferncia de datos
- Elemento de procesamento entrada/saída
- Unidade de Crossbar
- Procesador de servicio

Configuración do VPP300E no CESGA

Arquitectura Vectorial-Paralela

A arquitectura vectorial paralela do VPP300 combina tres tecnoloxías de procesamento paralelo, como se amosa na figura1.

- Procesamento paralelo dos datos a través do procesamento vectorial. Cada elemento de procesamento (PE) do ordenador realiza procesamento vectorial concorrente, de tal forma que se poden realizar varias operacións vectoriais en modo paralelo. Cada PE pode realizar procesamento vectorial cun rendemento pico de 2.4 GFLOPS.

- Procesamento paralelo das instruccións utilizando palabras de instrucción longas (LIW). A unidade de cada PE utiliza unha arquitectura cun conxunto de instruccións reducido (RISC) baseado en LIW para permitir procesamento paralelo a nivel de instruccións. En cada unidade escalar, pódense realizar ata tres instruccións de modo concorrente. Isto proporciona procesamento escalar de alta velocidade de ata 428 millóns de operacións por segundo (MOPS).

- Procesamento paralelo utilizando os PEs. Para mellora-lo rendemento do sistema, pódese utiliza-lo procesamento paralelo do sistema baseado en almacenamento distribuído. Para iso os PEs encóntranse conectados a través dunha rede crossbar, cunha capacidade de 615 MB/s bidireccional.

Almacenamento en memoria.

Os requirimentos necesarios en supercomputación obrigan á utilización de memorias cun grande ancho de banda. Isto provocou que durante moitos anos se utilizase RAM estática (SRAM) como medio de almacenamento en memoria debido á súa alta velocidade de acceso. Sen embargo, debido á escala das simulacións que se realizan nos cálculos científicos e tecnolóxicos é necesaria unha gran cantidade de almacenamento en memoria, o que produce que as SRAMs non sexan tan apropiadas debido ó seu baixo nivel de integración. Por outro lado, as memorias RAM dinámicas (DRAM) poden integrarse a grande escala, pero presentan unha velocidade de acceso considerablemente inferior. Polo tanto, parece difícil conseguir un almacenamento de gran capacidade e grande ancho de banda.

Para solucionar este problema, o VPP300 utiliza memoria DRAM síncrona (SDRAM). A memoria SDRAM presenta a mesma densidade de integración cá memoria DRAM, pero tamén presenta unha velocidade de acceso máis rápida debido a que emprega sincronización mediante reloxo. Coa utilización de SDRAM, cada PE pode ter unha memoria de almacenamento principal de ata 2 Gbytes.

Entrada/saída paralela

No VPP300 existen dous buses de entrada/saída de alta velocidade entre a canle e cada PE. Estes buses maximizan o rendemento de entrada/saída de cada procesador. Ademais, pódense engadir máis procesadores de entrada/saída co fin de aumenta-lo rendemento do procesamento paralelo.

Tecnoloxía de metal-óxido (CMOS.)

Debido ó desenvolvemento que sufriu a tecnoloxía CMOS con avances nos procesos de fabricación de semiconductores, é posible aplicar esta tecnoloxía que ata agora só se utilizaba en estacións de traballo e PCs, a ordenadores destinados ó cálculo intensivo, como é o caso do VPP300. Este ordenador utiliza tecnoloxía de 0.35 mm, cun tempo de propagación de sinal entre portas de 70 ps.

Arquitectura aberta.

Para a conexión de dispositivos externos utilízanse estándares coñecidos baseados en tecnoloxías SCSI, WIDE SCSI, FDDI, HIPPI E ATM. Isto facilita a actualización e a conectabilidade do sistema, ademais de diminuí-los custos.

Configuración do Sistema.

O equipamento do VPP300 consta dos seguintes elementos:

Elementos de procesamento (PE). A figura 2 mostra a configuración hardware de cada PE. A táboa 1 mostra as características principais dun PE. Cada PE consta das seguintes unidades:

- Unidade escalar (SU). A Unidade escalar executa as instruccións escalares e manexa as interrupcións.

- Unidade vectorial (VU). A unidade vectorial executa as instruccións vectoriais de alta velocidade. A VU posúe varias pipelines de execución de instruccións e un rexistro vectorial de alta capacidade.

- Unidade de almacenamento masivo (MSU). A unidade de almacenamento masivo utilízase para almacena-los programas e os datos. A MSU proporciona as grandes cantidades de almacenamento que precisa a unidade vectorial.

- Unidade de transferencia de datos (DTU). A unidade de transferencia de datos procesa as comunicacións de datos entre os PEs a través dunha rede crossbar e sincroniza a transferencia de datos.

- Elemento de procesamento de entrada/saída (IOPE). O IOPE consiste en controladores e adaptadores para conecta-las unidades 1) a 4) anteriores, as canles de control de entrada/saída e os distintos dispositivos de entrada/saída.

- Unidade de Crossbar (XB). A unidade de crossbar encárgase de transferi-los datos entre os PEs utilizando a DTU.

- Procesador de servicio (SVP). O procesador de servicio é un ordenador independente da cabina, dedicado ó control do proceso de acendido e do diagnóstico e mantemento do sistema.

Na figura 3 móstrase un esquema da conexión dos distintos PEs a través do crossbar.

Descrición Detallada do Hardware.

Unidade escalar (SU).

Arquitectura LIW.A arquitectura LIW permite o procesamento paralelo a nivel de instruccións. Cada instrucción LIW contén dous ou máis campos para as operacións que se van executar. As operacións asígnanse a palabras de instrucción mediante o compilador. Debido a que as instruccións se executan en serie sen modificación, a cantidade de hardware necesario redúcese e aumenta a velocidade de procesamento. A táboa 2 describe a execución dunha operación LIW.

Técnicas para o incremento de velocidade. As principais características da unidade escalar destinadas ó incremento da velocidade do procesamento son:

1. En cada palabra de instrucción de 64 bits pode haber de unha a tres operacións escalares ou unha operación vectorial.

2. Só se pode asignar unha dirección relativa do program counter (PC) como dirección de destino para unha operación de salto condicional. Esta restricción incrementa a velocidade para resolve-la dirección de destino dun salto e para a precarga das instruccións de destino do salto.

3.A capacidade de execución asíncrona permite que se execute a instrucción seguinte sen necesidade de esperar a que se complete a anterior operación asíncrona (esta operación asíncrona precedente require polo menos dous ciclos). Polo tanto,

a secuencia de execución das operacións asíncronas pódese modificar sempre que as dependencias dos datos permanezan invariables.

4.Incorpóranse instruccións para realizar "trace scheduling": é unha técnica do compilador que mellora o rendemento ó despraza-las instruccións cara a unha cadea de instruccións sen saltos denominada "bloque básico".

Unidade vectorial (VU).

A unidade vectorial realiza o procesamento vectorial, consistente nun método dunha única instrucción e múltiples datos (SIMD).

A unidade vectorial recibe unha instrucción vectorial desde a unidade escalar e procesa a instrucción vectorial. Os datos vectoriais procésanse na unidade de operación en pipeline.
Existen sete pipelines de execución de instruccións:

- Suma e operación lóxica.
- Multiplicación.
- División.
- Carga.
- Almacenamento.
- Dous pipelines de máscara.

Utilizando estes pipelines, é posible executar dúas ou máis instruccións vectoriais de forma paralela. O rexistro vectorial ten unha capacidade de 128 Kbytes.

Cada serie de procesamento vectorial execútase do modo seguinte. En primeiro lugar, os datos que se encontran na área de almacenamento principal cárganse no rexistro vectorial utilizando o pipeline de carga. Os datos procésanse a continuación utilizando o pipeline de procesamento. Os resultados que se obteñen almacénanse no almacenamento principal utilizando o rexistro vectorial e o pipeline de almacenamento.

Funcións RAS. O rexistro vectorial posúe o mesmo mecanismo de chequeo e corrección de erros (ECC) que a área de almacenamento principal. Este mecanismo é capaz de corrixir completamente os erros que se poidan producir sobre un único bit e de detecta-los erros que se producen sobre 2 bits. Tamén é capaz de corrixi-los erros de 1 bit que se poden producir na área de almacenamento principal, no pipeline de carga, no de almacenamento e no rexistro vectorial. Esta característica mellora en gran medida a fiabilidade global do sistema.

Unidade de almacenamento principal (MSU).

Os elementos utilizados para o almacenamento na MSU están baseados en memorias SDRAM con tempos de acceso de 60 ns (16 Mbits por chip). Cada PE posúe 64 módulos de memoria. Estes módulos de memoria, xunto co MCM que contén a arquitectura CMOSLSI encóntranse montados sobre unha placa nai de alta densidade, que permite mellora-la transferencia de datos entre os módulos LSI e os módulos SDRAM. Desta forma, é posible alcanza-los 2 Gbytes por cada PE. Ó engadir novos PEs, a capacidade de almacenamento principal aumenta de modo escalar.

Técnicas para o procesamento de alta velocidade. As operacións dos módulos de SDRAM encóntranse sincronizadas co reloxo do sistema. Deste modo, o número de elementos utilizados para a transferencia de datos a altas velocidades reduciuse. O almacenamento principal encóntrase dividido en 512 unidades ás que se pode acceder de modo independente. Esta distribucións proporciona un alto ancho de banda para poder manexa-lo elevado número de peticións á memoria principal que realiza a unidade vectorial.

Funcións RAS (Reliability, Availability & Serviciability). A unidade de almacenamento principal ofrece as seguintes funcións RAS para mellora-la fiabilidade do sistema:

Capacidade ECC. A MSU posúe características ECC para a corrección completa de posibles erros en 1 bit e para a detección de erros en 2 bits. Debido a que o rexistro vectorial tamén soporta a característica de ECC, o código extra necesario para realiza-los chequeos almacénase xunto ós datos tanto na MSU coma no rexistro vectorial.

Características de vixilancia (Patrol). A característica de vixilancia utilízase para a recuperación da memoria SDRAM de erros intermitentes en 1 bit. Se se detectan erros en bit durante a lectura de datos desde o almacenamento principal, a característica de vixilancia corrixe e reescribe os datos.

Unidades de transferencia de datos (DTU)

A unidade de transferencia de datos encóntrase instalada no interior de cada PE. As DTUs realizan as tarefas de comunicación entre os PEs a través da rede crossbar de modo independente ás operacións que se están realizando dentro de cada PE, o que proporciona un alto nivel de eficiencia en cálculo paralelo. Cada DTU está formada por unha unidade de procesamento de datos e unha unidade encargada da sincronización entre PEs. As principais características da DTU son:

Transferencia de datos:

O envío dos datos pódese realizar en paralelo coa recepción dos datos. A velocidade de transferencia de datos é de 570 megabytes por segundo. As operacións de acceso ós datos almacenados para poder realiza-la transferencia destes datos pódense clasificar en catro tipos:

- Patróns continuos.
- Patróns equiespaciados.
- Patróns de matrices parciais.
- Patróns indirectos.

A velocidade de transferencia entre PEs pódese mellorar seleccionando o patrón de acceso apropiado. A DTU posúe a capacidade de translación de direccións para traslada-las direccións de memoria para os datos que hai que transferir e as direccións de PE para os destinos das transferencias. Isto proporciona o acceso a un espacio de direccións de memoria virtuais.

Sincronización entre elementos de procesamento.

Cada DTU posúe unha característica de sincronización entre PEs para poder sincronizar dous ou máis PEs cos outros PEs (figura 4). Esta característica permite transmiti-la información sobre o estado de progreso na execución do programa (sobre cada PE) a tódolos outros PEs, e avisa ó sistema de que os PEs se encontran sincronizados despois de recibi-la confirmación de que os PEs recibiron a información contida na mensaxe que se transmitiu.

Unidade de crossbar (XB).

A unidade crossbar encóntrase conectada á DTU de cada PE para permiti-lo control das comunicacións de datos entre os distintos PEs. No VPP300 esta unidade de crossbar encóntrase integrada en cada PE.
A velocidade de comunicación entre dous PEs calquera é de 570 megabytes por segundo gracias á unidade de crossbar. As principais características deste crossbar son:
1. A non ser que exista un PE remoto utilizando as comunicacións, a contención rara vez se produce debido a que se utiliza un switch no crossbar para realiza-las comunicacións (figura 5).
2. A "distancia" entre os PEs é a mesma para todos. Polo tanto, incluso aínda que os procesadores se encontren seleccionados e agrupados de modo arbitrario, as características da rede non varían. Esta característica proporciona a capacidade de executar dous ou máis programas paralelos de modo eficiente.

Canles de entrada/saída, controladores e adaptadores

Pódense conectar dous tipos de canles (canle VME directo e canal SBUS) ó procesador de entrada/saída (IOPE). Ó bus de cada unha dúas das canles pódenselle conectar diversos tipos de controladores e de adaptadores utilizando un interface estándar (figura 6).

Canle VME

A canle VME é a responsable de controla-la conexión entre cada PE e a canle. Débense conectar cinco tipos de controladores para o control da entrada/saída á canle VME. As principais características da canle VME son:

1. Soporte para a función de transferencia de bloques de 64 bits, equivalente á función do VME64.

2. Soporte estándar do controlador de interrupcións e da función de arbitraxe do bus.

3. Soporte estándar da característica de monitorización de sincronización do bus VME e da característica de chequeo de paridade.

Canle SBUS

Débense conectar dous tipos de adaptadores de control de entrada/saída á canle do SBUS. As principais características do SBUS son:

1. Soporte da función de transferencia estendida de 64 bits.

2. Soporte da función de chequeo de paridade.

3. Soporte para a función de transferencia de acceso directo a memoria, DMA (transferencia en refachos de ata 64 bits).

Procesador de servicio (SVP)

O procesador de servicio realiza varias funcións operativas a través dos interfaces de conexión co hardware do sistema principal. As principais características do SVP son:

1. Control de acendido. O SVP encárgase de acender e de apagar todo o sistema.

2. Control da configuración. Se os PEs fallan, a configuración do equipo pode modificarse sen necesidade de dete-lo sistema, desconectando os PEs que presentan erros, utilizando simplemente un comando do sistema operativo.

3. Monitoreo do sistema principal. Para unha recuperación rápida en caso de que se produza un erro, a información sobre o erro é anunciada e transmitida ó centro de mantemento a través dunha liña de comunicación.

4. Función de operación automática. A función de operación automática sopórtase para a carga inicial do sistema operativo de acordo cun conxunto de comandos previamente definido e para o apagado do sistema utilizando un comando do propio sistema operativo.

Configuración do VPP300E no CESGA

A configuración do VPP300E no CESGA está formado por:

- 6 Elementos de Procesamento (PE), de 2.4 GFLOPS de potencia pico por PE, o que proporciona un total de 14.4 GFLOPS de potencia de cálculo.
- Catro canles VME de 80 MB/s cada un deles, cun total de 12 Slots.
- Catro canles SBUS de 100 MB/s con catro slots cada un deles.
- Dous adaptadores para conexión ATM.
- Dous adaptadores HIPPI para a conexión da cabina de discos RAID.
- Dous controladores de rede local Ethernet con conexión AUI e 10Base2.
- Tres controladores Wide SCSI2 cun total de oito canles de 20 MB/s.
- Un controlador SCSI2 con dúas canles de 4 MB/s.
- Un controlador RS232C con catro liñas de 38.4400 bps.
- Unha consola para o arranque e diagnóstico do sistema (SVP).
- Unha unidade de cinta DAT de 4 mm e 2 GB.
- Catro discos de 9 GB cada un instalados na cabina principal.
- Vinte discos de 9 GB cada un instalados na cabina de discos.
- Unha cinta de cartuchos con autocargador.
- 1 cabina de discos GEN5L con tecnoloxía RAID cunha capacidade de 72 GB.
- 2 Gbytes de memoria por elemento de procesamento, proporcionando un total de 12 Gbytes de memoria.
- 260 Gbytes de almacenamento en disco.

A distribución dos nodos móstrase na figura 7 Creáronse dous grupos IPL (Initial Program Load), agrupando os procesadores en dous grupos de 3 procesadores. Desta forma créanse dous sistemas completamente independentes e que poden configurarse de forma separada. Á súa vez, dentro de cada IPL, tódolos procesadores posúen as mesmas características operativas, o que facilita a configuración do equipo.

Dentro de cada IPL existe un procesador primario ou mestre, encargado de acceder ós discos nos que se encontra o sistema operativo e de distribuílo ós demais procesadores do seu mesmo IPL, ós que se lles denomina procesadores secundarios (secondary PE, S-PE). Dentro do IPL principal, o nodo que actúa de mestre denomínase primary PE, xa que ademais de actuar de mestre para o seu PIL, actúa de referencia para que o resto dos procesadores primarios dos outros IPL (neste caso, só un) comecen o proceso de arranque. O procesador primario do outro grupo IPL recibe o nome de IPL group master PE (IMPE).

Polo que respecta ó sistema de entrada e saída a disco, existen dous PEs de entrada/saída, que ademais de encontrarse dispoñibles para o cálculo vectorial encárganse das tarefas de escritura e lectura de datos ós sistemas de discos duros. Estes dous PEs repártense o traballo doutros 2 PEs cada un, tal como se presenta na figura 8. As peticións de entrada/saída dos procesadores transmítense a través do crossbar ata os PEs de entrada/saída, que á súa vez conectan con cada unha das cabinas de discos ou coa cabina de discos que se encontra en RAID 5, o Gen5 LE. A conexión cos discos das cabinas realízase a través dun interface Wide SCSI de 20 MB/s, mentres que a conexión ó Gen5 faise a través dun interface HIPPI con capacidade de ata 100 MB/s.

Á súa vez, un dos procesadores o (PE0) encóntrase dispoñible para o proceso interactivo (conexións telnet e execución de programas en modo interactivo), mentres que o resto dos procesadores están reservados para executar procesos en cola (batch), polo que o seu acceso está restrinxido para os usuarios.