VPP300
- Modificado (22.12.2005)

Ordenador Vectorial Fujitsu VPP300

El servidor VPP300E dejara de prestar servicio en el mes de Diciembre del 2003. Aquellos usuarios que todavía continúen utilizando este sistema, deben migrar sus aplicaciones a alguno de los servidores ya disponibles, como el HPC320 o el cluster SVG. Para mas información, pueden dirigirse a .

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

Arquitectura Vectorial-Paralela

Almacenamiento en memoria

Configuración del Sistema:

- Unidad escalar
- Unidad vectorial
- Unidad de almacenamiento principal
- Unidad de transferncia de datos
- Elemento de procesamiento entrada/salida
- Unidad de Crossbar
- Procesador de servicio

Configuración del VPP300E en el CESGA

Arquitectura Vectorial-Paralela

La arquitectura vectorial paralela del VPP300 combina tres tecnologías de procesamiento paralelo, como se muestra en la figura 1.

- Procesamiento paralelo de los datos a través del procesamiento vectorial. Cada elemento de procesamiento (PE) del ordenador realiza procesamiento vectorial concurrente, de tal forma que se pueden realizar varias operaciones vectoriales en modo paralelo. Cada PE puede realizar procesamiento vectorial con un rendimiento pico de 2.4 GFLOPS.

- Procesamiento paralelo de las instrucciones utilizando palabras de instrucción largas (LIW). La unidad escalar de cada PE utiliza una arquitectura con un conjunto de instrucciones reducido (RISC) basado en LIW para permitir procesamiento paralelo a nivel de instrucciones. En cada unidad escalar, se pueden realizar hasta tres instrucciones de modo concurrente. Esto proporciona procesamiento escalar de alta velocidad de hasta 428 millones de operaciones por segundo (MOPS).

- Procesamiento paralelo utilizando los PEs. Para mejorar el rendimiento del sistema, se puede utilizar el procesamiento paralelo del sistema basado en almacenamiento distribuido. Para ello, los PEs se encuentran conectados a través de una red crossbar, con una capacidad de 570 MB/s bidireccional.

Almacenamiento en memoria.

Los requerimientos necesarios en supercomputación obligan a la utilización de memorias con un gran ancho de banda. Esto ha provocado que durante muchos años se ha utilizado RAM estática (SRAM) como el medio de almacenamiento en memoria debido a su alta velocidad de acceso. Sin embargo, debido a la escala de las simulaciones que se realizan en los cálculos científicos y tecnológicos es necesaria una gran cantidad de almacenamiento en memoria, lo que produce que las SRAMs no sean tan apropiadas debido a su bajo nivel de integración. Por otro lado, las memorias RAM dinámicas (DRAM) pueden integrarse a gran escala, pero presentan una velocidad de acceso considerablemente inferior. Por tanto, parece difícil conseguir una almacenamiento de gran capacidad y gran ancho de banda.

Para solucionar este problema, el VPP300 utiliza memoria DRAM síncrona (SDRAM). La memoria SDRAM presenta la misma densidad de integración que la memoria DRAM, pero también presenta una velocidad de acceso más rápida debido a que emplea sincronización mediante reloj. Con la utilización de SDRAM, cada PE puede tener una memoria de almacenamiento principal de hasta 2 Gbytes.

Entrada/salida paralela

En el VPP300 existen dos buses de entrada/salida de alta velocidad entre el canal y cada PE. Estos buses maximizan el rendimiento de entrada/salida de cada procesador. Además, se pueden añadir más procesadores de entrada/salida con el fin de aumentar el rendimiento del procesamiento paralelo.

Tecnología de metal-óxido (CMOS.)

Debido al desarrollo que ha sufrido la tecnología CMOS con avances en los procesos de fabricación de semiconductores, es posible aplicar esta tecnología que hasta ahora sólo se utilizaba en estaciones de trabajo y PCs, a ordenadores destinados al cálculo intensivo, como es el caso del VPP300. Este ordenador utiliza tecnología de 0.35 mm, con un tiempo de propagación de señal entre puertas de 70 ps.

Arquitectura abierta.

Para la conexión de dispositivos externos se utilizan estándares conocidos basados en tecnologías SCSI, WIDE SCSI,

FDDI, HIPPI y ATM. Esto facilita la actualización y conectabilidad del sistema, además de disminuir los costes.

Configuración del Sistema.

El equipamiento del VPP300 consta de los siguientes elementos:

- Elementos de procesamiento (PE). La figura 2 muestra la configuración hardware de cada PE. La tabla 1 muestra las características principales de un PE. Cada PE consta de las siguientes unidades:

- Unidad escalar (SU). La unidad escalar ejecuta las instrucciones escalares y maneja las interrupciones.

- Unidad vectorial (VU). La unidad vectorial ejecuta las instrucciones vectoriales de alta velocidad. La VU posee varias pipelines de ejecución de instrucciones y un registro vectorial de alta capacidad.

- Unidad de almacenamiento masivo (MSU). La unidad de almacenamiento masivo se utiliza para almacenar los programas y los datos. La MSU proporciona las grandes cantidades de almacenamiento que precisa la unidad vectorial.

- Unidad de transferencia de datos (DTU). La unidad de transferencia de datos procesa las comunicaciones de datos entre los PEs a través de una red crossbar y sincroniza la transferencia de datos.

- Elemento de procesamiento de entrada/salida (IOPE). El IOPE consiste en controladores y adaptadores para conectar las unidades 1) a 4) anteriores, los canales de control de entrada/salida y los distintos dispositivos de entrada/salida.

- Unidad de Crossbar (XB). La unidad de crossbar se encarga de transferir los datos entre los PEs utilizando la DTU.

- Procesador de servicio (SVP). El procesador de servicio es un ordenador independiente de la cabina, dedicado al control del proceso de encendido y del diagnóstico y mantenimiento del sistema.

En la figura 3 se muestra un esquema de la conexión de los distintos PEs a través del crossbar.

Descripción Detallada del Hardware.

Unidad escalar (SU).

Arquitectura LIW. La arquitectura LIW permite el procesamiento paralelo a nivel de instrucciones. Cada instrucción LIW contiene dos o más campos para las operaciones que se van a ejecutar. Las operaciones se asignan a palabras de instrucción mediante el compilador. Debido a que las instrucciones se ejecutan en serie sin modificación, la cantidad de hardware necesario se reduce y aumenta la velocidad de procesamiento. La figura 1.55 describe la ejecución de una operación LIW.

Técnicas para el incremento de la velocidad. Las principales características de la unidad escalar destinadas al incremento de la velocidad de procesamiento son:

1. En cada palabra de instrucción de 64 bits puede haber de una a tres operaciones escalares ó una operación vectorial.

2. Sólo se puede asignar una dirección relativa del program counter (PC) como dirección de destino para una operación de salto condicional. Esta restricción incrementa la velocidad para resolver la dirección de destino de un salto y para la precarga de las instrucciones de destino del salto.

3. La capacidad de ejecución asíncrona permite que se ejecute la instrucción siguiente sin necesidad de esperar a que se

complete la anterior operación asíncrona (esta operación asíncrona precedente requiere al menos dos ciclos). Por tanto, la secuencia de ejecución de las operaciones asíncronas se puede modificar siempre que las dependencias de los datos permanezcan invariables.

4. Se incorporan instrucciones para realizar "trace scheduling": es una técnica del compilador que mejora el rendimiento al desplazar las instrucciones hacia una cadena de instrucciones sin saltos denominada "bloque básico".

Unidad vectorial (VU).

La unidad vectorial realiza el procesamiento vectorial, consistente en un método de una única instrucción y múltiples datos (SIMD).

La unidad vectorial recibe una instrucción vectorial desde la unidad escalar y procesa la instrucción vectorial. Los datos vectoriales se procesan en la unidad de operación en pipeline. Existen siete pipelines de ejecución de instrucciones: Existen siete pipelines de ejecución de instrucciones:

- Suma y operación lógica.
- Multiplicación.
- División.
- Carga.
- Almacenamiento.
- Dos pipelines de máscara.

Utilizando estos pipelines, es posible ejecutar dos o más instrucciones vectoriales de forma paralela. El registro vectorial tiene una capacidad de 128 Kbytes. El registro de máscara tiene una capacidad de 2 Kbytes.

Cada serie de procesamiento vectorial se ejecuta del modo siguiente. En primer lugar, los datos que se encuentran en el área de almacenamiento principal se cargan en el registro vectorial utilizando el pipeline de carga. Los datos se procesan a continuación utilizando el pipeline de procesamiento. Los resultados que se obtienen se almacenan en el almacenamiento principal utilizando el registro vectorial y el pipeline de almacenamiento.

Funciones RAS. El registro vectorial posee el mismo mecanismo de chequeo y corrección de errores (ECC) que el área de almacenamiento principal. Este mecanismo es capaz de corregir completamente los errores que se puedan producir sobre un único bit y de detectar los errores que se producen sobre 2 bits. También es capaz de corregir los errores de 1 bit que se pueden producir en el área de almacenamiento principal, en el pipeline de carga, en el de almacenamiento y en el registro vectorial. Esta característica mejora en gran medida la fiabilidad global del sistema.

Unidad de almacenamiento principal (MSU).

Los elementos utilizados para el almacenamiento en la MSU están basados en memorias SDRAM con tiempos de acceso de 60 ns (16 Mbits por chip). Cada PE posee 64 módulos de memoria. Estos módulos de memoria, junto con el MCM que contiene la arquitectura CMOS LSI se encuentran montados sobre una placa madre de alta densidad, que permite mejorar la transferencia de datos entre los módulos LSI y los módulos de SDRAM. De esta forma, es posible alcanzar los 2 Gbytes por cada PE. Al añadir nuevos PEs, la capacidad de almacenamiento principal aumenta de modo escalar.

Técnicas para el procesamiento de alta velocidad. Las operaciones de los módulos de SDRAM se encuentran sincronizadas con el reloj del sistema. De este modo, el número de elementos utilizados para la transferencia de datos a altas velocidades se ha reducido. El almacenamiento principal se encuentra dividido en 512 unidades a las que se

puede acceder de modo independiente. Esta distribución proporciona un alto ancho de banda para poder manejar el elevado número de peticiones a la memoria principal que realiza la unidad vectorial.

Funciones RAS (Reliability, Availability & Serviciability). La unidad de almacenamiento principal ofrece las siguientes funciones RAS para mejorar la fiabilidad del sistema:

Capacidad ECC. La MSU posee características ECC para la corrección completa de posibles errores en 1 bit y para la detección de errores en 2 bits. Debido a que el registro vectorial también soporta la característica de ECC, el código extra necesario para realizar los chequeos se almacena junto a los datos tanto en la MSU como en el registro vectorial.

Característica de vigilancia (Patrol). La característica de vigilancia se utiliza para la recuperación de la memoria SDRAM de errores intermitentes en 1 bit. Si se detectan errores en 1 bit durante la lectura de datos desde el almacenamiento principal, la característica de vigilancia corrige y reescribe los datos.

Unidades de transferencia de datos (DTU)

La unidad de transferencia de datos se encuentra instalada en el interior de cada PE. Las DTUs realizan las tareas de comunicación entre los PEs a través de la red crossbar de modo independiente a las operaciones que se están realizando dentro de cada PE, lo que proporciona un alto nivel de eficiencia en cálculo paralelo. Cada DTU está formada por una unidad de procesamiento de datos y una unidad encargada de la sincronización entre PEs. Las principales características de la DTU son:

Transferencia de datos:

El envío de los datos se puede realizar en paralelo con la recepción de los datos. La velocidad de transferencia de datos es de 570 megabytes por segundo:

- Patrones continuos.
- Patrones equiespaciados.
- Patrones de matrices parciales.
- Patrones indirectos.

La velocidad de transferencia entre PEs se puede mejorar seleccionando el patrón de acceso apropiado. La DTU posee la capacidad de traslación de direcciones para trasladar las direcciones de memoria para los datos a transferir y las direcciones de PE para los destinos de las transferencias. Esto proporciona el acceso a un espacio de direcciones de memoria virtuales.

Sincronización entre elementos de procesamiento.

Cada DTU posee una característica de sincronización entre PEs para poder sincronizar dos o más PEs con los otros PEs (figura 4). Esta característica permite transmitir la información sobre el estado de progreso en la ejecución del programa (sobre cada PE) a todos los otros PEs, y avisa al sistema que los PEs se encuentran sincronizados después de recibir la confirmación de que los PEs han recibido la información contenida en el mensaje que se ha transmitido.

Unidad de crossbar (XB).
Cada PE contiene un registro de máscara que indica qué grupo de PE va a ser sincronizado. Utilizando el registro de máscara, un programa se puede ejecutar en un grupo de PEs arbitrario. Esto permite que se ejecuten de modo eficiente dos o más programas paralelos de manera simultánea.
La velocidad de comunicación entre dos PEs cualesquiera es de 570 megabytes por segundo gracias a la unidad de crossbar. Las principales características de este crossbar son:
1. A menos que exista un PE remoto utilizando las comunicaciones, la contención rara vez se produce debido a que se utiliza un switch en el crossbar para realizar las comunicaciones (figura 5).
2. La "distancia" entre los PEs es la misma para todos. Por tanto, incluso aunque los procesadores se encuentren seleccionados y agrupados de modo arbitrario, las características de la red no varían. Esta característica proporciona la capacidad de ejecutar dos o más programas paralelos de modo eficiente..

Canales de entrada/salida, controladores y adaptadores

Se pueden conectar dos tipos de canales (canal VME directo y canal SBUS) al procesador de entrada/salida (IOPE). Al bus de cada uno de los canales se le pueden conectar diversos tipos de controladores y de adaptadores utilizando un interface estándar (figura 6).

Canle VME

El canal VME es el responsable de controlar la conexión entre cada PE y el canal. Se deben conectar cinco tipos de controladores para el control de la entrada/salida al canal VME. Las principales características del canal VME son:

1. Soporte para la función de transferencia de bloques de 64 bits, equivalente a la función del VME64.

2. Soporte estándar del controlador de interrupciones y de la función de arbitraje del bus.

3. Soporte estándar de la característica de monitorización de sincronización del bus VME y de la característica de chequeo de paridad.

Canal SBUS

Se deben conectar dos tipos de adaptadores de control de entrada/salida al canal del SBUS. Las principales características del SBUS son:

1. Soporte de la función de transferencia extendida de 64 bits.

2. Soporte de la función de chequeo de paridad.

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

Procesador de servicio (SVP)

El procesador de servicio realiza varias funciones operativas a través de los interfaces de conexión con hardware del sistema principal. Las principales características del SVP son:

1. Control de encendido. El SVP se encarga de encender y de apagar todo el sistema.

2. Control de la configuración. Si los PEs fallan, al configuración del equipo puede modificarse sin necesidad de detener el sistema, desconectando los PEs que presentan errores, utilizando simplemente un comando del sistema operativo.

3. Monitoreo del sistema principal. Para una recuperación rápida en caso de que se produzca un error, la información sobre el error es anunciada y transmitida al centro de mantenimiento a través de una línea de comunicación.

4. Función de operación automática. La función de operación automática se soporta para la carga inicial del sistema operativo de acuerdo con un conjunto de comandos previamente definido y para el apagado del sistema utilizando un comando del propio sistema operativo.

Configuración del VPP300E en el CESGA

La configuración del VPP300E en el CESGA está formado por:

- 6 Elementos de Procesamiento (PE), de 2.4 GFLOPS de potencia pico por PE, lo que proporciona un total de 14.4 GFLOPS de potencia de cálculo.
- Cuatro canales VME de 80 MB/s cada uno de ellos, con un total de 12 Slots.
- Cuatro canales SBUS de 100 MB/s con cuatro slots cada uno de ellos.
- Dos adaptadores para conexión ATM.
- Dos adaptadores HIPPI para la conexión da cabina de discos RAID.
- Dos controladores de red local Ethernet con conexión AUI e 10Base2.
- Tres controladores Wide SCSI2 con un total de ocho canales de 20 MB/s.
- Un controlador SCSI2 con dos canales de 4 MB/s.
- Un controlador RS232C con cuatro líneas de 38.4400 bps.
- Una consola para el arranque y diagnóstico del sistema (SVP).
- Una unidad de cinta DAT de 4 mm y 2 GB.
- Cuatro discos de 9 GB cada uno instalados en la cabina principal.
- Veinte discos de 9 GB cada uno instalados en la cabina de discos.
- Una cinta de cartuchos con autocargador.
- 1 cabina de discos GEN5L con tecnología RAID con una capacidad de 72 GB.
- 2 Gbytes de memoria por elemento de procesamiento, proporcionando un total de 12 Gbytes de memoria.

- 260 Gbytes de almacenamiento en disco.

La distribución de los nodos se muestran en la figura 7. Se han creado dos grupos IPL (Initial Program Load), agrupando los procesadores en dos grupos de 3 procesadores. De esta forma se crean dos sistemas completamente independientes y que pueden configurarse de forma separada. A su vez, dentro de cada IPL, todos los procesadores poseen las mismas características operativas, lo que facilita la configuración del equipo.

Dentro de cada IPL existe un procesador primario o maestro, encargado de acceder a los discos en los que se encuentra el sistema operativo y de distribuirlo a los demás procesadores de su mismo IPL, a los que se les denomina procesadores secundarios (secondary PE, S-PE). Dentro del IPL principal, el nodo que actúa de maestro se denomina primary PE, ya que además de actuar de maestro para su IPL, actúa de referencia para que el resto de los procesadores primarios de los otros IPL (en este caso, sólo uno) comiencen el proceso de arranque. El procesador primario del otro grupo IPL recibe el nombre de IPL group master PE (IMPE).

En lo que respecta al sistema de entrada y salida a disco, existen dos PEs de entrada/salida, que además de encontrarse disponibles para el cálculo vectorial se encargan de las tareas de escritura y lectura de datos a los sistemas de discos duros. Estos dos PEs se reparten el trabajo de otros 2 PEs cada uno, tal y como se representa en la figura 8. Las peticiones de entrada/salida de los procesadores se transmiten a través del crossbar hasta los PEs de entrada/salida, quienes a su vez conectan con cada una de las cabinas de discos ó con la cabina de discos que se encuentra en RAID 5, el Gen5 LE. La conexión con los discos de las cabinas se realiza a través de un interface Wide SCSI2 de 20 MB/s, mientras que la conexión al Gen5 se realiza a través de un interface HIPPI con capacidad de hasta 100 MB/s.

A su vez, uno de los procesadores (el PE0) se encuentra disponible para el proceso interactivo (conexiones telnet y ejecución de programas en modo interactivo), mientras que el resto de los procesadores están reservados para ejecutar procesos en cola (batch), por lo que su acceso está restringido para los usuarios.