Fujitsu AP3000

El servidor de calculo AP3000 dejo de prestar servicios de cálculo el día 1 de Octubre. Excepcionalmente, durante un período transitorio sería posible el acceso al sistema para la realización de cálculos que no puedan ejecutarse en ninguno de los otros sistemas disponibles. Los datos de usuario almacenados en el AP3000 seguirán disponibles para todos los usuarios a través del sistema de almacenamiento y del servidor de cálculo HPC4500. Para más información, pueden dirigirse a sistemas@cesga.es.

Ordenador	Procesador	CPU´S	Memoria	Potencia pico
Fujitsu AP3000	ESCALAR	20	2,5 GB	12 GFLOPS

El AP3000 es un ordenador paralelo DM-MIMD capaz de aprovechar las aplicaciones ya existentes. Para ello, el AP3000 utiliza estaciones UltraSPARC como nodos, y por tanto puede ejecutar una amplia variedad de software que ya se encuentra disponible para este tipo de estaciones de trabajo. Además, posee una arquitectura de memoria distribuida que proporciona un nivel de escalabilidad muy superior al que se puede alcanzar mediante arquitecturas de multiprocesadores simétricos (SMP).

Características.

Entre las características más importantes del AP3000 se encuentran:

1. Implementación de alto rendimiento utilizando procesamiento paralelo multinodo: habitualmente, en el procesamiento paralelo, las comunicaciones entre los nodos afectan en gran modo al rendimiento general del sistema. Para poder alcanzar un alto rendimiento en procesamiento paralelo, es necesario un sistema de comunicaciones entre nodos que presente una baja latencia y un alto ancho de banda. Para conseguir aumentar la velocidad de transmisión, el AP3000 utiliza una red de comunicación de alta velocidad denominada AP-Net, basada en los anteriores desarrollos que ya se utilizaron en la arquitectura del AP1000. Para conseguir un nivel de latencia reducido y un elevado ancho de banda en las comunicaciones entre nodos, el AP3000 utiliza un esquema de encaminamiento de mensajes similar al utilizado en el AP1000. Para que el nivel de latencia sea bajo, es importante no sólo aumentar la velocidad de transmisión de datos en la red, sino también reducir de forma significativa el tiempo necesario para establecer y configurar el envío de los mensajes. Por ello, en el AP3000 se soporta un sistema de comunicaciones a nivel de usuario de forma que la comunicación de mensajes puede ser activada directamente sin ningún tipo de ayuda por parte del sistema operativo.

2. Alto throughput para los programas existentes: el AP3000 utiliza estaciones de trabajo ya existentes como nodos, de tal forma que se pueden utilizar directamente aplicaciones que todavía no han sido preparadas para ser utilizadas en el procesamiento paralelo. Para manejar programas destinados al procesamiento distribuido con una alta velocidad, es necesario utilizar interfaces de comunicación que sean rápidos y compatibles con las redes de área local estándar. Por tanto, las operaciones tales como el acceso a ficheros utilizando NFS ó IP (Internet Protocol) se encaminan a través de la red AP-Net para acelerar la velocidad de transmisión de datos.

3. Facilidad en el control y mantenimiento del sistema. Los grandes sistemas (basados en más de 100 estaciones de trabajo) resultan extremadamente difíciles de administrar. Esto obliga a que los administradores de sistema deban recibir todas las facilidades posibles que les permitan controlar de forma simultánea el encendido de los nodos, la instalación de los nodos, el monitoreo del estado de operación, y realizar otro tipo de tareas relacionadas. De igual modo, se deben soportar las funciones que permitan controlar el sistema de forma automática de acuerdo con ciertos planes operativos prefijados.

4. Soporte para el aumento del número de nodos y de canales de entrada/salida: el AP3000 soporta la capacidad de ampliar, de un modo sencillo, el cluster de estaciones de trabajo que lo forma, así como la ampliación de su red de comunicación de alta velocidad. Para la red AP-Net se utiliza una red toroidal bidimensional con alto nivel de expansión, y su escalabilidad le permite soportar desde 4 hasta 1024 nodos.

Número de nodos	Desde 4 hasta 1024
Tipo de nodos	U170, U200, U300, P250, P300
Capacidad de memoria	Desde 128 Mbytes hasta 2 Tbytes
Discos duros internos	Desde 8.4 Gbytes hasta 4.2 Tbytes
Red interna	Ap-net (200 mbytes/s bidireccional)
Redes externas conectables	Ethernet, Fast Ethernet, FDDI, ATM,...
Dispositivos externos conectables	Arrays de discos, librerias de cintas,...
Sistema operativo	Solaris

Tabla 1.-Especificaciones del AP3000

Configuración del sistema

Nodos

La figura 1 muestra la configuración hardware de los nodos. La tabla 1.15 muestra las especificaciones del AP3000 y la tabla 1.16 muestra las especificaciones de los nodos del AP3000.

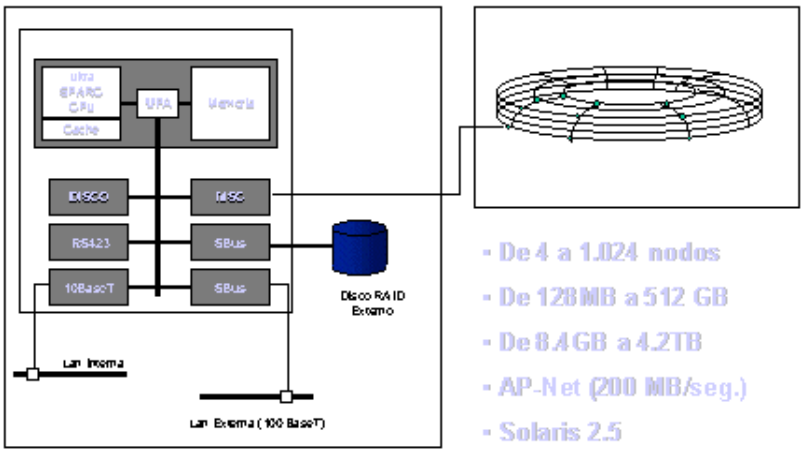

Figura 1. Configuración hardware dos nodos do AP3000 e rede de conexión.

Comunicaciones

A continuación se describe la arquitectura de comunicaciones implementada en el AP3000. La tarjeta MSC (message controller) se encuentra conectada a cada nodo a través de un SBus (bus de entrada/salida) para su conexión con la AP-Net. La tarjeta MSC está formada por un chip controlador de mensajes (MSC) y una memoria buffer. El MSC incluye controladores DMA para la transferencia de datos con la AP-Net.
La AP-Net está basada en una topología toroidal bidimensional y consiste en un conjunto de controladores de enrutamiento (RTCs) encargados de dirigir los mensajes. Entre las características de la red AP-Net se encuentran:

1. Alta velocidad de transmisión de datos a través del puerto: 200 Mbytes/s. El método de encaminamiento es estático. Los mensajes se encaminan primero en la dirección del eje X y a continuación en la dirección del eje Y.

2. Wormhole routing. El wormhole-routing divide los datos (mensajes) que se van a transmitir en pequeños trozos denominados "flits", cada uno de ellos formado por varios bytes. Los nodos de encaminamiento transmiten los mensajes en forma de "flits". Los "flits" de la cabecera de los mensajes determinan el camino que debe seguir el mensaje para llegar a su destino, y los datos que le siguen se envían por el mismo camino.

3. Canal de comunicaciones virtual dual. El hardware AP-Net soporta caminos de comunicación virtuales, denominados canales, de forma que los datos puede ser transferidos de modo independiente entre los nodos utilizando canales duales. Uno de los canales duales se utiliza para la comunicación IP utilizada por el sistema, mientras que el otro canal se utiliza por el usuario para el procesamiento paralelo del software de aplicaciones. Cada canal tiene caminos lógicos de comunicación diferentes para evitar que se produzca "deadlocking" y para manejar las peticiones de mensajes y responder a los mensajes de forma separada. Además, cada camino se encuentra duplicado para evitar el "deadlocking" en la topología toroidal. Por tanto, existen un total de ocho canales de comunicación lógicos sobre un único camino físico de comunicación.

4. Sincronización de barreras entre nodos. En un sistema de procesamiento paralelo, a pesar de que los nodos operan de modo independiente, el sistema entero debe permanecer sincronizado en los pasos que realiza. El AP3000 consigue la sincronización de barreras entre nodos distribuyendo y recogiendo los mensajes de sincronización de la red. El hardware RTC se encarga de distribuir y recoger los mensajes de sincronismo

5. Reliability, Availability & Serviceability. Para poder alertar sobre posibles errores en la red, el proceso de monitoreo en SYSCNTL busca errores en los RTCs. Cuando se producen errores, se emite un mensaje acerca del error que se recibe en la estación de control. Los RTCs chequean a qué nodos se emiten los mensajes. En caso de que se intente enviar un mensaje a un nodo que se encuentre fuera de los grupos definidos, o que se reciba un mensaje incorrecto de una fuente externa, el RTC informa de que se ha producido un error.

Interface de Comunicación

El MSC es el hardware utilizado para soportar la comunicación entre nodos. Posee un controlador de comunicaciones con dos canales para el sistema y dos canales para el usuario. Además de las operaciones convencionales de SEND y RECV, que transfieren mensajes a través de los buffers de envío y recepción, el MSC también soporta el acceso directo a la memoria de los nodos remotos, así como las funciones PUT y GET, CSI (compare and swap instruction) y FOP (fetch and operation).

1. La función SEND trasmite los datos de la memoria local hasta un nodo específico. Los datos que se transmiten se escriben en el buffer de recepción de mensajes del nodo al que se envía el mensaje. El buffer se controla a través de mecanismos hardware.

2. PUT copia los datos que se encuentran en el nodo local hasta la memoria del nodo remoto. GET copia los datos que se encuentran en el nodo remoto hacia el nodo local. De esta forma,

las funciones PUT y GET proporcionan un mecanismo de comunicación efectivo cuando los datos que se van a transmitir entre los nodos se encuentran previamente determinados. Esto es debido a que, al contrario de lo que sucede en el caso de SEND y RECV, no existe la necesidad de copiar los datos en el receptor.

3. Las funciones CSI y FOP se utilizan para el acceso exclusivo a la memoria de sistemas remotos. Estos mecanismos pueden ser utilizados para el control exclusivo de una base de datos.

Interface del usuario

El MSC posee la capacidad de encolar las instrucciones de transmisión de datos tales como PUT, GET y SEND. Esta característica permite que las peticiones de transmisión de datos se procesen de modo separado, de forma que los cálculos y las transmisiones se pueden ejecutar de modo simultáneo.

El controlador de comunicaciones de canal dual en cada MSC está supervisado por un dispositivo de comunicaciones del sistema y un dispositivo de comunicaciones a nivel de usuario. El dispositivo de comunicaciones a nivel de sistema está instalado en el sistema operativo Solaris para permitir la comunicación IP. La comunicación a nivel de usuario está implementada a través de una librería de comunicaciones utilizada para el acceso directo al hardware de comunicaciones del MSC.

Los usuarios pueden utilizar las características de comunicación a alta velocidad utilizando librerías de paso de mensajes estándar como MPI y PVM.

Configuración del AP3000 en el CESGA.

El ordenador AP3000 instalado en el CESGA está formado por 16 nodos U300, de los cuales 4 poseen dos procesadores por nodo, contabilizando un total de 20 procesadores.

La capacidad de memoria del equipo instalado en el CESGA es de 2.5 GB de memoria (128 MB por procesador), y el almacenamiento en disco totaliza 89 GB (4.2 GB por nodo y un array de 25 GB en el nodo 0).

A pesar de que existen 16 nodos, la configuración de las colas sólo permite la ejecución de trabajos paralelos de hasta 12 procesadores, debido a que 4 nodos se encuentran disponibles para aplicaciones de usuario y procesos interactivos, así como compilación y edición de programas.

Tipo de Nodo	U170	U200	U300	P250	P300
Número de CPUs	1	1 ó 2	1 ó 2	1	1
Procesador	UltraSPARC-I (167 MHz)	UltraSPARC-I (200 MHz)	UltraSPARC-II (300 MHz)	UltraSPARC-II (248 MHz)	UltraSPARC-II (296 MHz)
SPECint95	6.26	7.72	12.1	10.0	12.1
SPECfp95	9.06	11.4	13.3	14.9	18.1
Memoria Caché	L1 - 32KB L2 - 512KB	L1 - 32KB L2 - 1MB	L1 - 32KB L2 - 8MB	Interna. 32KB Externa. 1MB	Interna. 32KB Externa. 1MB
Memoria	64MB-1GB	128MB-2GB	128MB-2GB	128MB-2GB	128MB-2GB
Disco duro interno	2.1GB-4.2GB	4.2GB-8.4GB	4.2GB-8.4GB	4.2GB-13.2GB	4.2GB-13.2GB
Número de slots disponibles	2(S Bus)	2(S Bus)	2(S Bus)	2(PCI)	2(PCI)

Táboa 2. Especificacións dos nodos do AP3000