AP3000

- Actualizado (01.12.2005)

O Fujitsu AP3000

O servidor de cálculo AP3000 deixou de prestar servicios de cálculo o día 1 de Outubro. Excepcionalmente, durante un período transitorio será posible o acceso ó sistema para a realización de cálculos que non poidan executarse en ningún dos outros sistemas dispoñibles. Os datos de usuario almacenados no AP3000 seguirán dispoñibles para tódolos usuarios a través do sistema de almacenamento e do servidor de cálculo HPC4500. Para máis información, poden dirixirse a .

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

O AP3000 é un ordenador paralelo DM-MIMD capaz de aproveita-las aplicacións xa existentes. Para iso, o AP3000 utiliza estacións UltraSPARC como nodos, e polo tanto pode executar unha ampla variedade de software que xa se encontra dispoñible para este tipo de estacións de traballo. Ademais, posúe unha arquitectura de memoria distribuída que proporciona un nivel de escalabilidade moi superior ó que se pode alcanzar mediante arquitecturas de multiprocesadores simétricos (SMP).

Características

Configuración do sistema.

- Nodos.

- Comunicacións

- Interface de Comunicación.

- Interface de Usuario.

Configuración do AP3000 no CESGA.

Características

Entre as características máis importantes do AP3000 encóntranse:
Implementación de alto rendemento utilizando procesamento paralelo multinodo: habitualmente, no procesamento paralelo, é necesario un sistema de comunicacións entre nodos que presente unha baixa latencia e un alto ancho de banda. Para conseguir aumenta-la velocidade de transmisión, o AP3000 utiliza unha rede de comunicación de alta velocidade denominada AP-Net, baseada nos anteriores desenvolvementos que xa se utilizaron na arquitectura do AP1000. Para conseguir un nivel de latencia reducido e un elevado ancho de banda nas comunicacións entre nodos, o AP3000 utiliza un esquema de encarreiramento de mensaxes similar ó utilizado no AP1000. Para que o nivel de latencia sexa baixo, é importante non só aumenta-la velocidade de transmisión de datos na rede, senón tamén reducir de forma significativa o tempo necesario para establecer e configura-lo envío das mensaxes. Para iso, no AP3000 sopórtase un sistema de comunicacións a nivel de usuario de forma que a comunicación de mensaxes poida ser activada directamente sen ningún tipo de axuda por parte do sistema operativo.

Alto throughput para os programas existentes: o AP3000 utiliza estacións de traballo xa existentes como nodos, de tal forma que se poden utilizar directamente aplicacións que aínda non foron preparadas para ser utilizadas no procesamento paralelo. Para manexar programas destinados ó procesamento distribuído cunha alta velocidade, é necesario utilizar interfaces de comunicación que sexan rápidos e compatibles coas redes de área local estándar. Polo tanto, as operacións tales como o acceso a ficheiros utilizando NFS ou IP (Internet Protocol) encamíñanse a través da rede AP-Net para acelera-la velocidade de transmisión de datos.

Facilidade no control e mantemento do sistema. Os grandes sistemas (baseados en máis de 100 estacións de traballo) resultan extremadamente difíciles de administrar. Isto obriga a que os administradores de sistema deban recibir tódalas facilidades posibles que lles permitan controlar de forma simultánea o acendido dos nodos, a instalación dos nodos, a observación no monitor do estado de operación, e realizar outro tipo de tarefas relacionadas. De igual modo, débense soporta-las funcións que permitan controla-lo sistema de forma automática, de acordo con certos plans operativos prefixados.

Soporte para o aumento do número de nodos e de canles de entrada/saída: o AP3000 soporta a capacidade de ampliar, dun modo sinxelo, o cluster de estacións de traballo que o forma, así como a ampliación da súa rede de comunicación de alta velocidade. Para a rede AP-Net utilízase unha rede toroidal bidimensional con alto nivel de expansión, e a súa escalabilidade permítelle soportar desde 4 ata 1024 nodos.

Número de nodos
Desde 4 hasta 1024

Tipo de nodos	
U170, U200, U300, P250, P300	
Capacidade de memoria	
Desde 128 Mbytes hasta 2 Tbytes	
Discos duros internos	
Desde 8.4 Gbytes hasta 4.2 Tbytes	
Rede 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.-Espacificacións do AP3000

Configuración do sistema

Nodos

A figura 1 mostra a configuración hardware dos nodos. A táboa 1 mostra as especificacións do AP3000 e a táboa 2 mostra as especificacións dos nodos do AP3000.

Comunicacións

A continuación descríbese a arquitectura de comunicacións implementada no AP3000. A tarxeta MSC (message controller) encóntrase conectada a cada nodo a través dun Sbus (bus entrada/saída) para a súa conexión coa AP-Net. A tarxeta MSC inclúe controladores DMA para a transferencia de datos coa AP-Net.

A AP-Net está baseada nunha topoloxía toroidal bidimensional e consiste nun conxunto de controladores de enrutamento (RTCs) encargados de dirixi-la mensaxe. Entre as características da rede AP-Net encóntranse:

1. Alta velocidade de transmisión de datos a través do porto: 200 Mbytes/s. O método de encamiñamento é estático. As mensaxes encamíñanse primeiro na dirección do eixe X e a continuación na dirección do eixe Y.

2. Wormhole routing. O wormhole-routing divide os datos (mensaxes) que se van transmitir en pequenos anacos denominados "flits", cada un deles formado por varios bytes. Os nodos de encamiñamento transmiten as mensaxes en forma de "flits". Os "flits" da cabeceira das mensaxes determinan o camiño que debe segui-la mensaxe para chegar ó seu destino, e os datos que lle seguen envíanse polo mesmo camiño.

3. Canle de comunicacións virtual dual. O hardware AP-Net soporta camiños de comunicación virtuais, denominados canles, de forma que os datos poden ser transferidos de modo independente entre os nodos utilizando canles duais. Unha das canles duais utilízase para a comunicación IP utilizada polo sistema, mentres que a outra canle utilízase polo usuario para o procesamento paralelo do software de aplicacións. Cada canle ten camiños lóxicos de comunicación diferentes para evitar que se produza "deadlocking" na topoloxía toroidal. Polo tanto, existen un total de oito canles de comunicación lóxicas sobre un único camiño físico de comunicación.

4. Sincronización de barreiras entre nodos. Nun sistema de procesamento paralelo, a pesar de que os nodos operan de modo independente, o sistema enteiro debe permanecer sincronizado nos pasos que realiza. O AP3000 consegue a sincronización de barreiras entre nodos distribuíndo e recollendo as mensaxes de sincronización da rede. O hardware RTC encárgase de distribuír e de recolle-las mensaxes de sincronismo.

5. Reliability, Availability & Serviceability. Para poder alertar sobre posibles erros na rede, o proceso de monitoreo en SYSCNTL busca erros nos RTCs. Cando se producen erros, emítese unha mensaxe acerca do erro que se recibe na estación de control. Os RTCs chequean a qué nodos se emiten as mensaxes. En caso de que se intente enviar unha mensaxe a un nodo que se encontre fóra dos grupos definidos, ou que se reciba unha mensaxe incorrecta dunha fonte externa, o RTC informa de que se produciu un erro.

Interface de Comunicación

O MSC é o hardware utilizado para soporta-la comunicación entre nodos. Posúe un controlador de comunicacións con dúas canles para o sistema e dúas canles para o usuario. Ademais das operacións convencionais de SEND e RECV, que transfiren mensaxes a través dos buffers de envío e recepción, o MSC tamén soporta o acceso directo á memoria dos nodos remotos, así como as funcións PUT e GET, CSI (compare and swap instruction) e FOP (fetch and operation).

1. A función SEND transmite os datos da memoria local ata un nodo específico. Os datos que se transmiten escríbense no buffer de recepción de mensaxes do nodo ó que se envía a mensaxe. O buffer contrólase a través de mecanismos hardware.

2. PUT copia os datos que se encontran no nodo local ata a memoria do nodo remoto. GET copia os datos que se encontran no nodo remoto ata o nodo local. Desta forma, as funcións PUT e GET proporcionan un mecanismo de comunicación efectivo cando os datos que se van transmitir entre os nodos se encontran previamente determinados. Isto é debido a que, ó contrario do que sucede no caso de SEND e RECV, non existe a necesidade de copia-los datos no receptor.

3. As funcións CSI e FOP utilízanse para o acceso exclusivo á memoria de sistemas remotos. Estes mecanismos poden ser utilizados para o control exclusivo dunha base de datos.

Interface do usuario

O MSC posúe a capacidade de encola-las instruccións de transmisión de datos tales como PUT, GET e SEND. Esta característica permite que as peticións de transmisión de datos se procesen de modo separado, de forma que os cálculos e as transmisións pódense executar de modo simultáneo.

O controlador de comunicacións de canle dual en cada MSC está supervisado por un dispositivo de comunicacións do sistema e un dispositivo de comunicacións a nivel de usuario. O dispositivo de comunicacións a nivel de sistema está instalado no sistema operativo Solaris para permiti-la comunicación IP. A comunicación a nivel de usuario está implementada a través dunha librería de comunicacións utilizada para o acceso directo ó hardware de comunicacións do MSC.

Os usuarios poden utiliza-las características de comunicación a alta velocidade utilizando librerías de paso de mensaxes estándar como MPI e PVM.

Configuración do AP3000 no CESGA

O ordenador AP3000 instalado no CESGA está formado por 16 nodos U300, dos cales 4 posúen dous procesadores por nodo, contabilizando un total de 20 procesadores.

A capacidade de memoria do equipo instalado no CESGA é de 25 GB de memoria (128 MB por procesador), e o almacenamento en disco totaliza 89 GB (4.2 GB por nodo e un array de 25 GB no nodo 0).

A pesar de que existen 16 nodos, a configuración das colas só permite a execución de traballos paralelos de ata 12 procesadores, debido a que 4 nodos se encontran dispoñibles para aplicacións de usuario e procesos interactivos, así como compilación e edición de programas.