Sistemas Multiprocesador de Memoria Compartida Comerciales

Florentino Eduardo Gargollo Acebrás, Pablo Lorenzo Fernández, Alejandro Alonso Pajares y Andrés Fernández Bermejo

Escuela Politécnia de Ingeniería de Gijón, Campus de Viesques, Universidad de Oviedo,Asturias, España
secretariageneral@uniovi.es,
http://www.epigijon.uniovi.es/

Resumen Actualmente la informática a nivel usuario ha dejado de lado el uso de arquitecturas monoprocesadoras (salvo en los PCs "low cost" y en los llamados Netbooks), pasando a utilizar arquitecturas con chips de al menos 2 núcleos.
Este cambio producido en los últimos años en el mundo PC ya se dio en la supercomputación hace más de dos décadas con el fin de lograr alcanzar los "Grandes Retos" de la ciencia.
En consecuencia, es este campo el responsable directo del nacimiento de los multiprocesadores; los cuales se pueden dividir en dos grandes tipos: Multiprocesadores de Memoria Compartida (MMC) y Multiprocesadores de Memoria Distribuida (MMD).
A lo largo de este documento, analizaremos los MMC indicando sus características generales, así como los tres tipos de multiprocesadores más representativos (SMP, PVP y DSM), detallando para cada uno de ellos sus características y un ejemplo de su uso en un supercomputador.

Palabras clave Multiprocesador, memoria compartida, procesadores vertoriales, procesadores simétricos, memoria distribuida compartida.

1. Descripción general y características comunes de los SM de memoria compartida comerciales

A lo largo de esta sección se realizará una breve introducción a los multiprocesadores y a los distintos sistemas que éstos pueden utilizan para compartir la memoria.

1.1. ¿Qué es un multiprocesador?

Se denomina multiprocesador a un sistema que cuenta con más de un microprocesador, funcionando de modo paralelo e independiente del resto, para la ejecución de una o varias tareas, bajo el control de un único sistema operativo. Son, pues, sistemas MIMD[1], en los cuales "varias unidades funcionales realizan

[1] Multiple Instruction, Multiple Data

diferentes operaciones sobre diferentes datos"[1]. Una de las características más interesantes de estos sistemas es el uso de memoria compartida, mediante el cual todos los procesadores disponen de un espacio de direccionamiento común.

Atendiendo a la forma en la que la memoria está distribuida, se puede clasificar a los Multiprocesadores como Multiprocesadores de Memoria Compartida (MMC en adelante) y como Multiprocesadores con memoria distribuida (MMD en adelante).

1.2. Tipos de Multiprocesadores

Aunque el presente documento se centra en los MMC, es adecuado ofrecer una pequeña introducción a los MMD.

Multiprocesadores con Memoria Compartida En los MMC, la memoria se organiza en uno o varios módulos, compartidos por todos los procesadores a través de distintos tipos de interconexión (tratados más adelante), con un acceso constante. A este tipo de arquitectura se le conoce como UMA[2]. El acceso a los módulos por parte de los procesadores se realiza en paralelo, pero cada módulo sólamente puede atender una petición en cada instante de tiempo.

Multiprocesadores con Memoria Distribuida Este tipo de Multiprocesadores distribuye la memoria de manera que dentro de cada procesador posee uno o varios módulos de memoria propia y está conectado mediante una red de interconexión al resto de procesadores. De esta manera, cada procesador podrá acceder tanto a su memoria local, como a la memoria remota de cualquiera del resto de procesadores. Este tipo de arquitectura se denomina NUMA[3].

1.3. Características de los MMC

Entre las características de los MMC se encuentran:

- Tiempos de acceso a memoria uniformes, ya que todos los procesadores se encuentran igualmente comunicados con la memoria principal
- Las lecturas y escrituras de cada uno de los procesadores tienen exactamente las mismas latencias
- La programación es mucho más fácil que en los MMD, debido a que la gestión de la memoria de cada módulo es transparente para el programador.
- Al acceder simultáneamente a la memoria se producen colisiones y esperas, lo que es un problema.
- Debido a la organización de la arquitectura, es poco escalable en número de procesadores, debido a que puede surgir un cuello de botella si se aumenta el número de CPU's.

[2] Uniform Memory Access
[3] Non-Uniform Memory Access

2. SMP (Multiprocesador simétrico)

Los sistemas SMP utilizan una modalidad de procesamiento en paralelo en la que "todos los procesadores son tratados como iguales"[2]. Los SMP están basados en el modelo de memoria de acceso uniforme, en donde la memoria física está uniformemente compartida por todos los procesadores, lo cual implica que todos ellos posean los mismos tiempos de acceso a todas las palabras de memoria. Estos sistemas poseen una red de interconexión entre los distintos procesadores y la memoria, habitualmente en forma de bus, aunque también pueden tener otro tipo de sistema de interconexión. Cabe destacar que en los sistemas SMP todos los procesadores tienen el mismo acceso a los periféricos, lo cual implica que todos los procesadores tienen la misma capacidad para ejecutar programas tal como el Kernel o las rutinas de entrada y salida. Otra característica importante es que todo el sistema está controlado por un mismo sistema operativo que posibilita la cooperación entre los procesadores y sus programas, debido a este alto nivel de cooperación se clasifican como sistemas fuertemente acoplados. La arquitectura de estos multiprocesadores sigue un esquema como el mostrado en la figura 1, en la página 3.

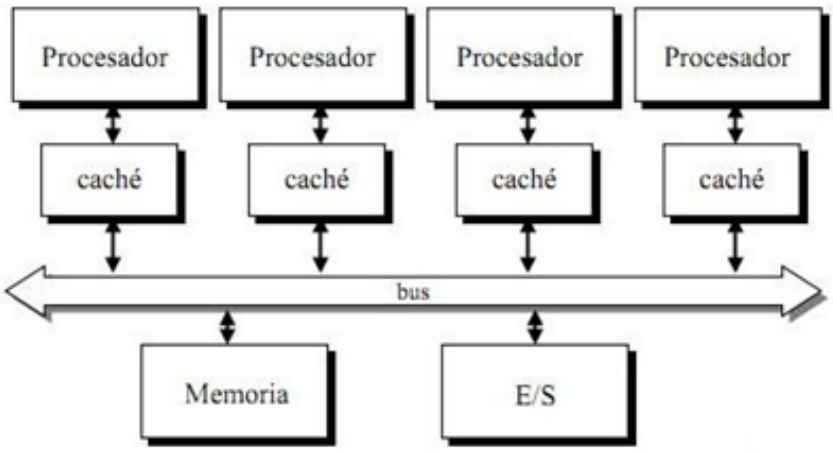

Figura 1. Arquitectura de memoria SMP

El principal problema que se presenta en los SMP en cuanto a su arquitectura de memoria es la coherencia de caché, ya que los distintos procesadores pueden mantener su propia caché, por lo tanto si uno de ellos modifica un dato de la memoria compartida otros pueden mantener copias en su propia cache incoherentes. Para solucionar este problema hay diferentes estrategias hardware y software, aunque este aspecto merecería un estudio en profundidad queda fuera

del alcance de este documento. Los SMP no sólo se encuentran en entornos de supercomputación, si no que los PCs con procesadores multinúcleo aplican esta arquitectura, tratando a cada núcleo como un procesador independiente.

2.1. Ejemplos

Existen actualmente muchos ejemplos de multiprocesadores simétricos de memoria compartida en el mercado. En la tabla 1, página 4 se muestran algunas de las características y precios de algunos de modelos reales del mercado:

Procesador	**Núcleos**	**Frecuencia**	**Caché**	**Memoria**	**Precio**	**Año**
SPARC T3	16	1.65 GHz	6 MB L2	128 GB DDR3	18600$	2010
Intel Xeon W3690	6	3.46 GHz	12 MB L3	24 GB DDR3	999$	2011
MIPS32 1074k	4	1.5 GHz	32+32KB L1	-	-	2010
ARM Cortex A15	4	2 GHz	-	-	-	-
AMD Opteron 6100	12	2.5GHz	12MB L3	-	1500$	2010

Cuadro 1. Tabla comparativa de modelos SMP en el mercado. Nótese que el precio del SPARC T3 incluye el rack con todos los accesorios, discos duros, memoria, etc.

2.2. Arquitectura de ejemplo: Supercomputador Cray Jaguar XT5-HE

El Centro Nacional de Ciencias de la Computación (NCCS) del Laboratorio Nacional de Oak Ridge (Tennessee, Estados Unidos) posee uno de los supercomputadores incluidos en la lista Top500, en la que figuran los computadores de mayores capacidades del mundo. Se trata del Cray Jaguar XT5-HE, que se puso en funcionamiento en 2009 y estuvo situado en el primer puesto de la lista Top500 durante el segundo semestre de 2009 y el primero de 2010. Actualmente se encuentra en el segundo puesto de la lista. El Jaguar XT5-HE es un supercomputador que está compuesto por dos particiones, la XT4 y la XT5. La partición XT5 se compone de 18688 núcleos de computación con 2 procesadores AMD Opteron de 6 núcleos (2,6GHz) cada uno y 16GB de memoria, lo que hace un total de 224256 cores, con una memoria total de 300TB (DDR2 800) y un rendimiento que se sitúa en los 2,3 petaflopss. La partición XT4 está formada por 7832 nodos de computación, cada uno de ellos con un procesador AMD Opteron quad core (2,1GHz) y 8GB de memoria, lo que hace un total de 31238 cores y 62TB de memoria RAM. El Jaguar XT5-HE está disponible para su utilización (las dos particiones) en el campo de la investigación por parte de cualquier organización, previo registro en el NCCS y solicitud de uso. Además, el Oak Ridge National Laboratory ofrece visitas a sus instalaciones.

Procesador utilizado: AMD Opteron Un ejemplo de sistemas multiprocesador de memoria compartida que siguen el esquema SMP son los procesadores de la serie Opteron de AMD. En la figura 2, página 5 se presenta, a modo de ejemplo, un esquema de la arquitectura de un procesador AMD Opteron Quad Core:

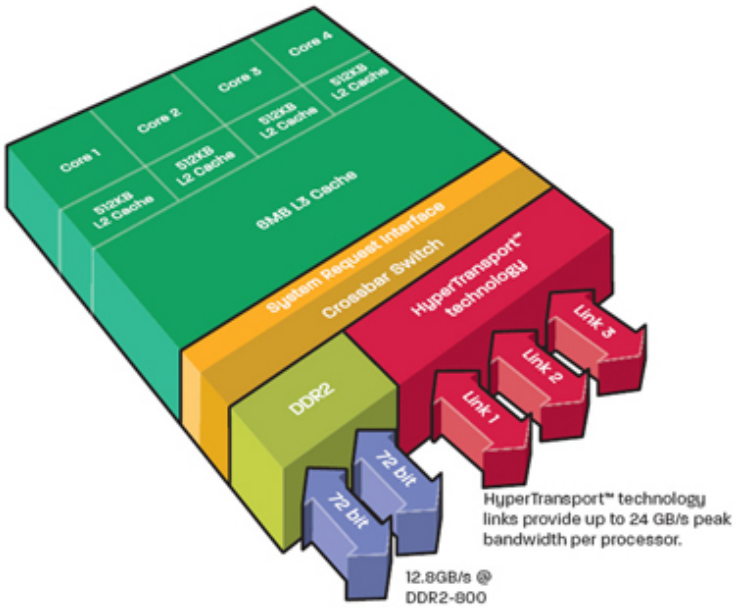

Figura 2. Arquitectura del procesador AMD Opteron Quad Core

Las características principales a nivel de arquitectura de estos procesadores son:

- 4 núcleos independientes (con frecuencia variable e independiente en cada núcleo).
- Caché L1 + L2 (512 KB) independientes en cada núcleo.
- Caché L3 de 6 MB compartida entre los 4 núcleos.
- Controladora de memoria DDR2 independiente del bus principal para un acceso más rápido a la memoria principal.
- Sistema Hyper Transport que sustituye a los buses antiguos y permite la conexión de hasta 3 dispositivos con un mayor ancho de banda (hasta 8GB/s).

Se puede observar que aunque la memoria caché de nivel 1 y 2 son independientes en cada procesador, la memoria caché de nivel 3 ya aparece como compartida, como una manera de reducir la latencia de la memoria compartida principal.

3. PVP: Procesador Vectorial Paralelo

Los procesadores vectoriales cuentan con una micro-arquitectura orientada al procesamiento de vectores, además de un repertorio de instrucciones máquina que implementan operaciones en donde tanto los operandos como el resultado son vectores.

Esto permite que los procesadores vectoriales apliquen una instrucción sobre un conjunto de datos, vector, en vez de aplicarla sobre un único dato, por lo tanto una única instrucción puede representar una importante carga de trabajo, lo que reduce el requisito de ancho de banda para instrucciones. En procesadores no vectoriales la búsqueda y decodificación de instrucciones representa a menudo un cuello de botella, denominado cuello de botella de Flynn.

El paralelismo viene de la independencia de los elementos de las matrices entre sí, aunque esto no ocurre siempre si es habitual, por lo que las operaciones de unas matrices con otras pueden realizarse en paralelo, o al menos en el mismo cauce de instrucciones sin que se produzca un conflicto entre los datos.

No todas las tareas son susceptibles de este tipo de procesamiento, pero en general, esta arquitectura resulta especialmente indicada entre otros para problemas físicos y matemáticos, los cuales en su mayoría se pueden expresar fácilmente mediante matrices.

Las máquinas multiprocesadoras que utilizan procesadores vectoriales pueden ser vistas como un caso especial de las máquinas SMP, especialmente en cuanto al sistema de memoria compartida. Esta combinación de procesadores vectoriales, de gran potencia y hechos a medida, con arquitecturas SMP, hacen que los sistemas PVP consigan grandes prestaciones para aplicaciones científicas y matemáticas.

Un problema de los sistemas PVP es la escalabilidad, debido a las colisiones en la red de interconexión, el número de procesadores que pueden componer el sistema está limitado. Como solución para poder escalar estos sistemas se suelen usar dos niveles, un nivel más profundo de multiprocesadores vectoriales, con memoria compartida, y un nivel más alto que representa una estructura de multicomputadores, lo cual implica memoria distribuida.

3.1. Ejemplos

Los PVP están diseñados para operar con grandes vectores de datos al mismo tiempo. Su compilador realiza una vectorización de los bucles de código para partir el trabajo en bloques y de esta manera operar a la vez distintas partes del problema.

Debido a su coste, en la actualidad no existen demasiadas implementaciones de este tipo de procesadores en computadores. Las principales prestaciones de los supercomputadores que operan con procesadores vectoriales son las mostradas en la tabla 2, página 7.

Computador	Procesadores por nodo	Máximo de nodos	Tamaño máximo de memoria compartida (nodo)	Rendimiento de pico del procesador	Rendimiento de pico del nodo
NEC SX-6	8	-	64Gb	8 GFLOPS	64GFLOPS
NEC SX-8	8	64	128Gb	16GFLOPS	128GFLOPS
NEC SX-9	16	512	1TB	16GFLOPS	>100GFLOPS
Cray SV1	32	-	-	-	-

Cuadro 2. Tabla comparativa de modelos PVP en el mercado.

3.2. Arquitectura de ejemplo: Earth Simulator

El Earth Simulador es un supercomputador de la Agencia Japonesa para la Tecnología y Ciencia Marina y Terrestre. Es un computador NEC SX-9/E/1280M160, que utiliza 5112 procesadores NEC a 3200Mhz organizados mediante la técnica NEC Vector de multiprocesamiento paralelo. El esquema de conexión de los distintos procesadores es el mostrado en la figura 3 en la página 7

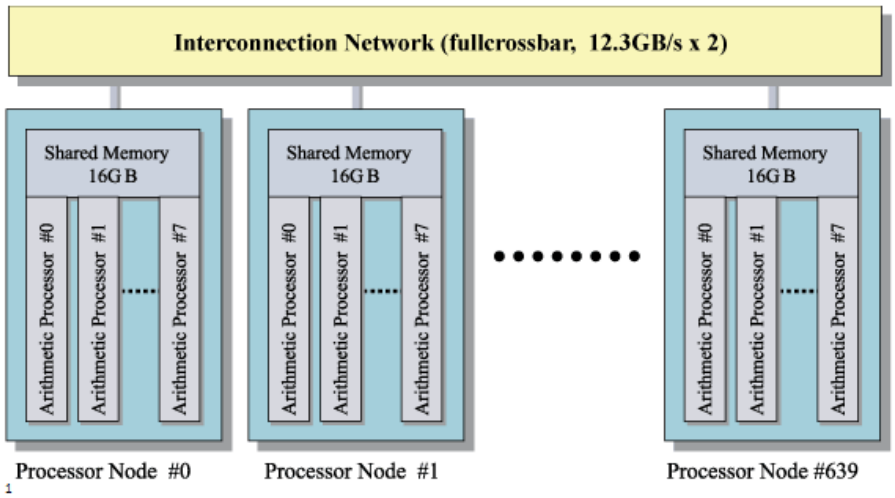

Figura 3. Arquitectura del supercomputador Earth Simulator[3]

Los procesadores se organizan nodos de procesamiento, conteniendo cada uno 8 procesadores vectoriales que comparten 16Gb de memoria. Los nodos de procesamiento están unidos mediante una red de interconexión conmutada cruzada.

Las características principales del computador, así como las tasas de rendimiento de los procesadores y los nodos son los mostrados en la tabla 3 en la página 8

Rendimiento de pico del procesador aritmético	8Gflops	**Número total de procesadores aritmétivos**	5120
Rendimiento de pico del nodo de procesamiento	64Gflops	**Número total de nodos de procesamiento**	640
Memoria compartida de cada nodo de procesamiento	16Gb	**Rendimiento de pico total**	40Tflops
		Memoria principal total	10TB

Cuadro 3. Tabla de características del supercomputador Earth Simulator

El área de trabajo del supercomputador es la simulación del sistema terrestre completo mediante el procesamiento de modelos de datos de enorme tamaño sobre el presente y el pasado, para ayudar a predecir situaciones futuras en la tierra. Algunas aplicaciones realizadas son estudios sobre el cambio climático, el análisis de coexistencia entre el humano y el resto del medio ambiente y la simulación de condiciones atmosféricas. El supercomputador es utilizado tanto por el gobierno, como por distintas empresas que pagan por su tiempo de uso. También dedica tiempo de uso para otros proyectos de investigación que deben ser anteriormente aprobados por la comisión del Earth Simulator. Estos últimos proyectos constituyen una gran parte del trabajo del computador, y se componen de varias decenas de nuevos proyectos anuales.

4. DSM: Memoria compartida distribuida

Como hemos visto anteriormente, los sistemas de memoria compartida tradicionales SMP, utilizan un mismo espacio de memoria compartido entre todos los procesadores. La comunicación entre la memoria y los procesadores generalmente se realiza mediante un bus, el cual puede llegar a suponer un cuello de botella en el acceso a memoria si el número de procesadores es suficientemente alto. Las arquitecturas NUMA (Non-Uniform Memory Architecture) intentan aliviar este cuello de botella, acercando parte de la memoria a cada procesador, aunque esto deriva en que el acceso a la memoria remota es más lento que el

acceso a la memoria local. Desafortunadamente, tanto los sistemas SMP como NUMA tienen un precio elevado, lo que motiva el uso de los sistemas distribuidos, relativamente más baratos. No obstante, todas las comunicaciones y sincronizaciones deben hacerse mediante el paso de mensajes, ya que cada sistema tiene su propia memoria local, separada del resto. En general resulta más fácil la programación para sistemas con un solo procesador o sistemas multiprocesadores con un bloque de memoria compartida, que mediante el paso de mensajes, de ahí el nacimiento de los sistemas DSM (Distributed Shared Memory), que no es más que una técnica para simular un espacio común de direcciones entre multicomputadores. Existen varias alternativas para conseguir simular este espacio común de direcciones a partir de la arquitectura de memoria distribuida, por ejemplo mediante el uso de caches, más rápido y caro, mediante el uso de memoria virtual con modificaciones en el software, más lento y barato, o también mediante soluciones hibridas entre hardware y software. Como es de esperar los sistemas DSM también plantean varios problemas que hay que tener en cuenta y para los cuales se proponen distintas soluciones, una vez más el problema de la coherencia es uno de los principales focos de conflicto, aunque su estudio queda fuera del alcance de este documento.

4.1. Ejemplos

Las principales prestaciones de los supercomputadores DSM son las mostradas en la tabla 4 en la página 10.

4.2. Arquitectura de ejemplo: Pleiades

El Pleiades es un supercomputador de la División de Supercomputación Avanzada de la NASA el cual se encuentra situado en el campus de investigación Ames, en California[7]. Pleiades, actualmente en la posición 11 del TOP500, soporta una variedad de proyectos científicos y de ingeniería, incluyendo modelos para evaluar la predicción del clima decenal para el Grupo Intergubernamental de Expertos sobre el Cambio Climático; el diseño de futuros vehículos espaciales; modelos detallados de los halos de materia oscura y la evolución de las galaxias. Este supercomputador se trata de un SGI Altix ICE 8200EX, que utiliza 21504 procesadores Intel EM64T Xeon organizados en 10752 nodos de 2 procesadores por nodo. De estos 10752, 3584 son nodos con procesadores Xeon X5670 (Westmere) de 6 núcleos y 1280 con procesadores Xeon X5570 (Nehalem) de 4 núcleos; en ambos casos trabajan a 2.93GHz. El esquema de los nodos “Westmere” es el mostrado en la figura 4 de la página 11 (Para los “Nehalem” el esquema de conexión es el mismo pero con 4 cores).

Los procesadores se comunican mediante la Tecnología de Interconexión de Intel llamada QuickPath. Esta tecnología permite comunicar procesadores con controladores de memoria integrados. Así mismo, los nodos se comunican entre si utilizando para ello InfiniBand (bus de comunicaciones serie de alta velocidad, diseñado tanto para conexiones internas como externas) dado que todos los nodos están conectados siguiendo una topología de hipercubo de 11 dimensiones

Computador	**Número de procesadores**	**Tamaño máximo de memoria compartida**	**Cache**	**Rendimiento de pico teórico del procesador**	**Rendimiento de pico teórico del computador**
SGI ORIGIN 3800	16 a 512	2GB a 1TB	Cada procesador: 8MB cache L2, 32 KB cache L1 de datos, 32 KB cache L1 de instrucciones	800 Megaflops	400 Gigaflops
SGI Altix 4700	4864*2cores	39 TB	32 KB de cache L1 , 1 MB of L2 instruction cache, 256 KB de cache L2 de datos y 9 MB de cache L3 por cada núcleo	12.8 Gflop/s	62.3 TFlops
Cray X1[4]	4096	16GB por nodo	2MB "Ecache" compartidos entre cada cuatro procesadores	12.8 Gflop/s	-
ASURA[5]	1024 procesadores en 128 clusters	256Mb por nodo, 32GB en el sistema completo	Cache compartida, dividida en varios niveles jerárquicos (no especificados)	9.6 Gflop/s	-

Cuadro 4. Tabla comparativa de prestaciones de distintos supercomputadores DSM

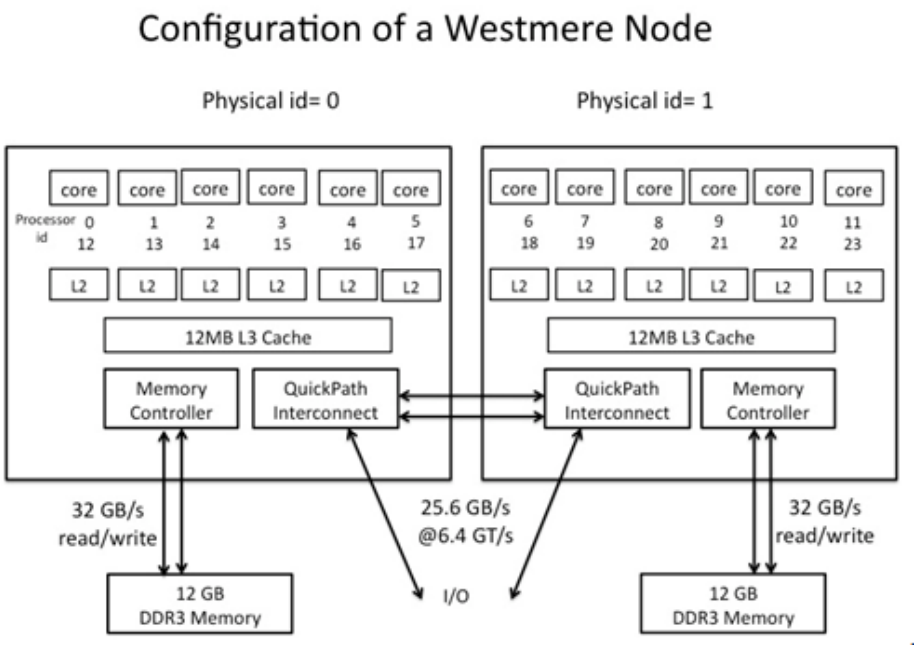

Figura 4. Arquitectura de Westmere[6]

parcial. En cuanto a las características principales del Pleiades, así como las tasas de rendimiento de los procesadores y los nodos son los mostrados en la tabla 5 de la página 11.

Numero de Racks (Bastidores)	168
Número Total de Nodos	10752
Número Total de Núcleos	100352
Memoria Principal	164 TB
Rendimiento máximo con High-Performance LINPACK	772.7 Tflops/s
Rendimiento máximo Teórico	973.291 Tflops/s

Cuadro 5. Tabla con las características de Pleiades

Referencias

1. Free Online Dictionary of Computing, "Multiple Instruction/Multiple Data" `http://foldoc.org/MIMD`
2. Francisco Armando, Dueñas Rodríguez, "Symmetric Multiprocessing (SMP)" Universidad La Salle. Cancún. `http://www.monografias.com/trabajos6/symu/symu.shtml`

3. Earth Simulator Center Architecture, http://www.jamstec.go.jp/es/en/images/system_b.gif
4. Thomas H. Dunigan, Jr., Jeffrey S. Vetter, James B. White III, Patrick H. Worley Oak Ridge National Laboratory, "Performance Evaluation of the Cray X1 Distributed Shared Memory Architecture".
5. Mori, Saito, Goshima, Yanagihara, Tanaka, Fraser, Joe, Nitta, Tomita, .A distributed shared memory multiprocessor: "ASURA - Memory and cache architectures" sc, pp.740-749, Proceedings of the 1993 ACM/IEEE conference on Supercomputing, 1993
6. NASA, "Westmere's architecture", http://www.nas.nasa.gov/Users/Documentation/Ice/hardware_pleiades.html
7. NASA, "Pleiades", http://www.nas.nasa.gov/Resources/Systems/pleiades.html