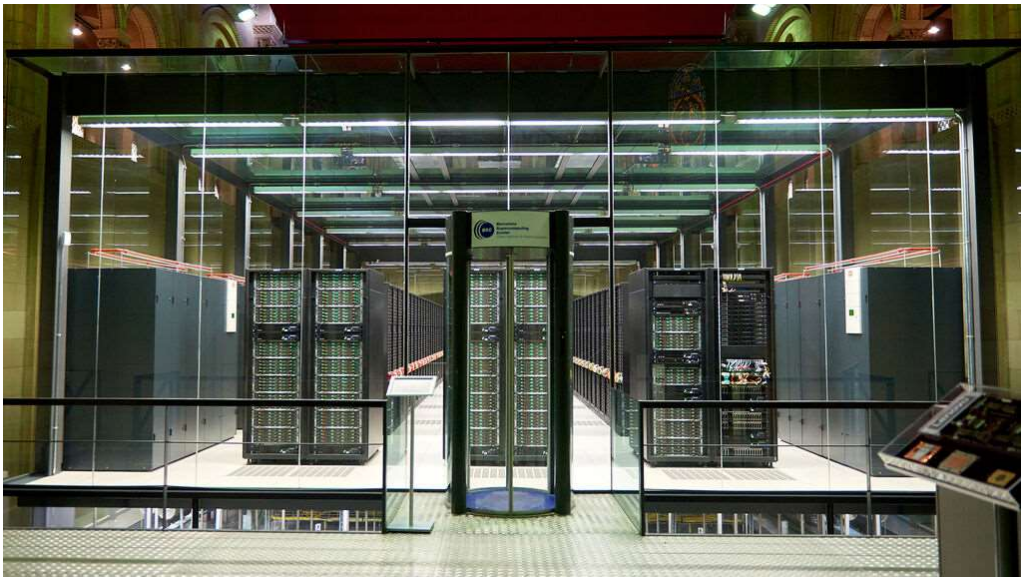
UNIVERSIDAD
DE BURGOS

Mare Nostrum 4

Arquitectura Avanzada de Computadores



Eduardo Mora González

ÍNDICE

1. Introducción

"*Todo el mundo puede construir una CPU rápida. El truco está en construir un sistema rápido*". Así resumía *Seymour Cray*, el conocido como padre de la supercomputación, el reto de desarrollar las computadoras más potentes del mundo. Este reto ha implicado a muchos países a obtener este supercomputador y ponerse en cabeza a nivel mundial, pero antes de entrar en más detalle debemos definir algunos conceptos.

Un **Supercomputador** es un ordenador con capacidades de cálculo muy superiores a las comunes y están orientadas a fines específicos. La mayoría de los supercomputadores se componen de unidades menos potentes, pero trabajando de forma conjunta con un objetivo común, aumentando tanto la potencia del conjunto como su rendimiento **[1]**.

Como podemos ver en la definición, los supercomputadores son varias máquinas trabajando de forma conjunta a la vez, de lo que sale el concepto de computación paralela.

La **computación paralela** se basa en la teoría de *"Dividir los problemas grandes en varios pequeños y solucionarlos simultáneamente*" esto permite ejecutar más instrucciones en menos tiempo **[2]**.

Antes hemos mencionado una frase de *Seymour Cray* y esta persona creó la primera supercomputadora de la historia: El ***CDC 6600*** diseñada en 1965 y fabricada por *Control Data Corporation*.

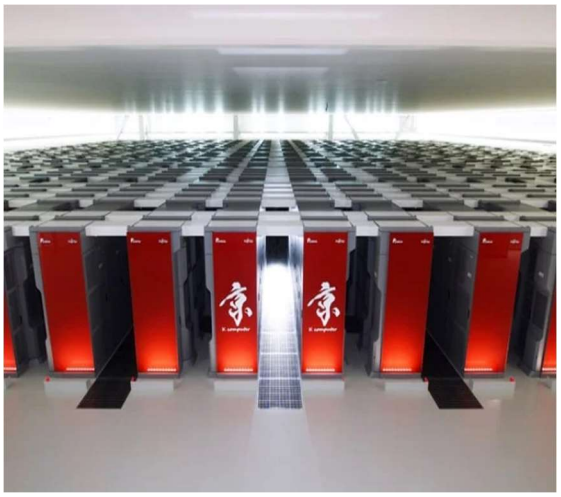
El uso de esta supercomputadora fue principalmente para la investigación de la física de alta energía nuclear.

El ***CDC 6600*** posee una CPU de 60 bits y 10 unidades periféricas de procesamiento y se utiliza un marcador para el *plotting* de las órdenes **[3]**.

Actualmente, El superordenador japonés ***Fugaku*** lidera el ranking de los más potentes del mundo, este encuentra en el Centro de Ciencia Computacional de Japón.

La supercomputadora ***Fugaku*** está construida con el microprocesador *Fujitsu* A64FX. Esta CPU se basa en la arquitectura ARM versión 8.2A y adopta las extensiones vectoriales escalables para supercomputadoras.

Fugaku utiliza un «sistema operativo ligero multinúcleos» llamado *IHK/McKernel*. El sistema operativo utiliza tanto Linux y el núcleo ligero *McKernel* funcionando simultáneamente, lado a lado.



La infraestructura en la que se ejecutan ambos núcleos se denomina Interfaz para núcleos heterogéneos. Las simulaciones de alto rendimiento se ejecutan en *McKernel*, con *Linux* disponible para todos los demás servicios compatibles con *POSIX* **[4]**.

En España existe la **Red Española de Supercomputación (RES)** que es una infraestructura distribuida que consiste en la interconexión de 12 supercomputadores con el objetivo de ofrecer recursos de computación de alto rendimiento a la comunidad científica. La RES está coordinada por el **Barcelona Supercomputing Center (BSC)**.

Los supercompudares que forman la RES son:

- ***MARENOSTRUM & MINOTAURO*** en Barcelona Supercomputing center.
- ***FINISTERRAE II*** en el Centro de Supercomputación de Galicia.
- ***LAPALMA*** en el Instituto de Astrofísica de Canarias.
- ***ALTAMIRA*** en el Instituto de Física de Cantabria.
- ***PICASSO*** en la Universidad de Málaga.
- ***TIRANT*** en la Universidad de Valencia.
- **CAESARAUGUSTA** en el Instituto de Biocomputación y Física de Sistemas Complejos.

- ***CALÉNDULA*** en la Fundación Centro de Supercomputación de Castilla y León.
- ***PIRINEUS*** en *Consorci de Serveis Universitari* de Catalunya.
- ***LUSITANIA*** en *Cénits*.
- ***CIBELES*** en la universidad Autónoma de Madrid.

En el desarrollo de este trabajo se hablará del ***MareNostrum 4*** el cual ha sido denominado el supercomputador más diverso del mundo por la heterogeneidad de su arquitectura **[6]**.

2. MareNostrum 4

MareNostrum es el nombre genérico que utiliza el BSC para referirse a las diferentes actualizaciones de su supercomputador más emblemático y el más potente de España. Hasta el momento se han instalado cuatro versiones.

En marzo del 2004, el Gobierno español e *IBM* firmaron un acuerdo para construir uno de los computadores más rápidos de Europa el ***MareNostrum* 1** cuya potencia de cálculo era de 42,35 Teraflops (42,35 billones de operaciones por segundo).

En noviembre de 2006 su capacidad se incrementó, debido a la gran demanda por parte de los proyectos científicos. La capacidad de cálculo de ***MareNostrum* 2** se aumentó a 94,21 Teraflops, lo que supone el doble de su capacidad anterior. Paso de tener 4.812 procesadores a tener 10.240.

Con la actualización de 2012-2013, ***MareNostrum* 3** obtuvo un rendimiento máximo de 1,1 Petaflops. Tenía 48.896 procesadores Intel Sandy Bridge en 3056 nodos, incluyendo 84 Xeon Phi 5110P en 42 nodos, más de 115 TB de memoria principal y 2 PB de almacenamiento en disco GPFS [7].

A finales de junio de 2017 entró en operación el ***MareNostrum* 4** el cual hablaremos a continuación en detalle.

2.1. Arquitectura

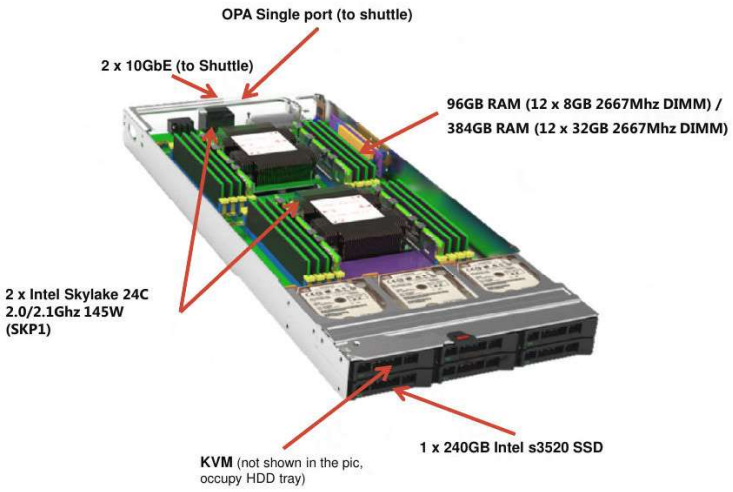
MareNostrum 4 es una supercomputadora basada en procesadores *Intel Xeon Platinum* de la generación *Skylake*. Es un sistema Lenovo compuesto por Racks de Computación SD530, un Intel Omni-Path de alto rendimiento interconexión de red y ejecutando SuSE Linux Enterprise Server como sistema operativo.

Su rendimiento *Linpack Rmax* actual es 6.2272 operaciones de Peta. Este bloque de uso general consta de 48 racks que albergan 3456 nodos con un total de 165,888 núcleos de procesador y 390 Terabytes de memoria principal. Los nodos de computación están equipados con:

- 2 sockets CPU Intel Xeon Platinum 8160 con 24 núcleos cada uno a 2,10 GHz para un total de 48 núcleos por nodo.
- L1d 32K; Caché L1i 32K; Caché L2 1024K; Caché L3 33792K.
- 96 GB de memoria principal 1.880 GB / núcleo, 12x 8GB 2667Mhz DIMM (216 nodos de memoria alta,10368 núcleos con 7,928 GB / núcleo), algunos nodos disponen de 384 GB de memoria principal en vez de 96 GB.
- Adaptador PCI-E Intel Omni-Path HFI Silicon 100 Series de 100 Gbit / s.
- Ethernet de 10 Gbit.
- SSD local de 240 GB disponible como almacenamiento temporal durante los trabajos.

Los procesadores admiten instrucciones de vectorización conocidas como SSE, AVX hasta AVX 5121 **[8]**.

En la imagen se puede apreciar la distribución de todos los elementos que contienen los nodos de cómputo **[9]**.



2.2. Interconexión

Los 3,456 nodos de cómputo están interconectados a través de una red de alta velocidad: Intel *Omni-Path* (OPA). Los diferentes nodos están conectados vía cables de fibra óptica y switches *Intel Omni-Path Director Class*.

Seis racks en ***MareNostrum*** están dedicados a elementos de la red, los cuales permiten la conexión entre los diferentes nodos gracias a la red OPA. Las principales características de un switch *Omni-Path Director Class* son **[9]**:

- Hasta 786 x 100GbE puertos en 20U (+1U Shelf)
- 12 x *hot swap PSUs* (N+N)
- *Hot swap fan modules*
- 2 x *Management modules*
- 8 x *Double spine modules (non-blocking)*
- Hasta 24 x 32 *port leaf modules* (19 occupied – 608 ports)
- Cada leaf module contiene 2 ASICs
- Consumo energético de 9.4kW

2.3. Almacenamiento

MareNostrum 4 dispone de una capacidad de almacenamiento en disco de 14 *Petabytes* y está conectado a las infraestructuras Big Data del BSC-CNS que tienen una capacidad total de 24,6 *Petabytes*. Todos sus componentes están conectados entre ellos a través de una red de alta velocidad *Omnipath*. Como sus antecesores, MareNostrum 4 también está conectado a los centros de investigación y universidades europeas a través de las redes *RedIris* y *Geant*.

Además, cada nodo tiene un disco duro de estado sólido local que se puede utilizar como espacio temporal local para almacenar archivos temporales durante la ejecución de uno de sus trabajos. La cantidad de espacio dentro del sistema de archivos es de aproximadamente 200 GB. Todos los datos almacenados en estos discos duros locales en los nodos informáticos no estarán disponibles en los nodos de inicio de sesión y estos se limpiarán automáticamente una vez que finalice el trabajo **[8]**.

2.4. Herramientas de administración y gestión de trabajos

Slurm es la utilidad utilizada para el soporte de procesamiento por lotes, por lo que todos los trabajos deben ejecutarse a través de ella.

Algunos aspectos importantes del sistema son:

- Todos los trabajos que soliciten 48 o más núcleos utilizarán automáticamente todos los nodos solicitados en modo exclusivo.
- La cantidad máxima de trabajos en cola (en ejecución o no) es 366.

Existen varias colas presentes en las máquinas y diferentes usuarios pueden acceder a diferentes colas. Todas las colas tienen diferentes límites en la cantidad de núcleos para los trabajos y la duración. La configuración estándar y los límites de las colas son los siguientes:

Queue	Maximum numer of nodes (cores)	Maximum wallclock
Debug	16 (768)	2 h
Interactive	(max 4 cores)	2 h
BSC	50 (2400)	48 h
RES Class A	200 (9600)	72 h
RES Class B	200 (9600)	48 h
RES Class C	21 (1008)	24 h
PRACE	400 (19200)	72 h

Slurm informará el estado de los trabajos iniciados. Si todavía están esperando para entrar en ejecución, serán seguidos por el motivo. ***Slurm*** usa códigos para mostrar esta información, los más relevantes son **[8]**:

- **COMPLETED (CD):** El trabajo ha completado la ejecución.
- **COMPLETANDO (CG):** El trabajo está terminando, pero algunos procesos aún están activos.
- **FAILED (F):** el trabajo terminó con un código de salida distinto de cero.
- **PENDIENTE (PD):** el trabajo está esperando la asignación de recursos. El estado más común después de ejecutar "*sbatch*", se ejecutará eventualmente.
- **PREEMPTED (PR):** el trabajo se canceló debido a que otro trabajo lo reemplazó.

- **EN EJECUCIÓN (R):** el trabajo está asignado y en ejecución.
- **SUSPENDIDO (S):** un trabajo en ejecución se ha detenido con sus núcleos liberados a otros trabajos.
- **DETENIDO (ST):** un trabajo en ejecución se ha detenido con sus núcleos retenidos.

2.5. Eficiencia energética

"Nosotros estamos consumiendo para todas las instalaciones 1,7 megavatios o, lo que es lo mismo, 1,5 millones de euros. Si ahora vamos a pasar a consumir 12 megavatios la multiplicación es fácil" **[10]**, esto decían el BSC para concienciar de que debían buscar una forma más eficiente energéticamente hablando para el ***MareNostrum 4***.

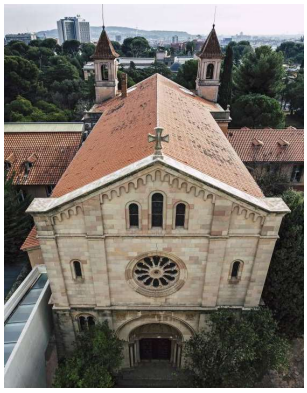
"La nueva maquinaria del supercomputador MareNostrum-4 de Barcelona, que ha incorporado una nueva tecnología para hacerlo más potente, ha sido calificado como el más "verde" de Europa, según la lista 'Green 500', que mide la eficiencia energética de los superordenadores más potentes del mundo" **[11]**, con este titular de periódico podemos comprobar que el propósito mencionado anteriormente se ha cumplido ya que se ha creado el ordenador con más eficiencia energética de Europa.

Esta eficiencia energética se debe a que la máquina cuenta con un bloque de propósito general (que tiene 48 racks con más de 3.400 nodos equipados con chips Intel Xeon) y una memoria central de 390 Terabytes. Su potencia pico será de más de 11 Petaflops/s, o lo que es lo mismo, será capaz de realizar más de 11.000 billones de operaciones por segundo, diez veces más que el ***MareNostrum 3***. Aunque su potencia será diez veces mayor que la de su antecesor, su consumo energético solamente aumentará un 30% y pasará a ser de 1,3 MWatt/año **[12]**.

2.6. Instalaciones de refrigeración

La refrigeración del ***MareNostrum 4*** combina el enfriamiento por aire con el uso de agua fría en la parte trasera de los racks, las torres o armarios que albergan las placas base.

Pero crear una instalación de refrigeración no fue una tarea fácil. Las peculiaridades del edificio (una capilla) les obligaron a pensar en soluciones creativas, no solo para meter los armarios con los procesadores por la única puerta de madera que hay disponible, sino para ubicar los sistemas de refrigeración y los transformadores eléctricos necesarios para que la máquina funcione.



La temperatura dentro del espacio acristalado tiene que mantenerse en torno a los 24 grados, pero hace falta mucho más que la refrigeración por aire para que eso sea posible se tuvo que instalar el sistema de refrigeración por la puerta trasera, eso significa que unas tuberías llevan agua fría hasta la parte de atrás de los armarios, para que puedan eliminar el calor que expulsan.

Toda la maquinaria que mantiene el ordenador en buenas condiciones (los equipos que generan el agua fría, los transformadores, los generadores eléctricos...) está fuera, bajo el suelo y en una caseta insonorizada. En el caso de que se vaya la luz, unas baterías se conectan automáticamente **[13].**

3. Conclusiones

La realización de este trabajo me ha resultado muy interesante al descubrir al ***MareNostrum 4*** un supercomputador desconocido para mí a nivel técnico, ya que había oído hablar de el pero nunca me he parado a estudiarlo a fondo.

Además, este trabajo me ha llevado a la curiosidad de indagar sobre otros supercomputadores y las diferencias que tienen con este estudiado.

Finalmente, viendo la evolución de la informática, cada día me sorprendo mas de las cosas que se pueden hacer con los computadores y los avances que estos beneficiaran a la humanidad.

Referencias

1. http://www.cenits.es/faq/preguntas-generales/que-es-un-supercomputador
2. https://arquitecturadecomputadora.wordpress.com/2013/06/07/computacion-paralela/#:~:text=Es%20una%20t%C3%A9cnica%20de%20programaci%C3%B3n,Paralelismo%20a%20nivel%20de%20bit
3. http://agusjejeje.blogspot.com/2012/08/cdc-6600.html
4. https://retina.elpais.com/retina/2020/06/22/innovacion/1592824062_713254.html
5. https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.geeky-gadgets.com%2Ffugaku-supercomputer-14-05-2020%2F&psig=AOvVaw3xxXxIi9WgvPMhO4zJdAMY&ust=1600764071761000&source=images&cd=vfe&ved=0CAIQjRxqFwoTCOio2IDt-esCFQAAAAAdAAAAABAE
6. https://www.res.es/es
7. https://www.bsc.es/es/marenostrum/marenostrum#:~:text=MareNostrum%204&text=Son%20los%20siguientes%3A,de%20Energ%C3%ADa%20de%20los%20EE.
8. https://www.bsc.es/user-support/mn4.php
9. https://www.bsc.es/es/marenostrum/marenostrum/informacion-tecnica
10. https://retina.elpais.com/retina/2019/01/31/tendencias/1548924916_561984.html
11. https://www.elperiodico.com/es/sociedad/20180625/el-supercomputador-marenostrum-4-considerado-el-mas-verde-de-europa-6906263
12. https://www.datacentermarket.es/proyectos/noticias/1094141032709/el-supercomputador-marenostrum-4-sera-12-veces-mas-potente-que-marenostrum-3.1.html
13. https://netlogyc.com/vivamus-laoreet-turpis-leo/