MareNostrum4 es la denominación para la nueva infraestructura de supercomputación del Barcelona Supercomputing Center, MareNostrum4 (abreviado como MN4) estará formado por los siguientes componentes:

- Almacenamiento centralizado (ya licitado en el concurso CONSU2016008OP Lote 1)
- Clusters de cómputo

Dentro de los clusters de cómputo, se pedirán 2 tipos de clusters, los cuales han de presentar tecnologías complementarias:

- Cluster de cómputo de propósito general (abreviado como CPG)
- Clusters de cómputo con tecnologías emergentes (abreviado como CTE)

Estos componentes combinados servirán para la ejecución óptima de los diversos códigos científicos de supercomputación. El cluster de propósito general se encargará de la ejecución de la gran mayoría de aplicaciones científicas. Mientras que los clusters de tecnologías emergentes (emerging technologies) permitirán la ejecución de ciertas aplicaciones de producción de supercomputación como aplicaciones de cognitive computing, Deep Learning y Big Data; como la valoración de nuevas arquitecturas para la instalación en 2019-2020 de una actualización tecnológica del superordenador principal del BSC-CNS.

MareNostrum4 debe ser el sistema de supercomputación que sustituya a MareNostrum3 adquirido por el BSC, siendo una de sus funciones principales proporcionar servicio a los investigadores científicos europeos y españoles, a través de los recursos aportados a PRACE (http://www.prace-ri.eu/) y la RES (http://www.res.es).

Este pliego técnico establece los requerimientos y puntos de mejora para la adquisición de los clusters de cómputo de MareNostrum4 para el BSC-CNS.

Se establecerá el 30 de Junio de 2020 como fin de proyecto de MareNostrum4.
Cualquier referencia a GPFS se entenderá también como Spectrum Scale, siendo este el nuevo nombre para la misma tecnología.

Clusters cómputo MareNostrum4

A nivel de clusters de cómputo, MareNostrum 4 deberá disponer de un cluster de cómputo con procesadores de propósito general para poder sacar el máximo rendimiento al mayor número de aplicaciones actuales del ecosistema de supercomputación.

Por otro lado, se deberá de proveer de varios clusters con tecnologías emergentes (arquitecturas diferentes a la propuesta de propósito general), capaz de ejecutar aplicaciones de producción de supercomputación y que , al menos algunos de ellos, sea especialmente adecuados para cognitive computing y Deep Learning.

Los clusters de cómputo propuestos deberán satisfacer las siguientes condiciones generales:

- Cluster de Propósito General (CPG)

  Se deberá de proveer de un cluster de cómputo con un rendimiento pico mínimo de 9.5 Petaflops (siempre que se haga referencia a Petaflops o Teraflops en este documento serán de doble precisión). Todo el rendimiento de este cluster deberá venir proporcionado únicamente por procesadores de propósito general. No se considera como procesadores de propósito general ningún tipo de aceleradores, tales como los de la familia Xeon Phi ni los aceleradores gráficos.

- Clusters de Tecnologías Emergentes (CTE)

  Se deberán proveer diversos clusters, con un mínimo de dos, basados en tecnologías emergentes (plataformas diferentes a la presentada como de propósito general) con diferentes procesadores, aceleradores, combinaciones de estos, …. Estos clusters serán independientes del cluster de propósito general, aunque deberán poder usar el sistema de almacenamiento descrito en el Lote1 del concurso CONSU02016008OP. Se podrán presentar tantos clusters como se deseen, cada uno con tecnologías diferentes, con el objetivo de permitir al BSC la evaluación de las tecnologías, procesadores y aceleradores, que puedan ser utilizadas en los sistemas pre-Exascale más potentes en los años 2018-2020. Se valorará la diversidad en diferentes tecnologías de procesadores/aceleradores que cada CTE disponga, su capacidad de cálculo proporcionada por la evolución final de cada CTE, y las actualizaciones de los clusters durante la evolución del proyecto; de forma que cada fase de los CTE debe tener la potencia adecuada para conseguir estos objetivos.

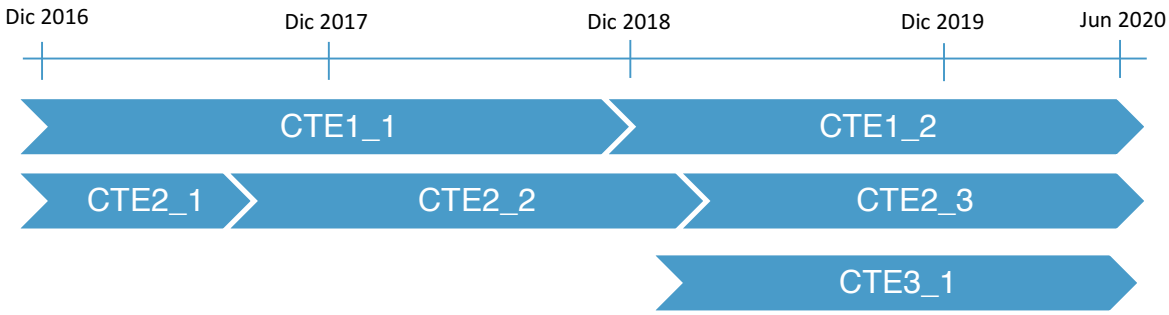


Figura descriptiva CTE y sus evoluciones tecnológicas, en la que se representan por ejemplo 3 CTE diferentes, con 2, 3 y 1 actualizaciones tecnológicas, respectivamente

- Minimizar el tiempo de pérdida de servicio al realizar la transición de MN3 a MN4

  Para poder dar un servicio continuado, se deben presentar soluciones que minimicen la disrupción de servicio, haciendo explícito en la documentación el tiempo esperado de

pérdida de servicio al hacer la transición de MN3 a MN4, y la capacidad de cálculo disponible en cada momento

- Limitación de consumo eléctrico

  Hasta Enero de 2018, todos los clusters se instalarán en la capilla de Torre Girona. Entre todos ellos no deberán superar un consumo eléctrico de 1.3 MW con carga CPD. Consideramos esta carga CPD como equivalente al 70% del consumo eléctrico máximo de los equipos instalados ejecutando HPL. A este consumo máximo, se debe restar el consumo de los 6 racks del sistema de ficheros (descrito en el Lote 1 del concurso CONSU02016008OP). Esta limitación de consumo viene determinada por la capacidad de refrigeración de la capilla. Las actualizaciones y nuevos CTE posteriores a Enero de 2018 se podrán ubicar fuera de la capilla de Torre Girona, en zona próxima a la misma, sin consideración de limitación de consumo eléctrico.

- Limitación de espacio

  El espacio disponible en la capilla es de 120 m2. En este espacio se debe tener en cuenta el espacio ya ocupado por el sistema de ficheros del proyecto Marenostrum4 descrito en el Lote 1 del concurso CONSU02016008OP. Para las actualizaciones y nuevos CTE posteriores a Enero de 2018 no se considerará limitación de espacio.

- Modificaciones de infraestructura

  Cualquier modificación de infraestructura actual para la instalación y funcionamiento adecuado del superordenador MareNostrum4 debe estar incluida en el proyecto a entregar. Para las actualizaciones y nuevos CTE posteriores a Enero de 2018 no se considerará modificaciones ni preparación de infraestructura, a excepción de las modificaciones requeridas para conectar electricidad y refrigeración a los racks entregados.

- Capacidades técnicas

  Para las tareas de instalación, configuración y posterior mantenimiento del superordenador, la empresa licitadora deberá disponer de un equipo con las capacidades y conocimientos mínimos necesarios para poder realizar adecuadamente la ejecución de este contrato (haber instalado clusters previamente por encima de 1000 nodos por cluster). Se deberá describir el número de personas de dicho equipo y el perfil de las mismas en este pliego técnico, asegurando así el correcto desarrollo e implantación del contrato/proyecto.

- Fechas de producción

  El cluster de propósito general deberá estar en producción antes del 1 de julio de 2017. Como mínimo, la primera evolución de un CTE debe estar en producción en 2016.

A continuación, pasamos a describir en detalle los requerimientos mínimos y los deseables técnicos para los clusters de cómputo de MareNostrum4.
En las siguientes tablas, los campos se identifican por las letras R y D, cuyo significado es:
R- Representa que lo anunciado es un requerimiento que se debe cumplir en la solución presentada, en el caso de no hacerlo la oferta quedará desclasificada.
D- Representa un requerimiento deseable a tener y que se valorará positivamente aquellas soluciones que lo incorporen.

1.- Hardware Cluster de propósito general

A continuación, pasamos a describir en detalle los requerimientos mínimos y mejoras sobre el hardware del cluster de propósito general (CPG) para el proyecto MareNostrum4.

1.1.- Descripción Hardware

Ref	Descripción
R1	Cluster de propósito general formado por el número de nodos de cómputo necesarios para proporcionar un mínimo de 9.5 PFlops pico. No se considera de propósito general aceleradores gráficos (GPU) o la familia de procesadores Xeon Phi (KNC,KNL, …).
R2	Se consideran 2 tipos de nodo de cómputo que podrá haber en el CPG: - Nodo cómputo normal - Nodo cómputo fat Cada nodo de cómputo sólo podrá pertenecer a uno de los 2 tipos, y todos los nodos de cómputo de un tipo deberán ser idénticos.
R3	Un nodo de cómputo normal deberá tener las siguientes características técnicas mínimas: - 2 Chips o sockets de propósito general por nodo - 2 GB/core de memoria principal volátil - La configuración de memoria presentada deberá ser equilibrada desde/hacia todos los cores de un mismo socket a la memoria (DIMMs misma velocidad y tamaño) y la frecuencia de acceso a memoria deberá ser la más alta que la familia de los procesadores ofertados permita. - Los buses que interconectan los sockets de un nodo deberán ser equilibrados y tener el máximo ancho de banda que la familia de los procesadores ofertados permita, la cantidad de estos buses será evaluado.
R4	Un mínimo del 5% de nodos totales del cluster serán de tipo fat, dichos nodos serán idénticos a los nodos normales a excepción que contarán con un mínimo de 8 GB/core. A destacar que, deberán seguir cumpliendo que la configuración de memoria presentada deberá ser equilibrada desde/hacia cada socket y la frecuencia de acceso a memoria deberá ser la más alta que la familia de los procesadores ofertados permita. Estos nodos deberán estar ubicados de forma consecutiva y conectados a la misma isla de red de baja latencia (ver apartado 1.2).
D5	Se valorará una proporción superior del 5% de número de nodos fat en el cluster, con un máximo de 10%.
R6	Todos los nodos de cómputo (normal ó fat) deberán incorporar un almacenamiento local con una capacidad mínima de: - 5 veces la memoria principal de los nodos normales en el caso de ser

Ref	Descripción
	almacenamiento local basado en tecnología HDD ó - 2 veces la memoria principal de los nodos normales en el caso de ser almacenamiento local basado en tecnología SSD
D7	Se valorará que la inclusión de almacenamiento interno sea basado en SSD y que la capacidad en SSD sea superior a 2 veces la memoria principal
D8	Se valorará la inclusión de nodos con tecnologías de memoria no volátil como puede ser 3DXpoint o similares. Se valorará el número de nodos y la cantidad de memoria proporcionada, así como la fecha de disponibilidad.
R9	Todos los nodos de cómputo deberán tener como mínimo las siguientes interfaces de red para conectarse con el resto de componentes del cluster: - Tarjeta para conexión a una red de baja latencia para el tráfico de datos al filesystem HPC del BSC (GPFS) y para aplicaciones MPI de un ancho de banda mínimo teórico de 100Gbits/nodo. - Una interfaz 1 Gbit Ethernet (red Interna cluster, gestión out-of-line) - Una Interfaz 1 Gbit Ethernet para el tráfico control GPFS, a conectar a la red de control de GPFS del lote 1 del concurso CONSU02016008OP.
R10	En caso de que los nodos estén empaquetados en un chasis: - La interfaz de gestión out-of-line podrá ser compartida por todos los nodos del chasis - En el caso de tener switch interno para alguna de las otras redes deberá cumplir el número de interfaces por nodo.
R11	Todo nodo de cómputo deberá de ofrecer los buses independientes suficientes para poder soportar las conexiones a las diversas redes que se describen anteriormente, sin ser ningún factor limitante.
R12	Se requiere un esquema de bloques de los nodos de cómputo ofertados con los anchos de banda entre los diferentes componentes de un nodo (máximo y útiles expresados en GB/s): procesadores, memoria, diversos buses PCI-Express, cualquier componente I/O..
D13	Se valorará el esquema de bloques de la placa base presentado
R14	Se deberán proveer 5 nodos, a usarse como login nodes. Estos 5 logins deberán ser idénticos que los nodos de cómputo normales, a excepción que deberán contar con una interfaz extra de 10Gbit ethernet para permitir la conexión a la VLAN pública del BSC.
R15	Todos los nodos del cluster y logins deberán disponer de un sistema de administración remoto (out-of-band), el cual debería permitir como mínimo: poder realizar el power on/off, coger la consola, monitorización del entorno (Temperatura, consumo, …), generación de alarmas, detección de problemas hardware/firmware, led de identificación, etc.
R16	A la hora de calcular la potencia de cálculo del cluster proporcionado, sólo se tendrá en cuenta los nodos de cómputo (excluyendo los logins como cualquier servidor de gestión del cluster).

Ref	Descripción
D17	Se valorará la mejora en potencia de cálculo pico total en PFlops respecto al mínimo requerimiento, con un límite superior de 11 PFlops
R18	Se deberán incluir todos aquellos servidores para la gestión del cluster. Entre otras cosas este hardware deberá hacerse responsable de la gestión de imágenes de sistema operativo, servicios básicos para el cluster como head servers, DHCP, NTP, DNS, …; el sistema de colas, monitorización, etc. Estos servidores deberán de disponer del hardware necesario para realizar las tareas asignadas a nivel de: cpu, memoria, almacenamiento, interfaces de red, rendimiento, etc. Se deberá rellenar la tabla número 2 para cada tipo de server de gestión ofrecido. Se deberán de proveer como mínimo los siguientes servidores físicos: - Estructura jerárquica de servidores para la gestión de imágenes y sub-partes del cluster de cómputo (2 nodos centrales (head nodes) y N servidores de segundo nivel) - 2 Servidores de monitorización (mínimo: 128GB RAM, RAID SSD con 4TB neto y alto ancho de banda a red) - 2 Servidores de sistema de colas - 2 Servidores de monitorización de red de baja latencia - 2 Servidores con 128GB de memoria principal cada uno mínimo para máquinas virtuales con servicios no críticos - Almacenamiento centralizado (mínimo 10TB neto) con sus servidores asociados. Este almacenamiento guardará las imágenes de sistema operativo de los nodos de cómputo. Será montado en los servidores de clustering y exportado via nfs-root a los nodos de cómputo. Los head nodes, los 2 servidores de máquinas virtuales y todos los servidores de monitorización deberán tener una interfaz 1 Gbit de red extra para la conexión a las VLANs del BSC.
R19	Se deberá enrackar un Keyboard-Video-Mouse de 1U con acceso a la consola gráfica de todos los servers de administración mediante un switch de consolas. Todos los servidores de gestión deberán disponer de un sistema de administración remoto (out-of-band), el cual debería permitir como mínimo: poder realizar el power on/off, coger la consola gráfica en remoto via web, monitorización del entorno (Temperatura, consumo, …), generación de alarmas, detección de problemas hardware/firmware, etc.
D20	Se valorará mejoras sobre los mínimos, el hardware presentado como el diseño para los servidores de la gestión del cluster.
R21	Todos los servidores y servicios que conformen la administración del cluster deberán estar completamente redundados en modo de alta disponibilidad, no deberá de existir elementos que sean un único punto de fallo, tanto a nivel hardware como a nivel software.
R22	Se requiere que se rellene la siguiente tabla (Tabla 1- Descripción hardware Nodos CPG), en ella se especifican los valores mínimos a cumplir, y se deberá

Ref	Descripción
	indicar los valores ofertados.
D23	Se valorará la mejora en cualquiera de las entradas con valor mínimo de la tabla 1. Y en las entradas que no haya valor mínimo se compararán los valores ofertados por cada solución. No se valorará en este punto mejoras ya valoradas anteriormente.
R24	El firmware de los nodos deberá registrar, por ejemplo, en el sistema de gestión out-of-line, cualquier fallo recuperable o irrecuperable de cualquier de los componentes (especialmente de los DIMMs de memoria). De la misma manera, deberá tener un linde definido de errores recuperables de tal manera que genere una alarma recomendando la sustitución de aquel componente de forma proactiva antes del fallo irrecuperable.
R25	Todos los racks de cómputo del CPG deberán ser idénticos a nivel de elementos hardware incluidos por rack. Como por ejemplo y sin estar limitado a: número de nodos de cómputo, número de switches, orden de enrackado, cableado interno y número de conexiones que entran y salen de él.
R26	Se deberá entregar un nodo de cómputo extra, no se deberá enrackar, idéntico a los proporcionados para el CPG (con todos sus componentes) para poder enseñar en las visitas.

Tabla1 – Descripción hardware nodos CPG

Concepto	Valor mínimo	Valor ofertado
Características nodo de cómputo		
Número chips o sockets por nodo	2	
Modelo procesador		
Ancho de banda (GB/s) entre procesadores		
Cores por procesador ofertado		
Frecuencia nominal de cada core		
Frecuencia turboboost de cada core		
Frecuencia (modo vectorial) de cada core		
FLOPs pico por ciclo de cada core del procesador		
GFLOP pico por procesador		
GFLOP pico por nodo		
Consumo típico por procesador (max TDP)		
GFlop pico por procesador / max TDP		
Tecnología y frecuencia memoria RAM		
Frecuencia real funcionamiento memoria		
Capacidad almacenamiento local		
Tecnología almacenamiento local	HDD / SSD	
Interfaz y bandwidth de acceso a almacenamiento		
RPM disco duro interno (en caso HDD)		
IOPS almacenamiento local		
Número de nodos con memoria no volátil		
Capacidad memoria no volátil por nodo		
Fecha de entrega de la memoria no volátil		

Interfaces 10GE incorporadas por nodo		
Interfaces 1 GE incorporadas por servidor		
Número de interfaces red baja latencia		
Tecnología interfaces red baja latencia		
Ancho banda a red de baja latencia	100 Gbit/s	
Nodo cómputo normal		
Número de nodos de cómputo normal		
Memoria RAM por core ofertada	2	
Memoria RAM ofrecida por nodo		
Número DIMMs y tamaño por DIMM		
Nodo cómputo fat		
Número de nodos de cómputo fat		
% número de nodos fat respecto al total	5%	
Memoria RAM por core ofertada	8	
Memoria RAM ofrecida por nodo		
Número DIMMs y tamaño por DIMM		
Características globales cluster cómputo (sin considerar los logins)		
Número nodos (normal+fat)		
TB Memoria RAM total (normal+fat)		
Almacenamiento interno total (normal+fat)		
PFlop pico nodos propósito general (normal+fat)	9.5	
Número de nodos de cómputo por rack		
Número de racks de cómputo cluster CPG		

Tabla2 – Descripción hardware por Servidor de gestión

Concepto	Valor mínimo	Valor ofertado
Características servidores de gestión		
Número de servidores gestión tipo A		
Servicio proporcionado servidor tipo A		
Número y modelo de procesador		
Memoria RAM		
Configuración DIMMs por servidor		
Almacenamiento compartido para la gestión del cluster (si aplica)		
Almacenamiento interno por servidor (# discos, tamaño y tecnología)		
Controladora RAID (si aplica)		
Interfaces 1 Gbit Ethernet por servidor		
Interfaces 10 Gbit Ethernet por servidor		
Interfaces 40 Gbit Ethernet por servidor		

1.2.- Switches y redes

A continuación, se detallan los requisitos comunes para todas las redes del cluster de propósito general y en las consiguientes tablas los requisitos específicos para cada una de las redes.

Ref	Descripción
R1	Se deberán de proveer de esquemas de conexionado físico de cada una de las redes que conforman el CPG. También se debe describir el ancho de banda disponible a cada nivel de las redes y la latencia introducida por cada elemento hardware. Cada una de las redes descritas debe ser completamente disjunta a nivel físico.
R2	Todos los switches de cualquier red deberán tener doble fuente de alimentación, y redundancia a nivel de ventiladores. Todos estos componentes deberán ser modulares y poderse cambiar en caliente, sin la parada del switch en cuestión. Si dentro de una misma red se proveen switches de fabricantes diferentes y se detecta cualquier incompatibilidad a la hora de conectarlos entre ellos (GBIC, fibra, limitación funcionalidades, rendimiento), el licitante deberá sustituir los switches necesarios para que todos sean del mismo fabricante para eliminar la incompatibilidad.
R3	Para cada una de las redes y una vez todos los componentes conectados deberá existir un 5% de puertos libres por cada nivel, a excepción del nivel más bajo de cada red.
R4	Se requiere que se rellene la tabla (Tabla 3- Descripción hardware Switches y redes CPG), en ella se especifican los valores mínimos a cumplir. En el caso de proporcionar más de un tipo de switch por red, se deberán rellenar los datos de la tabla 3 por cada tipo de switch proporcionado.
D5	Se valorará la mejora en cualquiera de las entradas con valor mínimo. Y en las entradas que no haya valor mínimo se compararán los valores ofertados por cada solución. No se valorará en este punto mejoras ya valoradas o mencionadas de forma aparte en otra entrada.
R6	Todos los switches de segundo nivel de cualquiera de las redes deberán ser redundantes entre ellos, pudiendo evitar cualquier punto único de fallo. Debería poder caer un equipo y realizar su sustitución sin ningún tipo de corte o afectación.
R7	Todos los elementos de red ofertados deberán tener un “end of life” comercial mínimo hasta la finalización del proyecto Marenostrum 4.
R8	En todas las redes que hayan conexiones de velocidades diferentes, los switches deberán incorporar los buffers necesarios para ofrecer los rendimientos line-rate entre las diferentes velocidades.

Ref	Descripción
Red Interna cluster	
R9	Se deberá proveer del hardware necesario (switches, cables, etc.) para poder establecer la red interna del cluster con tecnología 1/10 Gigabit Ethernet. Todos los cables y fibras de esta red física que vayan a la misma velocidad deberán ser del mismo color y de un color diferente a las otras redes de la máquina, de tal manera que puedan distinguirse visualmente.
R10	Todos los puertos de cada tipo de un mismo switch deberán ser line-rate entre ellos sin ningún tipo de sobre-suscripción.
R11	Requerimientos de funcionalidades de los switches de esta red: - Soporte Jumbo Frames (MTU > 9000) - Line-rate Nivel 2 switching - Definición de Access-list a nivel 2 y nivel 3 - Spaning-tree (MSTP y RSTP) - Capacidad para filtrar los paquetes BPDU a nivel de puerto físico del equipo - Port mirroring - Broadcast storm control - QoS - Snmp - SSH - Minimum 256 VLANs - LACP (Soporte hash LACP L3 + L4) - Flowcontrol - Soporte de más de 10000 MACs en la tabla de forwarding - 802.1Q - Fuentes redundantes y hot-swap - Ventiladores redundantes y hot-swap - MC-LAG (Multi-Chassis Link Aggregation Group) ó VLT (Virtual Link Trunking) (requerimiento para switches de segundo nivel)
R12	En esta red física se configurarán 2 dominios de broadcast diferentes (2 VLANs): - 1 VLAN => Red Interna cluster (DHCP,Boot, …) - 1 VLAN => Red gestión de elementos del cluster (IPMI, Switches, racks, …) que sólo será visible desde los servidores de gestión y será inaccesible desde los logins o cualquier nodo de cómputo
R13	En esta red física se conectará: - Cada nodo de cómputo con una interfaz 1 Gbit Ethernet (Red cluster) - La interfaz de acceso mediante IPMI a cada uno de los nodos de cómputo. (Puede usarse la misma interfaz de 1 Gbit del nodo si se soporta con VLAN tagging) - Dos interfaces por cada servidor de gestión del cluster de 10 y/ó 40 Gbit Ethernet (VLAN interna cluster, VLAN gestión out-of-line) - Cualquier interfaz de gestión de cualquiera de los componentes del

Ref	Descripción
	cluster (racks, IPMI servers, puertas frías, PDU, switches, etc.)
R14	Los servidores de servicio deberán conectarse al segundo nivel de switches de esta red en modo linerate a 10 y/ó 40 Gbit Ethernet mediante bonding. El primer nivel de switches de esta red deberá introducir una contención de aproximada de 2:1, por ejemplo. Switches de primer nivel con 48 puertos de 1 Gbit con 2 uplinks de 10 Gbit Ethernet al nivel superior. La sobresuscripción a niveles superiores deberá venir determinada por las necesidades del diseño presentado. Tanto la sobresuscripción como el nivel de switches deberá ser común y equilibrado desde cualquiera de los nodos de cómputo.
D15	Se valorará el diseño de la red presentado teniendo en cuenta conceptos como: - La redundancia en la caída de enlaces (up-links) entre switches. - Redundancia en la conexión de los diversos elementos a la red de management (servidores de servicio, nodos de cómputo, etc.) - La óptima o mejor distribución de la conexión de los elementos a los diferentes switches teniendo en cuenta los patrones de tráfico que esta red va a soportar y la sobresuscripción de la red presentada
Red de control de GPFS	
R16	Se deberá proveer del hardware necesario (switches, cables, etc.) para poder establecer la red para el tráfico de control de GPFS con tecnología 1/10 Gigabit Ethernet a cada uno de los nodos de cómputo del cluster y logins. Todos los cables y fibras de esta red física que vayan a la misma velocidad deberán ser del mismo color y de un color diferente a las otras redes de la máquina, de tal manera que puedan distinguirse visualmente.
R17	Todos los puertos de cada tipo de un mismo switch deberán ser line-rate entre ellos sin ningún tipo de sobre-suscripción.
R18	Requerimientos de funcionalidades de los switches de esta red: - Soporte Jumbo Frames (MTU > 9000) - Line-rate Nivel 2 switching - Line-rate Nivel 3 routing - Definición de Access-list - Routing (dinámico y estático) - Spaning-tree (MSTP y RSTP) - Capacidad para filtrar los paquetes BPDU a nivel de puerto físico del equipo - Port mirroring - Broadcast storm control - QoS - Snmp - SSH - Minimum 256 VLANs - LACP (Soporte hash LACP L3 + L4) - Flowcontrol

Ref	Descripción
	- Soporte de más de 10000 MACs en la tabla de forwarding - 802.1Q - Fuentes redundantes y hot-swap - Ventiladores redundantes y hot-swap - MC-LAG (Multi-Chassis Link Aggregation Group) ó VLT (Virtual Link Trunking) (requerimiento para switches de segundo nivel)
R19	En esta red física se configurará 1 dominio de broadcast (1 VLAN): Red de control de GPFS. (Misma VLAN que la de mismo nombre del Lote 1 del concurso CONSU02016008OP)
R20	En esta red física se conectará: - Cada nodo de cómputo y login con una interfaz 1 Gbit Ethernet (Red de control de GPFS) Esta red se deberá conectar a nivel superior hacia la red de control de GPFS definida en el Lote1 del concurso CONSU02016008OP. Esta red deberá tener una topología de estrella, y los switches centrales o del nivel más alto de esta topología son los que se deberán conectar, mediante bondings a los switches de más alto nivel de los propuestos en el lote 1 del concurso CONSU02016008OP.
R21	Desde cada nodo de cómputo a la red de control de GPFS del Lote1 del concurso CONSU02016008OP deberá haber un máximo de sobresuscripción de 16:1. El primer nivel de switches de esta red deberá introducir una contención de aproximada de 2:1, por ejemplo. Switches de primer nivel con 48 puertos de 1 Gbit con 2 uplinks de 10 Gbit Ethernet al nivel superior. Tanto la sobresuscripción como el nivel de switches deberá ser equilibrado e igual desde cualquiera de los nodos de cómputo.
D22	Se valorará el diseño de la red presentado teniendo en cuenta conceptos como: - La redundancia en la caída de enlaces (up-links) entre switches. - Redundancia en la conexión de los diversos elementos a la red de management (servidores de servicio, nodos de cómputo, etc.) - La óptima o mejor distribución de la conexión de los elementos a los diferentes switches teniendo en cuenta los patrones de tráfico que esta red va a soportar y la sobresuscripción de la red presentada
D23	Se valorará que se implementen las 2 redes físicas ethernet (Red interna gestión y Red de control de GPFS), mediante una única red física basada en 10 Gbit Ethernet definiendo las 3 VLANs encima de esa red física y con las conexiones necesarias de cada red. En este caso, obviamente no aplicaría el requisito R1 de este apartado de tener 3 redes físicas disjuntas. En el caso de incluir esta mejora D23, no aplica los requerimientos de bloqueos especificados en la entrada R21. Éstos cambiarían a: “Desde cada nodo de cómputo a la red de control de GPFS del Lote1 del concurso CONSU02016008OP deberá haber un máximo de sobresuscripción de 128:1 y

Ref	Descripción
	que el primer nivel de switches de esta red deberá introducir una contención máxima de 8:1, por ejemplo: Switches de primer nivel con 48 puertos de 10 Gbit con 2 uplinks de 40 Gbit Ethernet al nivel superior". Aun así, se seguirá aplicando los requerimientos de redundancia y equilibrio de la entrada R21, y se valorará el diseño presentado tal como indica D22.
Red Interconexión MPI / GPFS datos RDMA	
R24	Se deberá proveer del hardware necesario (switches, cables, etc., su esquema y etiquetado) para poder establecer la red interna de alto rendimiento y muy baja latencia sobre la cual se va enviar: - Comunicaciones MPI - Tráfico de datos GPFS RDMA Esta red deberá ofrecer un mínimo de 100 Gbits por link Todos los cables y fibras de esta red física deberán ser del mismo color y de un color diferente a cualquier otra red de la máquina. La única excepción puede ser los cables de cobre EDR u OPA que sólo se fabriquen en color negro. Para el resto de cableado se deberá cumplir ese requisito intra o inter rack.
R25	Todos los nodos de cómputo y logins deberán estar conectados a esta red de interconexión, como los servidores de monitorización de esta red.
R26	Dicha red a nivel de cómputo deberá ser no bloqueante en grupos o islas no menores de 20.000 cores, dichos grupos o islas deberán ser iguales en número, es decir, múltiplos del número total de nodos. Los nodos entre islas podrán tener un factor máximo bloqueante de 2:1.
R27	A esta red también se conectarán los elementos del almacenamiento GPFS del BSC que se describen en el lote 1 del concurso CONSU02016008OP (MADDR y MM servers). Dicha conexión deberá ser directa (sin el uso de routers) y distribuida uniformemente entre los switches del nivel superior sin sobre suscripción hacia cada isla, según se expresa en el lote1 del concurso CONSU02016008OP. Haciendo que el rendimiento sea uniforme desde cualquier nodo de cómputo al almacenamiento y maximizando la alta disponibilidad en caso de fallo de cualquier switch. Se deberá de proveer de todo aquel hardware y servicios extra (switches, fibras, tareas de cableado) necesario para implementar estas conexiones de manera que nunca sea un factor limitante para poder sacar el máximo rendimiento al almacenamiento del Lote 1 del concurso CONSU02016008OP, especialmente si se propone una tecnología diferente. Como referencia, el Lote 1 tiene una conectividad de 47 links de tecnología OPA o 94 de tecnología EDR.
D28	Se valorará como mejora: - La reducción del bloqueo entre islas en la red presentada - El mayor número de cores por isla - El número mínimo de switches para la creación de la red - Minimizar el número de saltos por switches entre cualquiera de los nodos del cluster, máximo permitido 4.

Ref	Descripción
	- Routing adaptativo en la red según congestión, etc.
D29	Se deberá presentar el esquema de conexionado propuesto para esta red el cual también será valorado, a nivel de redundancia, uniformidad, etc. También se valorará interoperabilidad entre diferentes generaciones tecnológicas (backward and forward compatibility), así como la posibilidad de la capacidad de soportar diferentes arquitecturas (ARM, Intel, Power, etc)
R30	Todos los switches de la red de baja latencia deberán poder ser gestionables desde la red ethernet interna del cluster en la VLAN de gestión de dispositivos.

Tabla 3 – Descripción hardware switches y redes CPG

Concepto	Valor mínimo	Valor ofertado
Red Interna cluster		
Número de switches proporcionados		
Marca switch		
Modelo switch		
Número de puertos 1GE por switch		
Número de puertos 10GE por switch		
Número de puertos 40GE por switch		
Número de puertos libres		
Latencia introducida por el switch		
Red control de GPFS		
Número de switches proporcionados		
Marca switch		
Modelo switch		
Número de puertos 1GE por switch		
Número de puertos 10GE por switch		
Número de puertos 40GE por switch		
Número de puertos libres		
Latencia introducida por el switch		
Red MPI / GPFS datos RDMA		
Número de switches proporcionados		
Marca switch		
Modelo switch		
Número de puertos por switch		
Tecnología de conexión		
Ancho de banda por puerto		
Número de puertos libres		
Latencia introducida por el switch		
Contención Red MPI / GPFS datos RDMA		
Número de nodos por isla sin contención		
Número de cores por isla sin contención	20000	
Número de islas del CPG		
Contención (máxima) entre islas	2:1	

2.- Hardware clusters de cómputo de tecnologías emergentes

En la siguiente tabla describimos los requerimientos y deseables del hardware de los clusters de tecnologías emergentes.

2.1.- Descripción Hardware

Ref	Descripción
R1	Se deberán proveer como mínimo dos clusters basados en tecnologías emergentes (plataformas diferentes a la presentada como de propósito general) con diferentes procesadores, aceleradores, combinaciones de estos, …. Estos clusters serán independientes del cluster de propósito general, aunque deberán poder usar el sistema de almacenamiento descrito en el Lote1 del concurso CONSU02016008OP. Se podrán presentar tantos clusters como se deseen, cada uno con tecnologías diferentes, con el objetivo de permitir al BSC la evaluación de las tecnologías, procesadores y aceleradores, que puedan ser utilizadas en los sistemas pre-Exascale más potentes en los años 2018-2020. Cualquier CTE o su correspondiente evolución deberá tener, en el momento de puesta en producción, la tecnología más avanzada de su familia y deberá seguir siendo así al menos durante los siguientes 6 meses de entrar en producción. Cada evolución tecnológica de un CTE o CTE nuevo deberá como mínimo ofrecer una tecnología nueva en el tipo de procesador/acelerador ofertado. Y será opcional la inclusión de actualizaciones tecnológicas en: interconexión de baja latencia, memoria volátil y no volátil, empaquetamiento en rack, etc.
R2	Los diversos CTE deberán ser completamente independientes del cluster de propósito general en el sentido que cualquier mantenimiento hardware/software de cualquier de los CTEs no deberá afectar en nada al cluster de propósito general y viceversa.
R3	Todos los clusters de tecnologías emergentes deberán ser capaces de usar y montar el sistema de ficheros paralelo descrito en el Lote 1 del concurso CONSU02016008OP.
R4	Los CTE podrán compartir elementos de administración como elementos de red, servidores de administración siempre y cuando no sea un factor limitante tecnológico o de rendimiento. Por otro lado, tal como se indica en R2, en ningún caso los CTE podrán compartir elementos hardware/software con el CPG.
R5	La tecnología ofrecida en estos clusters deberá ser diferente a la ofertada en el cluster de propósito general y equivalente a arquitecturas disponibles en 2018-2020, por ejemplo, basado en nuevas arquitecturas incluyendo, pero no limitado, ARM, Power, GPGPU, Xeon Phi. Estos clusters deberán ofrecer una arquitectura/tecnología que el BSC no tenga en otro cluster en producción. (Los recursos de HPC del BSC se pueden consultar en las siguientes direcciones: http://www.bsc.es/marenostrum-support-services/mn3

Ref	Descripción
	http://www.bsc.es/marenostrum-support-services/other-hpc-facilities).
R6	Cada CTE y cada una de sus actualizaciones deberá tener la potencia de cálculo adecuada para poder evaluar dicha tecnología de futuro y su evolución, pudiendo realizar ejecuciones de producción. Como mínimo dos CTE, en su actualización o evolución tecnológica final que no incluye las fases iniciales del mismo, deberán ofrecer un mínimo de potencia pico de 500 TFlops.
D7	Se valorarán los PetaFlops (PFlops) pico ofrecidos por los CTE en la última de sus actualizaciones o evolución tecnológica final que no incluye las fases iniciales del mismo. Para cada CTE sólo se valorarán las potencias de pico superiores a 500 Tflops.
D8	Se valorará la diversidad en diferentes tecnologías de procesadores/aceleradores que cada CTE disponga/ofrezca, considerando sólo aquellos que sean relevantes en relación a los sistemas pre-Exascale más potentes previstos en los años 2018-2020, de forma que el BSC pueda evaluar adecuadamente cada una de estas tecnologías durante la duración del proyecto MN4. Se incluye en esta valoración las diversas evoluciones de cada CTE de acuerdo con las disponibilidades tecnológicas.
D9	Se valorará, para cada uno de los CTE considerando sólo aquellos que sean relevantes en relación a los sistemas pre-Exascale más potentes previstos en los años 2018-2020, la posibilidad de realizar co-diseño con los propietarios de cada una de las tecnologías presentadas. Se debe describir el alcance y las características de ese co-diseño para cada uno de los CTE.
R10	Cualquier tarea de desenraque, cableado o modificación de los CTE para las diversas evoluciones técnicas o actualizaciones deberá estar incluido.
R11	Los CTE deberán poder ejecutar aplicaciones de producción especialmente acondicionadas a la tecnología ofrecida. La empresa licitadora deberá aportar experiencia en la compilación, adaptación y optimización de aplicaciones de producción a la tecnología ofrecida en cada CTE. Proporcionar lista de aplicaciones que se beneficiarían de cada CTE y el rendimiento esperado comparándolo con arquitecturas actuales.
R12	La configuración de memoria presentada debería ser equilibrada desde/hacia todos los cores de un mismo socket a la memoria (DIMMs misma velocidad y tamaño) y la frecuencia de acceso a memoria deberá ser la más alta que la familia de los procesadores ofertados permita. Así mismo, la capacidad de memoria proporcionada debe estar equilibrada para poder evaluar y ejecutar las aplicaciones asociadas a cada tecnología. Los buses que interconectan los sockets de un nodo deberán ser equilibrados y tener el máximo ancho de banda que la familia de los procesadores ofertados permita, la cantidad de estos buses será evaluado.
D13	Se valorará la fecha de puesta en producción de cada CTE y de cada una de sus evoluciones.

Ref	Descripción
	Se puede indicar con la fecha de instalación y la fecha de "General Availability" o el intervalo de tiempo entre ambas fechas, considerando mejor disponer del equipo antes de "General Availability".
R14	Todos los nodos de cómputo deberán incorporar un almacenamiento local.
R15	Todos los nodos de cómputo deberán tener como mínimo las siguientes interfaces de red para conectarse con el resto de componentes del cluster: - Tarjeta para conexión a una red de baja latencia para el tráfico de datos al filesystem HPC del BSC (GPFS) y para aplicaciones MPI de un ancho de banda mínimo teórico de 100Gbits/nodo. - Una interfaz 1 Gbit Ethernet (red Interna cluster y gestión out-of-line) - Una interfaz de mínimo 1/10 Gbit Ethernet (red control GPFS) En caso de no poder configurar la interfaz out-of-line en la interfaz de 1 Gbit, se deberá conectar un enlace extra de 1 Gbit para tal efecto. En caso de que los nodos estén empaquetados en chasis: - La interfaz de gestión out-of-line podrá ser compartida por todos los nodos del chasis - En el caso de tener switch interno deberá cumplir con el número de interfaces por nodo antes descritas.
R16	Todo nodo de cómputo deberá de ofrecer los buses independientes suficientes para poder soportar las conexiones a las diversas redes que se describen anteriormente, sin ser ningún factor limitante.
R17	Se requiere un esquema de bloques de los nodos de cómputo ofertados con los anchos de banda entre los diferentes componentes de un nodo (máximo y útiles expresados en GB/s): procesadores, memoria, diversos buses PCI-Express, cualquier componente I/O.
R18	Cada CTE/evolución deberá proveer 1 login node. Este login deberá ser idéntico a los nodos de cómputo de dicho CTE/evolución, a excepción que deberá contar con una interface ethernet adicional para permitir la conexión a la VLAN pública del BSC. La última evolución de cada CTE deberá tener 2 login nodes.
R19	Todos los nodos de cada CTE y sus logins deberán disponer de un sistema de administración remoto (out-of-band), el cual debería permitir como mínimo: poder realizar el power on/off, coger la consola, monitorización del entorno (Temperatura, consumo, …), generación de alarmas, detección de problemas hardware/firmware, led de identificación, etc.
R20	Cada CTE/evolución deberá incluir todos aquellos servidores para la gestión del cluster. Entre otras cosas este hardware deberá hacerse responsable de la gestión de imágenes de sistema operativo, servicios básicos para el cluster como DHCP, NTP, DNS, …; el sistema de colas, monitorización, etc. Estos servidores deberán de disponer del hardware necesario para realizar estas tareas a nivel de: cpu, memoria, almacenamiento, interfaces de red, etc. En la documentación se deberá especificar las características de estos servidores.

Ref	Descripción
	El servidor de administración principal deberá incorporar una interfaz ethernet adicional para la conexión a la VLAN para su administración.
D21	Se valorará el hardware como el diseño presentado para los servidores de la gestión de cada CTE.
R22	Todos los servidores y servicios que conformen la administración del cluster deberán estar completamente redundados, no deberá de existir elementos que sean un único punto de fallo, tanto a nivel hardware como a nivel software.
R23	Por cada cluster CTE/evolución, se requiere que se rellene la tabla (Tabla 4-Descripción hardware CTE), y especificar en ella los valores ofertados. Para cada evolución se deberá indicar sólo los valores introducidos por aquella evolución, sin contar las evoluciones anteriores.
D24	Se valorará la mejora en cualquiera de las entradas. Se compararán los valores ofertados por cada solución. No se valorará en este punto mejoras ya valoradas anteriormente.
R25	El firmware de los nodos deberá registrar, por ejemplo, en el sistema de gestión out-of-line, cualquier fallo recuperable o irrecuperable de cualquier de los componentes (especialmente de los DIMMs de memoria). De la misma manera, deberá tener un linde definido de errores recuperables de tal manera que genere una alarma recomendando la sustitución de aquel componente de forma proactiva antes del fallo irrecuperable.

Tabla4 – Descripción hardware cluster CTE

<NOMBRE TECNOLOGIA EMERGENTE>	Evolución 1	Evolución 2	Evolución N
Características cluster CTE			
Nombre modelo procesador/acelerador			
Número de nodos de cómputo			
PFlop Pico			
TB Memoria RAM total cluster			
Almacenamiento interno total			
Fecha de entrada en producción			
Fecha de salida de producción			
“General availability” ó intervalo en meses desde GA a producción			
Número de racks de cómputo cluster CTE			
Número de nodos de cómputo por rack			
Características nodo de cómputo CTE			
Número chips o sockets procesador por nodo			
Modelo procesador			
Ancho de banda (GB/s) entre procesadores			
Cores por procesador ofertado			
Frecuencia nominal de cada core			

FLOPs por ciclo de cada core del procesador			
GFLOP pico por procesador			
Consumo típico por procesador (max TDP)			
Número de aceleradores por nodo (si aplica)			
Modelo acelerador (si aplica)			
Tecnología conexión CPU a GPU (si aplica)			
Ancho de banda (GB/s) de CPU a GPU (si aplica)			
GFLOP pico por acelerador (si aplica)			
Consumo típico por acelerador (TDP) (si aplica)			
Total GFLOP pico por nodo de cómputo			
Tecnología y frecuencia memoria RAM			
Frecuencia real funcionamiento memoria			
Número DIMMs y tamaño por DIMM			
GB Memoria RAM por nodo ofrecida			
Capacidad almacenamiento local			
Tecnología almacenamiento local			
Interfaz y bandwidth de acceso a almacenamiento			
RPM disco duro interno (en caso HDD)			
IOPS almacenamiento local			
Interfaces >10GE incorporadas por nodo			
Interfaces 10GE incorporadas por nodo			
Interface gestión out-of-line			
Interfaces 1 GE incorporadas por nodo			
Número de interfaces red baja latencia			
Tecnología interfaces red baja latencia			
Ancho banda a red de baja latencia			

2.2.- Switches y redes

Cada CTE debe estar formado por mínimo 2 o 3 redes físicas, una red interna de cluster, red de control GPFS y una red de baja latencia. A continuación, se detallan los requisitos comunes para todas las redes y en las consiguientes tablas los requisitos específicos para cada red.

Ref	Descripción
R1	Se deberán de proveer los esquemas de conexionado físico de cada una de las redes para cada CTE propuesto. Cada una de las redes descritas debe ser completamente disjunta a nivel físico.
R2	Todos los switches de cualquier red deberán tener doble fuente de alimentación, y redundancia a nivel de ventiladores. Todos estos componentes deberán ser modulares y poderse cambiar en caliente, sin la parada del switch en cuestión.
R3	Se requiere que se rellene la tabla (Tabla 5- Descripción hardware Switches y redes CTE). En el caso de proporcionar más de un tipo de switch por red, se deberán rellenar los datos de la tabla 5 por cada tipo de switch proporcionado. Para cada evolución se deberá indicar sólo los switches introducidos por aquella evolución, sin contar las evoluciones anteriores. En el caso de usar algún switch de una evolución anterior se deberá actualizar el número de puertos ocupados/libres.
D4	Se valorará la mejora en cualquiera de las entradas de la tabla 5. Se compararán los valores ofertados por cada solución. No se valorará en este punto mejoras ya valoradas o mencionadas de forma aparte de otra entrada. Para las redes de control de GPFS e interconexión MPI, cualquier valor de ancho de banda por link y nodo de cómputo requerido son los valores mínimos iniciales los cuales se deben incrementar de forma proporcional con la potencia de cálculo que cada nueva evolución o CTE incorpore.
R5	Todos los switches de segundo nivel de cualquiera de las redes deberán ser redundantes entre ellos, pudiendo evitar cualquier punto único de fallo.

Ref	Descripción
Red Interna cluster	
R6	Se deberá proveer del hardware necesario (switches, cables, etc.) para poder establecer la red interna del cluster con tecnología 1/10 Gigabit Ethernet. Todos los cables y fibras de esta red física que vayan a la misma velocidad deberán ser del mismo color y de un color diferente a las otras redes de la máquina, de tal manera que puedan distinguirse visualmente.
R7	Todos los puertos de cada tipo de un mismo switch deberán ser line-rate entre ellos sin ningún tipo de sobre-suscripción.

Ref	Descripción
R8	Requerimientos de funcionalidades de los switches de esta red: - Soporte Jumbo Frames (MTU > 9000) - Line-rate Nivel 2 switching - Line-rate Nivel 3 routing - Definición de Access-list - Routing (dinámico y estático) - Spaning-tree (MSTP y RSTP) - Capacidad para filtrar los paquetes BPDU a nivel de puerto físico del equipo - Port mirroring - Broadcast storm control - QoS - Snmp - SSH - Minimum 256 VLANs - LACP (Soporte hash LACP L3 + L4) - Flowcontrol - Soporte de más de 5000 MACs en la tabla de forwarding - 802.1Q - Fuentes redundantes y hot-swap - Ventiladores redundantes y hot-swap - MC-LAG (Multi-Chassis Link Aggregation Group) ó VLT (Virtual Link Trunking) al menos en el segundo nivel de la red
R9	En esta red física se configurarán 2 dominios de broadcast diferentes (2 VLANs): - 1 VLAN => Red interna cluster (DHCP, Boot, …) - 1 VLAN => Red gestión de elementos del cluster (IPMI, Switches, racks, …) que sólo será visible desde los servidores de gestión y será inaccesible desde los logins o cualquier nodo de cómputo. Esta VLAN puede necesitar enlaces a 1 Gbit en esta misma red física.
R10	En esta red se conectará: - Cada nodo de cómputo y login con una interfaz 1 Gbit Ethernet (Red cluster) - La interfaz de acceso mediante IPMI a cada uno de los nodos de cómputo. (Puede usarse la misma interfaz de 1 Gbit del nodo si se soporta con VLAN tagging, y si no se necesitará un enlace extra 1 Gbit) - Mínimo 2 enlaces de 10 Gbit para los servidores de gestión del cluster (VLAN interna cluster, VLAN gestión out-of-line) - Cualquier interfaz de gestión de cualquiera de los componentes del cluster (racks, IPMI Servers, puertas frías, PDU, switches, etc.)
R11	El primer nivel de switches de esta red deberá introducir una contención de 2:1, por ejemplo. Switches de primer nivel con 48 puertos de 1 Gbit con 2 uplinks de 10 Gbit Ethernet al nivel superior. La sobresuscripción a niveles superiores deberá venir determinada por las necesidades del diseño presentado.

Ref	Descripción
	Tanto la sobresuscripción como el nivel de switches deberá ser común desde cualquiera de los nodos de cómputo.
D12	Se valorará el diseño de la red presentado teniendo en cuenta conceptos como: - La redundancia en la caída de enlaces (up-links) entre switches. - Redundancia en la conexión de los diversos elementos a la red de management (servidores de servicio, nodos de cómputo, etc.) - La óptima o mejor distribución de la conexión de los elementos a los diferentes switches teniendo en cuenta los patrones de tráfico que esta red va a soportar y la sobresuscripción de la red presentada
Red de control GPFS	
R13	Se deberá proveer del hardware necesario (switches, cables, etc.) para poder establecer la red para el tráfico de control de GPFS con tecnología 1/10 Gigabit Ethernet a cada uno de los nodos de cómputo del cluster y logins. Todos los cables y fibras de esta red física que vayan a la misma velocidad deberán ser del mismo color y de un color diferente a las otras redes de la máquina, de tal manera que puedan distinguirse visualmente.
R14	Todos los puertos de cada tipo de un mismo switch deberán ser line-rate entre ellos sin ningún tipo de sobre-suscripción.
R15	Requerimientos de funcionalidades de los switches de esta red: - Soporte Jumbo Frames (MTU > 9000) - Line-rate Nivel 2 switching - Line-rate Nivel 3 routing - Definición de Access-list - Routing (dinámico y estático) - Spaning-tree (MSTP y RSTP) - Capacidad para filtrar los paquetes BPDU a nivel de puerto físico del equipo - Port mirroring - Broadcast storm control - QoS - Snmp - SSH - Minimum 256 VLANs - LACP (Soporte hash LACP L3 + L4) - Flowcontrol - Soporte de más de 5000 MACs en la tabla de forwarding - 802.1Q - Fuentes redundantes y hot-swap - Ventiladores redundantes y hot-swap MC-LAG (Multi-Chassis Link Aggregation Group) ó VLT (Virtual Link Trunking) al menos en el segundo nivel de la red
R16	En esta red física se configurará 1 dominio de broadcast (1 VLAN): Red de control de GPFS. (Misma VLAN que la de mismo nombre del Lote 1 del

Ref	Descripción
	concurso CONSU02016008OP)
R17	En esta red física se conectará: - Cada nodo de cómputo y login con una interfaz 1/10 Gbit Ethernet (Red de control de GPFS) Esta red se deberá conectar a nivel superior hacia la red de control de GPFS definida en el Lote1 del concurso CONSU02016008OP. Esta red deberá tener una topología de estrella, y los switches centrales o del nivel más alto de esta topología son los que se deberán conectar, mediante bondings a los switches de más alto nivel de los propuestos en el lote 1 del concurso CONSU02016008OP.
R18	Desde cada nodo de cómputo a la red de control de GPFS del Lote1 del concurso CONSU02016008OP deberá haber un máximo de sobresuscripción de 16:1. El primer nivel de switches de esta red deberá introducir una contención de aproximada de 2:1, por ejemplo. Switches de primer nivel con 48 puertos de 1 Gbit con 2 uplinks de 10 Gbit Ethernet al nivel superior. Tanto la sobresuscripción como el nivel de switches deberá ser equilibrado e igual desde cualquiera de los nodos de cómputo.
D19	Se valorará el diseño de la red presentado teniendo en cuenta conceptos como: - La redundancia en la caída de enlaces (up-links) entre switches. - Redundancia en la conexión de los diversos elementos a la red de management (servidores de servicio, nodos de cómputo, etc.) - La óptima o mejor distribución de la conexión de los elementos a los diferentes switches teniendo en cuenta los patrones de tráfico que esta red va a soportar y la sobresuscripción de la red presentada
D20	Se valorará que se implementen las 2 redes físicas ethernet (Red interna gestión y Red de control de GPFS), mediante una única red física basada en 10 Gbit Ethernet o superior definiendo las 3 VLANs encima de esa red física y con las conexiones necesarias de cada red. En este caso, obviamente no aplicaría el requisito R1 de este apartado de tener redes físicas disjuntas. En el caso de incluir la mejora D20, no aplica los requerimientos de bloqueos especificados en la entrada R18. Éstos cambiarían a: “Desde cada nodo de cómputo a la red de control de GPFS del Lote1 del concurso CONSU02016008OP deberá haber un máximo de sobresuscripción de 128:1 y que el primer nivel de switches de esta red deberá introducir una contención máxima de 8:1, por ejemplo: Switches de primer nivel con 48 puertos de 10 Gbit con 2 uplinks de 40 Gbit Ethernet al nivel superior”. Aun así, se seguirá aplicando los requerimientos de redundancia y equilibrio de la entrada R18, y se valorará el diseño presentado tal como indica D19.
Red Interconexión MPI / GPFS datos RDMA*	
R21	Se deberá proveer del hardware necesario (switches, cables, etc., su esquema

Ref	Descripción
	y etiquetado) para poder establecer la red interna de alto rendimiento y baja latencia sobre la cual se va enviar: - Comunicaciones MPI - Tráfico de datos GPFS RDMA (si fuera posible) * Esta red deberá ofrecer un mínimo de 100 Gbits por link Todos los cables y fibras de esta red física deberán ser del mismo color y de un color diferente a cualquier otra red de la máquina. La única excepción puede ser los cables de cobre que sólo se fabriquen en color negro. Para el resto de cableado se deberá cumplir ese requisito intra o inter rack.
R22	Todos los nodos de cómputo y logins deberán estar conectados a esta red de interconexión.
R23	* En el caso de no poder conectar la red de baja latencia a la que se propone en el lote 1 del concurso CONSU02016008OP (MADDR y MM servers), el tráfico de GPFS datos se enviaría por la red de control GPFS. En tal caso, la red de control de GPFS de forma obligatoria debería estar formada por links de 10 Gbit Ethernet o superior por cada nodo de cómputo. El deseable D20 dejaría de valorarse debido a que sería de obligado cumplimiento.
R24	Si se pudiera conectar las dos redes de baja latencia (cluster CTE y almacenamiento lote 1 del concurso CONSU02016008OP): Dicha conexión deberá ser directa (sin el uso de routers) y distribuida uniformemente entre los switches del nivel superior sin sobresuscripción hacia cada CTE. Haciendo que el rendimiento sea uniforme desde cualquier nodo de cómputo al almacenamiento y maximizando la alta disponibilidad en caso de fallo de cualquier switch. Se deberá de proveer de todo aquel hardware y servicios extra (switches, fibras, tareas de cableado) necesario para implementar estas conexiones, especialmente si se propone una tecnología diferente a la del Lote 1 del concurso CONSU02016008OP.
R25	La red deberá ser no-bloqueante full fat-tree entre todos los nodos de cada CTE
R26	Todos los switches de la red de baja latencia deberán poder ser gestionables desde la red ethernet interna del cluster en la VLAN de gestión de dispositivos de cada CTE.

Tabla 5 – Descripción hardware switches y redes CTE

<TECNOLOGIA EMERGENTE>	Evolución 1	Evolución 2	Evolución N
Red Interna cluster			
Número de switches proporcionados			
Marca switch			
Modelo switch			
Número de puertos 1GE por switch			
Número de puertos 10GE por switch			
Número de puertos 40GE por switch			
Número de puertos libres			
Latencia introducida por el switch			
Red Control GPFS			
Número de switches proporcionados			
Marca switch			
Modelo switch			
Número de puertos 1G por switch			
Número de puertos 10G por switch			
Número de puertos 40G por switch			
Número de puertos libres			
Ancho de banda por nodo computo			
Ancho banda por nodo / TF peak nodo			
Red MPI / GPFS datos RDMA			
Número de switches proporcionados			
Marca switch			
Modelo switch			
Número de puertos por switch			
Tecnología de conexión			
Ancho de banda por puerto (Gbit)			
Ancho banda por nodo / TF peak nodo			

3.- Adecuación infraestructura

En este apartado se expresan los requerimientos relacionados con las modificaciones de la infraestructura de capilla: eléctrica, hidráulica, aire acondicionado, seguridad, extinción de incendios, sistema de gestión o como cualquier otra que se deba modificar y adecuar para albergar los clusters de cómputo que se engloban en el proyecto MareNostrum4.

Ref	Descripción
Requerimientos adecuación infraestructura	
R1	La instalación y puesta en marcha de MareNostrum4 deberá ser "llave en mano" y deberá incluir cualquier obra, instalación, modificación de la infraestructura actual de capilla de Torre Girona para su instalación y operación óptima.
R2	Se deberá presentar un proyecto explicando las modificaciones necesarias en la infraestructura actual para albergar el MareNostrum4. Dichas modificaciones deberán cumplir con todos los requerimientos y certificaciones legales pertinentes según legislación vigente. Adicionalmente a esos requisitos, se deberá proveer: - Diagrama de Gantt específico con las tareas de adecuación de la infraestructura descritos en este apartado con los tiempos estimados y los cortes de servicio esperados - Descripción en detalle a nivel técnico de cada una de las tareas de la adecuación - Cálculo a nivel eléctrico y frigorífico de las modificaciones sugeridas para el óptimo funcionamiento de MareNostrum4, teniendo en cuenta el límite de la climatización existente de 1300KW por capacidad de los intercambiadores y dimensión de tubería de distribución.
R3	El proyecto deberá incluir la puesta en marcha de las instalaciones, incluyendo protocolos de pruebas y curso de formación sobre las modificaciones realizadas al equipo técnico de mantenimiento del BSC.
D4	El plan de modificaciones presentado será valorado tanto a nivel técnico como por los mínimos cortes de producción que produzcan, y el mayor reaprovechamiento de materiales e infraestructura existente. De la misma manera, se valorará cualquier mejora presentada en las instalaciones para su futura ampliación.
R5	Se deberá realizar cualquier modificación necesaria del sistema de gestión de infraestructura actual (BMS – Niagara System) para adaptarse a los cambios devenidos de la instalación de MareNostrum4.
R6	Se deberá reemplazar cualquier elemento actual de la infraestructura por mal funcionamiento o deterioro, debido a las obras de adecuación a realizar.
R7	La empresa licitadora deberá hacerse cargo de coordinar con las empresas de mantenimiento de las instalaciones ya existentes (en mantenimiento actualmente hasta 26 de marzo del 2018) y proporcionar los servicios de

Ref	Descripción
	mantenimiento preventivo de la infraestructura actualizada de la capilla de Torre Girona, una vez modificada para albergar MareNostrum4, mientras dure el proyecto/arrendamiento del superordenador MareNostrum4. Corresponderá a la empresa licitadora la interlocución única de todo el mantenimiento con el BSC.
R8	En caso de incidencia urgente en la infraestructura se deberá personar un técnico competente en cuestión de 2 horas.
R9	Se incluirán los años de garantía adicionales necesarios, en todos los nuevos equipamientos instalados, para cubrirlos hasta el fin del proyecto de MareNostrum4. Garantía de equipamientos basada en la legislación en vigor condicionándose al buen uso y mantenimiento de las instalaciones.
R10	Se deberá re-etiquetar cualquier elemento de la infraestructura (CETACs, magnetotérmicos, etc.) para adecuarse a MareNostrum4. Mirando de mantener la nomenclatura de racks existente en el BSC: - Cx – Rack de cómputo - IBx ó OPAx – Rack de red de baja latencia - Mx – Rack de Management (bajo SAI)
R11	Se deberán de sustituir todas las baldosas de suelo técnico de la urna de capilla por unas nuevas. Dichas baldosas deberán ser compatibles con la estructura “Heavy-Duty” instalada. Las baldosas deberán cumplir con las características siguientes: - Lado 600 mm +/- 0,2 - Diagonal 848,5 mm +/- 0,3 - Espesor sin recubrimiento 30 mm +/- 0,2 - Resistencia eléctrica inferior a 10^7 ohmios - Resistencia por encima de 2000 Kg/ m2 El BSC deberá poder elegir el acabado superior de las baldosas para que esté acorde con la estética de la instalación. Una vez todas las baldosas sean cortadas, troqueladas e instaladas, se deberá de proveer de un 5% extra de baldosas del mismo modelo para futuros cambios. Aparte de baldosas cerradas se deberán de proveer baldosas del mismo tipo troqueladas para garantizar el paso de flujo de aire, según la necesidad de la máquina a instalar. También se deberán de proveer 6 baldosas de vidrio para poder mostrar el falso suelo.
R12	El BSC proveerá aquella información necesaria de la infraestructura actual de capilla de Torre Girona a las empresas interesadas a licitar para la preparación de este apartado. En https://bts.bsc.es/uwjE7mBD y con las credenciales Username: uwjE7mBD Password: bS;_9Z3g Se puede encontrar: • Descripción general de la infraestructura de la capilla

Ref	Descripción
	• mapas de esquemas eléctricos y mecánicos en alta resolución formato PDF
R13	Se minimizará el tamaño de los troquelados/cortes de las baldosas para el paso de cables y tubos, para minimizar el escape de aire frío del falso suelo y por visibilidad. Los troquelados/cortes, siempre que se pueda, se harán contra un lateral de las baldosas para poder retirar la baldosa sin tener que descablear los racks. Cualquier corte de baldosa se deberá proteger con material “armaflex encolado” para evitar el corte y deterioro de fibras y cables de cobre
R14	Se deberá de conectar todos los nuevos racks a la red de tierra equipotencial existente. O la modificación dicha red si hiciera falta.
R15	Reparación del solado y pintado con pintura plástica impermeable antipolvo Procolor o similar, del mismo RAL existente.
R16	Sustitución de las baldosas diferentes existentes a las originales en sala técnica, instalación y nivelación de las baldosas existentes en buen estado recuperadas en URNA a sala técnica para su correcto mantenimiento. Retirada de las baldosas existentes y estructura de falso suelo dañada o deteriorada, a un vertedero oficial para el reciclaje de residuos. Retirada de las baldosas en buen estado a un espacio que determine el BSC.
R17	Retirada de cualquier instalación y material fuera de uso, por las modificaciones necesarias de la infraestructura a un vertedero oficial para el reciclaje de residuos.
R18	Necesidades de seguridad y salud de la obra, protecciones individuales, protecciones colectivas, elementos de higiene, casetas de obra para el acopio de materiales, vestuarios, W.C., etc. Se deberá valorar todo lo necesario para la correcta ejecución y funcionamiento de la obra, según la normativa vigente.
R19	Contratación del coordinador de seguridad y salud, durante la ejecución de la obra, para asegurar el cumplimiento de los medios de prevención de riesgos laborales, según normativa vigente e indicación del BSC.
R20	Limpieza final de obra, incluyendo la recogida diaria de residuos para el reciclaje a un vertedero oficial.
R21	Mientras se realice la modificación de la infraestructura, se puede dar el caso de que en la urna de capilla siga en funcionamiento y producción algunos racks de la primera fila (Almacenamiento Lote 1 del concurso CONSU02016008OP); como alguno de los racks de los CTE ya entregados. Dentro del plan de adecuación se tiene de contemplar dicha circunstancia, para su aislamiento durante las obras.
R22	Con la finalización del proyecto de adecuación se deberá presentar la siguiente documentación: - Memoria descriptiva de las instalaciones reformadas - Planos de todas las instalaciones (Instalación de Media Tensión, Instalación de Baja Tensión, Instalación de Climatización, Instalación de Protección Contra

Ref	Descripción
	Incendios “PCI”, Instalación de Sistema de Gestión “BMS”, Instalación de Detención de fugas de agua, Instalación de cableado de red, Instalación de falso suelo) en formato: AutoCAD 2000 o superior y PDF. - Planos Asbuilt de las instalaciones (Instalación de Media Tensión, Instalación de Baja Tensión, Instalación de Climatización, Instalación de Protección Contra Incendios “PCI”, Instalación de Sistema de Gestión “BMS”, Instalación de Detención de fugas de agua, Instalación de cableado de red, Instalación de falso suelo) en formato: en papel, AutoCAD 2000 o superior y PDF. - Hojas de cálculo de las instalaciones, (Instalación de Media Tensión, Instalación de Baja Tensión, Instalación de Climatización, Instalación de Protección Contra Incendios “PCI”). - Manual de mantenimiento para cada instalación que incluya como mínimo: Descripción de la instalación, operaciones de mantenimiento, plan de mantenimiento, precauciones que deben tomarse al realizar dicho mantenimiento, y certificado de garantía del contratista de toda la instalación.
R23	Antes de empezar con la adecuación de la infraestructura, la empresa contratista del concurso se deberá hacer cargo de la catalogación, recogida de todo el cableado hacia cada rack de cómputo y preparación para envío de todos los racks actuales que existen en la urna de capilla (a excepción de los racks que conformen el almacenamiento del Lote1 del concurso CONSU02016008OP). No se podrá reutilizar ningún componente de estos racks actuales para la nueva solución. Los racks actuales que existen en capilla son: 37 racks de computo de tipo idataplex (1200 kg por rack) 4 racks standard 42"" (800kg por rack aprox.) 8 racks standard 42"" (800kg por rack aprox.)" De la misma manera, cubrirá también el transporte a un almacén en Madrid designado por el BSC y su estancia en dicho almacén durante 3 meses El contacto del almacén, donde el BSC ya tiene almacenado otros materiales, es: HTM Dpto. Comercial Calle los Frailes, 52 28814 Daganzo de Arriba Madrid Tfno: +34 902 052 591 Fax: +34 902 877 629 comercial@htmsl.com www.hightechmovements.com Es responsabilidad de la empresa licitadora preguntar este coste.

Ref	Descripción
R24	En el momento de la retirada de todo el material de MareNostrum3 de la urna. Se deberá mover los patch panels de fibra y cobre situados en el rack M2, reinstalándolos en los nuevos racks que vayan en esa posición. En el caso de rotura se deberán refusionar las fibras o cables rotos.

4.- Operacional

En este apartado se describen los requerimientos operacionales relacionados con los clusters de cómputo (propósito general y tecnologías emergentes) del proyecto MareNostrum4.

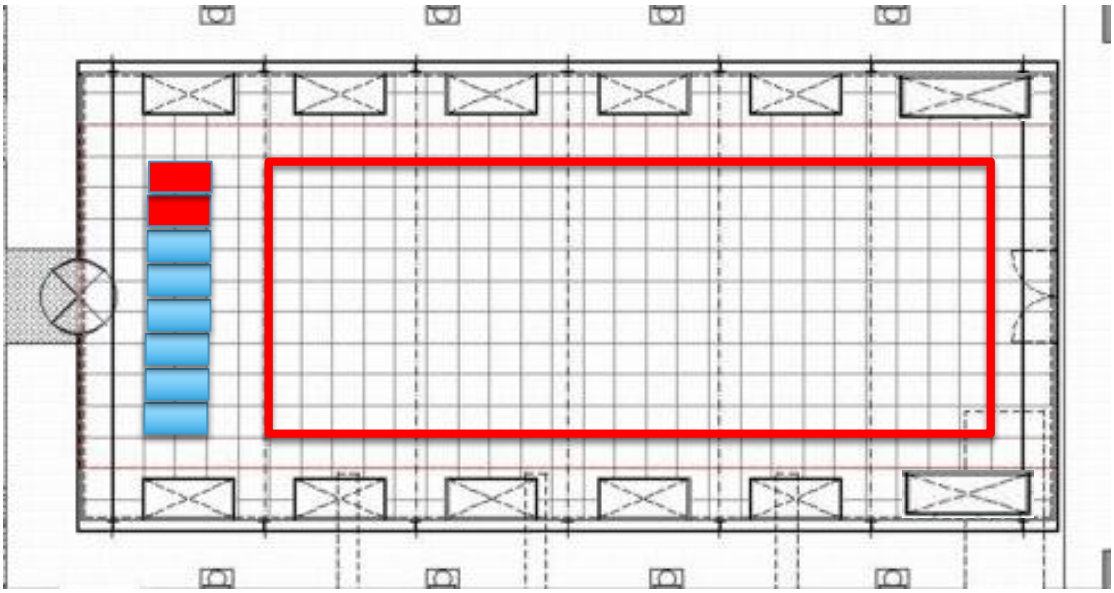
Ref	Descripción
Requerimientos operacionales	
R1	Los racks deberán venir incluidos con la solución y deberán incorporar las PDU's adecuadas para conectar todos los equipos de la solución y proporcionar redundancia en la circunstancia de la caída del 50% de las PDUs de cada rack, redundancia N+N, sin ninguna pérdida de rendimiento. En el caso de ofrecer PDUs monitorizables ó gestionables se tendrá de integrar dentro de la red interna del cluster VLAN de gestión.
R2	Los racks deberán incorporar refrigeración dentro del rack que elimine como mínimo el 95% del calor generado en caso de puerta trasera o mínimo del 80% en el caso de direct-liquid cooling. En caso de puertas traseras gestionables deberá de conectarse a la red interna del cluster VLAN de gestión.
R3	Una vez instalados los clusters de cómputo, la temperatura de la urna de MareNostrum deberá ser lo más constante posible en todo su volumen, no podrán existir "puntos calientes" que afecten a la temperatura de entrada de los nodos de cómputo (inlet temperature). Dadas todas las temperaturas inlet de todos los nodos de cómputo en funcionamiento no podrá existir una diferencia superior a 8ºC entre la mínima y la máxima.
R4	La capacidad frigorífica que la infraestructura de la urna de capilla del BSC es capaz de proporcionar es de un máximo de 1300kW (CRAHs + HXB). Se debe demostrar que los clusters planificados de instalar en la urna de capilla pueden refrigerarse con dicha capacidad con carga de CPD que se considera al 70% del pleno rendimiento (ejecución HPL).
R5	El peso de cada rack no deberá nunca superar más de 2000 Kg. x m2. Los racks deberán poder entrar en la urna de capilla de forma vertical.
R6	Se deberá presentar en la documentación un esquema frontal con la ocupación de los racks de los diversos equipos presentados en la solución. En ella se deberá claramente especificar el hardware ofertado, como las U's que ocupa cada componente de la solución. También se deberán de especificar por cada tipo de rack el número de cables/fibras que sale de cada rack para cada una de las redes definidas.
R7	Todos los nodos de cómputo y resto de componentes de la solución deben disponer de fuentes de alimentación redundadas N+N
R8	Se deberá presentar esquema de conexionado eléctrico interno de cada tipo de rack de la solución. Mostrando la redundancia de todos los elementos a nivel de alimentación a través de diferentes grupos de PDU.

Ref	Descripción
	Se deberá realizar la conexión de elementos a PDUs para que haya una repartición uniforme entre las diversas fases eléctricas.
R9	Se deberá presentar un floor plan de toda la solución, indicando el espacio ocupado. Teniendo en cuenta que el espacio máximo son 120m2 descontando una fila de 8 racks de 42U estándar. Mirar esquema núm. 1. Los clusters de tecnologías emergentes y sus evoluciones planificadas hasta enero 2018 se deberán instalar en la urna de capilla con las siguientes condiciones: - No deberán superar los 3 racks standard en total - Cada actualización/evolución de CTE deberá ocupar un máximo de medio rack (21U) A partir de enero 2018 los CTE y sus evoluciones planificadas se podrán instalar en otro CPD fuera de la capilla de Torre Girona, en zona próxima a la misma, sin consideración de limitación de consumo eléctrico ni de espacio de ningún tipo. Aquellos CTE que sean instalados fuera de la capilla de Torre Girona se deberá indicar por cada CTE/Evolución: - m^2 de ocupación, incluyendo zona de servicio - Consumo típico CPD esperado (70% de la ejecución de HPL) - Tipo de conexiones requeridas: eléctrica, refrigeración, conexiones de red para acceso a sistema de ficheros, etc. En el floor plan se deberá indicar los racks necesarios para el cumplimiento de los mínimos y cuales son parte de las mejoras ofrecidas. En el proyecto de implantación, el floor plan de la urna propuesto podrá ser modificado por el BSC, sin que cualquier sobrecoste del cambio lo tenga que asumir el BSC.
D10	Se valorará los m^2 mínimos usados por el cluster de propósito general según la distribución de racks presentada para cumplir los mínimos establecidos (9.5 PFlops) y respetando los espacios mínimos de servicio de todos los elementos, y dejando baldosas registrables entre filas de racks. Mirar esquema núm. 1.
D11	Se valorará del diseño del floor plan de los racks en capilla teniendo en cuenta su disposición para la visibilidad de las partes más destacadas desde la entrada de la exclusa, la parte trasera de la sala y el puente de las visitas encima de la urna. Se valorará una proyección en 3D o dibujo de cómo quedaría la máquina con el floor plan propuesto.
R12	Según esquema núm. 1. La primera fila se podrá usar hasta 2 racks estándar 19” los cuales dispondrán de alimentación a SAI (máximo de 15 KW por rack, en media) y donde será obligatorio situar los siguientes elementos: - Todo elemento de gestión (servers, switches centrales de la red interna y red de control de GPFS) del cluster de propósito general para su funcionamiento - Nodos de logins del cluster de propósito general
R13	Se deberá presentar otro floor plan mostrando como se piensa realizar el cableado de cada una de las redes entre racks por el falso suelo, por las

Ref	Descripción
	bandejas “rejiband” asociadas (cualquier cambio de diseño de las bandejas actuales deberá estar descrito en el apartado 3 de adecuación de la infraestructura).
R14	Todo movimiento de racks dentro de la urna de capilla se deberá hacer con la colocación de láminas para no marcar las nuevas baldosas de suelo técnico.
R15	Todo conexionado entre racks se deberá realizar a través del falso suelo, no se permitirá la tirada de cables entre racks colindantes o por la parte superior del rack. El cableado dentro del rack deberá ser ordenado y nunca salir del espacio que determina la planta del rack.
R16	El consumo máximo de la solución instalada en la urna de capilla con carga CPD (considerado como el 70% del máximo consumo ejecutando HPL) (Cluster de Propósito General, Sistema de ficheros y todos los CTE instalados hasta Enero 2018) no podrá ser superior a 1.3 MW
R17	Se deberá poder integrar con el sistema de monitorización de los clusters los valores del entorno de la sala. Temperatura de los nodos, humedad, etc. Pudiendo definir alertas y avisos en casos de cualquier problema
R18	La máquina deberá disponer de un sistema de monitorización de la temperatura de tal manera que provoque la parada controlada del sistema en caso de temperatura muy alta.
R19	Se exigirá en la instalación cableado (fibra,eléctrico,cobre Ethernet, etc.) ordenado, elegante y vistoso debido a que quedará a la vista. Todos los racks deberán instalarse sin puerta frontal.
R20	Todo cable o fibra que forme parte de la misma red y tecnología deberá ser del mismo color en toda la máquina y entre cualquiera de los diferentes componentes hardware que formen esa red. Cada red física deberá usar un color diferente entre ellas.
R21	Todo componente de la solución (rack, server, switch, cable, fibra, …) deberá ir debidamente etiquetado, para ser identificado físicamente de forma única según nomenclatura que se establezca entre el BSC y la empresa instaladora. En los cables y fibras se deberá indicar origen y destino de la conexión.
R22	La solución deberá incluir el montaje en racks de toda la solución, además de la recogida de todos los materiales sobrantes de la instalación.
R23	Cada uno de los racks se deberá entregar con todos sus componentes enrackados y con el cableado intra-rack completamente realizado y completamente optimizado para la refrigeración de todos los componentes y el fácil acceso a los diversos componentes para su sustitución. Todos los nodos de cómputo deberán haber pasado un burn- in test en fábrica para evitar los DOA (Dead on Arrival).
R24	Se deberá presentar un diagrama de Gantt especificando y describiendo las tareas y el tiempo estimado en las mismas sobre la instalación de los clusters de cómputo (CPG y CTE). Este diagrama de Gantt será complementario al que se pide en el apartado 3 sobre la adecuación de la infraestructura. Este

Ref	Descripción
	diagrama deberá cubrir desde la llegada del hardware hasta la puesta en producción de cualquiera de los clusters, según las indicaciones y requerimientos expresados en el apartado 7 de condiciones de aceptación.
R25	Cualquier U vacia en cualquiera de los racks deberá taparse frontalmente con tapas ciegas.

Esquema núm. 1- Planta Capilla



- Los racks azules pertenecen al storage del BSC ampliado, según lote 1 del concurso CONSU02016008OP
- Los racks rojos son los racks con acceso a SAI a poder usar por el CPG (servidores de gestión, almacenamiento de gestión, logins, switches centrales red interna y control de GPFS, …)
- El área roja marcada es la zona a poder usar para instalar el resto de componentes del superordenador que no requieren SAI (nodos de cómputo, switches red MPI, switches leaf de las otras redes, CTE (hasta Enero 2018), …)

5.- Software

En este apartado se describe el software a proporcionar en los diversos clusters de cómputo del proyecto MareNostrum4 (CPG y CTE). Si algún componente sólo es para alguno de los dos tipos de clusters se expresará explícitamente sino se entiende que afecta todos los clusters, y que se debe proporcionar por separado para cada uno de los clusters ofrecidos.

Ref	Descripción
R1	El sistema operativo deberá ser UNIX like y compatible con el X/Open Standard POSIX 1003 (IS/IEC 9945). El sistema operativo deberá ser Linux, todos los componentes deberán llevar la misma versión de sistema operativo. Dicho sistema operativo deberá proporcionar soporte Enterprise y estar soportado por cualquiera del resto de componentes del software stack de la máquina: Sistema de clustering, sistema de colas, sistema de ficheros, compiladores, drivers, etc.
R2	El Linux proporcionado deberá tener una versión de kernel que soporte nativamente mediante módulo las siguientes herramientas de traceo: - RAPL - LTTng - PEBS
R3	Se deberá aportar también todo el software necesario para la gestión de todos los componentes que formen la solución: Switches, etc.
R4	Cada cluster deberá incorporar un software de clustering como, por ejemplo, xCAT, que realice la gestión de todos los elementos del cluster y los servicios básicos del mismo. Dicho software de clustering deberá ofrecer y/o implementar entre otras características: - Una única imagen de sistema operativo para los nodos de computación que pueda ser mantenida y que los cambios se distribuyan de forma automática a todos los nodos del cluster. - Arranque y parada de los nodos de cómputo - El arranque completo de la máquina debe realizarse en menos de 20 minutos - Los nodos de cómputo del CPG deben arrancar por red, teniendo su rootfs en remoto ya sea via NFS u otra metodología, como el modo statelite de xCAT. - Los diferentes servidores que proporcionan los servicios de clustering deben estar configurados en alta disponibilidad, el fallo de uno no se debe ver reflejado en el funcionamiento normal del sistema, ni en ninguno de los nodos de cómputo de los que sea responsable. - Definición mediante reglas y/o expresiones regulares de los diversos DNS, IPs y alias del cluster, bajo las premisas y requerimientos propuestos por el equipo técnico del BSC, y la población automática de la configuración de DNS, /etc/hosts, etc.

Ref	Descripción
	- Consulta de valores del entorno de los nodos de cómputo, como puede ser: temperatura, velocidad ventiladores, voltajes, etc. mediante un comando de forma centralizada - Eliminación o sustitución de nodos del cluster - Recolección y filtrado de las alarmas de todos los componentes de hardware del cluster mediante SNMP traps, posibilidad definir acciones dependiendo de los traps recibidos. - Consulta centralizada de los eventos históricos registrados en el sistema out-of-line por cada nodo de cómputo: Power on/off, errores hardware preventivos, etc. - Consulta y generación/actualización automática del inventario de hardware de todos los nodos de forma centralizada. (Números de serie, modelos de dimms, tarjetas, etc.) - Definición de diversos grupos de nodos de cómputo, posibilidad de lanzar comandos de forma paralela mediante la herramienta de clustering a dichos grupos. - Comando para consultar/cambiar la configuración del BIOS/UEFI (Boot device, HyperThreading/SMT configuration, IPMI IP, etc.) de los nodos de cómputo de forma centralizada y paralela. - Estructura jerárquica de administración, con 2 servidores centrales y varios servidores que se encargan de la gestión de un subconjunto del cluster. Visión única del cluster desde los servidores centrales. - Gestión centralizada de consolas y recolección de logs - Toda operativa de la herramienta del clustering deberá ofrecerse como mínimo por línea de comandos - Discovery y auto-configuración de nodos de cómputo en el cluster según reglas y puerto de switch.
R5	Con el sistema operativo se debe incluir todo el entorno de programación para la arquitectura de la máquina, como mínimo deberá incluir C, C++, Java, Fortran. A parte del entorno de programación Open-Source proporcionado por el sistema operativo, se deberá de proporcionar el entorno de programación específico para la arquitectura del procesador proporcionada. Para los nodos de los clusters de tecnologías emergentes se deberá de proveer los lenguajes y el entorno de programación adecuado para poder programarlos mediante los paradigmas estándar, según su arquitectura: por ejemplo, para las aceleradoras Nvidia, el soporte para CUDA, OpenACC y OpenCL, y para otros aceleradores, cualquier lenguaje propio más OpenCL. En todos los casos se deberá dar soporte para lenguajes de programación C, C++ y Fortran.
R6	Los compiladores de los diversos procesadores ofertados deberán venir con licencias flotantes con tantas licencias como logins existentes de ese tipo.
R7	Se deberán de proporcionar las librerías numéricas (secuenciales y paralelas) proporcionadas por el fabricante de los procesadores debidamente optimizadas para cada arquitectura. Como pueden ser MKL ó ESSL/pESSL. Se

Ref	Descripción
	deberán aportar actualizaciones de librerías con nuevos sistemas.
R8	También se deberá proporcionar los compiladores, librerías y/o las herramientas necesarias para el uso paralelo de la arquitectura mediante paradigmas estándares como OpenMP ó MPI. Para OpenMP deberá soportar la versión 3.1, y para MPI se deberá soportar completamente el estándar MPI versión 3.0. A parte, de la versión open-source se deberá proveer de una implementación especializada en la arquitectura propuesta en el caso que exista, como por ejemplo Intel MPI, Spectrum MPI o similares.
R9	Las librerías paralelas para el uso de MPI deben ser optimizadas para el uso de la red de baja latencia ofertada para cada uno de los clusters ofrecidos.
R10	Cualquier de los softwares anteriores mencionados deberán ser compatibles con las herramientas de traceo que desrrolla el BSC. (https://www.bsc.es/computer-sciences/performance-tools)
R11	Se deberá incluir un software de sistema de colas por cada cluster que permita el envío de trabajos batch a la máquina y su uso normal de producción, coordinado con el sistema de gestión del cluster, como por ejemplo Slurm. Dicho sistema deberá soportar como mínimo: - Ejecución prólogo, epilogo y spawn de procesos paralelos, escalable a decenas de miles de cores por job - Configuración de prioridades basadas en fair-share. Pudiendo definir más de 2 niveles dentro del árbol de fair-share y pudiendo asignar la cuota de horas asignadas a un proyecto como último valor del árbol de fair-share - Definición de reservas puntuales y regulares, sin la necesidad de especificar la lista exacta de nodos o parar el scheduling para su creación - Accounting por job a nivel de walltime y de consumo eléctrico - El sistema de colas deberá ser compatible con las herramientas de monitorización de HPC del BSC, como, por ejemplo, Slurm. - El sistema de colas deberá soportar y gestionar los diversos recursos de las tecnologías existentes en los clusters de tecnologías emergentes - Sistema de plugins para poder añadir características como el lanzamiento de Jobs gráficos X11, integración con elasticsearch/grafana - Alocatación de Jobs teniendo en cuenta la topología de la red de baja latencia - Debe ser capaz de limitar los recursos a usar dentro de un nodo mediante limites o cgroups. - Poder cambiar la frecuencia de funcionamiento de los procesadores por job, para hacer power-aware scheduling
R12	Se deberán incluir el software y las licencias de GPFS cliente para todos los nodos de cómputo y logins ofertados para todos los clusters de cómputo, para poder conectarse al almacenamiento descrito en el Lote 1 del concurso

Ref	Descripción
	CONSU02016008OP.
R13	Se requerirá la inclusión de debuggers paralelos, como pueden ser DDT o Totalview, con licencia de uso con un mínimo de 1024 cores. Dichas licencias deberían ser flotantes a poder ser usadas desde cualquier cluster y deberá soportar las tecnologías ofrecidas tanto en CPG como CTE.
R14	En el proyecto se deberá de incluir la instalación de un sistema de monitorización por cluster ofertado, deberá estar basado en tecnología compatible con la que usa actualmente en el BSC, como por ejemplo ganglia. Dicho sistema deberá recoger métricas de todos los elementos físicos y lógicos de los nodos de cómputo (uso cpu, ocupación memoria, GPFS, uso de las redes, etc.). La misma herramienta deberá poder mostrar gráficas históricas de subgrupos o globales del cluster sobre cualquier métrica, pudiendo configurar hora inicio y fin.
R15	Se deberá incluir software de gestión y monitorización de la red de interconexión MPI que permita de forma centralizada: - Localización de “soft failures”. - Links con failure rates por encima de lo deseado - Alarmas y detección de errores graves dentro de la red - Mostrar la carga de tráfico a nivel real por cada link y a nivel global - Poder seleccionar un subconjunto de nodos y realizar una monitorización de los mismos
R16	En el proyecto se deberá incluir un sistema de alertas por cluster, como, por ejemplo, nagios o similar. Que compruebe la disponibilidad de todos los componentes de administración del cluster y genere alertas vía email.

6.- Mantenimiento y soporte

Ref	Descripción
R1	El arrendador o empresario asumirá durante el plazo de vigencia del contrato de arrendamiento la obligación del mantenimiento del objeto del mismo (hardware y software). Delante de fallos hardware se deberán reparar con una respuesta en 4 horas dentro de las horas de oficina (08:00 – 17:00) y con un servicio de soporte de Next Business Day. En caso de incidencias muy críticas que impliquen una afectación global de la producción de los clusters, se deberá proveer un seguimiento continuo 24x7 hasta la resolución de la incidencia.
R2	La empresa licitadora se hará cargo de la reparación y sustitución durante el periodo del proyecto MareNostrum4 de cualquier componente hardware de los clusters de cómputo.
D3	Una vez acabado el proyecto de MareNostrum4, se valorará la extensión de la garantía / mantenimiento tanto en años de duración como cobertura.
D4	Se valorará que la licencia del sistema operativo sea del tipo: Site License, para que cubra sistemas operativos para otras máquinas del propio BSC o de la RES.
R5	El proyecto de instalación incluirá la comprobación del buen funcionamiento, integración y óptimo rendimiento de la solución.
R6	Se exigirá un trabajo en equipo con el departamento de operaciones del BSC, para la coordinación de todas las tareas de este pliego. Cualquier plan o toma de decisión se deberá verificar con el departamento de operaciones del BSC antes de llevarla a cabo.
R7	Se proporcionará (dentro de período del proyecto MareNostrum4): - Acceso a todo el software upgrade (incluyendo sistemas operativos, clientes GPFS y firmware) de todos los componentes de la solución - Punto único de suporte para el aviso de problemas e incidencias de cualquier componente que componga la solución
R8	Se exigirá soporte pro-activo, notificando y recomendado subidas de versión tanto de software como de firmware de cualquier componente de la solución.
R9	Se deberá entregar al final de la instalación una documentación digital en la que se describa: - Descripción general de los componentes de la solución - Esquema de conexionado físico e IPs - Valores de configuración empleados - Explicación del proceso de instalación y tareas realizadas - Explicación procedimientos para: Puesta en marcha, y disaster recover
R10	Toda la instalación y desarrollo del proyecto se deberá hacer on-site en las instalaciones del BSC bajo la supervisión del grupo de sistemas del BSC. En ningún caso se permitirá el acceso externo o remoto para la configuración o instalación de la solución presentada.

	Durante el mantenimiento no se permite el acceso remoto y toda modificación se debe hacer on-site.
D11	Se valorará la existencia de un remanente de stock de piezas de recambio on-site para la pronta resolución de problemas hardware.
R12	Se deberán ofrecer formación durante la instalación de la solución, que cubran: - Conceptos básicos - Administración básica y procedimientos básicos de configuración - Optimización de la solución - Solución de problemas
R13	En la implantación de la solución presentada se exigirá la participación activa y presencial (si se requiere) de los expertos de cada uno de los componentes que forman la solución: - Responsables de hardware/ desarrolladores de firmware - Desarrolladores o responsables técnicos de software de clustering - Desarrolladores o responsables técnicos de sistema de colas - Desarrolladores o responsables técnicos de redes o switches ethernet - Desarrolladores o responsables técnicos de red MPI ofertada - Desarrolladores o responsables técnicos de compiladores, entornos de ejecución paralela Teniendo la posibilidad el personal del BSC poder intercambiar emails de forma directa con dichas personas con el fin de solucionar cualquier problema que surja durante el desarrollo e instalación de la máquina.
R14	El equipo técnico encargado de la instalación hardware y software deberá disponer de la formación y capacidades técnicas para la realización de este tipo de instalaciones, ya que es imprescindible para la correcta ejecución del contrato. Con lo que deberán disponer experiencia en la instalación de clusters de la misma envergadura, es decir, como mínimo de unos 1000 nodos físicos por cluster. Se deberá aportar documentación que lo acredite, incluyendo listado de personal, Curriculum Vitae y funciones. Se propone la tabla 6 como ejemplo de la información mínima a proporcionar.
R15	De la misma manera, para las tareas de mantenimiento hardware del superordenador una vez en producción, la empresa licitadora deberá disponer de un equipo de personas suficiente para el desarrollo y asistencia en la cercanía de Barcelona, con posibilidad de personarse en las dependencias del BSC en menos de 2 horas. Se deberá describir el número de personas, Curriculum Vitae de dicho equipo y el perfil técnico o responsabilidades de las mismas, el cual será evaluado. Se propone la tabla 6 como ejemplo de la información mínima a proporcionar.
D16	La información y curriculums de los equipos técnicos proporcionados serán valorados, así como las instalaciones por encima de 1000 nodos que hayan participado (se considerará más favorable las instalaciones con mayor número de nodos).
R17	Para la realización del cálculo de los diversos diagramas de Gantt se deberá considerar jornadas de trabajo de 8 horas diarias de lunes a viernes.

Tabla 6.- Ficha persona de Equipo técnico

Concepto	Valor
Nombre	
Empresa	
Perfil o especialidad	
Perteneciente a Equipo de instalación o equipo de mantenimiento hardware	
Años trabajando en la empresa actual	
Listado de otras empresas donde ha trabajado	
Instalaciones realizadas (>1000 nodos x cluster):	
- Cliente donde se realizó	
- Número de nodos de la instalación	
- Año realización instalación	

7.- Condiciones de aceptación

En este apartado se listan las condiciones a cumplir para la aceptación de cada uno de los clusters de cómputo (Propósito general y de tecnologías emergentes), para considerar que están listos para su puesta en producción, como las condiciones para no incurrir en penalización.

Ref	Descripción
R1	Se deberá demostrar el rendimiento de cómputo y la escalabiliad con benchmarks sintéticos como son: HPL, IMB (Pallas Benchmark) y Stream. Se deberá aportar código, compilación y experiencia en ejecución de dichos benchmarks. Se comprobará la correcta ejecución como que muestre el rendimiento esperado según exprese el fabricante de la cpu como el de la red de baja latencia. Estos benchmarks se deberán ejecutar en un subconjunto de los nodos del CPG como por cada CTE/Evolución.
R2	Para aceptar el cluster de propósito general se deberá ejecutar con un mínimo de 1024 cores toda la benchmark suite del BSC que contiene entre otros programas: AMBER, GROMACS, NAMD, WRF, NEMO y VASP, con sus respectivos inputs. Estas ejecuciones deberán ejecutar correctamente y con una escalabilidad superior al 60% hasta 1024 cores. En la aceptación de la máquina se realizarán ejecuciones con un número de cores de hasta 1024, por cada una de las ejecuciones se deberá proporcionar una escalabilidad respecto a las anteriores (ejecutadas con menos cores) de como mínimo un 60%. Esta eficiencia en escalabilidad se medirá con el speedup de la ejecución. El test de aceptación se evaluará respecto a la ejecución con menos cores posible, según la configuración de GB por core que cada máquina proponga, debido al tamaño de los inputs. El código fuente de estos benchmarks está disponible en internet, y se debe aportar la experiencia de compilación y ejecución de los benchmarks. Los códigos, optando siempre por la última versión estable disponible, se pueden obtener desde las páginas web oficiales de cada código: http://ambermd.org/ http://www.gromacs.org/ http://www.ks.uiuc.edu/Research/namd/ http://www.wrf-model.org/index.php http://www.nemo-ocean.eu/ https://www.vasp.at Los inputs que se deben utilizar están disponibles en https://bts.bsc.es/uwjE7mBD y hacer login con las credenciales Username: uwjE7mBD Password: bS;_9Z3g
R3	Para comprobar su funcionamiento óptimo también se deberá ejecutar un conjunto de benchmarks standard de tecnologías BigData que proporcionará el BSC basados en tecnologías hadoop, spark y cassandra entre otras

	tecnologías. Los benchmarks de Big Data a realizar serán los benchmarks descritos en el repositorio del BSC de benchmarks de Big Data: aloja.bsc.es y el benchmark suite para spark (https://github.com/SparkTC/spark-bench)
R4	Una vez acabada la instalación se deberá de comprobar que todos los requerimientos de operativa (apartado 4) establecidos en este pliego se cumplen. Como, por ejemplo, sin estar limitados a: - Redundancia eléctrica allí donde se requiera - Adecuación y optimización de la infraestructura - Repartición de las fases eléctricas - Cableado óptimo para el flujo de aire y refrigeración de todos los componentes - Etc.
R5	Se deberá comprobar el funcionamiento óptimo del superordenador con el sistema de almacenamiento del BSC, especialmente con total compatibilidad a nivel de red de baja latencia.
R6	Se deberá comprobar el funcionamiento óptimo de todos los componentes de la solución y demostrar empíricamente que cumplen los rendimientos (GB/s, IOPS, PFlops, …) ofertados, como la totalidad de las funcionalidades descritas en este pliego, tanto en el apartado de hardware como de software de los clusters.
R7	El cluster de propósito general deberá estar en producción para los usuarios y habiendo pasado todas las condiciones de aceptación (apartado 7) antes del 1 de julio de 2017.
D8	Se valorará una mejora sobre la fecha prevista de puesta en producción del cluster de propósito general, respaldado por el diagrama de Gantt requerido en el apartado “Operacional”
R9	Se deberá haber entregado al final de la instalación de cada cluster la documentación descrita en el apartado 6 de este pliego, sobre la instalación y administración del sistema en formato Office.
R10	El cluster CPG deberá demostrar su estabilidad para producción, para tal efecto, se lanzarán 100 jobs al sistema de colas con la misma ejecución de un código real, que sea altamente estable y probado en MareNostrum3, de la lista descrita en R2. Cada ejecución usará un mínimo de 1024 cores y un múltiplo de los cores de cada nodo de cómputo, de forma que los nodos queden completamente ocupados. Cada job tendrá una duración mínima de 2 horas. Los jobs se deberán distribuir por toda la máquina, (se configurará que cada job se ejecute en 1 isla o sin bloqueo en la red MPI). Después de la prueba se deberá cumplir lo siguiente: - Se deberán haber ejecutado y finalizado correctamente más del 98% de los Jobs - La variabilidad en el tiempo de ejecución de todos los Jobs no podrá ser superior al 7% En caso de no cumplir con alguno de los requisitos se deberá realizar un análisis por el no cumplimiento y subsanarlo, antes de volver a intentarlo.

R11	Para la última evolución de cada uno de los CTE también se deberá realizar una prueba de estabilidad, similar a la descrita en el R10 de este apartado 7, adaptada a cada CTE. Se lazarán 100 jobs al sistema de colas con la misma ejecución, con una duración mínima de 2 horas y una cantidad de cores significativa y múltiple de los que cada nodo de cómputo tenga, de forma que los nodos queden completamente ocupados. Después de la prueba se deberá cumplir lo siguiente: - Se deberán haber ejecutado y finalizado correctamente más del 98% de los Jobs - La variabilidad en el tiempo de ejecución de todos los Jobs no podrá ser superior al 7% En caso de no cumplir con alguno de los requisitos se deberá realizar un análisis por el no cumplimiento y subsanarlo, antes de volver a intentarlo.